Combining Single Nucleotide Statistics and Multiple Gene Features for Human Promoter Recognition

Wenxuan Xu
School of Computer Science and Technology & Provincial Key laboratory for Computer Information Processing, Soochow University, Suzhou 215006, Jiangsu, China
Email: rifflexiansen@qq.com

Li Zhang
School of Computer Science and Technology & Collaborative Innovation Center of Novel Software Technology and Industrialization, Soochow University, Suzhou 215006, Jiangsu, China
Email: zhangliml@suda.edu.cn

Zhao Zhang
School of Computer Science and Technology & Provincial Key laboratory for Computer Information Processing, Soochow University, Suzhou 215006, Jiangsu, China
Email: cszzhang@suda.edu.cn

Fanzhang Li
School of Computer Science and Technology & Collaborative Innovation Center of Novel Software Technology and Industrialization, Soochow University, Suzhou 215006, Jiangsu, China
Email: lfzh@suda.edu.cn

***Abstract*—The prediction and recognition of promoter in human genome play an important role in DNA sequence analysis. This paper proposes a system for human promoter recognition combining single nucleotide statistics and multiple gene features. In this system, a gene dataset can first be divided into two subsets by single nucleotide statistics. Secondly, multiple gene features are extracted for each subset, including DNA rigidity, word-based feature and CpG-island. Specially, we choose frequencies of n-mers as the word-based features by extracting the most informative and discriminative n-mers that distinguish promoter sequence regions from other DNA sequences regions using Kullback–Leibler divergence. Finally, multiple support vector machines are adopted to construct a human promoter recognition system corresponding to these features. Experimental result shows that our method has high sensitivity and specificity.**

Keywords*—CpG-island, DNA Rigidity, Human promoter recognition, Kullback-Leibler Divergence, Nucleotide Statistics, Support Vector Machines

I. Introduction

In genetics, a promoter is a region of DNA that initiates transcription of a particular gene. It contains gene transcription start sites (TSSs) and controls biological activities of genes [1]. It determines the direction, speed and accuracy of DNA transcription. The promoter recognition plays an important role in studying the regulation of human gene expression. Thus, it is a very important task that how to quickly and accurately recognize human promoter at present.

Since Fickett and Hatzigeorgiou published the first review paper on promoter recognition algorithms in 1997[2], the recognition technology of human promoter has been developed rapidly. More and more researchers use the knowledge of bioinformatics to predict and recognize the promoter with the help of computer technology, which is low cost, less time-consuming and leads to more reliable results.

One of the key problems in promoter recognition is how to extract the most informative and discriminative features to differentiate the categories of promoters from non-promoters. Signal, context and structure features are the three types of features which can be used to recognize core-promoter regions essentially. CpG-islands feature is widely used in many recognition algorithms as one of the signal features [3], [4], [5]. Because DNA sequences are always seen as the collections of documents, the statistical features based on the unit DNA words called n-mers are also used to predict and recognize promoters which belong to the context features [6]. Kullback-Leibler (KL) divergence is a meaningful statistical measure, which can measure the difference between two probability distributions. We can use the Kullback-Leibler divergence to select n-mers which can reduce the search space.

In addition to the simple application of the signal features and context features, more and more attention is paid to DNA three-dimensional structures. As an important structure feature which is derived from DNA three-dimensional structures [14], DNA flexibility has been proposed to be an effective feature for promoter recognition [15]. These structure features can provide important supplemental information for promoter recognition, which are different from signal features and context features.

Besides feature extraction, another important task is to select appropriate classifiers to differentiate categories of promoters from non-promoters based on selected features. A lot of methods in machine learning are applied to promoter recognition, such as support vector machine (SVM) [7], markov model [8], [9], relevance vector machine [10], linear and quadratic discriminant analysis [11], [12] and neural network [13]. In our system, we take SVM as the classifier for promoter recognition. SVM is a supervised classification method and it has been proved to be a good promoter recognition algorithm. SVM deals with the large number of high-dimensional and complex data much better than other statistical or machine learning methods in most promoter recognition algorithms.

There are many softwares of promoter prediction and recognition based on the above classification algorithms, such as Neural Network Promoter Prediction (NNPP) [16], PromoterExplorer [17], Dragon Promoter Finder [18], Prometheus [19], Promoter 2.0[20], and Increment of Diversity with Quadratic Discriminant analysis (IDQD) [21]. With the progress in the efficient algorithms applied in the softwares mentioned above, the performance can be improved by a secondary processing in features by autoencoders. The present development trend of promoter recognition is considering the promoter, the coding exons and the introns of genomic regions at the same time for the reason that the properties of promoter regions are considerably different from those of other genomic regions, such as exons, introns, 3'UTRs and intergenic regions.

In this paper, a system for human promoter recognition is proposed. We consider the promoter, the coding exons and the introns of DNA sequences at the same time. In this system, a gene dataset can first be divided into two subsets by using single nucleotide statistics. Second, multiple gene features are extracted for each subset, including DNA rigidity, word-based feature and CpG-island. Specially, we use Kullback–Leibler (KL) divergence to select the most informative and discriminative n-mers features to identify the promoters and non-promoters in large genomic sequences. Finally, multiple support vector machines are adopted to construct a human promoter recognition system corresponding to these features.

The contribution of this paper is to combine single nucleotide statistics (or DNA rigidity) and multiple gene features for promoter recognition. In addition, a classification system of multiple SVMs based on these features is presented. The rest of this paper is organized as follows. Section II introduces three kinds of features and a classification method based on multiple SVMs. We show experimental results in Section III and conclude this paper in Section IV.

II. Our Approach

A. Feature Extraction

Multiple gene features are extracted for each subset, including DNA rigidity, word-based feature and CpG-island.

1) DNA Rigidity

DNA three-dimensional structure features are characterized by the local angular parameters (twist, roll, and tilt) as well as the translational parameters (shift, slide, and rise). Sequence-dependent DNA rigidity is an important physical property derived from DNA three-dimensional structure [22]. The general DNA rigidity patterns in human promoters have been examined and used for computational promoter prediction [23].

We use statistical mechanics models to obtain DNA rigidity profiles. We take trinucleotide model to calculate the rigidity features of human gene sequences. Trinucleotide parameter values of the trinucleotide model are provided in [24]. We use 6-mers (6 bases long sequence) to calculate the characteristic value of each base site in sequence. 6-mers rigidity values r are calculated by adding four overlapping trinucleotide parameter values：

$$r = \sum_{i=1}^{4} t_i \tag{1}$$

where i is the position index and t_i is the rigidity parameter of each trinucleotide at position i. We calculate the 6-mer rigidity values from the starting position of the sequence. For example the 7-mer TATAAAA has the rigidity value at the first position T:

$$r_T = t_{TAT} + t_{ATA} + t_{TAA} + t_{AAA}$$

and the rigidity value at the second position A:

$$r_A = t_{ATA} + t_{TAA} + t_{AAA} + t_{AAA}$$

If the sequence is L in length, its rigidity profile is L−5 in length based on 6-mers. Therefore, we can calculate the rigidity profile vector $\mathbf{R} = \left[r_1, ..., r_{L-5}\right]^T$ for any given sequences based on the conversion table of 32 unique trinucleotide rigidity parameters [25].

Let the rigidity profile set be $X_D = \{\mathbf{f}_i^D\}_{i=1}^n$, where n is the number of genes, $\mathbf{f}_i^D = \left[\, f_{i1}^D, f_{i2}^D, \cdots, f_{i(L-5)}^D \right]^T \in \mathbb{R}^{L-5}$ is the ith gene rigidity feature vector with $f_{ij}^D = \sum_{k=j}^{i+3} t_k$, $j = 1, 2, \cdots, L-5$, and t_k is the rigidity parameter of each trinucleotide at position k.

2) Word-based features based on Kullback-Leibler Divergence

DNA sequences can be considered as the collections of documents consisting of the letter A (adenine), C (cytosine), G (guanine) and T (thymine). Each letter represents a nucleotide. n consecutive nucleotides are called an n-mer or n-word. There are 4^n n-mers. The frequency distribution of n-mers has an important biological significance.

Because of the existence of a large number of zero in frequency of n-mers, the computation of tractable search space of feature extraction will become very complex for classifiers. Thus, the computation needs to be simplified for maintaining the most information in promoters. Kullback-Leibler divergence is a meaningful statistical measure, which can measure the difference between two probability distributions. We can use the Kullback-Leibler divergence to select n-mers which can reduce the search space.

Let $\mathbf{f}_{pr}$ be the frequency of n-mers in promoters and $\mathbf{f}_{np}^a (a = 1, 2, 3)$ be respectively the frequency of n-mers in there kinds of non-promoters where a=1 represents exon, a=2 represents intron and a=3 represents 3'-UTR. Kullback-Leibler divergence is defined as follows：

$$D_a\left(\mathbf{f}_{pr}, \mathbf{f}_{np}^a\right) = \sum_{i=1}^{4^n} d_i^a = \sum_{i=1}^{4^n} f_{pr}\left(i\right) \ln \frac{f_{pr}(i)}{f_{np}^a(i)} \tag{2}$$

where

$$d_i^a = f_{pr}\left(i\right) \ln \frac{f_{pr}(i)}{f_{np}^a(i)}, \quad i = 1, \cdots, 4^n \tag{3}$$

We sort d_i^a, $i = 1, \cdots, 4^n$, in descending order and form a new vector $\mathbf{d}^a = \left[d_1^a, \cdots, d_{4^K}^a\right]^T \in \mathbb{R}^{4^n}$. To obtain the most m_a informative and discriminative n-mer, we define the following optimization problem:

$$\min \quad \frac{\sum_{i=1}^{m_a} d_i^a}{D_a\left(\mathbf{f}_{pr}, \mathbf{f}_{np}^a\right)} - \theta \tag{4}$$

$$\text{subject to} \quad \frac{\sum_{i=1}^{m_a} d_i^a}{D_a\left(\mathbf{f}_{pr}, \mathbf{f}_{np}^a\right)} \geq \theta$$

$$a = 1, 2, 3$$

where $\theta > 0$ is a threshold, say 0.98.

The salient features are the frequency of the first m_a n-mer which are used to recognize the promoters and the a th non-promoters. Then the salient features can be extracted from the gene. Let the salient feature sets be $X_r^K = \{\mathbf{f}_i^{KL}\}_{i=1}^n$, where $\mathbf{f}_i^{KL} \in \mathbb{R}^{m_a}$, *r*=1 represents the salient features distinguishing promoter and exon, *r*=2 represents the salient features distinguishing promoter and intron, *r*=3 represents the salient features distinguishing promoter and 3'-UTR.

3) CpG-island

The CpG island is a region of DNA longer than 200bp enriched with phosphodiesterase-linked cytosine (C) and guanine (G) pairs [23], where the global frequency of GC content (C+G) is greater than 50% and the ratio of expected to observed CG dinucleotide（Obs/Exp）is greater than 60%. More specifically, it shows that CpG islands can be found around promoters in about half of mammals according to the statistics of DNA data available [26]. The CpG-island features make the proportion rise to 72% [7].

Extensive researches have shown that about 60% of human gene promoters are associated with CpG islands [27]. So the CpG-island is an important feature for human promoter recognition. In this paper, we use two important CpG-island features including the global frequency of GC content (GC_con) and the ratio of expected to observed CG dinucleotide (o/e):

$$GC\_con = \frac{n_C + n_G}{L} \tag{5}$$

$$o/e = \frac{n_{CG} * L}{n_C * n_G} \tag{6}$$

where *L* is the length of a DNA sequence and n_C, n_G and n_{CG} are numbers of C, G and CG in the DNA sequence, respectively. The CpG feature set can be expressed as $\left\{\mathbf{f}_i^{CpG}\right\}_{i=1}^n$, where $\mathbf{f}_i^{CpG} \in \mathbb{R}^2$, $f_{i1}^{CpG} = GC\_con$ and $f_{i2}^{CpG} = o/e$.

B. Nucleotide Statistics

It's particularly difficult to extract the common features, which have obvious biological significance, because the genetic data is very complex, high-dimensional and contains a huge amount of information. The composition of nucleotides in different DNA fragments is different. In addition, the composition of nucleotides at different positions in the same fragment is also different.

Therefore, the contents of A, G, C and T are very important statistical features. In this paper, we call a DNA sequence C-prefer, in which the content of C is greater than G. Otherwise, a DNA sequence is called G-prefer if the content of C is not greater than G. According to the content of C and G, all the genes can be divided into two categories, in which feature extraction is independently performed.

To implement the classification of C-prefer and G-prefer genes, we use the ratio of contents of C and G. Let the subset of C-prefer DNA sequences be X_C with $|X_C| = n_1$, and the subset of G-prefer DNA sequences be X_G with $|X_G| = n_2$. Of course, $n_1 + n_2 = n$. For the two subsets X_C and X_G, we separately extract features mentioned above.

C. Recognition

1) SVM

Support Vector Machine (SVM) is a universal learner based on statistical learning theory proposed by Vapnic et al. [28]. SVM can implement the structural risk minimization rule to improve the generalization ability of learners. The decision model of SVM is constructed by limited training samples and the independent test set can still get the minimum error. In SVM, kernel functions are used to map the original samples into a high-dimensional feature space, in which the original sample could be linearly separable.

Given a set of training samples $(\mathbf{x}_i, y_i), i = 1, \cdots, n$, where $\mathbf{x}_i \in \mathbb{R}^d$, d is the number of features, $y_i \in \{-1, +1\}$ is the class label of $\mathbf{x}_i$, the goal of SVM is to find a hyper plane which maximizes the margin. The dual programming of SVM can be described as:

$$\max \ \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \tag{7}$$

$$\text{subject to} \quad \sum_{i=1}^n \alpha_i y_i = 0, \quad \alpha_i \geq 0$$

where $k(\mathbf{x}_i, \mathbf{x}_j)$ is a kernel function, α_i is the Lagrange multiplier. When $0 < \alpha_i < C$, the corresponding $\mathbf{x}_i$ is called the non-bounded support vector. If $\alpha_i = C$, then the corresponding $\mathbf{x}_i$ is called the bounded support vector. Once we solve the programming (7), we can make a decision for an unseen sample $\mathbf{x}$, or

$$f(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^n \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + b\right) \tag{8}$$

where $\operatorname{sgn}(\bullet)$ denotes the sign function, and b is the threshold of model which can be computed by $b = y_{sv} - \sum_{i=1}^n \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}_{sv})$ with a non-bounded support vector $(\mathbf{x}_{sv}, y_{sv})$.

2) Recognition Algorithm based on multiple SVMs

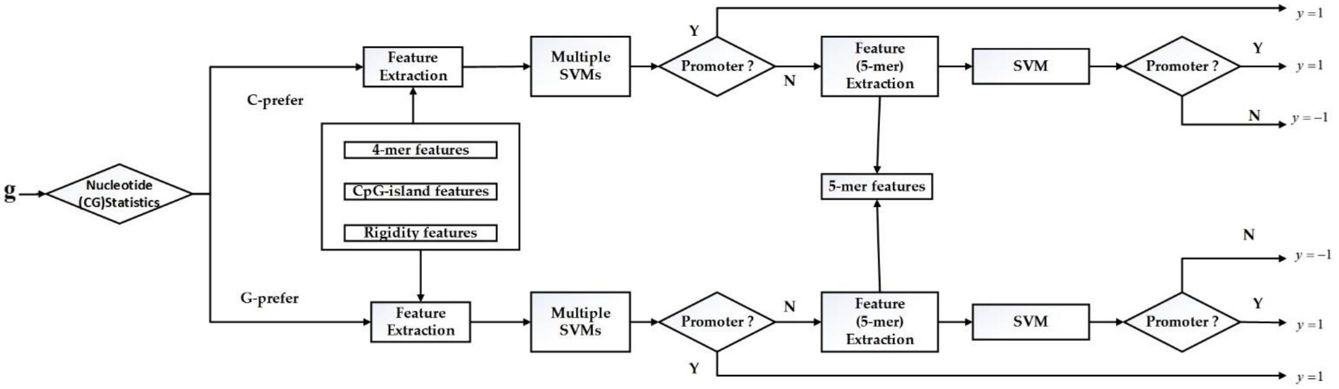


Fig. 1. The framework of our method

SVM has proven to be a good algorithm for promoter recognition [29]. However, the performance of SVM still depends on the features. Here we design multiple gene features for promoter recognition. If we use one SVM to train these multiple gene features, bad result would be obtained. Thus, a recognition algorithm based on multiple SVM is presented in this subsection.

The framework of the proposed algorithm is shown in Fig. 1. There are three stages in this framework, including single nucleotide statistics, multiple feature extraction, and multiple SVMs. In the following, we describe our method from training and test phases.

a) Training phase

Given a set of training gene sequences $X=\{\mathbf{g}_i,l_i\}_{i=1}^n$, where $\mathbf{g}_i\in\mathbb{R}^L$, L is the length of sequences, $l_i\in\{\text{"promoter","exon","intron","3'-UTR"}\}$ is the class label of the gene $\mathbf{g}_i$, and n is the number of genes. Our goal is to recognize promoters, thus we also can represent the set of training gene sequences as $X'=\{\mathbf{g}_i,y_i\}_{i=1}^n$, where $y_i\in\{-1,+1\}$. $y_i=+1$ means that $\mathbf{g}_i$ is a promoter, and $y_i=-1$ means that $\mathbf{g}_i$ is an exon or intron or 3'-UTR.

In the training phase, all genes are first divided into two categories, or the C-prefer subset and the G-prefer subset. The contents of C and G nucleotide are counted for each training gene. Then according to the ratio of contents of C and G, we divide them into two subsets, or the subset of C-prefer DNA sequences be X_C with $|X_C|=n_1$, and the subset of G-prefer DNA sequences be X_G with $|X_G|=n_2$. Of course, $n_1+n_2=n$.

Second, we separately extract multiple features from these two subsets X_C and X_G. The extracted features include rigidity, CpG-island, and 4-mer features. The subsets of rigidity feartures can be obtained by using the way mentioned in Section 2.1.1. By introducing the class labels y_i, we can get two training subsets $X_C^r=\left\{\mathbf{x}_i^r,y_i\right\}_{i=1}^{n_1}$ and $X_G^r=\left\{\mathbf{x}_i^r,y_i\right\}_{i=1}^{n_2}$, where $\mathbf{x}_i^r\in\mathbb{R}^{(L-5)}$. For the two subsets of rigidity features, we can construct two different SVMs (7) and denote them by SVM-C-Rigidity $f_C^r(\mathbf{x}^r)$ and SVM-G-Rigidity $f_G^r(\mathbf{x}^r)$, respectively.

In order to reduce time complexity， we use the Kullback-Leibler divergence to select 4-mer as the component features, respectively. Note that it requires extracting salient features between promoter and non-promoter, or exon, intron, and 3'-UTR. Thus, we can get six feature subsets for the 4-mer features, $\left(X_C^{4mer}\right)_a=\{\mathbf{x}_i^{4mer},y_i\}_{i=1}^{n_{1a}}$ and $\left(X_G^{4mer}\right)_a=\{\mathbf{x}_i^{4mer},y_i\}_{i=1}^{n_{2a}}$, $a=1,2,3$, where n_{11}, n_{12} and n_{13} are respectively the number of promoter and exton, the number of promoter and intron , the number of promoter and 3'-UTR in the C-prefer subset, and n_{21}, n_{22} and n_{23} are respectively the number of promoter and exton, the number of promoter and intron , the number of promoter and 3'-UTR in the G-prefer subset. For the six subsets, we construct six SVMs and denote them by SVM-C-4mer-a $f_C^{4mer-a}(\mathbf{x}^{4mer})$ and SVM-G-4mer-a $f_G^{4mer-a}(\mathbf{x}^{4mer})$, respectively. $a=1$ represents the SVM model distinguishing promoter and exon, $a=2$ represents the SVM model distinguishing promoter and intron, and $a=3$ represents the SVM model distinguishing promoter and 3'-UTR.

The subsets of CpG-island features can be obtained by using the method mentioned in Section 2.1.3. Similarly, we can get two training subsets $X_C^{CpG}=\left\{\mathbf{x}_i^{CpG},y_i\right\}_{i=1}^{n_1}$ and $X_G^{CpG}=\left\{\mathbf{x}_i^{CpG},y_i\right\}_{i=1}^{n_2}$, where $\mathbf{x}_i^{CpG}\in\mathbb{R}^2$. For the two subsets of CpG-island features, we can also construct two different SVMs (7) and denote them by SVM-C-CpG $f_C^{CpG}(\mathbf{x}^{CpG})$ and SVM-G-CpG $f_G^{CpG}(\mathbf{x}^{CpG})$, respectively.

Multiple SVM models can provide a decision for a gene by applying the majority voting rule. To improve classification performance, we use the 5-mer features to deal with recognized non-promoter. Similar to the 4-mer features, we can also have six training feature subsets for the 5-mer features, $\left(X_C^{5mer}\right)_a=\{\mathbf{x}_i^{5mer},y_i\}_{i=1}^{n_{1a}}$ and $\left(X_G^{5mer}\right)_a=\{\mathbf{x}_i^{5mer},y_i\}_{i=1}^{n_{2a}}$, $a=1,2,3$. For these six training subsets, we also construct six SVMs and denote them by SVM-C-5mer-a $f_C^{5mer-a}(\mathbf{x}^{5mer})$, and SVM-G-5mer-a

$f_G^{5mer-a}(\mathbf{x}^{5mer})$, respectively.

Totally, we have 16 SVM models, which are independently trained by applying 16 feature subsets.

b) Test phase

We can predict an unseen gene $\mathbf{g}$ and assign an estimated class label $\hat{y}$ to it when the 16 SVMs are well trained. If $\hat{y}=1$, the gene $\mathbf{g}$ is a promoter; otherwise, it is a non-promoter. In the following, we discuss how to assign the class label for $\mathbf{g}$.

First, the gene $\mathbf{g}$ could be C-prefer or G-prefer according to its nucleotide statistics. Then, we extract its rigidity $\mathbf{x}_{C/G}^{r}$, CpG-island $\mathbf{x}_{C/G}^{CpG}$, and 4-mer features $\mathbf{x}_{C/G}^{4mer}$, respectively, where the subscript C/G denotes $\mathbf{g}$ belonging to the C-prefer or G-perfer subset. Next, Classifier models $f_{C/G}^{4mer-a}(\mathbf{x}_{C/G}^{4mer})$ lead to three outputs $f_{C/G}^{4mer-1}(\mathbf{x}_{C/G}^{4mer})$, $f_{C/G}^{4mer-2}(\mathbf{x}_{C/G}^{4mer})$, and $f_{C/G}^{4mer-3}(\mathbf{x}_{C/G}^{4mer})$. We use the majority voting rule to decide whether $\mathbf{g}$ is a promoter or not and obtain the estimated label $\hat{y}_{C/G}^{4mer}$ by using the 4-mer features. For $\mathbf{x}_{C/G}^{r}$ and $\mathbf{x}_{C/G}^{CpG}$, SVM-C/G-Rigidity and SVM-C/G-CpG are used to predict the estimated label $\hat{y}_{C/G}^{r}$ and $\hat{y}_{C/G}^{CpG}$, respectively. For the three estimated labels $\hat{y}_{C/G}^{4mer}$, $\hat{y}_{C/G}^{r}$ and $\hat{y}_{C/G}^{CpG}$, we again use the majority voting rule to decide whether $\mathbf{g}$ is a promoter or not and obtain the estimated label $\hat{y}_{C/G}$ by these three kinds of features.

If $\hat{y}_{C/G}=+1$, then we decide that $\mathbf{g}$ is a promoter and let the final estimated label $\hat{y}=+1$. If $\hat{y}_{C/G}=-1$, we extract the 5-mer features $\mathbf{x}_{C/G}^{5mer}$ of the gene $\mathbf{g}$ and use SVM models $f_{C/G}^{5mer-a}(\mathbf{x}_{C/G}^{5mer})$ to generate three outputs $f_{C/G}^{5mer-1}(\mathbf{x}_{C/G}^{5mer})$, $f_{C/G}^{5mer-2}(\mathbf{x}_{C/G}^{5mer})$, and $f_{C/G}^{5mer-3}(\mathbf{x}_{C/G}^{5mer})$. The majority voting rule is used again to assign the estimated label $\hat{y}_{C/G}^{5mer}$ of $\mathbf{g}$.

If $\hat{y}_{C/G}^{5mer}=+1$, then we determine that $\mathbf{g}$ is a promoter and let the final estimated label $\hat{y}=+1$. If $\hat{y}_{C/G}^{5mer}=-1$, then we determine that $\mathbf{g}$ is a non-promoter and let the final estimated label $\hat{y}=-1$.

III. Experiments and Results

To implement SVM, we use the libsvm-2.89 toolbox written by Chih-Jen Lin (http：//www.csie.ntu.edu.tw/~cjlin). We choose the radial basis function (RBF) kernel function. $k(\mathbf{x}_i,\mathbf{x})=\exp(-\gamma\,||\,\mathbf{x}_i-\mathbf{x}\,||^2)$ with kernel parameter $\gamma>0$. There are two parameters C and γ. In order to find the best parameters for our problem, we use 10-fold cross-validation to select the optimal parameters C and γ.

A. Datasets

An experiment of a recognition algorithm using statistical pattern recognition methods requires a large number of the promoters and the non-promoters with accurate annotation. In this paper, we focus on differentiating short [−200, +50] bps promoter and non-promoter sequences in the same length around the transcription start point(TSS) which are defined by the DBTSS database [31] from other genomic regions, and other alternative TSSs related to tissue specific gene expression are not considered. We use 30, 964 promoter sequences [−200, +50] bps around the TSSs from the DBTSS to be the training and test sets because DBTSS provides the best combination of coverage and quality at present. In order to accurately estimate *n*-mer frequency. We construct non-promoter sets by randomly extracting 10, 000 exons and 10, 000 introns with 251 bps in length from the EID database, and 10, 000 3'UTR sequences with 251 bps in length from the UTRdb database. We randomly select 8000 samples from promoter, exon, intron and 3'-UTR sets, respectively. Among 8000 samples, 4000 samples are considered as the training ones and the rest are test ones for each class. Thus, we have 4000 training and test samples respectively. In both training and test sets, the ratio of promoter, exon, intron and 3'UTR is 1:1:1:1. The sampling process would be repeated 10 times. In this paper, the promoter is taken as the class of +1 and the non-promoter is the class of -1. The positive and negative samples are unbalanced.

For the selected training samples, we extract the 4-mer and 5-mer, CpG-island, rigidity features for them. In our experiments, m_a takes value in the interval $[200,250]$ in the 4-mer features, and $[490,560]$ in the 5-mer features. For the test samples, we extract the 4-mer, CpG-island, rigidity features for them. For samples which are first determined as non-promoters, we also extract the 5-mer features.

B. Evaluation Measures

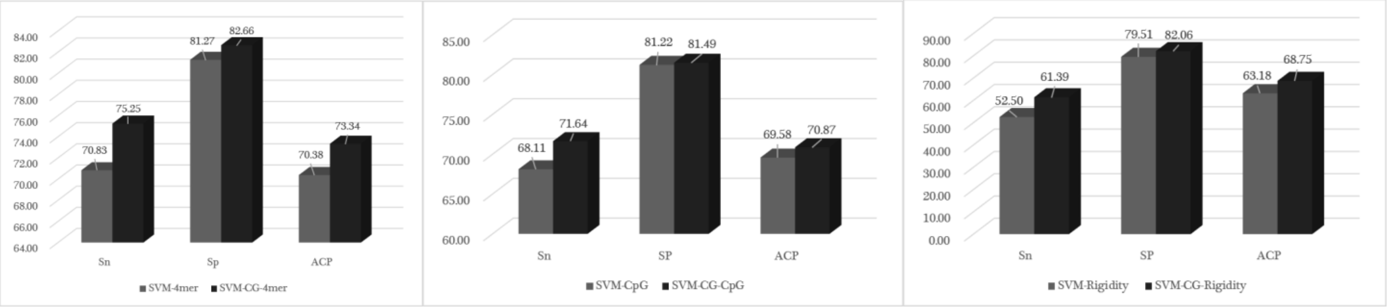
In this paper, evaluation measures proposed by Bajic [30] can be used to assess our algorithm, which are sensitivity S_n, specificity S_p and averaged conditional probability ACP and are defined as follows:

$$S_n=\frac{TP}{TP+FN} \tag{9}$$

$$S_p=\frac{TN}{TN+FP} \tag{10}$$

$$ACP=\frac{1}{4}\left(\frac{TP}{TP+FN}+\frac{TP}{TP+FP}+\frac{TN}{TN+FP}+\frac{TN}{TN+FN}\right) \tag{11}$$

where TP represents the number of positive sample identified



(a) 4-mer features (b) CpG-island features (c) Rigidity features

Fig. 2. Comparison of single feature models with and without single nucleotide statistics, (a) 4-mer, (b) CpG-island, and (c) rigidity.

correctly, *TN* represents the number of negative sample identified correctly, *FP* represents the number of negative sample which is identified as positive samples and *FN* denotes the number of positive samples not to be identified correctly.

C. Efficiency evaluation of single nucleotide statistics

In the first step of our framework, genes are divided into two categories of C-prefer and G-prefer according to single nucleotide statistics. In this experiment, we try to validate the efficiency of single nucleotide statistics.

First, we observe the effect of single nucleotide statistics on single feature. We apply single nucleotide statistics before we use three kinds of features to train SVM models, respectively. Then we get SVM decision models, SVM-CG-4mer, SVM-CG-rigidity, and SVM-CG-CpG which all are hybrid decision models. For example, SVM-CG-4mer combines SVM-C-4mer-1, SVM-C-4mer-2, SVM-C-4mer-3, SVM-G-4mer-1, SVM-G-4mer-2, and SVM-G-4mer-3. We also train SVM models without single nucleotide statistics. Then we have we get SVM decision models, SVM-4mer, SVM-rigidity, and SVM-CpG where SVM-4mer is hybrid decision models of three models.

Fig. 2 shows the evaluation measures of 3 kinds of single feaures with and without single nucleotide statistics, respectively. We can see that the performance with single nucleotide statistics is higher than that without single nucleotide statistics.

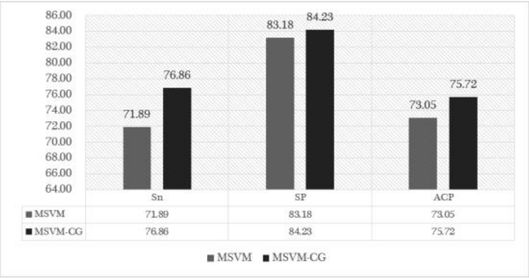


Fig. 3. Comparison of MSVM-CG and MSVM

Second, we observe the effect of single nucleotide statistics on multiple features. We use our method described in Section 2.3.2, called MSVM-CG here, and multiple SVMs without single nucleotide statistics, called MSVM. Note that the only difference between MSVM and MSVM-CG is the processing of C-prefer and G-prefer sets. Experimental results are shown in Fig. 3. We have the same conclusion as before.

In a nutshell, sensitivity S_n, specificity S_p and *ACP* are all improved when applying single nucleotide statistics. Thus, it is necessary to perform single nucleotide statistics for promoter recognition.

D. Efficiency evaluation of SVMs with 5-mer features

In the final step of our framework, 5-mer features are extracted and trained by SVMs. Now, we validate it is necessary to add this kind of features in order to improve the performance of our method. Let MSVM-CG be our method with 5-mer feature, and MSVM-CG\5mer be our method without 5-mer features. Table 1 gives the comparison of the two methods.

As show in Table 1, we find that the specificity of MSVM-CG is better than MSVM-CG\5mer, whose specificity is about 2% higher. While the ACP of MSVM-CG is also superior to MSVM-CG\5mer. But the sensitivity of MSVM-CG is inferior to it. Thus, the method with the 5-mer features can improve two evaluation measures, which indicates the importance of the 5-mer features.

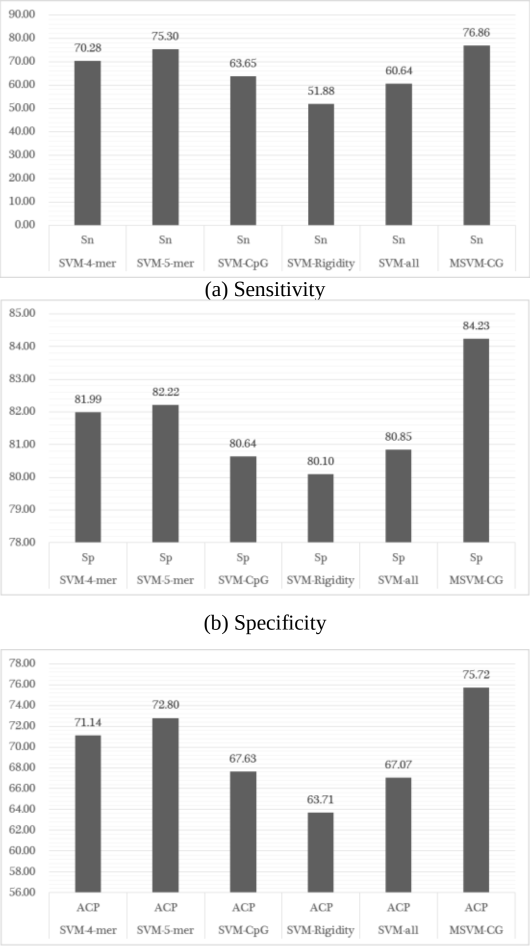
TABLE I. EFFICIENCY EVALUATION OF SVMS WITH 5-MER FEATURES

	MSVM-CG\5mer	**MSVM-CG**
Sn	76.86	79.19
Sp	84.23	83.46
ACP	75.72	75.42

E. Efficiency evaluation of MSVM-CG

As mentioned above, we use one SVM to train all extracted features, or the 4-mer, CpG-island, and rigidity. These features could be connected to new feature vectors which can be directly trained by single SVM. For short, we denote this method as SVM-all. MSVM-CG, SVMs with four kinds of single features (or SVM-4mer, SVM-5mer, SVM-CpG, SVM-Rigidity), and SVM-all are compared here. Sensitivity, specificity and ACP of the experiment in the same human genome are shown in Fig. 4, respectively.

We can see clearly from these figures, MSVM-CG can achieve the best performance, 76.85% sensitivity, 84.23% specificity and 75.72% *ACP*. The sensitivity of SVM-Rigidity is only about 50% and its specificity is only about 80.1% for rigidity features. SVM-5mer is better than SVM-all, SVM-4mer, SVM-CpG, and SVM-Rigidity, but inferior to MSVM-CG. The performance of SVM-all is worse than SVMs with single features, such as the 4-mer ones, which also indicate that our scheme dealing with multiple features is correct.



(a) Sensitivity

(b) Specificity

(c) ACP

Fig. 4. Comparison of five methods, (a) Sensitivity, (b) Specificity, and (c) ACP.

IV. Conclusion

In this paper, a novel method for promoter recognition is proposed. The promoter, the coding exons, the introns and 3'-UTR of DNA sequences are considered at the same time. In our method, a gene dataset can first be divided into two subsets by using single nucleotide statistics. Second, multiple gene features are extracted for each subset, including DNA rigidity, word-based feature and CpG-island. Word-based features are extracted from four kinds of sequences in human genome. Specially, we use Kullback–Leibler (KL) divergence to select the most informative and discriminative features to identify the promoters and non-promoters in large genomic sequences. Finally, multiple support vector machines (SVMs) are independently adopted to classify these features. Experiments based on human genome are performed. In order to better assess the performance of our method, we analyze single classifier with the single features and signal classifier with multiple features. The experimental results show that the sensitivity and specificity of our method reach 76% and 84%, respectively.

Since the genetic data is very complex and high-dimensional, we only perform experiments on limited samples and limited kinds of features. Therefore, in the future research, more representative training data should be used to extracted features and more feature extraction methods should be considered.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 61373093, 61033013, and 61271301, by the Natural Science Foundation of Jiangsu Province of China under Grant Nos. BK20140008, BK2011284 and BK201222725, by the Natural Science Foundation of the Jiangsu Higher Education Institutions of China under Grant No.13KJA520001, and by the Qing Lan Project.

References

[1] BAJIC V B, CHONG A, SEAH S H, et al. An intelligent system for vertebrate promoter recognition[J]. IEEE Intelligent systems, pp. 64-70, 2002.

[2] Fickett JW, Hatzigeorgiou AG. "Eukaryotic promoter recognition," Genome Res., pp. 861-878, 1997 Sep.

[3] IOSHIKHES I P, ZHANG M Q. "Large-sale human promoter mapping using CpG islands [J]". Nat Genet, pp. 61-63, 2000.

[4] DAVULURI R, GROSSE I, ZHANG M Q. "Computational identification of promoters and first exons in the human genome [J]". Nat Genet, pp. 412-417, 2001.

[5] Ponger L, MOUCHIROUD D. "CpGProD: Identifying CpG islands associated with transcription start sites in large genomic mammalian sequences [J]". Bioinformatics, pp. 631-633, 2002.

[6] Scherf M, Klingenhoff A, Werner T. "Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. JMol Biol, pp. 599–606, 2000.

[7] Saxonov S, Berg P, Brutlag D L et al. "A genome-wide anal-ysis of CpG dinucleotides in the human genome distin-guishes two distinct classes of promoters," Proc Natl Acad Sci USA, 103(5), pp. 1412-1417, 2006.

[8] Solovyev V.V., Shahmuradov I.A., "PromH: Promoters identification using orthologous genomic sequences," Nucleic Acids Res., 31(13), pp. 3540-3545, 2003 Jul 1.

[9] Ohler U., Stemmer G., Harbeck S., Niemann H., "Stochastic segment models of eukaryotic promoter regions," Pac Symp Biocomput., pp. 380-391, 2000, 5.

[10] Down T.A., Hubbard T.J., "Computational detection and location of transcription start sites in genomic DNA," Genome Res, 12(3), pp. 458-461, 2002 Mar.

[11] Davuluri R.V., Grosse I., Zhang M.Q., "Computational identification of promoters and first exons in the human genome," Nat Genet, 29(4), pp. 412-417, 2001 Dec

[12] Ioshikhes I.P., Zhang M.Q., "Large-scale human promoter mapping using CpG islands," Nat Genet, 26(1), pp. 61-63, 2000 Sep

[13] Bajic V.B., Seah SH, Chong A, Krishnan SP, Koh JL, Brusic V, "Computer model for recognition of functional transcription start sites in RNA polymerase II promoters of vertebrates," J Mol Graph Model., 21(5), pp. 323-332, 2003 Mar

[14] Fujii S, Kono H, Takenaka S, Go N, Sarai A, "Sequence-dependent DNA deformability studied using molecular dynamics simulations," Nucleic Acids Res., 35(18), pp. 6063-74, 2007.

[15] Pedersen AG, Baldi P, Chauvin Y, Brunak S. "DNA structure in human RNA polymerase II promoters" J Mol Biol, 281(4), pp. 663-73, 1998 Aug 28.

[16] Burden S, Lin YX, Zhang R. "Improving promoter prediction for the NNPP2.2 algorithm: a case study using Escherichia coli DNA sequences", Bioinformatics, pp. 601 – 607, 2005.

[17] X. Xie, S. Wu, K.-M. Lam, and H. Yan, "PromoterExplorer: an effective promoter identification method based on the AdaBoost algorithm," Bioinformatics, vol. 22, pp. 2722–2728, 2006.

[18] Bajic VB, Seah SH, Chong A, Zhang G, Koh JL, Brusic V., "Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoters," Bioinformatics, 18(1), pp. 198-199, 2002 Jan.

[19] Gangal R, Sharma P. "Human pol II promoter prediction: time series descriptors and machinel earning," Nucleic Acids Research, 33(4), pp. 1332-1336, 2005 Mar 1.

[20] S. Knudsen, "Promoter 2.0: for the recognition of pol II promoter sequences," Bioinformatics, vol. 15, pp. 356–361, 1999.

[21] Zhao X, Pei Z, Liu J, Qin S, Cai L., " Prediction of nucleosome DNA formation potential and nucleosome positioning using increment of diversity combined with quadratic discriminant analysis," Chromosome Res., 18(7), pp. 777-85, 2010 Nov;.

[22] Goddard N.L., Bonnet G., Krichevsky O. and Libchaber A., "Sequence dependent rigidity of single stranded DNA," Phys Rev Lett., 85 (11), pp. 2400-3, 2000, Sep 11.

[23] Zeng J., Zhu S. and Yan H., "Towards accurate human promoter recognition: a review of currently used sequence features and classification methods," Brief Bioinform., 10(5), pp. 498-508, 2009, Sep.

[24] Brukner I, Sánchez R, Suck D, Pongor S. "Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides," EMBO J., 14(8):1812-8, 1995 Apr 18.

[25] Packer M.J., Dauncey M.P., and Hunter C.A., "Sequence-dependent DNA structure: tetranucleotide conformational maps," J. Mol. Biol. 295, 2000.

[26] Antequera F, Bird A. "Number of CpG islands and genes in human and mouse," Proc Natl Acad Sci USA, Scotland, 90(24), pp. 11995-11999, 1993 Dec 15.

[27] Cross SH, Clark VH, Bird AP. "Isolation of CpG islands from large genomic clones," Nucleic Acids Res, 27(10): 2099-2107, 1999 May 15.

[28] Vapnik V., Cortes C, " Support-vector networks," Machine Learning, , 20(3), pp. 273-297, 1995

[29] Gangal R, Sharma P. "Human pol II promoter prediction: time series descriptors and machine learning," Nucleic Acids Res., 33(4), pp. 1333-1336, 2005 Mar 1.

[30] Bajic VB. "Comparing the success of different prediction programs in sequence analysis : a review." Brief Bioinform, (3), pp. 214-228, 2000 Sep.

[31] R. Yamashita, Y. Suzuki, H. Wakaguri, K. Tsuritani, K. Nakai, and S. Sugano, "DBTSS: database of human transcription start sites, progress report 2006," Nucleic Acids Res., vol. 34, pp. 86–89, 2006.