Genomics 111 (2019) 1160–1166

Contents lists available at ScienceDirect

Genomics

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

Methods Paper

iPromoter-FSEn: Identification of bacterial σ^{70} promoter sequences using feature subspace based ensemble classifier

Md. Siddiqur Rahman, Usma Aktar, Md. Rafsan Jani, Swakkhar Shatabda*

Department of Computer Science and Engineering, United International University Madani Avenue, Satarkul, Badda, Dhaka 1212, Bangladesh

ABSTRACT

Sigma promoter sequences in bacterial genomes are important due to their role in transcription initiation. Sigma 70 is one of the most important and crucial sigma factors. In this paper, we address the problem of identification of σ^{70} promoter sequences in bacterial genome. We propose iPromoter-FSEn, a novel predictor for identification of σ^{70} promoter sequences. Our proposed method is based on a feature subspace based ensemble classifier. A large set of of features extracted from the sequence of nucleotides are divided into subsets and each subset is given to individual single classifiers to learn. Based on the decisions of the ensemble an aggregate decision is made by the ensemble voting classifier. We tested our method on a standard benchmark dataset extracted from experimentally validated results. Experimental results shows that iPromoter-FSEn significantly improves over the state-of-the art σ^{70} promoter sequence predictors. The accuracy and area under receiver operating characteristic curve of iPromoter-FSEn are 86.32% and 0.9319 respectively. We have also made our method readily available for use as an web application from: http://ipromoterfsen.pythonanywhere.com/server.

1. Introduction

RNA polymerase in cells opens up the double helical structure of DNA and starts the synthesis of RNA or transcription. For transcription to take place, RNA polymerase must bind near to the gene locations in the DNA. These are small regions near gene containing 100 to 1000 base-pairs and known as promoters. In bacterial genomes, RNA polymerase forms a RNA polymerase holoenzyme by combining itself with specific sigma factor depending on the environment and gene. Thus sigma factors play a very important role in recognizing promoter sequences in DNA. Different sigma factors are distinguished according to their molecular weights. The sigma 70 (σ^{70}) factor is the 'housekeeping' of 'primary' sigma factor as it is associated with transcription of most of the genes [31]. Knowledge about sigma promoter regions provides us with essential information about the gene transcription, regulation and helps to improve annotation of the genome.

Computational methods for promoter sequence prediction gained much popularity due to cost expensive and time consuming nature of molecular techniques [72]. Several in silico computational methods are being used in the literature. They include position weight matrices [25], phylogenetic foot-printing [30], machine learning methods [72], etc. Machine learning methods have so far formulated the problem as a binary classification problem and used a large variety of supervised learning algorithms like Support Vector Machines [43, 42], Artificial Neural Network [21, 53, 23], Markov Models [1], Hidden Markov Models [54], Random Forest [51] for promoter sequence identification.

One of the earliest methods for predicting promoter sequence was proposed in [21]. They used 80 known promoter sequences to train a neural network. A prediction method of σ^{70} promoter sequences was proposed in [29] using a sequence alignment based kernel. They performed their experiments on 683 experimentally validated *Escherichia coli* promoter sequences. Another analysis on σ^{70} sequenecs were done in [40] where they used position correlation scoring matrix. Gordon et al. proposed a position weighted matrix based method in [28].

An experimentally verified database Pro54DB was proposed in [41] that contained 210 experimentally verified σ^{54} sequences. Lin et al. [43] used support vector machines to predict σ^{54} promoter sequences and proposed iPro54-PseKNC. iPro70-PseZNC was proposed in [42] for σ^{70} promoter sequence detection using support vector machine and pseudo nucleotide composition. Liu et al. [51] proposed iPromoter-2 L which is a two layer promoter sequence detector. In the first layer, a classifier is used to identify the promoter sequences from non-promoter regions and then in the second layer it distinguishes between several types of sigma promoter sequences. They have used Random Forest classifier to predict different sigma factors: $\sigma^{24,}$ $\sigma^{28,}$ σ^{32} $\sigma^{38,}$ $\sigma^{54,}$ σ^{70}.

In this paper, we present iPromoter-FSEn a method for identification of bacterial σ^{70} promoter sequences using feature subspace ensemble. iPromoter-FSEn uses a large number of features mostly generated from the DNA sequence information. In many of the methods in literature, we have observed that feature selection are done on the dataset that often leads to overfitting of the data. iPromoter-FSEn divides the total number of features into three subsets, exclusive and overlapping of each

* Corresponding author.
E-mail address: swakkhar@cse.uiu.ac.bd (S. Shatabda).

https://doi.org/10.1016/j.ygeno.2018.07.011
Received 25 April 2018; Received in revised form 7 July 2018; Accepted 12 July 2018
Available online 29 July 2018

other and the subset of features along with the dataset labels are fed to three different classifiers for training. Each single weak classifier are trained using the subset of features and an weighted voting scheme is used for final ensemble classifier. We have tested the performance of iPromoter-FSEn with that of state-of-the-art classifiers on a standard benchmark dataset. It has significantly improved the accuracy on cross-validation tests. Improvement in other metrics are also noticeable and hence establishes the effectiveness of the feature subs-pacing technique. We have also made our method readily available for use as an web application from: http://ipromoterfsen.pythonanywhere.com/server.

2. Materials and methods

To develop a really useful sequence-based statistical predictor for a biological system as reported in a series of recent publications [39, 7, 48, 52, 70, 5, 24, 75, 69, 2], one should strictly observe the famous 5-step rules [13]; i.e., making the following five steps very clear: (i) how to construct or select a valid benchmark dataset to train and test the predictor; (ii) how to formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (iii) how to introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) how to properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (v) how to establish a user-friendly web-server for the predictor that is accessible to the public. In the rest of the paper, we are to describe how to deal with these steps one-by- one.

A system diagram of our proposed method, iPromoter-FSEn is depicted in Fig. 1. Instances or sample sequences from the training dataset is first fed into the feature extraction module and features are generated for each of he data samples. The feature space containing all the data samples and features are then sub-sampled. Thus the feature space is divided into several overlapping and non-overlapping subspaces. Each of these subspaces are then fed to different classifiers who treat these sub-sampled feature as training data to themselves. Predictions from each of these classifiers are then combined using an ensemble technique to find the final prediction.

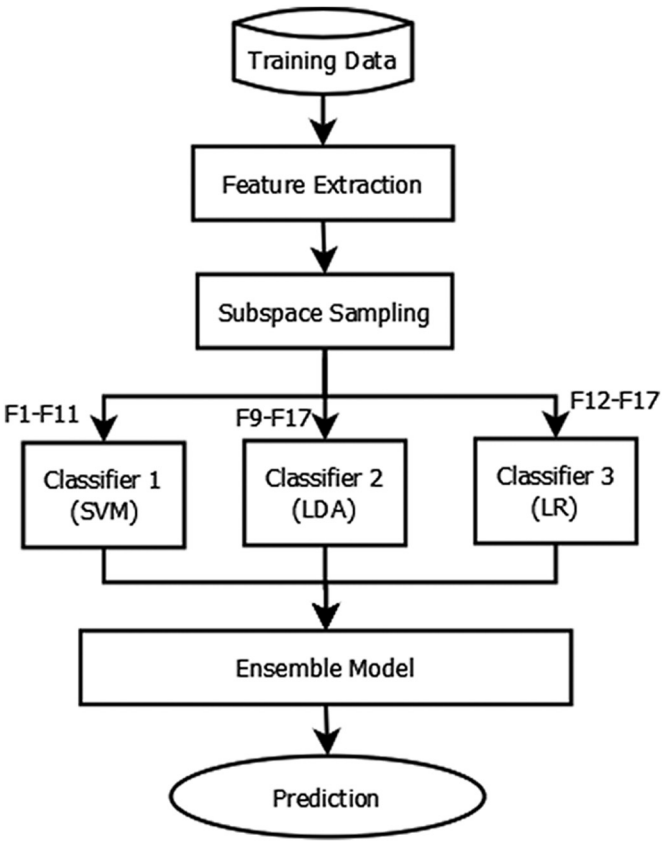

Fig. 1. System Diagram of iPromoter-FSEn.

2.1. Benchmark dataset

Selection of a standard benchmark dataset is very important step in establishing a prediction method. In this paper, we have used the σ^{70}promoter sequence dataset proposed in [42]. Lin et al. constructed this dataset by taking σ^{70}promoter sequences from *E. coli* genome. The promoter sequences were downloaded from RegulonDB [26]. All these sequences are experimentally validated. For any machine learning problem of binary classification, the dataset can be expressed as following:

$$\mathbb{S} = \mathbb{S}^{+} \cup \mathbb{S}^{-} \quad (1)$$

Here, $\mathbb{S}$ denotes the dataset and $\mathbb{S}^{+}$ is the set of positive samples or sequences that are included as σ^{70} promoter sequences and $\mathbb{S}^{-}$ is the set of negative samples. Here positive samples are all taken from the RegulonDB and their length is 81 base pairs. These 81 basepairs are taken as 60 base-pairs upstream and 20 base-pairs downstream from the transcription start site (TSS) in the middle. The negative dataset $\mathbb{S}^{-}$ was constructed by taking 81 base-pair sequences randomly taken from inter-genic and coding regions of *E. coli* genome. A small summary of the dataset is given in Table 1. Note that, 741 instances are in the positive dataset, $\mathbb{S}^{+}$ and 1400 sequences are in the negative dataset and the dataset is thus slightly imbalanced.

2.2. Fearure extraction

With the explosive growth of biological sequences in the post-genomic era, one of the most important but also most difficult problems in computational biology is how to express a biological sequence with a discrete model or a vector, yet still keep considerable sequence-order information or key pattern characteristic. This is because all the existing machine-learning algorithms can only handle vector but not sequence samples, as elucidated in a comprehensive review [16]. However, a vector defined in a discrete model may completely lose all the sequence-pattern information. To avoid completely losing the sequence-pattern information for proteins, the pseudo amino acid composition was proposed. Ever since the concept of Chou's PseAAC [44] was proposed, it has been widely used in nearly all the areas of computational proteomics [17]. Because it has been widely and increasingly used, recently three powerful open access soft-wares, called 'PseAAC-Builder', 'propy', and 'PseAAC-General', were established: the former two are for generating various modes of Chou's special PseAAC; while the 3rd one for those of Chou's general PseAAC, including not only all the special modes of feature vectors for proteins but also the higher level feature vectors such as 'Functional Domain' mode, 'Gene Ontology' mode, and 'Sequential Evolution' or 'PSSM' mode. Encouraged by the great successes of using PseAAC to deal with protein/peptide sequences, the concept of PseKNC (Pseudo K-tuple Nucleotide Composition) [3] was developed for generating various feature vectors for DNA/RNA sequences that have proved very useful as well [4]. Particularly, recently a very powerful web-server called 'Pse-in-One' [45] and its updated version 'Pse-in-One2.0' [49] have been established that can be used to

Table 1
Summary of the dataset taken from [41].

Class Label	Number of Instances	Sequence Length of Instances
Positive or Promoters	741	81
Negative or Non-Promoters	1400	81

generate any desired feature vectors for protein/peptide and DNA/RNA sequences according to the need of users' studies.

Genome or code of life is supposed to the source of all information. This has suggested many authors to use sequence based features only. Moreover, these features are very simple and easy to generate compared to structure, physico-chemical properties based or other derived features that are calculation intensive. In this paper, we too have used only sequence based feature generation techniques. In this section, we provide details of the features for representation of each promoter sequences in the dataset.

A DNA sequence $D \in \mathbb{S}$ of length L is a sequence of nucleotides. Formally, as below:

$$D = N_1, N_2, N_3, \cdots, N_L \tag{2}$$

We generate the features from this given sequence using our feature extraction method as following:

$$\mathbb{F}(D) = [f_1, f_2, \cdots, f_n] \tag{3}$$

2.2.1. Statistical measure of nucleotides

In this feature group, we have used different statistics on nucleotides as features. Among them are: frequency or count of each of the nucleotides [4], total GC count [1], mean, variance and standard deviations of the counts [3], GC skew [27] along the sequence [18] and AT/GC ratio [1].

2.2.2. $k-$ mer composition

We have considered composition of different lengths of $k-$ mers. Composition is the normalized frequency of $k-$ mers. In this paper, we have considered $k-$ mers of length $k = 2, 3, 4, 5, 6$. Thus the total number of features is $4^2 + 4^3 + 4^4 + 4^5 + 4^6 = 5792$.

2.2.3. Gapped $k-$ mer composition

Gapped $k-$ mer composition is previously used in the literature of protein and DNA attribute selection. We have used a similar concept here using < mi > g < /mi > < mo > − < /mo > gapped dinucleotide composition. However, we have further extended the idea of gapped composition to tri-nucleotides. For tri-nucleotides the gap in between are of either in the pattern of XX_X or XX_X. We have used various patterns of gaps, by differing the gaps in between, g = 1,2,....,75.

2.2.4. Approximate signal pattern count

It has been observed in the analysis of the transcription start sites in *E. coli* that promoter like signals are densely distributed in specific regions [34, 56]. It suggested us to find specific signals or their approximate matches within the promoter searching window. We have thus incorporated the approximate count of those signal patterns. Seven such patterns are used in this paper. They are: TATAAT, TAATAT, TATAAA, AAATAT, TTGACA, ACAGTT and AACGAT. We have counted all exact or approximate occurrence of all these patterns and their cyclic right shifted versions for all except the last one. Approximate occurrence of the strings were taken into account if there is at least three matches with the candidate pattern.

2.2.5. Position specific occurences

Among this group of features are mean inter-nucleotide occurrence [4], DNase I parameter [1] and position specific information of nucleotides [18].

2.2.6. Distribution of nucleotides

In addition to these features, we have also taken the frequency count of four different types of nucleotides along the sequence. We have divided the sequence into s equal spaced partitions and for each partition taken the frequency of the nucleotides as feature. We have kept $s = 10$. Inspired by the segmented distribution of position specific scoring matrix (PSSM) [20] used in predicting subcellular localization of gram positive and gram negative bacterial proteins, here we have used segmented distribution of nucleotide counts. We have taken the indices of partial counts of 10% upto 50% with 5 equal intervals.

A summary of all features generated for iPromoter-FSEn is given in Table 2.

Table 2
Summary of feature extracted for iPromoter-FSEn.

Sn	Feature Name	# of Features	Feature Group
F1	Nucleotide Frquency	4	Nucleotide Statistics
F2	Nucleotide Mean, Variance, Stddev	3	
F3	G-C Skew and AT/GC ratio	82	
F4	2-mer composition	16	$k-$ mer Composition
F5	3-mer composition	64	
F6	4-mer composition	256	
F7	5-mer composition	1024	
F8	6-mer composition	4096	
F9	Gapped dinucleotide count	1184	Gapped $k-$ mer Composition
F10	Gapped trinucleotide count (XX_X)	9472	
F11	Gapped trinucleotide count (X_XX)	9472	
F12	Approximate Signal Pattern Count	37	Approximate Signal Pattern Count
F13	Inter-residue distance	4	Position Specific Occurences
F14	DNase I derived parameter	1	
F15	Position specific nucleotides	81	
F16	Distribution of nucleotides	40	Distribution of nucleotides
F17	Segmented Distribution of nucleotides	20	

2.3. Feature subspace ensemble

The main idea of the algorithm of iPromoter-FSEn is depicted in Fig. 1. Feature subspace based ensemble classifiers have been previously used in the literature [68, 33]. They have shown superior performances over the feature selection methods and ensembles based of a selected classifiers. Random Forest and other ensemble classifiers also try to select features randomly [33] and do not utilize the full feature space. On the other hand, boosting algorithms ensembles based on the sample space. Here we, are using a simple ensemble voting classifier that uses three different classifiers. Each of these classifiers are fed three subset of features from the the total feature space. In the training phase, these three classifiers are learnt using these feature subsets. In the testing phase, each classifier provides a decision and the ensemble voting classifier is used to decide the final prediction. Note that, there a number of ensemble classifiers that were applied in the context of prediction of various attributes of biological entities in the literature as in [65, 66, 67, 64, 36, 38, 37, 46, 61, 47, 50, 58, 60].

2.4. Classification algorithms

In our ensemble classifier, we have used three classifiers: Support Vector Machine (SVM), Linear Discriminant Analysis (LDA) and Logistic Regression (LR). In this section, we provide a brief discussion on these three classifiers. Support Vector Machine [19] tries to separate the samples using a maximum margin. Often to allow non-linearly separable data, kernel functions are used. In this paper, we have used radial basis function as a kernel that can effectively allow the the input data to be extended to infinite dimension. Another effective classifier is linear discriminant analysis [55] that is a linear combination of linear functions of predictors. Logistic regression [32] is a linear classifier that learns a straight line from the data samples to separate them.

2.5. Performance evaluation

To test our hypothesis and establish an effective prediction method of iPromoter-FSEn, we have performed several experiments on the dataset. The selection of algorithm according to performance comparison often depends on the particular measures and sampling methods used [22]. There are several sampling methods used to test and validate the performance of classification methods: use of independent test sets, cross-fold validation, jack knife tests etc. Jackknife test and k-fold cross validation is used in this paper in order to compare the performances. These tests are the most reliable and robust methods in absence of a separate independent test set [15, 14]. We have used a number of metrics for performance evaluation of our method and in comparison with other methods. They are: accuracy (acc), sensitivity (S_n), Specificity (S_p), F_1 Score and Mathew's Correlation coefficient (MCC). These metrics are defined as in the following equations:

$$Accuracy\,(Acc) = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Sensitivity\,(S_n) = \frac{TP}{TP + FN} \tag{5}$$

$$Specificity\,(S_p) = \frac{TN}{TN + FP} \tag{6}$$

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{7}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{8}$$

Here, all the symbols are meant for binary classification problem. TP, TN, FP and FN respectively denotes the number of true positives, true negatives, false positives and false negatives of the prediction. All the metrics except MCC are in the range [0,1]. The best classifier is the one with value 1 for each of the metrics and 0 will indicate a worst one. MCC is in the range $[-1, +1]$, where $+1$ indicates a best classifier and -1 indicates a worst one. Often, in the case of imbalanced datasets as this one and also in the cases where prediction of the probabilistic classifiers depend on threshold value, these performance measures do not reflect the real performance. In such cases, area under precision recall curve (auPR) and area under receiver operating characteristic curve (auROC) are considered. Receiver operating curve is the plot of true positive rate against false positive rate. The value of auPR and auROC are also in the range of [0,1], where 1 means a perfect classifier and 0.5 means a random classifier.

Please note that though these metrics are all used extensively in the literature of prediction of different attributes of biological entities, we should be careful when to apply them in case of multi-label prediction problems as they are suggested in [14]. Such systems are becoming more frequent in system biology [8, 9, 73, 10, 74, 6], system medicine [11, 12] and biomedicine [59].

3. Results and discussion

In this section, we describe the experimental results and analysis. All the programs were written in Python language and sci-kit Learn [57] library. Each experiments were run 10 times and average results are reported.

The first set of experiments were done to show the effectiveness of the individual feature groups. As shown in Table 2, we have 17 different feature groups extracted for iPromoter-FSEn. We have trained using three different classifiers using each of these different groups. Three classifiers are: Support Vector Machine (SVM), Linear Discriminant Analysis (LDA) and Logistic Regression (LR). We have found that the best performing feature group is feature group 11 which is gapped tri-nucleotide count (X_XX). This group have the highest accuracy of 84.48% in cross validation using support vector machines. For logistic regression too this feature is the best performing one. However, in the case of LDA, the best performing feature is feature 16 with 81.49% accuracy. This is the nucleotide distribution feature. The worst

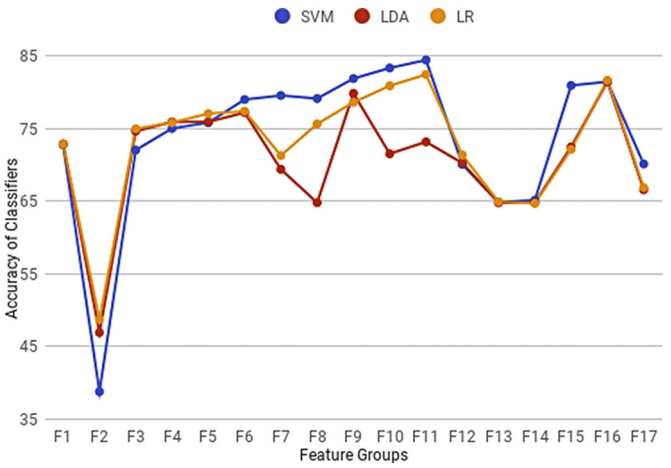

Fig. 2. Plot of accuracies achieved by different classifiers on different feature groups.

Table 3
Comparison of performance of ensemble classifier with single weak classifiers.

Method	S_n	S_p	Acc	MCC	auROC	auPR	F_1
SVM	73.31%	90.91%	84.76%	0.6589	0.9196	0.8649	0.7698
LDA	71.82%	87.855%	82.26%	0.6050	0.8904	0.8149	0.7389
Logistic Regression	86.72%	82.69%	84.10%	0.6715	0.9219	0.8763	0.7936
iPromoter-FSEn	76.69%	91.49%	86.32%	0.6950	0.9319	0.8879	0.7966

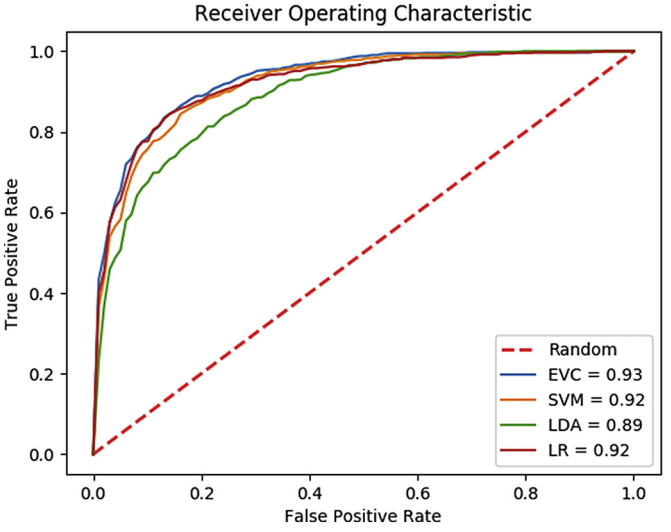

Fig. 3. Receiver Operating Characteristic (ROC) curve of the ensemble voting classifier (EVC) with comparison to single classifiers.

Table 4
Comparison of the performance of iPromoter-FSEn with that of other state-of-the-art methods.

Method	S_n	S_p	Acc	MCC	auROC	auPR	F_1
iPro70-PseZNC (All features)	74.63%	79.50%	77.81%	0.5274	–	–	–
iPro70-PseZNC (Optimized)	80.30%	86.79%	84.54%	0.6631	0.9088	–	–
Z-curve	74.6%	79.5%	77.8%	0.5270	0.848		
iPromoter-FSEn	76.69%	91.49%	86.32%	0.6950	0.9319	0.8879	0.7966
iPromoter-FSEn (jack knife)	76.52%	91.43%	86.27%	0.6924	0.9337	0.8870	0.7941

performing feature group was feature group 2 or nucleotide statistics. This feature group achieve only 38.81% accuracy using SVM. Performance in other classifiers were slightly better. A plot of accuracies achieved by three algorithms for different feature groups are given in Fig. 2.

In the feature analysis part with different classifiers, we have also used other measures like auROC, auPR, Sp, Sn, MCC and F1. Performance in auROC also follows the trend in accuracy. For details of these experiments, please refer to supplementary material. After the experiments we used an ensemble voting classifier with voting weights set to 0.42,0.20 and 0.38 for SVM, LDA and LR respectively based on their performances normalized on the dataset. After that, we used the ensemble voting classifier on this ensemble based on feature subspacing. We divided the feature into three groups. SVM used features 1–11, LDA used feature 9–17 and LR used features 12–17. Thus we have a overlapped feature subspaces for each of these classifiers.

In Table 3, we present a comparison of results achieved by the ensemble classifier used in iPromoter-FSEn with that of single classifiers. For this experiments, we have trained each of the single classifier with full set of features. In Table 3, best results among all classifiers are shown in bold faced fonts. We could notice that the ensemble classifier of iPromoter-FSEn outperforms all the single classifiers in all performance metrics except sensitivity. In the case of sensitivity, logistic regression is performing best. Note that, the accuracy here achieve by the ensemble is better than the best performing single classifier SVM by 1.56%. Note that the better values achieved in terms of auROC and auPR also indicates the effectiveness even though the dataset is slightly imbalanced. An analysis of the performances of the classifiers in terms of receiver operating characteristic curve is shown in Fig. 3.

We have also compared the performance of our algorithm with two other state-of-the-art algorithms. We have compared iPromoter-FSEn with: iPro70-PseZNC [42] and *Z*-Curve [71] method. The cross validation results achieved by all these algorithms and their variations are reported in Table 4. Note that, we have reported these results from that reported in [42]. Here too, bold faced values indicates the best results achieved by any algorithm. There are a few blank spaces in the table since F1, auPR and auROC were not reported in the literature by these methods. However, we believe that these methods are very important and specially in the case of promoter sequences where the dataset is slightly imbalanced. From the results reported in this table, we can clearly conclude that iPromoter-FSEn achieved higher accuracy, MCC, auROC and specificity compared to other methods. iPromoter-FSEn is second best in terms of sensitivity. Please note that, our features are very effective and does not require any feature selection which is required by iPro70-PseZNC. Our features with logistic regression however produces similar accurate results with better sensitivity as reported in Table 3.

Please note that all the results reported in Table 4 are average of 10 runs. Here each run on the data set was done as a *k*-fold cross fold validation with value of *k* set to 10. We have confirmed the significance of improvement in terms of MCC, accuracy and specificity of our method, iPromoter-FSEn with iPro70-PseZNC. Also note that, the results achieved by iPro70-PseZNC were after feature optimization which done within cross-folds risks overfit of data. In order to make sure our method is not overfitting the data, we also performed Jack Knife tests. These results are more robust and reported along with the cross-validation results in Table 4.

We have performed Wilcoxon Signed ranked test which is a non-parametric statistical test and that confirmed the significance of improvement. We have used the values in each individual runs within the cross-fold of the algorithms to compare between them. The accuracy of the ensemble classifier was tested with that of the individual weak classifiers. The ensemble classifier had *p*-value in each statistical test as $p = 0.00512 \leq 0.05$. The significance level was set to 0.05. Overall, it indicates the value is significant and thus rejects the null hypothesis and establishes the significance in improvement achieved by the ensemble classifier.

3.1. Web application

As pointed out in [14] and demonstrated in a series of recent publications [76, 63, 62, 35], user-friendly and publicly accessible web-servers represent the current direction for developing practically more useful prediction methods and computational tools. With a similar vision, we have implemented an web application based on the models learnt in this paper for iPromoter-FSEn. Our web application is redily available for use at: http://ipromoterfsen.pythonanywhere.com/server. The website contains a guideline for users with a user friendly interface.

4. Conclusion

In this paper, we have proposed iPromoter-FSEn a predictor for identification of σ^{70} promoter sequences. iPromoter-FSEn uses a large number of sequences based features and divides them into subspaces in order to be trained by an ensemble of three different classifiers. On standard benchmark dataset, iPromoter-FSEn significantly outperforms state-of-the-art methods. We believe our method has got potential for exploration and in future we wish to develop a web application and a database of promoter sequences and also extend the work for detection and prediction of other promoter sequences.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2018.07.011.

References

[1] Stéphane Audic, Jean-Michel Claverie, Detection of eukaryotic promoters using markov transition matrices, Comput. Chem. 21 (4) (1997) 223–227.
[3] Wei Chen, Tian-Yu Lei, Dian-Chuan Jin, Hao Lin, Kuo-Chen Chou, Pseknc: a flexible web server for generating pseudo k-tuple nucleotide composition, Anal. Biochem. 456 (2014) 53–60.
[4] Wei Chen, Hao Lin, Kuo-Chen Chou, Pseudo nucleotide composition or pseknc: an effective formulation for analyzing genomic sequences, Mol. BioSyst. 11 (10) (2015) 2620–2634.
[5] Xiang Cheng, Xuan Xiao, Kuo-Chen Chou, pLoc-mEuk: predict subcellular localization of multi-label eukaryotic proteins by extracting the key go information into general pseaac, Genomics 110 (1) (January 2018) 50–58.
[6] Xiang Cheng, Xuan Xiao, Kuo-Chen Chou, pLoc-mGneg: predict subcellular localization of gram-negative bacterial proteins by deep gene ontology learning via general pseaac, Genomics 110 (4) (July 2018) 231–239.
[7] Xiang Cheng, Xuan Xiao, Kuo-Chen Chou, pLoc-mHum: predict subcellular localization of multi-location human proteins via general pseaac to winnow out the crucial go information, Bioinformatics (2017) 1–9.
[8] Xiang Cheng, Xuan Xiao, Kuo-Chen Chou, pLoc-mPlant: predict subcellular

localization of multi-location plant proteins by incorporating the optimal go information into general pseaac, Mol. BioSyst. 13 (9) (2017) 1722–1727.
[9] Xiang Cheng, Xuan Xiao, Kuo-Chen Chou, pLoc-mVirus: predict subcellular localization of multi-location virus proteins via incorporating the optimal go information into general pseaac, Gene 628 (2017) 315–321.
[10] Xiang Cheng, Shu-Guang Zhao, Wei-Zhong Lin, Xuan Xiao, Kuo-Chen Chou, plocmAnimal: predict subcellular localization of animal proteins with both single and multiple sites, Bioinformatics 33 (22) (2017) 3524–3531.
[11] Xiang Cheng, Shu-Guang Zhao, Xuan Xiao, Kuo-Chen Chou, iATC-mISF: a multilabel classifier for predicting the classes of anatomical therapeutic chemicals, Bioinformatics 33 (3) (2016) 341–346.
[12] Xiang Cheng, Shu-Guang Zhao, Xuan Xiao, Kuo-Chen Chou, iATC-mHyb: a hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals, Oncotarget 8 (35) (2017) 58494.
[13] Kuo-Chen Chou, Some remarks on protein attribute prediction and pseudo amino acid composition, J. Theor. Biol. 273 (1) (2011) 236–247.
[14] Kuo-Chen Chou, Some remarks on protein attribute prediction and pseudo amino acid composition, J. Theor. Biol. 273 (1) (2011) 236–247.
[15] Kuo-Chen Chou, Some remarks on predicting multi-label attributes in molecular biosystems, Mol. BioSyst. 9 (6) (2013) 1092–1100.
[16] Kuo-Chen Chou, Impacts of bioinformatics to medicinal chemistry, Med. Chem. 11 (3) (2015) 218–234.
[17] Kuo-Chen Chou, An unprecedented revolution in medicinal chemistry driven by the progress of biological science, Curr. Top. Med. Chem. 17 (21) (2017) 2337–2358.
[18] Wikimedia Commons. Wikimedia Commons,The Free Media Repository, 2009. accessed 24-April-2018.(Online; File:pdb 2h27 ebi.jpg)
[19] Corinna Cortes, Vladimir Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[20] Abdollah Dehzangi, Rhys Heffernan, Alok Sharma, James Lyons, Kuldip Paliwal, Abdul Sattar, Gram-positive and gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into choùs general pseaac, J. Theor. Biol. 364 (2015) 284–294.
[21] Borries Demeler, Guangwen Zhou, Neural network optimization for *E. coli* promoter prediction, Nucleic Acids Res. 19 (7) (1991) 1593–1599.
[22] Janez Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (Jan) (2006) 1–30.
[23] Scheila de Avila e Silva, Franciele Forte, Ivaine T.S. Sartor, Tahila Andrighetti, Günther J.L. Gerhardt, Ana Paula Longaray Delamare, Sergio Echeverrigaray, DNA duplex stability as discriminative characteristic for *Escherichia coli* σ(54)-and σ(28)-dependent promoter sequences, Biologicals 42 (1) (2014) 22–28.
[24] Pengmian Feng, Hui Yang, Hui Ding, Hao Lin, Wei Chen, Kuo-Chen Chou, iDNA6mA-PseKNC: identifying dna n6-methyladenosine sites by incorporating nucleotide physicochemical properties into pseknc, Genomics (2018 Jan 31), https://doi.org/10.1016/j.ygeno.2018.01.005 (pii: S0888-7543(18)30009-0).
[25] J.W. Fickett, A.G. Hatzigeorgiou, Eukaryotic promoter recognition, Genome Res. 7 (9) (1997) 861–878.
[26] Socorro Gama-Castro, Heladia Salgado, Alberto Santos-Zavaleta, Daniela Ledezma-Tejeida, Luis Muñiz-Rascado, Jair Santiago García-Sotelo, Kevin Alquicira-Hernández, Irma Martínez-Flores, Lucia pannier, Jaime Abraham 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.
[27] Paul A. Ginno, Yoong Wearn Lim, Paul L. Lott, Ian Korf, Frédéric Chédin, Gc skew at the 5 and 3 ends of human genes links r-loop formation to epigenetic regulation and transcription termination, Genome Res. 23 (10) (2013) 1590–1600.
[28] J.J. Gordon, M.W. Towsey, J.M. Hogan, S.A. Mathews, P. Timms, Improved prediction of bacterial transcription start sites, Bioinformatics 22 (2) (2005) 142–148.
[29] Leo Gordon, Alexey Ya Chervonenkis, Alex J. Gammerman, Ilham A. Shahmuradov, Victor V. Solovyev, Sequence alignment kernel for recognition of promoter regions, Bioinformatics 19 (15) (2003) 1964–1971.
[30] Brian Grech, Stefan Maetschke, Sarah Mathews, Peter Timms, Genome-wide analysis of chlamydiae for promoters that phylogenetically footprint, Res. Microbiol. 158 (8–9) (2007) 685–693.
[31] Tanja M. Gruber, Carol A. Gross, Multiple sigma subunits and the partitioning of bacterial transcription space, Annu. Rev. Microbiol. 57 (1) (2003) 441–466.
[32] David W. Hosmer Jr., Stanley Lemeshow, Rodney X. Sturdivant, Applied Logistic Regression, vol 398, John Wiley & Sons, 2013.
[33] Weimin Huang, Yongzhong Yang, Zhiping Lin, Guang-Bin Huang, Jiayin Zhou, Yuping Duan, Wei Xiong, Random feature subspace ensemble based extreme learning machine for liver tumor detection and segmentation, Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE, IEEE, 2014, pp. 4675–4678.
[34] Araceli M. Huerta, Julio Collado-Vides, Sigma70 promoters in escherichia coli: specific transcription in dense regions of overlapping promoter-like signals, J. Mol. Biol. 333 (2) (2003) 261–278.
[35] Md Mofijul Islam, Sanjay Saha, Md Mahmudur Rahman, Swakkhar Shatabda, Dewan Md Farid, Abdollah Dehzangi, iProtGly-Ss: identifying protein glycation sites using sequence and structure based features, Proteins 86 (7) (July 2018) 777–789.
[36] Jianhua Jia, Zi Liu, Xuan Xiao, Bingxiang Liu, Kuo-Chen Chou, 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 (2015) 47–56.
[37] Jianhua Jia, Zi Liu, Xuan Xiao, Bingxiang Liu, Kuo-Chen Chou, iPPBS-Opt: a sequence-based ensemble classifier for identifying protein-protein binding sites by optimizing imbalanced training datasets, Molecules 21 (1) (2016) 95.
[38] Jianhua Jia, Zi Liu, Xuan Xiao, Bingxiang Liu, Kuo-Chen Chou, pSuc-Lys: predict lysine succinylation sites in proteins with pseaac and ensemble random forest approach, J. Theor. Biol. 394 (2016) 223–230.
[39] Yaser Daanial Khan, Nouman Rasool, Waqar Hussain, Sher Afzal Khan, Kuo-Chen Chou, iPhosT-PseAAC: identify phosphothreonine sites by incorporating sequence statistical moments into pseaac, Anal. Biochem. 550 (2018) 109–116.
[40] Qian-Zhong Li, Hao Lin, The recognition and prediction of < mi > σ < /mi > 70 promoters in escherichia coli k-12, J. Theor. Biol. 242 (1) (2006) 135–141.
[41] Zhi-Yong Liang, Hong-Yan Lai, Huan Yang, Chang-Jian Zhang, Hui Yang, Huan-Huan Wei, Xin-Xin Chen, Ya-Wei Zhao, Su Zhen-Dong, Wen-Chao Li, et al., Pro54DB: a database for experimentally verified sigma-54 promoters, Bioinformatics 33 (3) (2017) 467–469.
[42] H. Lin, Z.Y. Liang, H. Tang, W. Chen, Identifying sigma70 promoters with novel pseudo nucleotide composition, IEEE/ACM Trans. Comput. Biol. Bioinform. 10 (2017).
[43] Hao Lin, En-Ze Deng, Hui Ding, Wei Chen, Kuo-Chen 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.
[44] Sheng-Xiang Lin, Jacques Lapointe, Theoretical and experimental biology in one—a symposium in honour of professor kuo-chen choùs 50th anniversary and professor richard giegé's 40th anniversary of their scientific careers, J. Biomed. Sci. Eng. 6 (04) (2013) 435.
[45] Bin Liu, Fule Liu, Xiaolong Wang, Junjie Chen, Longyun Fang, Kuo-Chen Chou, Pse-in-one: a web server for generating various modes of pseudo components of dna, rna, and protein sequences, Nucleic Acids Res. 43 (W1) (2015) W65–W71.
[46] Bin Liu, Ren Long, Kuo-Chen 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.
[47] Bin Liu, Shanyi Wang, Ren Long, Kuo-Chen Chou, iRSpot-El: identify recombination spots with an ensemble learning approach, Bioinformatics 33 (1) (2016) 35–41.
[48] Bin Liu, Fan Weng, De-Shuang Huang, Kuo-Chen Chou, iRO-3wPseKNC: identify dna replication origins by three-window-based pseknc, Bioinformatics (2018) 1–8.
[49] Bin Liu, Hao Wu, Kuo-Chen Chou, 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 (04) (2017) 67.
[50] Bin Liu, Fan Yang, Kuo-Chen Chou, 2l-Pirna: a two-layer ensemble classifier for identifying piwi-interacting rnas and their function, Mol. Ther. Nucleic Acids 7 (2017) 267–277.
[51] Bin Liu, Fan Yang, De-Shuang Huang, Kuo-Chen Chou, iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based pseknc, Bioinformatics 34 (1) (2017) 33–40.
[52] Bin Liu, Fan Yang, De-Shuang Huang, Kuo-Chen Chou, iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based pseknc, Bioinformatics 34 (1) (2017) 33–40.
[53] A.V. Lukashin, V.V. Anshelevich, B.R. Amirikyan, A.I. Gragerov, M.D. Frank-Kamenetskii, Neural network models for promoter recognition, J. Biomol. Struct. Dyn. 6 (6) (1989) 1123–1133.
[54] Ronna R. Mallios, David M. Ojcius, David H. Ardell, An iterative strategy combining biophysical criteria and duration hidden markov models for structural predictions of chlamydia trachomatis σ^{66} promoters, BMC Bioinformatics 10 (1) (2009) 271.
[55] Sebastian Mika, Gunnar Ratsch, Jason Weston, Bernhard Scholkopf, Klaus-Robert Mullers, Fisher discriminant analysis with kernels, Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop, IEEE, 1999, pp. 41–48.
[56] Daniel G. Olson, Marybeth Maloney, Anthony A. Lanahan, Shuen Hon, Loren J. Hauser, Lee R. Lynd, Identifying promoters for gene expression in clostridium thermocellum, Metab. Eng. Commun. 2 (2015) 23–29.
[57] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al., Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830 (Oct).
[58] Wang-Ren Qiu, Shi-Yu Jiang, Bi-Qian Sun, Xuan Xiao, Xiang Cheng, Kuo-Chen Chou, iRNA-2methyl: identify RNA 2′-o-methylation sites by incorporating sequence-coupled effects into general pseknc and ensemble classifier, Med. Chem. 13 (8) (2017) 734–743.
[59] Wang-Ren Qiu, Bi-Qian Sun, Xuan Xiao, Zhao-Chun Xu, Kuo-Chen Chou, iPTM-mLys: identifying multiple lysine ptm sites and their different types, Bioinformatics 32 (20) (2016) 3116–3123.
[60] Wang-Ren Qiu, Bi-Qian Sun, Xuan Xiao, Zhao-Chun Xu, Jian-Hua Jia, Kuo-Chen Chou, iKcr-PseEns: Identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier, Genomics (2017 Nov 16), https://doi.org/10.1016/j.ygeno.2017.10.008 (pii: S0888-7543(17)30138-6).
[61] Wang-Ren Qiu, Xuan Xiao, Zhao-Chun Xu, Kuo-Chen Chou, iPhos-PseEn: identifying phosphorylation sites in proteins by fusing different pseudo components into an ensemble classifier, Oncotarget 7 (32) (2016) 51270.
[62] Farshid Rayhan, Sajid Ahmed, Swakkhar Shatabda, Dewan Md Farid, Zaynab Mousavian, Abdollah Dehzangi, M. Sohel Rahman, iDTI-ESBoost: identification of drug target interaction using evolutionary and structural features with boosting, Sci. Rep. 7 (1) (2017) 17731.
[63] Swakkhar Shatabda, Sanjay Saha, Alok Sharma, Abdollah Dehzangi, iPHLoc-ES: identification of bacteriophage protein locations using evolutionary and structural features, J. Theor. Biol. 435 (2017) 229–237.
[64] H.-B. Shen, Jie Yang, K.-C. Chou, Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction, Amino Acids 33 (1) (2007) 57–67.

[65] Hong-Bin Shen, Kuo-Chen Chou, Ensemble classifier for protein fold pattern recognition, Bioinformatics 22 (14) (2006) 1717–1722.
[66] Hong-Bin Shen, Kuo-Chen Chou, Gpos-PLoc: an ensemble classifier for predicting subcellular localization of gram-positive bacterial proteins, Protein Eng. Des. Sel. 20 (1) (2007) 39–46.
[67] Hong-Bin Shen, Kuo-Chen Chou, Hum-mploc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites, Biochem. Biophys. Res. Commun. 355 (4) (2007) 1006–1011.
[68] Hugo Silva, Hugo Gamboa, Ana Fred, One lead ecg based personal identification with feature subspace ensembles, International Workshop on Machine Learning and Data Mining in Pattern Recognition, Springer, 2007, pp. 770–783.
[69] Jiangning Song, Fuyi Li, Kazuhiro Takemoto, Gholamreza Haffari, Tatsuya Akutsu, Kuo-Chen Chou, Geoffrey I. Webb, PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural, and network features in a machine-learning framework, J. Theor. Biol. 443 (2018) 125–137.
[70] Jiangning Song, Yanan Wang, Fuyi Li, Tatsuya Akutsu, Neil D. Rawlings, Geoffrey I. Webb, Kuo-Chen Chou, iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites, Brief. Bioinform. (2018), https://doi.org/10.1093/bib/bby028.
[71] Kai Song, Recognition of prokaryotic promoters based on a novel variable-window z-curve method, Nucleic Acids Res. 40 (3) (2011) 963–971.
[72] Michael Towsey, Peter Timms, James Hogan, Sarah A. Mathews, The cross-species prediction of bacterial promoters using a support vector machine, Comput. Biol. Chem. 32 (5) (2008) 359–366.
[73] Xuan Xiao, Xiang Cheng, Genqiang Chen, Qi Mao, Kuo-Chen Chou, pLoc-mGpos: predict subcellular localization of gram-positive bacterial proteins by quasi-balancing training dataset and pseaac, Genomics (2018 May 26), https://doi.org/10.1016/j.ygeno.2018.05.017 S0888-7543(18)30260-X.
[74] Xuan Xiao, Xiang Cheng, Su Shengchao, Qi Mao, Kuo-Chen Chou, pLoc-mGpos: incorporate key gene ontology information into general pseaac for predicting subcellular localization of gram-positive bacterial proteins, Nat. Sci. 9 (09) (2017) 330.
[75] Hui Yang, Wang-Ren Qiu, Guoqing Liu, Feng-Biao Guo, H. Lin, iRSpot-Pse6NC: identifying recombination spots in *Saccharomyces cerevisiae* by incorporating hexamer composition into general pseknc, Int. J. Biol. Sci. 14 (8) (2018) 883–891, https://doi.org/10.7150/ijbs.24616.
[76] Rianon Zaman, Shahana Yasmin Chowdhury, Mahmood A. Rashid, Alok Sharma, Abdollah Dehzangi, Swakkhar Shatabda, Hmmbinder: DNA-binding protein prediction using hmm profile based features, Biomed. Res. Int. 2017 (2017).