Genomics xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Genomics

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

Original Article

Identification of prokaryotic promoters and their strength by integrating heterogeneous features

Hilal Tayara[a], Muhammad Tahir[a,b], Kil To Chong[a,c,*]

[a] Department of Electronics and Information Engineering, Chonbuk National University, Jeonju 54896, South Korea
[b] Department of Computer Science, Abdul Wali Khan University, Mardan 23200, Pakistan
[c] Advanced Electronics and Information Research Center, Chonbuk National University, Jeonju 54896, South Korea

ARTICLE INFO

Keywords:
Convolution neural network
Deep learning
DNA
iPSW(PseDNC-DL)
Promoter sites
Prompter strength

ABSTRACT

The promoter is a regulatory DNA region and important for gene transcriptional regulation. It is located near the transcription start site (TSS) upstream of the corresponding gene. In the post-genomics era, the availability of data makes it possible to build computational models for robustly detecting the promoters as these models are expected to be helpful for academia and drug discovery. Until recently, developed models focused only on discriminating the sequences into promoter and non-promoter. However, promoter predictors can be further improved by considering weak and strong promoter classification. In this work, we introduce a hybrid model, named iPSW(PseDNC-DL), for identification of prokaryotic promoters and their strength. It combines a convolutional neural network with a pseudo-di-nucleotide composition (PseDNC). The proposed model iPSW (PseDNC-DL) has been evaluated on the benchmark datasets and outperformed the current state-of-the-art models in both tasks namely promoter identification and promoter strength identification. The developed tool iPSW(PseDNC-DL) has been constructed in a web server and made freely available at https://home.jbnu.ac.kr/NSCL/PseDNC-DL.htm

1. Introduction

The promoter is a key element of DNA structure which regulates the transcription of particular genes in a particular cell. Gene expression regulation in prokaryotes is simple as compared to eukaryotic gene expression regulation. In former the two regulatory processes i.e., transcription and translation happened simultaneously while in latter case gene expression regulation is a more complex phenomenon, as initially DNA synthesis occurs followed by transcription and then translation. More importantly, in prokaryotes most of the genes are under the control of one operon, that regulates and transcribes most genes as one expression like clusters, for example in *Escherichia coli*, lac operon is required to transcribe several genes, while in eukaryotic organism each gene is regulated and transcribed individually [1,2]. Specific regions on chromosomes determine the fate of particular transcripts whether or how transcription might be initiated. Such sequences are termed as promoters, which are vital for gene expression regulation and controlling specific pathways. RNA polymerase (RNAP) and promoters share a flexible partnership in the initiation of transcription. One of the unique property of prokaryotic RNAP core enzyme (E) is its single form [3]. However, the onset of transcription is not facilitated by RNAP alone which could identify and attaches to promoter sequences. This identification of promoter requires regulatory proteins like σ-factors, binding temporarily with RNAP core enzyme constituting a holoenzyme (Eσ). The determination of transcription initiation site (TSS) and RNAP-promoter binding specificity is the key role of the holoenzyme. However, this ability of holoenzyme is dependent on various parameters like environmental conditions, stage of development and nutrition [4,5]. In *Escherichia coli*, based on the structure and function, σ-factors consist of multiple families i.e., σ70, σ54, σ38, σ32, σ28 and σ24 [6]. To date, there is no clear computational method in identifying the boundaries of the promoter region which is the main hindrance in promoter identification research. In this work, we examine the core promoter region ranging from -60 to $+21$ relative to TTS located at the $+1$ position.

The next-generation sequencing (NGS) add ease for the biologist to get effective information about gene expression regulation such as RNA-seq [7] or ChIP-seq [8] techniques; however, these methods inherit the pitfalls like cost-effectiveness and time consumption. The development of efficient computational models for identifying promoter region in prokaryotes genomes is a modern-day demand which must be answered effectively.

* Corresponding author at: Department of Electronics and Information Engineering, Chonbuk National University, Jeonju 54896, South Korea.
E-mail address: kitchong@jbnu.ac.kr (K.T. Chong).

https://doi.org/10.1016/j.ygeno.2019.08.009
Received 21 June 2019; Received in revised form 31 July 2019; Accepted 14 August 2019

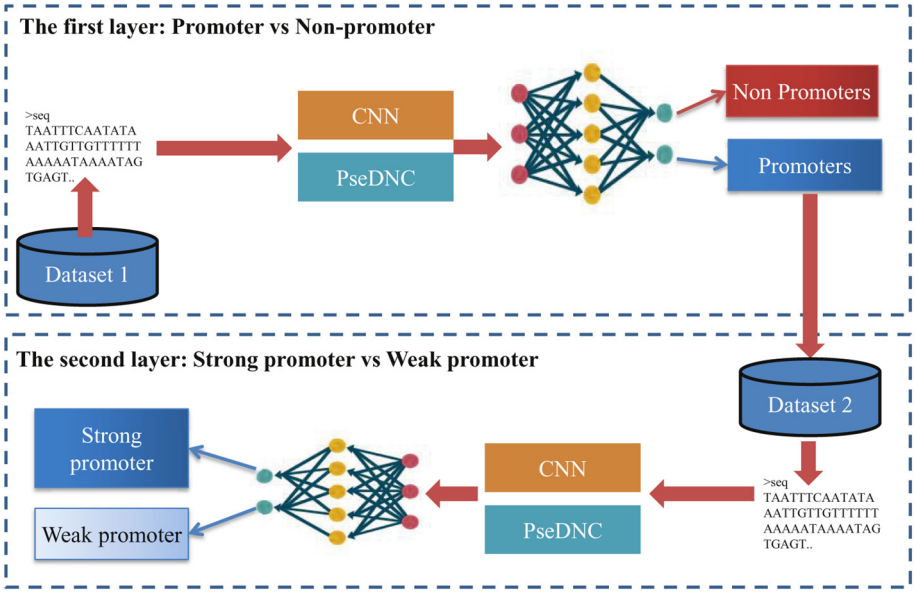

Fig. 1. Illustration of the two-step proposed model. The first layer discriminates the promoter sequences from non-promoter ones while the second layer decides the strength of the promoter sequences (strong or weak).

In this regards, Florquin et al. [9] introduced a structural model to predict the core promoter by using structural features of dinucleotides composition (DNC) and trinucleotides composition(TNC) to formulate the DNA sequences. Then, Li and Lin [10] produced a position correlation-scoring matrix (PCSM) method for the identification of sigma 70 promoters. For prokaryotic promoter recognition, Song [11] introduced two representative models with a variable-window *Z*-curve method. The model based on DNA duplex stability was proposed to predict sigma-54 and sigma-28 promoter samples for *Escherichia coli* [12]. In this connection, the computational model iPro54-PseKNC was introduced by Lin et al. to predict the sigma-54 promoters using pseudo-k-tuple nucleotide composition (PseKNC) feature method [13]. Liu et al. [6] introduced a two-layer computational model namely: "Promoter-2 L" to predict promoters and their six types using the multi-window-based PseKNC technique. Recently, Xiao et al. [14] introduced a two-layer predictor "PSW(2 L)-PseKNC" to identify promoters and their strength using the PseKNC method. In general, handcrafted features can be extracted using a powerful tools such as Type II PseKNC [15,16], repDNA [17], repRNA [18], Pse-Analysis [19], and BioSeq-Analysis [20]. However, the computational model may automatically extract the features from promoter sequences by deep learning approaches. The deep learning approaches produced very efficient outcomes in the domain of speech recognition [21], natural language processing [22], and image recognition [23–25]. Most recently, a number of prediction methods have been proposed based on deep learning for solving different biological tasks such as BiRen [26], CNNclust [27], DeepCpG [28], iDeepS [29], iRNA-PseKNC(2methyl) prediction model [30], alternative splicing sites prediction [31], branch point selection [32], splicing sites identification [33], etc.

Heterogeneous features have been integrated in different tools and shown efficient performance such as iRSpot-EL [34], HITS-PR-HHblits [35], ProtDec-LTR2.0 [36]. In this study, we introduce a novel two layers computational method called iPSW(PseDNC-DL) for the identification of promoters and their strength using deep learning methods and PseDNC feature extraction method. In the first layer, the model predicts whether a given DNA sample is a promoter or not; while in the second layer, the model identifies whether the predicted promoter is a strong promoter or a weak promoter. The proposed computational method has an efficient architecture for the identification of promoter and their strength using convolution neural networks (CNN) and outperforms the state-of-the-art machine learning methods. Previous methods prepared the features manually and then trained a classifier such as SVM. On the other hand, CNN learns features automatically from raw genomics sequences. In other words, deep learning based methods outperform machine learning based ones when there is a lack of problem understanding for features extraction which is the case of promoter identification task. A user-friendly web server has been constructed based on the developed tool iPSW(PseDNC-DL) and made freely accessible at https://home.jbnu.ac.kr/NSCL/PseDNC-DL.htm

In this paper, we follow the 5-step rules of Chou that have been followed in many publications such as [37–53]. These rules are preparing benchmark dataset, feature extraction, building a reliable predictor, cross validation and finally constructing an easy-to-use web-server.

2. Materials and methods

2.1. Benchmark datasets

In order to develop a useful computational model, we should select a reliable benchmark dataset to train and test the proposed model effectively. For this purpose, we download the *E.coli* benchmark datasets (http://www.jci-bioinfo.cn/iPSW(2 L)-PseKNC/images/Supp.pdf) from Xiao et al. [14]. This dataset is selected from the database RegulonDB [54] where all sequences are experimentally validated. It contains a positive subset of 3382 promoter sequences and a negative subset of 3382 non-promoter sequences. Furthermore, the positive subset contains 1591 strong promoter sequences and 1792 weak promoter sequences. These types of promoter strengths are based on their different levels in transcriptional activation and expression. In general, the strength of the promoters depends not only on the DNA sequence but also on the state of the cell. It worth noting that the weak promoters are as important as the strong ones as they can be considered as strong promoters at certain conditions of the cell. However, the main goal of this work is building a predictor based on the raw DNA sequences only. Therefore, the selected dataset contains only the sequences with strong evidence of promoter types (strong or weak) [54]. As a quality control, we use 5-fold cross-validation during the training process. The benchmark dataset is randomly split into five folds. Three folds are used for training, one for early stopping and the remaining fold for testing. Thus, the proposed model is trained 5 times and the reported results are the average performance of the 5-fold with standard error.

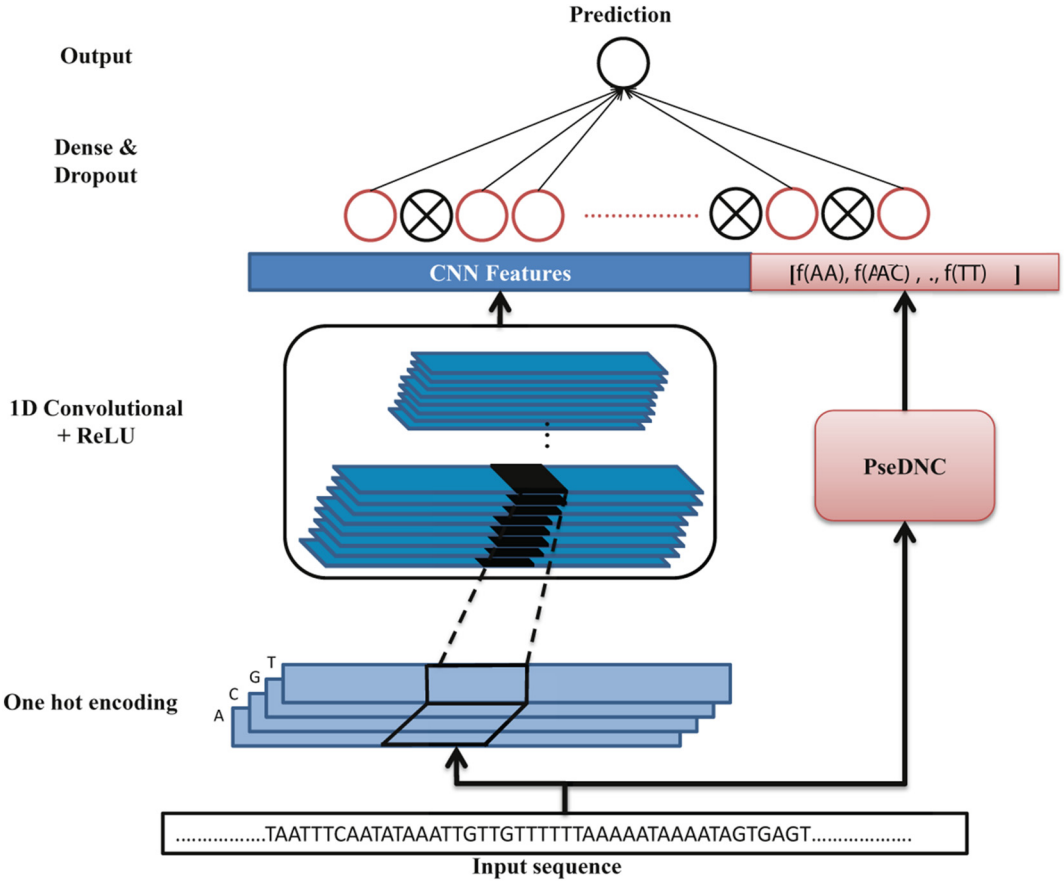

Fig. 2. Illustration of the proposed model iPSW(PseDNC-DL).

2.2. The proposed model

The proposed iPSW(PseDNC-DL) model, which is shown in Fig. 1, consists of two prediction layers where each layer contains deep learning and PseDNC. The first layer predicts whether the given sequence is a promoter sequence or not. If it is a promoter sequence it passes to the next layer in which the strength of the promoter is decided such as strong or weak.

The proposed model in each layer combines the extracted features from the convolution layers and PseDNC. The illustration of the proposed iPSW(PseDNC-DL) model is shown in Fig. 2.

Pseudo dinucleotide composition (PseDNC) is the simplest pseudo nucleic acid composition (NAC) and it is adopted in our study. The DNA sample S can be represented with L nucleic acid residues by:

$$S = R_1 R_2 R_3 ... R_L \tag{1}$$

where R_1 denotes the nucleic acid residue at first position of the sample, R_2 denotes the nucleic acid residue at the second position and so forth. We could mathematically describe each sample of DNA by its nucleic acid composition (NAC) such as

$$S = [f(T) f(G) f(C) f(A)]^{\mathbf{T}} \tag{2}$$

where $f(T)$, $f(G)$, $f(C)$ and $f(A)$ are the normalized occurrence frequencies of Thymine (T), Guanine (G), Cytosine (C), and Adenine (A), respectively, in the sample of DNA; the **T** is the transpose operator.

As shown in Eq. (2), through NAC, the information of whole sample is lost. Therefore, using the dinucleotide composition (DNC) to describe the sample of DNA results in a feature vector with $2^4 = 16$ elements such as

$$S = [f(TT) f(TG) f(TC) f(TA) ... f(AA)]^{\mathbf{T}} = [f_1 f_2 f_3 ... f_{16}]^T \tag{3}$$

where $f_1 = f(TT)$ is the normalized occurrence frequencies of TT in the sample of DNA; $f_2 = f(TG)$ is the normalized occurrence frequencies of TG in the sample of DNA, and so forth.

Convolution Neural Network is used to extract the important features from the raw DNA sequences automatically. We first apply the one-hot encoding for the input sample to permit successive convolution operations. The raw DNA sample is one-hot encoded and represented as a one-dimensional vector with four channels. The length of the vector is 81 nt and the four channels are A, C, G, and T and represented as (1000), (0100), (0010), and (0001), respectively. The configuration of the hyper-parameters of the CNN models in each layer is selected based on the grid search algorithm. The grid search algorithm is a hyperparameter optimization algorithm by which the best hyperparameters combination are selected from a manually defined subset of the hyperparameter space. The tuned hyperparameters are the dropout probability after convolution and dense layer, number of convolution layers, their filters, size of the filters, max pooling. The detailed configuration of the CNN models used in the first and the second layers of the promoter and promoter strength identification are given in Table 1.

In Table 1, the Conv1D(f,w,t) is a one-dimensional convolution operation with f filters with size of w and stride of t. All convolution layers are followed by rectified linear unit (ReLU) as an activation function. Concatenate operator is used to link together the learnt features from convolution layers and PseDNC features. Dropout(p) operator is used to occasionally remove intermediate values from the previous layer by randomly setting them to zero during training where p is the dropout probability. Dense(m) is the fully connected layer with m node. The last layer is the Sigmoid function which outputs prediction

Table 1
The configurations of the deep learning model for the identification of the promoters and their strength.

Model	Configuration	Output shape
1st layer: Promoter Identification	Input of raw DNA sequence	81 × 4
	Conv1D(16,7,1)	75 × 16
	Conv1D(32,7,1)	69 × 32
	Concatenta(PseDNC features and CNN features)	2224
	Dropout(0.7)	2224
	Dense(1)	1
	Sigmoid(1)	
2nd layer: Promoter strength Identification	Input of raw DNA sequence	81 × 4
	Conv1D(16,7,1)	75 × 16
	Conv1D(16,5,1)	71 × 16
	Concatenta(PseDNC features and CNN features)	1152
	Dropout(0.8)	1152
	Dense(1)	1
	Sigmoid(1)	1

probability. These operators are given mathematically as follows:

$$\mathrm{Conv}(R)_{jf} = ReLU\left(\sum_{s=0}^{S-1}\sum_{n=0}^{N-1} W_{sn}^{f} R_{j+s,n}\right) \tag{4}$$

$$f = w_{d+1} + \sum_{k=1}^{d} w_k z_k \tag{5}$$

$$f = w_{d+1} + \sum_{k=1}^{d} m_k w_k z_k \tag{6}$$

$$\mathrm{ReLU}(z) = max(0, z) \tag{7}$$

$$\mathrm{Sigmoid}(z) = \frac{1}{1+e^{-z}} \tag{8}$$

The convolution operation" Conv1D" is given in Eq. (4) where R represents the input of DNA sample, j and f represent the index of the output position and the index of the kernels, respectively. Every convolution filter W^f is represented by a weight matrix $S \times N$ where S is the filter size and N is the number of the input channels. The dense layer is given in Eq. (5) where z_k is the $1 \times$ d dimension feature vector, w_k is the weight of the z_k from the previous layer and w_{d+1} is an additive bias term. The dense layer is rewritten as Eq. (6) after adding dropout operator where m_k is sampled from Bernoulli distribution with probability of α. The activation functions ReLU and Sigmoid are given in Eqs. (7) and (8), respectively, where z is the input to these functions.

The proposed model has been constructed using Keras [55]. The weights were initialized using a random uniform in the range $[-0.05, 0.05]$. Adam optimizer is used with a learning rate of 0.001. The batch size and number of epochs are set to 16 and 40, respectively with early stopping based on validation loss.

2.3. Evaluation metrics

In this study, the following four evaluation parameters are used to study the performance of the computational method iPSW(PseDNC-DL) namely Accuracy (Acc), sensitivity (Sen), specificity (Spe), and Mathew's correlation coefficient (MCC). These parameters are represented by Chou's symbols [56–58] where r^+ represents the total number of the investigated promoters; r^- represents the number of the investigated non-promoter sequences; r_+^- represents the number of non-promoter sequences incorrectly classified as promoters; r_-^+ represents the number promoter sequences incorrectly classified as non-promoters.

$$ACC = 1 - \frac{r_-^+ + r_+^-}{r^+ + r^-} \tag{9}$$

$$Sn = 1 - \frac{r_-^+}{r^+} \tag{10}$$

$$Sp = 1 - \frac{r_+^-}{r^-} \tag{11}$$

$$MCC = \frac{1 - \frac{r_-^+ + r_+^-}{r^+ + r^-}}{\sqrt{\left(1 + \frac{r_+^- - r_-^+}{r^+}\right)\left(1 + \frac{r_-^+ - r_+^-}{r^-}\right)}} \tag{12}$$

Sensitivity and specificity individually denote the ability of the computational method to correctly predict the promoters and non-promoters. Accuracy calculates the correctness of our proposed prediction method for distinguishing promoters and non- promoter sites. MCC reflects the performance of the proposed prediction method on an imbalanced dataset, here the ratio of negative and positive samples are mostly the same. The range of MCC is $[-1,1]$ where -1 shows that the prediction completely does not agrees with the observation; 0 represents random prediction; 1 represents a perfect prediction. The area under the ROC curve (AUC) is used to evaluate the success rate of the proposed computational method. The AUC is an important indicator of the performance quality of the binary classifier.

3. Results and discussion

The proposed model iPSW(PseDNC-DL) combines the learned features from CNN and PseDNC. Therefore, in order to study the effects of adding hand-craft features (PseDNC) we build a model that contains CNN only. The parameters of the CNN models are selected using the grid search algorithm. Table 2 shows the comparison results of the proposed model iPSW(PseDNC-DL) with CNN model only. It is observed that adding PseDNC features improves all evaluation metrics for both layers (promoter identification layer and promoter strength identification layer). For more illustration Fig. 3 shows bar plots with standard errors of the performance of the proposed model iPSW(PseDNC-DL).

To further evaluate the performance of the iPSW(PseDNC-DL) model we compare it with the state-of-the-art model iPSW(2 L)-PseKNC [14]. Table 3 shows comparison results between iPSW(PseDNC-DL) and iPSW(2 L)-PseKNC [14]. The proposed model iPSW(PseDNC-DL) outperforms iPSW(2 L)-PseKNC in both tasks. In promoter identification task the iPSW (PseDNC-DL) improves the accuracy, specificity, sensitivity, and MCC by 1.97%, 1.94%,1.97%, and 3.94%, respectively. In promoter strength identification the proposed model iPSW(PseDNC-DL) improves the accuracy, sensitivity, and MCC by 1.15%, 3.58%, and 2.27%, respectively. These results show that combining CNN with PseDNC improves the performance of promoter and promoter strength

Table 2
Performance comparison between the proposed model iPSW(PseDNC-DL) that combines CNN and PseDNC features and a model that contains CNN only.

Model	Layer	Accuracy	Specificity	Sensitivity	MCC	auROC
CNN only	1st layer	84.45	85.54	83.33	0.6891	91.63
	2nd layer	72.00	78.46	64.80	0.4374	78.39
iPSW(PseDNC-DL)	1st layer	85.10	86.83	83.34	0.7024	92.50
	2nd layer	72.35	78.16	65.81	0.4440	78.97

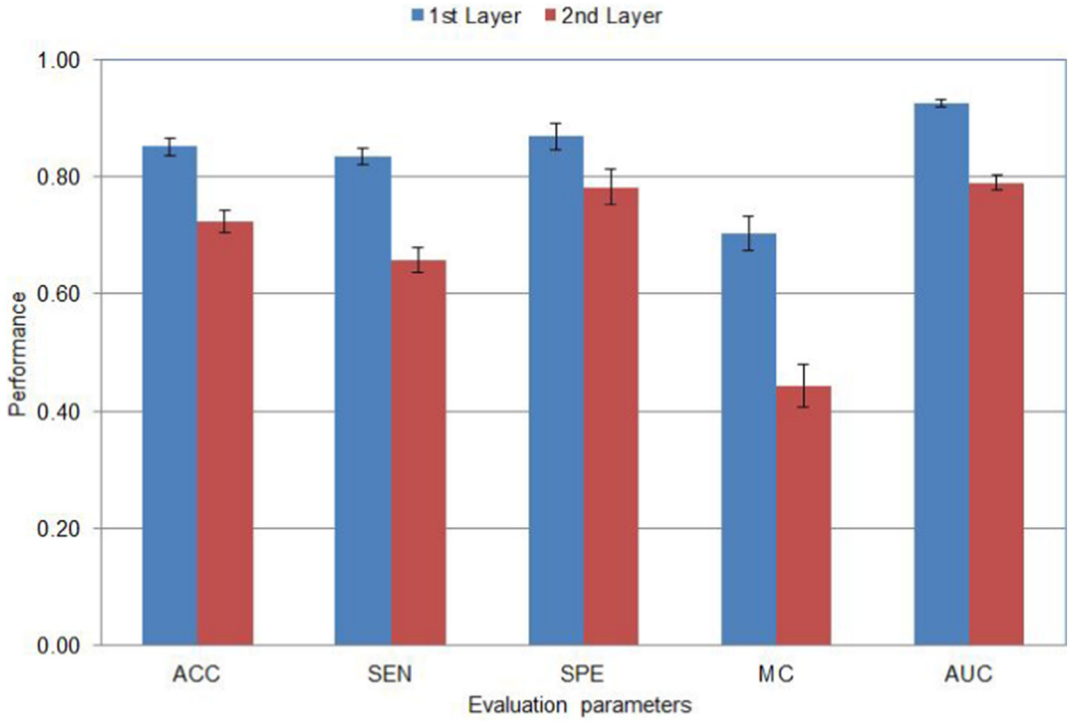

Fig. 3. Performance evaluation with standard error of the proposed model iPSW(PseDNC-DL) for both tasks, promoter and promoter strength identification.

Table 3
Performance comparison between the proposed model iPSW(PseDNC-DL) and the state-of-the-art model.

Model	Layer	Accuracy	Specificity	Sensitivity	MCC	AUC
iPSW(2 L)-PseKNC	1^{st} layer	0.8313	0.8489	0.8137	0.6630	0.9054
iPSW(PseDNC-DL)	1^{st} layer	0.8510	0.8683	0.8334	0.7024	0.9250
iPSW(2 L)-PseKNC	2^{nd} layer	0.7120	0.7917	06223	0.4213	0.7756
iPSW(PseDNC-DL)	2^{nd} layer	0.7235	0.7816	0.6581	0.4440	0.7897

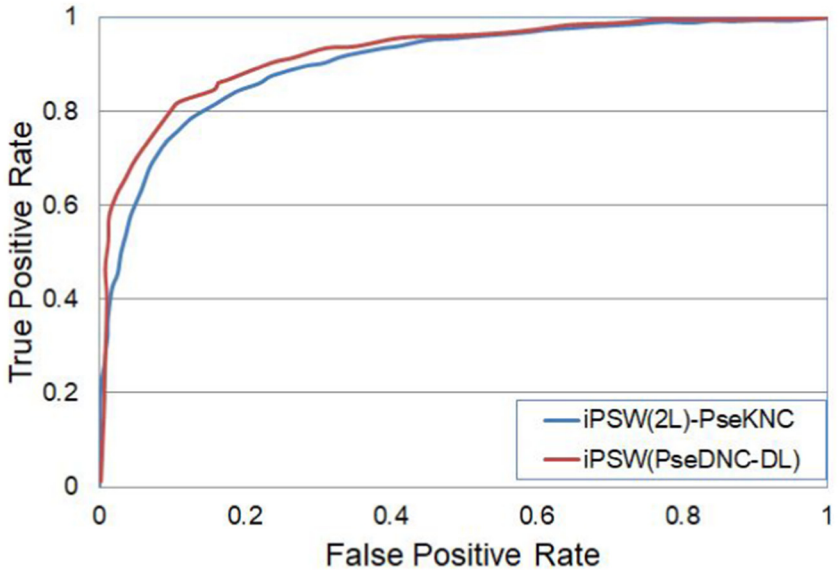

Fig. 4. The ROC curves of the proposed model iPSW(PseDNC-DL) and iPSW(2 L)-PseKNC for the promoter identification task.

identifications. It can be noticed that the MCC values of the second layer are not satisfying enough for both predictors. Thus, further studies should be considered such as improving the performance of the predictor and preparing larger datasets.

The AUC's of the proposed model iPSW(PseDNC-DL) and iPSW(2 L)-PseKNC of the promoter identification task is shown in Fig. 4. On the other hand, AUC's of the promoter strength identification task is shown in Fig. 5 for iPSW(PseDNC-DL) and iPSW(2 L)-PseKNC models.

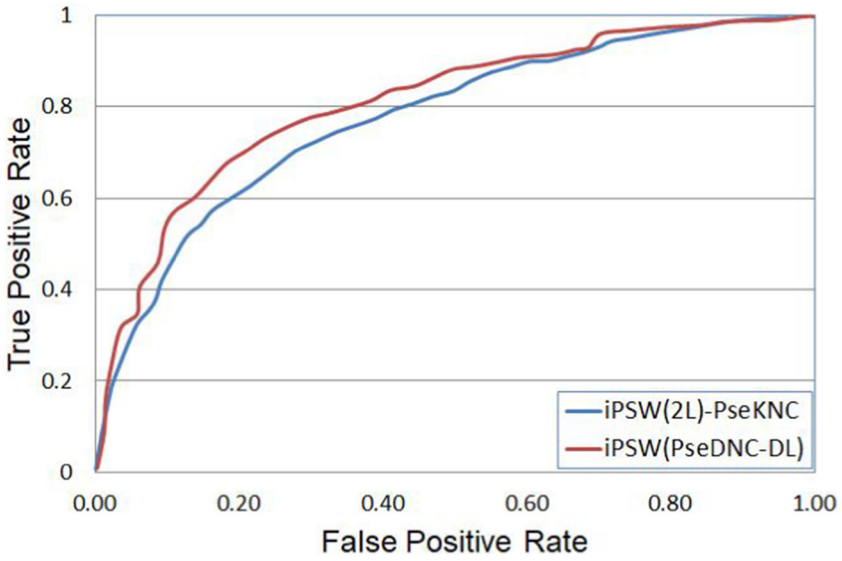

Fig. 5. The ROC curves of the proposed model iPSW(PseDNC-DL) and iPSW(2 L)-PseKNC for the promoter strength identification task.

Table 4
Comparison of the proposed promoter predictor iPSW(PseDNC-DL) with the other state-of-the-art predictors in using benchmark datasets provided by Liu et.al [6].

Model	Accuracy	MCC	Sensitivity	Specificity	Capacity
vw *Z*-curve [11]	0.8028	0.6098	0.7776	0.8280	No
PCSM [63]	0.7481	0.4980	0.7892	0.7070	No
iPro54 [13]	0.8045	0.6100	0.7776	0.8315	No
Stability [12]	0.7804	0.5615	0.7661	0.7948	No
iPromoter-2 L [6]	0.8168	0.6343	0.7920	0.8416	No
iPSW(2 L)-PseKNC [14]	0.8406	0.6811	0.8378	0.8434	Yes
iPSW(PseDNC-DL) [ours]	0.8566	0.7156	0.8972	0.8161	Yes

Finally, to further demonstrate the performance of the proposed model iPSW(PseDNC-DL) we compare it with other state-of-the-art promoter predictors on the benchmark dataset provided by Liu et.al [6]. The comparison results are given in Table 4 where capacity refers to the ability of the predictor to discriminate the strong and weak promoters. Thus, the comparison results show that the proposed model iPSW(PseDNC-DL) outperforms other state-of-the-art ones.

4. Web-server

In order to make the proposed tool accessible by other researchers, we developed an easy to use web-server at https://home.jbnu.ac.kr/NSCL/PseDNC-DL.htm. This step is followed by many researchers such

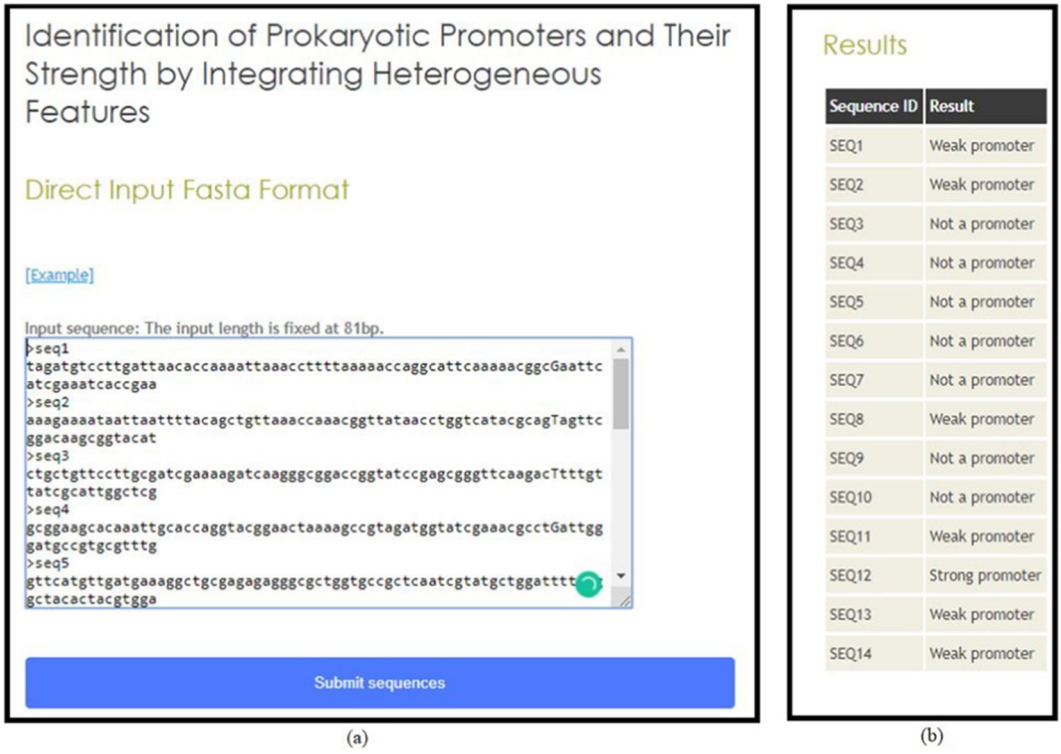

Fig. 6. Snippet of the developed web-server for the proposed tool iPSW(PseDNC-DL). (a) iPSW(PseDNC-DL) input, (b) iPSW(PseDNC-DL) output.

as [59–62]. The web-server is developed using Python and Flask library. It supports direct sequence input and uploading a file containing sequences for prediction. The length of each sequence should be 81 nt containing A, C, G, and T. In the case of uploading a file, the maximum number of sequences for prediction is 1000. Fig. 6 shows a snippet from the web-server where Fig. 6(a) shows an example of inserting sequences for prediction and Fig. 6(b) shows the output of the predictor.

5. Conclusions

Promoter identification is an important step for understanding gene transcription regulation for both drug discovery and academia. In this paper, a novel computational model has been proposed for identification of prokaryotic promoter and their strength using deep learning and pseudo dinucleotide composition. The proposed model has been evaluated on a benchmark dataset and outperformed the current state-of-the-art model in both tasks namely promoter identification and promoter strength identification. The developed tool iPSW(PseDNC-DL) has been constructed in a web server and made freely available at https://home.jbnu.ac.kr/NSCL/PseDNC-DL.htm

Funding

This research was supported by the Brain Research Program of the National Research Foundation (NRF) funded by the Korean Government (MSIT) (No. NRF-2017M3C7A1044815).

Declaration of Competing Interest

The authors declare no conflict of interest.

References

[1] M. Kozak, Initiation of translation in prokaryotes and eukaryotes, Gene 234 (2) (1999) 187–208.
[2] D. Sweetser, M. Nonet, R.A. Young, Prokaryotic and eukaryotic rna polymerases have homologous core subunits, Proc. Natl. Acad. Sci. 84 (5) (1987) 1192–1196.
[3] G.J. Schneider, R. Hasekorn, Rna polymerase subunit homology among cyanobacteria, other eubacteria and archaebacteria, J. Bacteriol. 170 (9) (1988) 4136–4140.
[4] S. Campagne, M.E. Marsh, G. Capitani, J.A. Vorholt, F.H. Allain, Structural basis for- 10 promoter element melting by environmentally induced sigma factors, Nat. Struct. Mol. Biol. 21 (3) (2014) 269.
[5] A. Feklistov, Rna polymerase: in search of promoters, Ann. N. Y. Acad. Sci. 1293 (1) (2013) 25–32.
[6] B. Liu, F. Yang, D.-S. Huang, K.-C. Chou, Ipromoter-2l: a two-layer predictor for identifying promoters and their types by multi-window-based pseknc, Bioinformatics 34 (1) (2017) 33–40.
[7] C. Trapnell, L. Pachter, S.L. Salzberg, Tophat: discovering splice junctions with rna-seq, Bioinformatics 25 (9) (2009) 1105–1111.
[8] T.S. Furey, Chip–seq and beyond: new and improved methodologies to detect and characterize protein–dna interactions, Nat. Rev. Genet. 13 (12) (2012) 840.
[9] K. Florquin, Y. Saeys, S. Degroeve, P. Rouze, Y. Van de Peer, Large-scale structural analysis of the core promoter in mammalian and plant genomes, Nucleic Acids Res. 33 (13) (2005) 4255–4264.
[10] Q.-Z. Li, H. Lin, The recognition and prediction of σ70 promoters in escherichia coli k-12, J. Theor. Biol. 242 (1) (2006) 135–141.
[11] K. Song, Recognition of prokaryotic promoters based on a novel variable-window z-curve method, Nucleic Acids Res. 40 (3) (2011) 963–971.
[12] S. e Silva, F. Forte, I.T. Sartor, T. Andrighetti, G.J. Gerhardt, A.P.L. Delamare, S. Echeverrigaray, Dna duplex stability as discriminative characteristic for escherichia coli σ54-and σ28-dependent promoter sequences, Biologicals 42 (1) (2014) 22–28.
[13] H. Lin, E.-Z. Deng, H. Ding, W. Chen, K.-C. Chou, ipro54-pseknc: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition, Nucleic Acids Res. 42 (21) (2014) 12961–12972.
[14] X. Xiao, Z.-C. Xu, W.-R. Qiu, P. Wang, H.-T. Ge, K.-C. Chou, Ipsw (2l)-pseknc: a two-layer predictor for identifying promoters and their strength by hybrid features via pseudo k-tuple nucleotide composition, Genomics. (2018), https://doi.org/10.1016/j.ygeno.2018.12.001 in press.
[15] F.-Y. Dao, H. Lv, F. Wang, C.-Q. Feng, H. Ding, W. Chen, H. Lin, Identify origin of replication in saccharomyces cerevisiae using two-step feature selection technique, Bioinformatics. 35 (2019) 2075–2083.
[16] C.-Q. Feng, Z.-Y. Zhang, X.-J. Zhu, Y. Lin, W. Chen, H. Tang, H. Lin, Iterm-pseknc: a sequence-based tool for predicting bacterial transcriptional terminators, Bioinformatics. 35 (2019) 1469–1477.
[17] B. Liu, F. Liu, L. Fang, X. Wang, K.-C. Chou, 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 (8) (2014) 1307–1309.
[18] B. Liu, F. Liu, L. Fang, X. Wang, K.-C. Chou, Reprna: a web server for generating various feature vectors of rna sequences, Mol. Gen. Genomics. 291 (1) (2016) 473–481.
[19] B. Liu, H. Wu, D. Zhang, X. Wang, K.-C. Chou, Pse-analysis: a python package for dna/rna and protein/peptide sequence analysis based on pseudo components and kernel methods, Oncotarget 8 (8) (2017) 13338.
[20] B. Liu, Bioseq-analysis: a platform for dna, rna and protein sequence analysis based on machine learning approaches, Brief. Bioinform. (2017), https://doi.org/10.1093/bib/bbx165.
[21] G. Hinton, L. Deng, D. Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, et al., Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups, IEEE Signal Process. Mag. 29 (6) (2012) 82–97.
[22] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural language processing (almost) from scratch, J. Mach. Learn. Res. 12 (Aug) (2011) 2493–2537.
[23] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[24] H. Tayara, K. Chong, Object detection in very high-resolution aerial images using one-stage densely connected feature pyramid network, Sensors 18 (10) (2018) 3341.
[25] H. Tayara, K.G. Soo, K.T. Chong, Vehicle detection and counting in high-resolution aerial images using convolutional regression neural network, IEEE Access 6 (2018) 2220–2230.
[26] B. Yang, F. Liu, C. Ren, Z. Ouyang, Z. Xie, X. Bo, W. Shu, Biren: predicting enhancers with a deep-learning-based model using the dna sequence alone, Bioinformatics 33 (13) (2017) 1930–1936, https://doi.org/10.1093/bioinformatics/btx105.
[27] G. Aoki, Y. Sakakibara, Convolutional neural networks for classification of alignments of non-coding rna sequences, Bioinformatics 34 (13) (2018) i237–i244, https://doi.org/10.1093/bioinformatics/bty228.
[28] C. Angermueller, H.J. Lee, W. Reik, O. Stegle, Deepcpg: accurate prediction of single-cell dna methylation states using deep learning, Genome Biol. 18 (1) (2017) 67.
[29] X. Pan, P. Rijnbeek, J. Yan, H.-B. Shen, Prediction of rna-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks, BMC Genomics 19 (1) (2018) 511.
[30] M. Tahir, H. Tayara, K.T. Chong, Irna-pseknc(2methyl): identify rna 2′-o-methylation sites by convolution neural network and chou's pseudo components, J. Theor. Biol. 465 (2019) 1–6, https://doi.org/10.1016/j.jtbi.2018.12.034 http://www.sciencedirect.com/science/article/pii/S0022519318306349.
[31] M. Oubounyt, Z. Louadi, H. Tayara, K.T. Chong, Deep learning models based on distributed feature representations for alternative splicing prediction, IEEE Access 6 (2018) 58826–58834, https://doi.org/10.1109/ACCESS.2018.2874208.
[32] I. Nazari, H. Tayara, K.T. Chong, Branch point selection in rna splicing using deep learning, IEEE Access (2018) 1, https://doi.org/10.1109/ACCESS.2018.2886569.
[33] H. Tayara, M. Tahir, K.T. Chong, Iss-cnn: identifying splicing sites using convolution neural network, Chemom. Intell. Lab. Syst. 188 (2019) 63–69, https://doi.org/10.1016/j.chemolab.2019.03.002 http://www.sciencedirect.com/science/article/pii/S0169743919300218.
[34] B. Liu, S. Wang, R. Long, K.-C. Chou, Irspot-el: identify recombination spots with an ensemble learning approach, Bioinformatics 33 (1) (2016) 35–41.
[35] B. Liu, S. Jiang, Q. Zou, Hits-pr-hhblits: protein remote homology detection by combining pagerank and hyperlink-induced topic search, Brief. Bioinform. (2018), https://doi.org/10.1093/bib/bby104.
[36] J. Chen, M. Guo, S. Li, B. Liu, Protdec-ltr2. 0: an improved method for protein remote homology detection by combining pseudo protein and supervised learning to rank, Bioinformatics 33 (21) (2017) 3473–3476.
[37] W. Chen, P. Feng, H. Ding, H. Lin, K.-C. Chou, Irna-methyl: identifying n6-methyladenosine sites using pseudo nucleotide composition, Anal. Biochem. 490 (2015) 26–33.
[38] J. Song, Y. Wang, F. Li, T. Akutsu, N.D. Rawlings, G.I. Webb, K.-C. Chou, Iprot-sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites, Brief. Bioinform. 20 (2019) 638–658.
[39] L. Cai, T. Huang, J. Su, X. Zhang, W. Chen, F. Zhang, L. He, K.-C. Chou, Implications of newly identified brain eqtl genes and their interactors in schizophrenia, Mol. Ther. Nucleic Acids 12 (2018) 433–442.
[40] K.-C. Chou, H.-B. Shen, Recent advances in developing web-servers for predicting protein attributes, Nat. Sci. 1 (2) (2009) 63.
[41] K.-C. Chou, Impacts of bioinformatics to medicinal chemistry, Med. Chem. 11 (3) (2015) 218–234.
[42] M. Tahir, M. Hayat, Inuc-stnc: a sequence-based predictor for identification of nucleosome positioning in genomes by extending the concept of saac and chou's pseaac, Mol. BioSyst. 12 (8) (2016) 2587–2593.
[43] M. Hayat, M. Tahir, Psofuzzysvm-tmh: identification of transmembrane helix segments using ensemble feature space by incorporated fuzzy support vector machine, Mol. BioSyst. 11 (8) (2015) 2255–2262.
[44] M. Tahir, M. Hayat, S.A. Khan, A two-layer computational model for discrimination of enhancer and their types using hybrid features pace of pseudo k-tuple nucleotide composition, Arab. J. Sci. Eng. (2017) 1–9.
[45] M. Tahir, M. Hayat, M. Kabir, Sequence based predictor for discrimination of

enhancer and their types by applying general form of chou's trinucleotide composition, Comput. Methods Prog. Biomed. 146 (2017) 69–75.
[46] M. Tahir, M. Hayat, S.A. Khan, Inuc-ext-psetnc: an efficient ensemble model for identification of nucleosome positioning by extending the concept of chou's pseaac to pseudo-tri-nucleotide composition, Mol. Gen. Genomics. 294 (1) (2019) 199–210.
[47] M. Kabir, M. Hayat, Irspot-gaensc: identifing recombination spots via ensemble classifier and extending the concept of chou's pseaac to formulate dna samples, Mol. Gen. Genomics. 291 (1) (2016) 285–296.
[48] S. Ahmad, M. Kabir, M. Hayat, Identification of heat shock protein families and j-protein types by incorporating dipeptide composition into chou's general pseaac, Comput. Methods Prog. Biomed. 122 (2) (2015) 165–174.
[49] M. Kabir, S. Ahmad, M. Iqbal, M. Hayat, Inr-2l: a two-level sequencebased predictor developed via chou's 5-steps rule and general pseaac for identifying nuclear receptors and their families, Genomics. (2019), https://doi.org/10.1016/j.ygeno.2019.02.006 in press.
[50] M. Waris, K. Ahmad, M. Kabir, M. Hayat, Identification of dna binding proteins using evolutionary profiles position specific scoring matrix, Neurocomputing 199 (2016) 154–162.
[51] M. Kabir, D.-J. Yu, Predicting dnase i hypersensitive sites via un-biased pseudo trinucleotide composition, Chemom. Intell. Lab. Syst. 167 (2017) 78–84.
[52] M. Tahir, H. Tayara, K.T. Chong, Ipseu-cnn: identifying rna pseudouridine sites using convolutional neural networks, Mol. Ther. Nucleic Acids 16 (2019) 463–470.
[53] M. Tahir, M. Hayat, Machine learning based identification of protein–protein interactions using derived features of physiochemical properties and evolutionary profiles, Artif. Intell. Med. 78 (2017) 61–71.
[54] S. Gama-Castro, H. Salgado, A. Santos-Zavaleta, D. Ledezma-Tejeida, L. Muniz-Rascado, J.S. García-Sotelo, K. Alquicira-Hernández, I. Martínez-Flores, L. Pannier, J.A. Castro-Mondragón, et al., Regulondb version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond, Nucleic Acids Res. 44 (D1) (2015) D133–D143.
[55] F. Chollet, et al., Keras: The Python Deep Learning Library, Astrophysics Source Code Library, 2018 Publication Date:06/2018.
[56] K.-C. Chou, Prediction of signal peptides using scaled window, Peptides 22 (12) (2001) 1973–1979, https://doi.org/10.1016/S0196-9781(01)00540-X http://www.sciencedirect.com/science/article/pii/S019697810100540X.
[57] W. Chen, P.-M. Feng, H. Lin, K.-C. Chou, Irspot-psednc: identify recombination spots with pseudo dinucleotide composition, Nucleic Acids Res. 41 (6) (2013) e68.
[58] Y. Xu, X.-J. Shao, L.-Y. Wu, N.-Y. Deng, K.-C. Chou, Isno-aapair: incorporating amino acid pairwise coupling into pseaac for predicting cysteine s-nitrosylation sites in proteins, PeerJ 1 (2013) e171.
[59] Z. Liu, X. Xiao, D.-J. Yu, J. Jia, W.-R. Qiu, K.-C. Chou, Prnam-pc: predicting n6-methyladenosine sites in rna sequences via physical–chemical properties, Anal. Biochem. 497 (2016) 60–67.
[60] B. Liu, R. Long, K.-C. Chou, Idhs-el: identifying dnase i hypersensitive sites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework, Bioinformatics 32 (16) (2016) 2411–2418.
[61] J. Wang, B. Yang, J. Revote, A. Leier, T.T. Marquez-Lago, G. Webb, J. Song, K.-C. Chou, T. Lithgow, Possum: a bioinformatics toolkit for generating numerical sequence feature descriptors based on pssm profiles, Bioinformatics 33 (17) (2017) 2756–2758.
[62] Z. Chen, P. Zhao, F. Li, A. Leier, T.T. Marquez-Lago, Y. Wang, G.I. Webb, A.I. Smith, R.J. Daly, K.-C. Chou, et al., Ifeature: a python package and web server for features extraction and selection from protein and peptide sequences, Bioinformatics 34 (14) (2018) 2499–2502.
[63] Q.-Z. Li, H. Lin, The recognition and prediction of σ70 promoters in escherichia coli k-12, J. Theor. Biol. 242 (1) (2006) 135–141.