Identification of Bacterial Sigma 70 Promoter Sequences Using Feature Subspace Based Ensemble Classifier

Usma Aktar
ID: 012 171 010
Department of Computer Science and Engineering
United International University

A thesis submitted for the degree of
MSc in Computer Science & Engineering

August 2018

Certificate

This thesis titled "Identification of Bacterial Sigma 70 Promoter Sequences Using Feature Subspace Based Ensemble Classifier" submitted by Usma Aktar, Student ID: 012 171 010, has been accepted as Satisfactory in fulfillment of the requirement for the degree of Master of Science in Computer Science and Engineering on 23th September 2018.

Board of Examiners

1._____ Supervisor

Dr. Swakkhar Shatabda
Associate Professor & Undergraduate Program Coordinator,
Department of Computer Science and Engineering,
United International University, Dhaka–1212, Bangladesh

2._____ Head Examiner

Prof. Dr. Md. Abul Kashem Mia
Professor & Dean SoSE,
Department of Computer Science and Engineering,
United International University, Dhaka–1212, Bangladesh

3._____ Examiner-I

Dr. Dewan Md. Farid
Associate Professor,
Department of Computer Science and Engineering,
United International University, Dhaka–1212, Bangladesh

4._____ Examiner-II

Ms. Rubaiya Rahtin Khan

Assistant Professor,

Department of Computer Science and Engineering,

United International University, Dhaka–1212, Bangladesh

5._____ Ex-Officio

Prof. Dr. Mohammad Nurul Huda

Professor & Coordinator - MSCSE,

Department of Computer Science and Engineering,

United International University, Dhaka–1212, Bangladesh

Declaration

This is to certify that the work entitled Identification of Bacterial Sigma 70 Promoter Sequences Using Feature Subspace Based Ensemble Classifier" is the outcome of the research carried out by me under the supervision of Dr. Swakkhar Shatabda, Associate Professor & Undergraduate Program Coordinator, Department of Computer Science and Engineering, United International University, Dhaka-1212, Bangladesh.

Usma Aktar
Student ID: 012 171 010,
MSCSE student, Department of Computer Science & Engineering
United International University, Dhaka–1212, Bangladesh

In my capacity as supervisor of the candidates thesis, I certify that the above statements are true to the best of my knowledge.

Dr. Swakkhar Shatabda
Associate Professor & Undergraduate Program Coordinator,
Department of Computer Science and Engineering,
United International University, Dhaka–1212, Bangladesh

Abstract

Sigma promoter sequences in bacterial genomes are important due to their role in transcription initiation. Sigma70 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`.

Published Papers

Work relating to the research presented in this thesis has been published/submitted by the author in the following peer-reviewed journals and conferences:

1. Md. Siddiqur Rahman, Usma Aktar, Md. Rafsan Jani, and Swakkhar Shatabda, "iPromoter-FSEn: Identification of bacterial σ^{70} promoter sequences using feature subspace based ensemble classifier," Genomics, 2018. DOI: `https://doi.org/10.1016/j.ygeno.2018.07.011`.
   [**Accepted**]
2. Md. Siddiqur Rahman, Usma Aktar, Md. Rafsan Jani, and Swakkhar Shatabda, "Identifying Sigma70 Promoters Using Multiple Windowing and Minimal Features."Molecular Genetics and Genomics, 2018. DOI: `https://doi.org/10.1007/s00438-018-1487-5`.
   [**Accepted**]

Acknowledgements

First of all, I want to thanks almighty Allah who keeps blessings on me to complete my research work. Then, I would like to express my deepest gratitudes to my supervisor Dr. Swakkhar Shatabda for his patience, continuous support and guidance in my thesis. I am very thankful to him for supporting and encouraging me during my thesis work. There is no limit to my joy in expressing my cordial gratitude to my friend Md. Siddiqur Rahman. His keen interest and encouragement were a great help throughout the research work. Also, I humbly extend my thanks to all concerned persons who co-operated with me in this regards. Finally, many thanks to my parents, sister, and brother for their unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. Without their encouragement and sacrifices, I would have not been able to carry out my work. Thank you all.

Contents

List of Figures

List of Tables

List of Algorithms

List of Abbreviations

DNA	**D**eoxyribo**N**ucleic **A**cid
RNA	**R**ibo**N**ucleic **A**cid
E. coli	***E**scherichia **c**oli*
TSS	**T**ranscript **S**tart **S**ite
PseAAC	**P**seudo **A**mino **A**cid **C**omposition
PSSM	**P**ositon **S**pecific **S**coring **M**atrix
PseKNC	**P**seudo **K**-tuple **N**ucleotide **Composition**
A	**A**denine
C	**C**ytosine
G	**G**uanine
T	**T**hymine
DNase	**D**eoxyribo**N**uclease
SVM	**S**upport **V**ector **M**achine
LR	**L**ogistic **R**egression
LDA	**L**inear **D**iscriminant **A**nalysis
EVC	**E**nsemble **V**ote **C**lassifier
Acc	**A**ccuracy
S_n	**Sn**sitivity
S_p	**Sp**ecificity

auPR	**A**rea **U**nder **P**recision-**R**ecall Curve
auROC	**A**rea **U**nder **R**eceiver **O**perating **C**haracteristic **C**urve
MCC	**M**athews **C**orrelation **C**oefficient
TP	**T**rue **P**ositive
TN	**T**rue **N**egative
FP	**F**alse **P**ositive
FN	**F**alse **N**egative
ROC	**R**eceiver **O**perating **C**haracteristic

Chapter 1

Introduction

DNA (Deoxyribonucleic Acid) transcription is a very important event and it is a part of the Central Dogma of biology. DNA has all the information that needs to be coded. But from DNA itself is not working because the data can actually work which should be moved to some form. It controls our entire cell by sitting inside the nucleus. In addition, it is a complex molecule that contains genetic information to build and maintain an organism. This information turns into an action when DNA transcription happens. Transcription is the process by which the information in DNA is copied into messenger RNA(mRNA) for protein production.

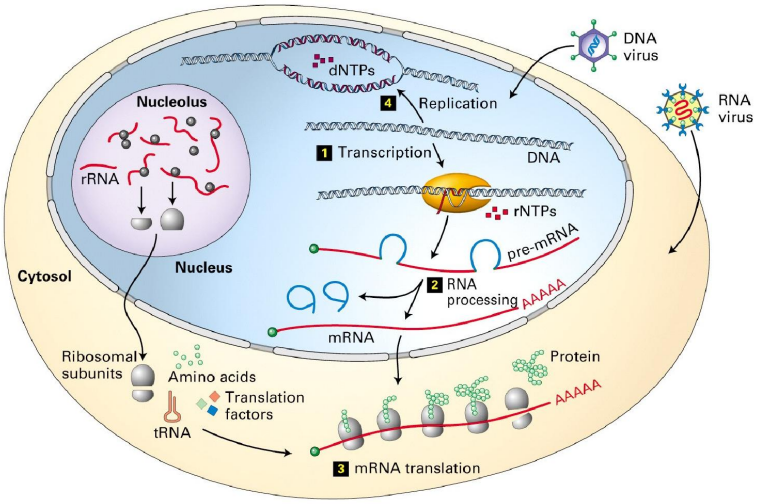

Figure 1.1: Cell biology, Charles Darwin, DNA, evolution, genetics, Gregor Mendel, protein. Image Source: [1]

Now, what DNA actually does? It provides messengers which go there and follow the task that the DNA has to do. And those messengers are the RNA. Therefore, RNA contains the information those are coded in the DNA. After that, RNA moves from the nucleus into the cytosol. Then, the RNA will be translated into protein products. Those proteins are the functional unit and they will do all the functions according to the law of DNA whatever is written inside the DNA. In Figure 1.1 is the overview of DNA transcription. In DNA, Promoters are binding sites for RNA polymerase. Promoters are the sequence of several nucleotides, which are essential for the initialization and regulation of gene transcription. Several key factors are involved in the process of gene transcription. In bacteria, sigma (σ) factor is a subunit of RNA polymerase and one of the most contributor for recognizing and binding promoter sequence during gene transcription (Figure 1.2).

The most important step in the process of DNA transcription is determining where a gene starts. Gene's start at sequences known as promoters. It is the enzyme RNA polymerase binds to the DNA chain to initiate transcription. In bacteria, RNA polymerase subunit called sigma (σ) factor helps promoter to recognize and bind (Figure 1.2) during gene transcription. There are seven such σ factor in bacterial DNA sequences. In this work, we will study only for identifying σ^{70} promoter. Most of the gene transcription in growing cells, σ^{70} factor is called 'primary' or 'housekeeping' sigma factor [7, 8].

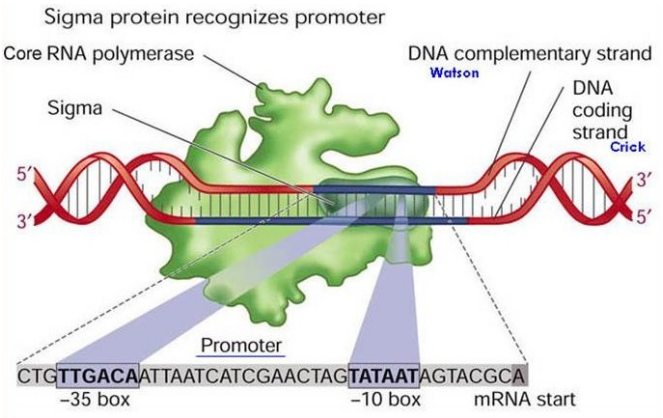

Figure 1.2: Recognization of promoter from DNA sequence by Sigma factor. Image Source: [2]

Although a vast of information provide by sequenced prokaryotic genomes, transcriptional regulatory networks are still poorly understood using the available genomic

information. Prediction of promoters correctly is more difficult. Computational methods for promoter sequence prediction gained much popularity due to cost expensive and time-consuming nature of molecular techniques [9]. There are many exciting works on identifying σ^{70} promoters, our idea in this work is to identify σ^{70} promoters using features subspace and also ensemble voting on multiple classifiers.

1.1 Motivation

Most of the previous work has done with a high performance to identify promoter sequence. As we have seen, most of them had to optimize the features by selecting feature using the different procedure such as a machine learning algorithm, $F1$ score or based on their predictive scores. The feature selection strategy typically chooses a small number of useful features from a large numbers of unnecessary or irrelevant features. This is broadly used to achieve the most important and short sized subsets of total features [10]. In the new era of genome sequence data sometime over-fitting may be happened due to the feature selection. Therefore, the main hypothesis in this work that motivates to further study on it is to use our all generated feature without any features selection. After that, we split our features into multiple feature groups. In addition, we send them into multiple popular machine learning classifier, three classifiers in this case, for the individual prediction. Finally, we vote on all of the classifier's decisions. In parallel, the main focusing part of this work is to solve a biological problem using computational methods as we have available data from the previous wet-experiment. There are mainly two reasons that make computational method popular. One is that the computational approach is faster than wet-experimental approach. Another positive site of computational approach is cost-effective.

However, sigma promoter sequences in bacterial genomes are important due to their role in transcription initiation. Sigma70 is one of the most important and crucial sigma factor. So, our results would be significantly helpful for many researchers in those fields especially, to genetic and cell Biologists. We expect our work will speed up the research and works in those areas and contribute to the advance of technology.

1.2 Objectives of the Thesis

Our objective mainly focusing on to develop a method for identifying σ^{70} promoter using feature subspaces for the multiple classifiers and ensemble voting on their decisions. For the performance evolution, we will perform a comparison test over the state-of-the-art σ^{70} promoter sequence predictors. Finally, we have also made our method readily available for use as a web application from: `http://ipromoterfsen.pythonanywhere.com/server`.

1.3 Thesis Contributions

In this thesis, we propose iPromoter-FSEn, a prediction model for identification of sigma70 promoter sequences. Our proposed method is based on a feature subspace based ensemble classifier. A large set 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 show that iPromoter-FSEn significantly improves over the state-of-the art sigma70 promoter sequence predictors. The accuracy and area under receiver operating characteristic curve of iPromoter-FSEn is 86.32% and 0.9319 respectively

1.4 Organization of the Thesis

The thesis is organised as follows:

Chapter 2 provides descriptions of various methods information those are used in the literature to identify σ^{70} promoter sequences. Most of the techniques proposed an interesting idea for identification of σ^{70} promoter sequences. This literature review description will help to represent the difference between previous methods to our method.

Chapter 3 presents a short description of our proposed method. After that, provide the details of the materials and methods of this method. Next, we define the idea of feature subspace ensemble and explain the classification algorithm. Finally,

we discuss performance evaluation to evaluate the performance of our proposed method.

Chapter 4 discusses experimental analysis and result. Comparison our result with another three state-of-the-art to show the performance of our method. We also describe a non-parametric statistical test. Finally, present our web application information.

Chapter 5 presents the conclusions and the future works.

Chapter 2

Related Work

In this chapter, we give descriptions of various methods information those are used in the literature to identify σ^{70} promoter sequences. Most of the techniques proposed an interesting idea for identification of σ^{70} promoter sequences. This literature review description will help to represent difference between previous methods to our method.

2.1 Literature Review

Computational methods for promoter sequence prediction gained much popularity due to cost expensive and time consuming nature of molecular techniques [9]. Several in silico computational methods are being used in the literature. They include position weight matrices [11], phylogenetic footprinting [12], machine learning methods [9], 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 [13, 14], Artificial Neural Network [15–17], Markov Models [18], Hidden Markov Models [19], Random Forest [20] for promoter sequence identification.

One of the earliest methods for predicting promoter sequence was proposed in [15]. They used 80 known promoter sequences to train a neural network. They used 80 known bacterial and phage promoter sequences combined with different numbers of random sequences. A pseudo-random number generating program was used to generate random sequences with equal composition A(Adenine), C(Cytosine), G(Guanine), T(Thymine). As a test set to evaluate their method, they used 30 plasmid and transposon promoter sequences with 1500 random sequences. They got 100% accuracies on pormoters and

98.4% on the random sequences with optimal parameters. However, these methods could have been very much affected by the choice of data.

A prediction method of σ^{70} promoter sequences was proposed in [21] using a sequence alignment based kernel. Several recognition methods they used for train and test dataset. They performed their experiments on 669 experimentally validated σ^{70} promoter sequences with known transcription start points of *Escherichia coli* as a positive dataset and two negative datasets of 709 examples each, taken from coding and non-coding regions of the same genome. This method used a function reflecting the quantitative measure of match between two sequences. The results show that their method performed well and achieved 16.5% average error rate on positive & coding negative data and 18.6% average error rate on positive & non-coding negative data.

Another analysis on σ^{70} promoter sequences were done in [22] where they used position correlation scoring matrix (PCSM) algorithm. The algorithm showed that the conservative hexamers segments in different sites play a key role of promoter regions performed on the 683 latest experimentally verified σ^{70} promoter sequences of *Escherichia coli* K-12.

Gordon et al. proposed a position weighted matrix based method in [23]. This method predicting the location of promoters and transcription start sites (TSSs) in *Escherichia coli*. However this method is known to result in large numbers of false positive predictions.

An experimentally verified database Pro54DB was proposed in [24] that contained 210 experimentally verified σ^{54} sequences. Lin et al. [13] used support vector machines to predict σ^{54} promoter sequences and proposed iPro54-PseKNC. iPro70-PseZNC was proposed in [14] for σ^{70} promoter sequence detection using support vector machine and pseudo nucleotide composition. Liu et al. [20] proposed iPromoter-2L 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: σ^{24} , σ^{28} , σ^{32} , σ^{38} , σ^{54} , σ^{70}.

2.2 Summary

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. To solve this problem we have proposed a method named iPromoter-FSEn to identify σ^{70} promoter sequences. In the next chapter, we will discuss about our proposed method iPromoter-FSEn in details.

Chapter 3

Proposed Method

In this chapter, first of all we give a short description of our proposed method. After that, we provide the details of the materials and methods of this method. Next, we define the idea of feature subspace ensemble and explain the classification algorithm. Finally, we discuss performance evaluation to evaluate the performance of our proposed method.

3.1 iPromoter-FSEn

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. iPromoter-FSEn divides the total number of features into three subsets, exclusive and overlapping of each 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 sub-spacing technique.

3.2 Materials and Methods

To develop a really useful sequence-based statistical predictor for a biological system as reported in a series of recent publications [20, 25–33], one should strictly observe the famous 5-step rules [34]; 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.

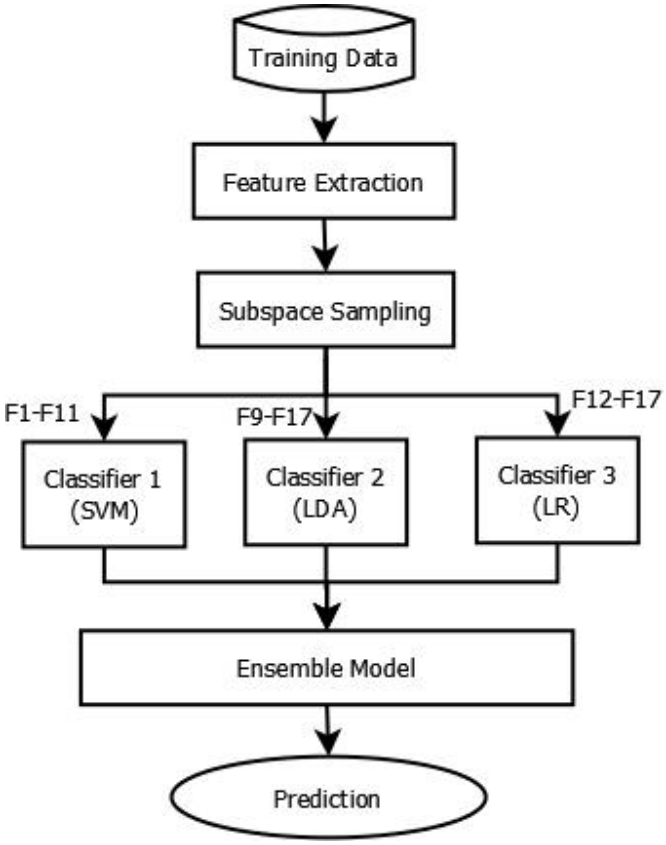

Figure 3.1: System Diagram of iPromoter-FSEn.

A system diagram of our proposed method, iPromoter-FSEn is depicted in Figure 3.1. 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.

3.2.1 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 [14]. Lin et al. constructed this dataset by taking σ^{70} promoter sequences from *E. coli* genome. The promoter sequences were downloaded from RegulonDB [35]. All these sequences are experimentally validated. For any machine learning problem of binary classification, the dataset can be expressed as following:

$$\mathbb{S} = \mathbb{S}^+ \cup \mathbb{S}^- \tag{3.1}$$

Class Label	Number of Instances	Sequence Length of Instances
Positive or Promoters	741	81
Negative or Non-Promoters	1400	81
Total Examples	2141	81

Table 3.1: Summary of the dataset taken from [14].

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 base pairs 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 3.1. Note that, 741 instances are in the positive dataset, $\mathbb{S}^+$ and 1400 sequences are in the negative dataset, $\mathbb{S}^-$ and the dataset is thus slightly imbalanced.

3.2.2 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 [36]. 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 Chous PseAAC [37] was proposed, it has been widely used in nearly all the areas of computational proteomics [38]. 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 Chous special PseAAC; while the 3rd one for those of Chous 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) [39] was developed for generating various feature vectors for DNA/RNA sequences that have proved very useful as well [40]. Particularly, recently a very powerful web-server called Pse-in-One' [41] and its updated version Pse-in-One2.0 [42] have been established that can be used to 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:

$$D = N_1, N_2, N_3, ..., N_L \tag{3.2}$$

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

$$\mathbb{F}(D) = [f_1, f_2, f_3, ..., f_n] \tag{3.3}$$

3.2.3 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 [43] along the sequence (81) and AT/GC ratio (1).

3.2.4 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 $4^2 + 4^3 + 4^4 + 4^5 + 4^6 = 5792$.

3.2.5 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 g–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 X_XX. We have used various patterns of gaps, by differing the gaps in between, g = 1,2,...,75.

3.2.6 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 [44, 45]. 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.

3.2.7 Position Specific Occurences

Among this group of features are mean inter-nucleotide occurrence (4), DNase I parameter (1) and position specific information of nucleotides (81).

3.2.8 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

Sn	Feature Name	# of Features	Feature Group
F1	Nucleotide Frequency	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	

Table 3.2: Summary of feature extracted for iPromoter-FSEn.

the segmented distribution of position specific scoring matrix (PSSM) [46] used in predicting subcellular localization of gram positive and gram negative bacterial pro-

teins, 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 3.2.

3.3 Feature Subspace Ensemble

The main idea of the algorithm of iPromoter-FSEn is depicted in Figure 3.1. Feature subspace based ensemble classifiers have been previously used in the literature [47, 48]. 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 [48] 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 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 [49–61].

3.4 Classification Algorithm

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 classifiers.

3.4.1 Support Vector Machine

Support Vector Machine [62] is one of the most popular and an exciting supervised machine learning algorithms. This algorithm is highly preferable as it produces remarkable accuracy with low computational power. It tries to separate the samples using a hyperplane with maximum margin. Often to allow non-linearly separable data, kernel functions are used. In this paper, we have tried many experiments to choose our kernel function for the support vector machine such as linear, sigmoid, radial basis,

polynomial etc. Finally, we have found the best result using radial basis function as a kernel that can effectively allow the input data to be extended to infinite dimension.

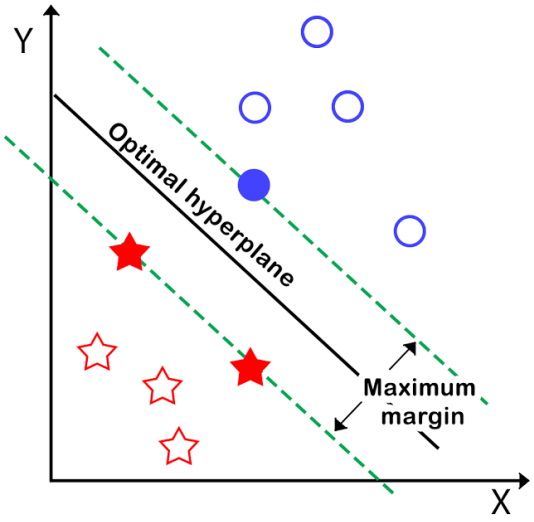

Figure 3.2: Support Vector Machine. Image Source: [3]

3.4.2 Logistic Regression

Logistic regression [63] is another supervised machine learning algorithm. It is a linear classifier that calculates the probability using a logical or sigmoid function to measure the relation between certain dependable variables and one or more distinct variables. The produced result of sigmoid function is binary or in the form of 0/1 or -1/1.

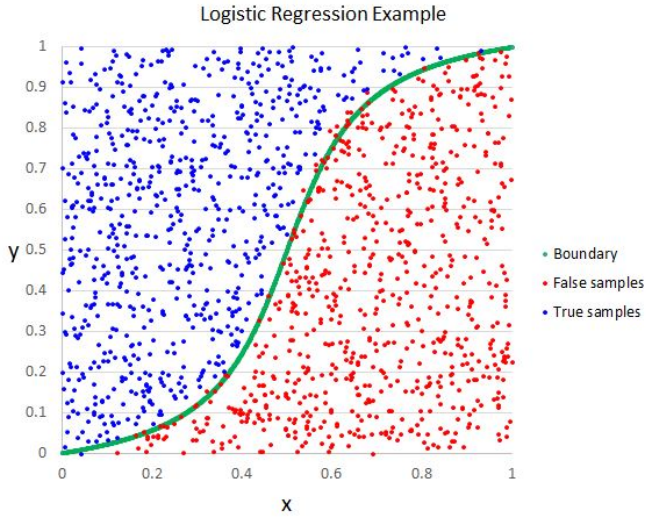

Figure 3.3: Logistic Regression. Image Source: [4]

3.4.3 Linear Discriminant Analysis

Another effective classifier is linear discriminant analysis that is a generalization of Fisher's linear discriminant [64]. This algorithm used in machine learning to find a linear combination of features that classify two or more classes of objects or events. The goal of this algorithm is to find the feature subspace that optimizes class separability. This technique used for dimensionality reduction that helps to reduce computational costs for given classification task and to avoid overfitting by minimizing the error in parameter estimation.

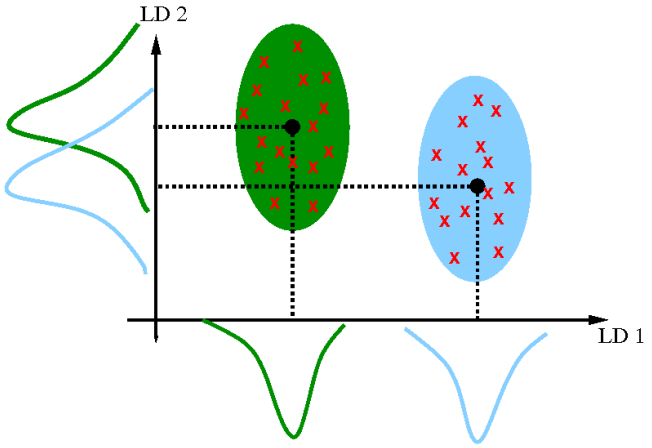

Figure 3.4: Linear Discriminant Analysis. Image Source: [5]

3.4.4 Ensemble Vote Classifier

The Ensemble classifier [49] is a meta-classifier that use multiple similar or conceptually different classifier to produce improved results. Ensemble methods usually produce more accurate solutions than a single(weak) classifier would. The three most popular methods for combining the predictions from different models are bagging, boosting and voting. We used a weighted majority voting ensemble method for final predictions based on individual weak classifiers decision. As weak classifiers, we used support vector machine (SVM), Logistic Regression(LR), and Linear Discriminant Analysis(LDA). Firstly these weak classifier takes the decision on training dataset after that ensemble vote classifier wraps those classifiers and average the predictions of the sub-classifiers to make predictions for new data.

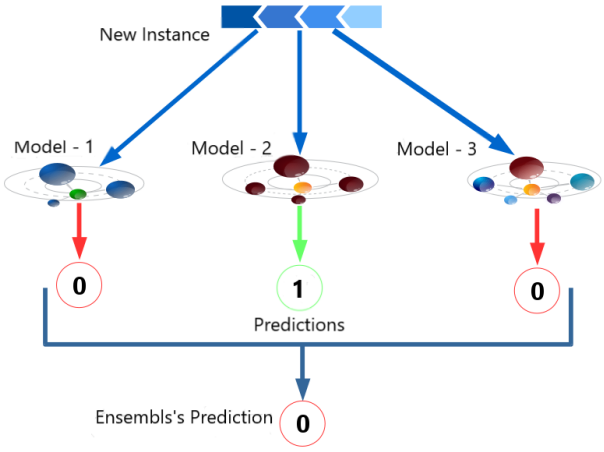

Figure 3.5: Ensemble Vote Classifier. Image Source: [6]

3.5 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 [65]. 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 [34, 66]. 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 Mathews Correlation coefficient (MCC). These metrics are defined as in the following equations:

$$Accuracy(Acc) = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.4}$$

$$Sensitivity(S_n) = \frac{TP}{TP + FN} \tag{3.5}$$

$$Specificity(S_p) = \frac{TN}{TN + FP} \tag{3.6}$$

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{3.7}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.8}$$

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.

3.6 Summary

In this chapter, we discussed about the materials and methods that follows the famous 5-step rules [34] to make our method a useful sequence-based statistical predictor for a biological system. And next chapter will discuss about the experimental results and analysis.

Chapter 4

Experimental Analysis

In this chapter, we describe the experimental results and analysis. All the programs were written in Python language and sci-kit Learn [67] library. Each experiments were run 10 times and average results are reported.

4.1 Experimental Results

The first set of experiments were done to show the effectiveness of the individual feature groups. As shown in Table 3.2, 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 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 Figure 4.1.

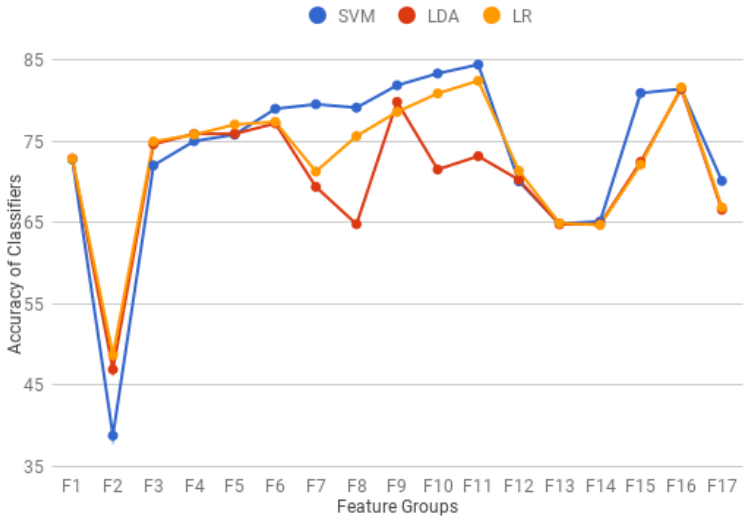

Figure 4.1: Plot of accuracies achieved by different classifiers on different feature groups.

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.

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**

Table 4.1: Comparison of performance of ensemble classifier with single weak classifiers.

In Table 4.1, 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 4.1, 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 Figure 4.2

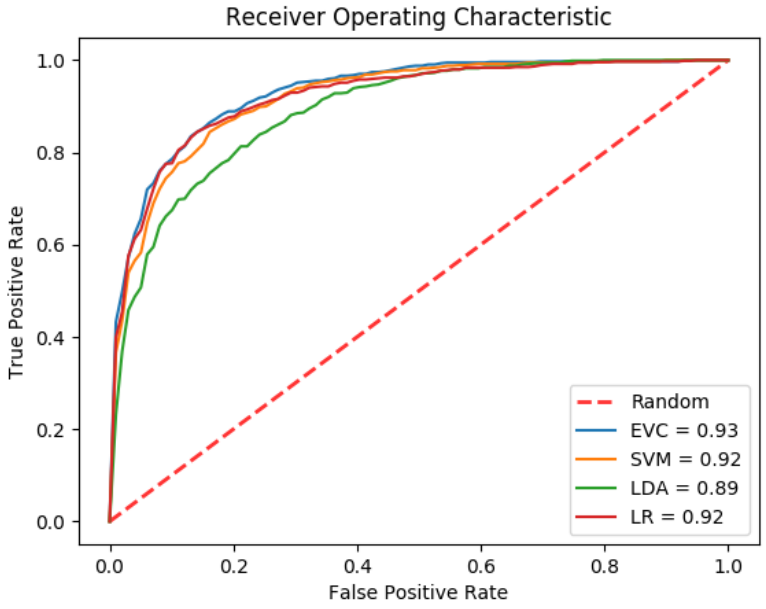

Figure 4.2: Receiver Operating Characteristic (ROC) curve of the Ensemble Vote Classifier (EVC) with comparison to single classifiers.

4.2 Comparison Results

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 [14] and Z-Curve [68] method. The cross validation results achieved by all these algorithms and their variations are reported in Table 4.2. Note that, we have reported these results from that reported in [14]. 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 4.1.

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

Table 4.2: Comparison of the performance of iPromoter-FSEn with that of other state-of-the-art methods.

4.3 Statistical Test

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.

4.4 Web Application

As pointed out in [34] and demonstrated in a series of recent publications [69–72], 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 readily available for use at:`http://ipromoterfsen.pythonanywhere.com/server`. The website contains a guideline for users with a user friendly interface.

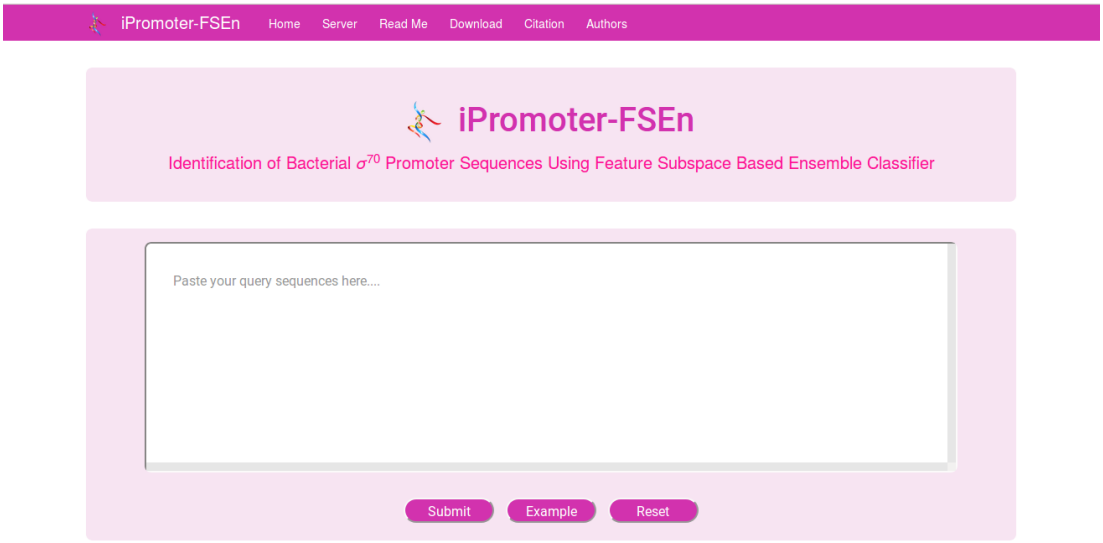

Figure 4.3: Web Application.

4.5 Summary

Please note that all the results reported in Table 4.2 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 4.2.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

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.

5.2 Future Work

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.

Bibliography

[1] darwinview.blogspot.com, "This view of life: July 2007," 2018, [Online; accessed 08-August-2018]. [Online]. Available: http://darwinview.blogspot.com/2007/07/ x, 1

[2] quantinsti.com, "Support-vector-machine," 2018, [Online; accessed 08-August-2018]. [Online]. Available: https://quantra.quantinsti.com/glossary/Support-Vector-Machine x, 16

[3] helloacm.com, "Logistic-regression-algorithm," 2018, [Online; accessed 08-August-2018]. [Online]. Available: https://helloacm.com/a-short-introduction-logistic-regression-algorithm/ x, 17

[4] www.apsl.net, "linear-discriminant-analysis," 2018, [Online; accessed 08-August-2018]. [Online]. Available: https://www.apsl.net/blog/2017/07/18/using-linear-discriminant-analysis-lda-data-explore-step-step/ x, 18

[5] www.datajango.com, "heterogeneous-ensemble-learning-hard-voting-soft-voting," 2018, [Online; accessed 08-August-2018]. [Online]. Available: http://www.datajango.com/heterogeneous-ensemble-learning-hard-voting-soft-voting/ x, 19

[6] T. M. Gruber and C. A. Gross, "Multiple sigma subunits and the partitioning of bacterial transcription space," *Annual Reviews in Microbiology*, vol. 57, no. 1, pp. 441–466, 2003. 2

[7] P. Smolen, D. A. Baxter, and J. H. Byrne, "Modeling transcriptional control in gene networksmethods, recent results, and future directions," *Bulletin of mathematical biology*, vol. 62, no. 2, pp. 247–292, 2000. 2

[8] M. Towsey, P. Timms, J. Hogan, and S. A. Mathews, "The cross-species prediction of bacterial promoters using a support vector machine," *Computational biology and chemistry*, vol. 32, no. 5, pp. 359–366, 2008. 3, 6

[9] M. Dash and H. Liu, "Feature selection for classification," *Intelligent data analysis*, vol. 1, no. 3, pp. 131–156, 1997. 3

[10] J. W. Fickett and A. G. Hatzigeorgiou, "Eukaryotic promoter recognition," *Genome research*, vol. 7, no. 9, pp. 861–878, 1997. 6

[11] B. Grech, S. Maetschke, S. Mathews, and P. Timms, "Genome-wide analysis of chlamydiae for promoters that phylogenetically footprint," *Research in microbiology*, vol. 158, no. 8-9, pp. 685–693, 2007. 6

[12] H. Lin, E.-Z. Deng, H. Ding, W. Chen, and K.-C. Chou, "ipro54-pseknc: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition," *Nucleic acids research*, vol. 42, no. 21, pp. 12 961–12 972, 2014. 6, 7

[13] H. Lin, Z.-Y. Liang, H. Tang, and W. Chen, "Identifying sigma70 promoters with novel pseudo nucleotide composition," *IEEE/ACM transactions on computational biology and bioinformatics*, 2017. 6, 7, 11, 23

[14] B. Demeler and G. Zhou, "Neural network optimization for e. coli promoter prediction," *Nucleic acids research*, vol. 19, no. 7, pp. 1593–1599, 1991. 6

[15] A. Lukashin, V. Anshelevich, B. Amirikyan, A. Gragerov, and M. Frank-Kamenetskii, "Neural network models for promoter recognition," *Journal of Biomolecular Structure and Dynamics*, vol. 6, no. 6, pp. 1123–1133, 1989.

[16] S. d. A. e Silva, F. Forte, I. T. Sartor, T. Andrighetti, G. J. Gerhardt, A. P. L. Delamare, and S. Echeverrigaray, "Dna duplex stability as discriminative characteristic for escherichia coli σ54-and σ28-dependent promoter sequences," *Biologicals*, vol. 42, no. 1, pp. 22–28, 2014. 6

[17] S. Audic and J.-M. Claverie, "Detection of eukaryotic promoters using markov transition matrices," *Computers & chemistry*, vol. 21, no. 4, pp. 223–227, 1997. 6

[18] R. R. Mallios, D. M. Ojcius, and D. H. Ardell, "An iterative strategy combining biophysical criteria and duration hidden markov models for structural predictions of chlamydia trachomatis σ 66 promoters," *BMC bioinformatics*, vol. 10, no. 1, p. 271, 2009. 6

[19] B. Liu, F. Yang, D.-S. Huang, and K.-C. Chou, "ipromoter-2l: a two-layer predictor for identifying promoters and their types by multi-window-based pseknc," *Bioinformatics*, vol. 34, no. 1, pp. 33–40, 2017. 6, 7, 10

[20] L. Gordon, A. Y. Chervonenkis, A. J. Gammerman, I. A. Shahmuradov, and V. V. Solovyev, "Sequence alignment kernel for recognition of promoter regions," *Bioinformatics*, vol. 19, no. 15, pp. 1964–1971, 2003. 7

[21] Q.-Z. Li and H. Lin, "The recognition and prediction of σ70 promoters in escherichia coli k-12," *Journal of theoretical biology*, vol. 242, no. 1, pp. 135–141, 2006. 7

[22] J. J. Gordon, M. W. Towsey, J. M. Hogan, S. A. Mathews, and P. Timms, "Improved prediction of bacterial transcription start sites," *Bioinformatics*, vol. 22, no. 2, pp. 142–148, 2005. 7

[23] Z.-Y. Liang, H.-Y. Lai, H. Yang, C.-J. Zhang, H. Yang, H.-H. Wei, X.-X. Chen, Y.-W. Zhao, Z.-D. Su, W.-C. Li *et al.*, "Pro54db: a database for experimentally verified sigma-54 promoters," *Bioinformatics*, vol. 33, no. 3, pp. 467–469, 2017. 7

[24] Y. D. Khan, N. Rasool, W. Hussain, S. A. Khan, and K.-C. Chou, "iphost-pseaac: Identify phosphothreonine sites by incorporating sequence statistical moments into pseaac," *Analytical biochemistry*, vol. 550, pp. 109–116, 2018. 10

[25] X. Cheng, X. Xiao, and K.-C. Chou, "ploc-mgneg: Predict subcellular localization of gram-negative bacterial proteins by deep gene ontology learning via general pseaac," *Genomics*, 2017.

[26] B. Liu, F. Weng, D.-S. Huang, and K.-C. Chou, "iro-3wpseknc: Identify dna replication origins by three-window-based pseknc." *Bioinformatics*, vol. 1, p. 8, 2018.

[27] X. Cheng, X. Xiao, and K.-C. Chou, "ploc-meuk: Predict subcellular localization of multi-label eukaryotic proteins by extracting the key go information into general pseaac," *Genomics*, vol. 110, no. 1, pp. 50–58, 2018.

[28] P. Feng, H. Yang, H. Ding, H. Lin, W. Chen, and K.-C. Chou, "idna6ma-pseknc: Identifying dna n6-methyladenosine sites by incorporating nucleotide physicochemical properties into pseknc," *Genomics*, 2018.

[29] W. Chen, P. Feng, H. Yang, H. Ding, H. Lin, and K.-C. Chou, "irna-3typea: Identifying three types of modification at rnas adenosine sites," *Molecular Therapy-Nucleic Acids*, vol. 11, pp. 468–474, 2018.

[30] J. Song, Y. Wang, F. Li, T. Akutsu, N. D. Rawlings, G. I. Webb, and K.-C. Chou, "iprot-sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites," *Briefings in bioinformatics*, 2018.

[31] H. Yang, W.-R. Qiu, G. Liu, F.-B. Guo, W. Chen, K.-C. Chou, and H. Lin, "irspot-pse6nc: identifying recombination spots in saccharomyces cerevisiae by incorporating hexamer composition into general pseknc," *International journal of biological sciences*, vol. 14, no. 8, p. 883, 2018.

[32] J. Song, F. Li, K. Takemoto, G. Haffari, T. Akutsu, K.-C. Chou, and G. I. Webb, "Prevail, an integrative approach for inferring catalytic residues using sequence, structural, and network features in a machine-learning framework," *Journal of theoretical biology*, vol. 443, pp. 125–137, 2018. 10

[33] K.-C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *Journal of theoretical biology*, vol. 273, no. 1, pp. 236–247, 2011. 10, 19, 20, 24

[34] S. Gama-Castro, H. Salgado, A. Santos-Zavaleta, D. Ledezma-Tejeida, L. Muñiz-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 research*, vol. 44, no. D1, pp. D133–D143, 2015. 11

[35] K.-C. Chou, "Impacts of bioinformatics to medicinal chemistry," *Medicinal chemistry*, vol. 11, no. 3, pp. 218–234, 2015. 12

[36] S.-X. Lin and J. Lapointe, "Theoretical and experimental biology in onea symposium in honour of professor kuo-chen chous 50th anniversary and professor richard giegés 40th anniversary of their scientific careers," *Journal of Biomedical Science and Engineering*, vol. 6, no. 04, p. 435, 2013. 12

[37] K.-C. Chou, "An unprecedented revolution in medicinal chemistry driven by the progress of biological science," *Current topics in medicinal chemistry*, vol. 17, no. 21, pp. 2337–2358, 2017. 12

[38] W. Chen, T.-Y. Lei, D.-C. Jin, H. Lin, and K.-C. Chou, "Pseknc: a flexible web server for generating pseudo k-tuple nucleotide composition," *Analytical biochemistry*, vol. 456, pp. 53–60, 2014. 12

[39] W. Chen, H. Lin, and K.-C. Chou, "Pseudo nucleotide composition or pseknc: an effective formulation for analyzing genomic sequences," *Molecular BioSystems*, vol. 11, no. 10, pp. 2620–2634, 2015. 12

[40] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, and K.-C. Chou, "Pse-in-one: a web server for generating various modes of pseudo components of dna, rna, and protein sequences," *Nucleic acids research*, vol. 43, no. W1, pp. W65–W71, 2015. 12

[41] B. Liu, H. Wu, and K.-C. 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," *Natural Science*, vol. 9, no. 04, p. 67, 2017. 12

[42] P. A. Ginno, Y. W. Lim, P. L. Lott, I. F. Korf, and F. Chédin, "Gc skew at the 5'and 3'ends of human genes links r-loop formation to epigenetic regulation and transcription termination," *Genome research*, pp. gr–158 436, 2013. 13

[43] A. M. Huerta and J. Collado-Vides, "Sigma70 promoters in escherichia coli: specific transcription in dense regions of overlapping promoter-like signals," *Journal of molecular biology*, vol. 333, no. 2, pp. 261–278, 2003. 13

[44] D. G. Olson, M. Maloney, A. A. Lanahan, S. Hon, L. J. Hauser, and L. R. Lynd, "Identifying promoters for gene expression in clostridium thermocellum," *Metabolic Engineering Communications*, vol. 2, pp. 23–29, 2015. 13

[45] A. Dehzangi, R. Heffernan, A. Sharma, J. Lyons, K. Paliwal, and A. Sattar, "Gram-positive and gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into chou s general pseaac," *Journal of theoretical biology*, vol. 364, pp. 284–294, 2015. 14

[46] H. Silva, H. Gamboa, and A. Fred, "One lead ecg based personal identification with feature subspace ensembles," in *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer, 2007, pp. 770–783. 15

[47] W. Huang, Y. Yang, Z. Lin, G.-B. Huang, J. Zhou, Y. Duan, and W. Xiong, "Random feature subspace ensemble based extreme learning machine for liver tumor detection and segmentation," in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*. IEEE, 2014, pp. 4675–4678. 15

[48] H.-B. Shen and K.-C. Chou, "Ensemble classifier for protein fold pattern recognition," *Bioinformatics*, vol. 22, no. 14, pp. 1717–1722, 2006. 15, 18

[49] H.-B. Shen and K.-C. Chou, "Gpos-ploc: an ensemble classifier for predicting subcellular localization of gram-positive bacterial proteins," *Protein Engineering, Design & Selection*, vol. 20, no. 1, pp. 39–46, 2007.

[50] H.-B. Shen and K.-C. Chou, "Hum-mploc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites," *Biochemical and biophysical research communications*, vol. 355, no. 4, pp. 1006–1011, 2007.

[51] H.-B. Shen, J. Yang, and K.-C. Chou, "Euk-ploc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction," *Amino Acids*, vol. 33, no. 1, pp. 57–67, 2007.

[52] J. Jia, Z. Liu, X. Xiao, B. Liu, and K.-C. Chou, "ippi-esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical

properties and wavelet transforms into pseaac," *Journal of theoretical biology*, vol. 377, pp. 47–56, 2015.

[53] J. Jia, Z. Liu, X. Xiao, B. Liu, and K.-C. Chou, "psuc-lys: predict lysine succinylation sites in proteins with pseaac and ensemble random forest approach," *Journal of theoretical biology*, vol. 394, pp. 223–230, 2016.

[54] J. Jia, Z. Liu, X. Xiao, B. Liu, and K.-C. Chou, "ippbs-opt: a sequence-based ensemble classifier for identifying protein-protein binding sites by optimizing imbalanced training datasets," *Molecules*, vol. 21, no. 1, p. 95, 2016.

[55] B. Liu, R. Long, and 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*, vol. 32, no. 16, pp. 2411–2418, 2016.

[56] W.-R. Qiu, X. Xiao, Z.-C. Xu, and K.-C. Chou, "iphos-pseen: identifying phosphorylation sites in proteins by fusing different pseudo components into an ensemble classifier," *Oncotarget*, vol. 7, no. 32, p. 51270, 2016.

[57] B. Liu, S. Wang, R. Long, and K.-C. Chou, "irspot-el: identify recombination spots with an ensemble learning approach," *Bioinformatics*, vol. 33, no. 1, pp. 35–41, 2016.

[58] B. Liu, F. Yang, and K.-C. Chou, "2l-pirna: a two-layer ensemble classifier for identifying piwi-interacting rnas and their function," *Molecular Therapy-Nucleic Acids*, vol. 7, pp. 267–277, 2017.

[59] W.-R. Qiu, S.-Y. Jiang, B.-Q. Sun, X. Xiao, X. Cheng, and K.-C. Chou, "irna-2methyl: Identify rna 2'-o-methylation sites by incorporating sequence-coupled effects into general pseknc and ensemble classifier," *Medicinal Chemistry*, vol. 13, no. 8, pp. 734–743, 2017.

[60] W.-R. Qiu, B.-Q. Sun, X. Xiao, Z.-C. Xu, J.-H. Jia, and K.-C. Chou, "ikcr-pseens: Identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier," *Genomics*, 2017. 15

[61] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995. 15

[62] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013, vol. 398. 16

[63] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Mullers, "Fisher discriminant analysis with kernels," in *Neural networks for signal processing IX, 1999. Proceedings of the 1999 IEEE signal processing society workshop.* Ieee, 1999, pp. 41–48. 17

[64] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, vol. 7, no. Jan, pp. 1–30, 2006. 19

[65] K.-C. Chou, "Some remarks on predicting multi-label attributes in molecular biosystems," *Molecular Biosystems*, vol. 9, no. 6, pp. 1092–1100, 2013. 19

[66] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011. 21

[67] K. Song, "Recognition of prokaryotic promoters based on a novel variable-window z-curve method," *Nucleic acids research*, vol. 40, no. 3, pp. 963–971, 2011. 23

[68] R. Zaman, S. Y. Chowdhury, M. A. Rashid, A. Sharma, A. Dehzangi, and S. Shatabda, "Hmmbinder: Dna-binding protein prediction using hmm profile based features," *BioMed research international*, vol. 2017, 2017. 24

[69] S. Shatabda, S. Saha, A. Sharma, and A. Dehzangi, "iphloc-es: Identification of bacteriophage protein locations using evolutionary and structural features," *Journal of theoretical biology*, vol. 435, pp. 229–237, 2017.

[70] F. Rayhan, S. Ahmed, S. Shatabda, D. M. Farid, Z. Mousavian, A. Dehzangi, and M. S. Rahman, "idti-esboost: Identification of drug target interaction using evolutionary and structural features with boosting," *Scientific reports*, vol. 7, no. 1, p. 17731, 2017.

[71] M. M. Islam, S. Saha, M. M. Rahman, S. Shatabda, D. M. Farid, and A. Dehzangi, "iprotgly-ss: identifying protein glycation sites using sequence and structure based features," *Proteins: Structure, Function, and Bioinformatics*, 2018. 24