Theory Biosci. (2011) 130:91–100
DOI 10.1007/s12064-010-0114-8

ORIGINAL PAPER

Eukaryotic and prokaryotic promoter prediction using hybrid approach

Hao Lin · Qian-Zhong Li

Received: 4 August 2010 / Accepted: 23 October 2010 / Published online: 3 November 2010

Abstract Promoters are modular DNA structures containing complex regulatory elements required for gene transcription initiation. Hence, the identification of promoters using machine learning approach is very important for improving genome annotation and understanding transcriptional regulation. In recent years, many methods have been proposed for the prediction of eukaryotic and prokaryotic promoters. However, the performances of these methods are still far from being satisfactory. In this article, we develop a hybrid approach (called IPMD) that combines position correlation score function and increment of diversity with modified Mahalanobis Discriminant to predict eukaryotic and prokaryotic promoters. By applying the proposed method to *Drosophila melanogaster*, *Homo sapiens*, *Caenorhabditis elegans*, *Escherichia coli*, and *Bacillus subtilis* promoter sequences, we achieve the sensitivities and specificities of 90.6 and 97.4% for *D. melanogaster*, 88.1 and 94.1% for *H. sapiens*, 83.3 and 95.2% for *C. elegans*, 84.9 and 91.4% for *E. coli*, as well as 80.4 and 91.3% for *B. subtilis*. The high accuracies indicate that the IPMD is an efficient method for the identification of eukaryotic and prokaryotic promoters. This approach can also be extended to predict other species promoters.

H. Lin (✉)
Key Laboratory for NeuroInformation of Ministry of Education,
Center of Bioinformatics, School of Life Science
and Technology, University of Electronic Science
and Technology of China, Chengdu 610054, China
e-mail: hlin@uestc.edu.cn

Q.-Z. Li
Laboratory of Theoretical Biophysics, School of Physical
Sciences and Technology, Inner Mongolia University,
Hohhot 010021, China

Keywords Promoter prediction · Position correlation score function · Increment of diversity · Modified Mahalanobis Discriminant

Introduction

Promoters are functional regions containing complex regulatory elements for determining the transcription initiation of genes. Thus, promoter prediction using computational techniques is important for discovering genes that are missed by gene predictors and devising experiments to understand transcriptional regulation (Abeel et al. 2008a, b). Although many methods have been developed for promoter prediction, performances of existing methods are still far from being satisfactory. It is necessary to develop more efficient approaches to accurately and rapidly predict promoters.

It is well known that prokaryotic and eukaryotic promoters use different DNA sequences to regulate gene expression. In prokaryotes, the transcription of most of genes is regulated by the σ^{70}-promoters. The σ^{70}-promoters commonly contain three basic regulatory elements (Hawley and McClure 1983): Pribnow box (or called TATA box) with consensus TATAAT around −10 bp upstream of transcription start site (TSS), −35 box with consensus TTGACA around −35 bp upstream of TSS and initiator (Inr) around TSS. In eukaryotes, all protein-coding genes and certain small nuclear RNAs are regulated by pol II promoters. The core regions of the pol II promoters usually contain several regulatory motifs (Pedersen et al. 1999; Bajic et al. 2004): TATA box around −25 bp upstream of TSS, initiator and downstream promoter element (DPE) around +30 bp downstream of TSS.

In the past 20 years, many algorithms, such as position weight matrix (PWM) (Prestridge 1995; Huerta and

Collado-Vides 2003; Li and Lin 2006; Akan and Deloukas 2008), hidden Markov model (HMM) (Pedersen et al. 1996; Ohler et al. 1999, 2001, 2002; Ohler 2006), artificial neural network (NN) (Horton and Kanehisa 1992; Pedersen and Engelbrecht 1995; Reese 2001; Burden et al. 2005), and support vector machine (SVM) (Gordon et al. 2003, 2006; Gangal and Sharma 2005; Anwar et al. 2008) have been developed to identify promoters in eukaryotic and prokaryotic genomes using CpG islands, hexamer compositions, transcription factor binding sites (TFBS) and other information. Consequently, some on-line available tools, such as Eponine (Down and Hubbard 2002), Corepromoter (Zhang 2005), CpGProD (Ponger and Mouchiroud 2002), Promoter2.0 (Knudsen 1999), FirstEF (Davuluri et al. 2001), promH (Solovyev and Shahmuradov 2003), Promoter Scan (Prestridge 1995), Dragon promoter finder (Bajic et al. 2002), NNPP2.2 (Reese 2001), McPromoter (Ohler 2006), ARTS (Sonnenburg et al. 2006), ProSOM (Abeel et al. 2008a, b), and RBF-TSS (Mahdi and Rouchka 2009) have been designed for the detection of promoters. Although contemporary methods have achieved great progress in promoter recognition, they were still limited in predictive performance (Anwar et al. 2008; Abeel et al. 2008a, b; Yang et al. 2008). The performances of these methods are unreliable with poor specificity or poor sensitivity. Phylogenetic footprinting takes advantage of relative conservation of motifs among related species (Janky and van Helden 2008; Satija et al. 2008). Grech et al. (2007, 2008) have developed phylogenetic footprinting programs to analyze promoters of *Chlamydia trachomatis*. But these motifs are short and not fully conserved which results in a lot of false positives. Moreover, some motifs are probably very common in certain species, but less in other species (Abeel et al. 2008a, b). These hamper the ability of a single program to predict promoters in different species using same model (Abeel et al. 2008a, b). Hence, further developments of prediction techniques are required to improve the predictive accuracy.

In this article, we present a hybrid approach based on modified Mahalanobis Discriminant (MD) to identify eukaryotic and prokaryotic promoters. Two kinds of algorithms that are position correlation score function (PCSF) (Li and Lin 2006; Gordon et al. 2006; Kielbasa et al. 2005) and increment of diversity (ID) (Laxton 1978) are proposed to describe signal features and composition features of sequences. At first, based on PWM, the PCSF is calculated for the description of regulatory signals and short functional elements. Subsequently, according to information theory, the increment of diversity (ID) algorithm is used to measure the similarity of oligonucleotides compositions in specific sub-region between test and training sequences. Finally, by use of PCSF and ID values as inputs, the modified MD is applied to predict promoters (Goni et al. 2007; Zhang et al. 2004; Shahmuradov et al. 2005; Levitsky and Katokhin 2003). We evaluate the performance of proposed approach for five species: *D. melanogaster*, *H. sapiens*, *C. elegans*, *E. coli*, and *B. subtilis*. Comparative results demonstrate that our approach can improve accuracies for promoter prediction.

Materials and methods

Materials

The 1886 *D. melanogaster* and 1787 *H. sapiens* pol II promoter sequences are extracted from Eukaryotic Promoter Database (EPD, Release 90, http://www.epd.isb-sib.ch/) (Schmid et al. 2006). The 598 *C. elegans* promoters are extracted from CEPDB (http://rulai.cshl.edu/cgi-bin/CEPDB/home.cgi). Total of 741 *E. coli* K-12 σ^{70} promoters and 270 *B. subtilis* σ^{43} promoters are extracted from RegulonDB (http://www.cifn.unam.mx/Computational_Genomics/regulondb/) (Salgado et al. 2004) and DBTBS (http://dbtbs.hgc.jp/) (Makita et al. 2004). The eukaryotic and prokaryotic promoters are 300-bp long regions from -249 to $+50$ bp flanking TSSs (TSSs are the 0th sites) and 81-bp long regions from -60 to $+20$ bp flanking TSSs (TSSs are the 0th sites).

The non-promoter sequences are randomly extracted from coding regions and non-coding regions. For *D. melanogaster*, 2859 coding sequences and 1799 introns constructed by Ohler et al. (2002) downloaded from website http://www.fruitfly.org/sequence/drosophila-datasets.html. For *H. sapiens*, 1800 coding sequences and 1800 introns are randomly selected from human DNA sequences (http://www.fruitfly.org/sequence/human-datasets.html). For *C. elegans*, 600 coding sequences and 600 introns are randomly extracted from Exon/Intron Database (EID) (Shepelev and Fedorov 2006). For prokaryotes, 1400 *E. coli* negative sequences (700 coding sequences and 700 convergent intergenic sequences), and 600 *B. subtilis* negative sequences (300 coding sequences and 300 convergent intergenic sequences) are extracted from GenBank. The lengths of negative sequences are 300 and 81 bp, respectively, for eukaryote and prokaryote. The hypothetical non-TSSs are located in the 250th position for eukaryotic sequences and in the 60th position for prokaryotic sequences, so the non-promoters have the same format as the promoter ones: [non-TSS -249 … non-TSS $+50$] for eukaryotic sequences and [non-TSS -60 … non-TSS $+20$] for prokaryotic sequences. Sequences having regions with other IUPAC code letters, such as “N,” “W,” “S” have been filtered out from both positive and negative datasets.

The definition for sequence conservation

For investigating the signal properties of promoter sequences, the conservation of oligonucleotide with length k-mer at the ith site can be calculated from following formula (Li and Lin 2006):

$$M_k(i) = \sum\nolimits_x [p_i(x) - p_\mathrm{e}]^2 \Big/ p_\mathrm{e} \tag{1}$$

where $p_i(x)$ and p_e denote the observed probability and expected probability of k-mer oligonucleotide x at the ith site, respectively. Two approaches can be used to calculate expected probability p_e: one is equal distribution of the k-mer oligonucleotide; another is the real k-mer oligonucleotide counts for each species. In this study, the first approach was used to calculate the p_e. For example, if $k = 1$, the expected probabilities of four bases is 0.25; and the observed probabilities of bases A, C, G, and T at the ith site denote as p_i(A), p_i(C), p_i(G), and p_i(T), respectively. The $M_1(i)$ denotes the conservation of bases at the ith site. It can be proved that the larger the $M_k(i)$ value, the more conserved the ith site. $M_k(i)$ equals to zero for random sequence.

PCSF

The PWM can be constructed by counting the frequencies of oligonucleotides in conserved sites of training sequences. The probability p_{xi} of an oligonucleotide x at the ith site can be formulated as (Li and Lin 2006; Wasserman and Sandelin 2004; Kielbasa et al. 2005):

$$p_{xi} = (n_{xi} + b_{xi})/(N_i + B_i) \tag{2}$$

where n_{xi} and b_{xi} are real counts and pseudocounts of k-mer oligonucleotide x at the ith site, respectively. N_i and B_i are total number of real counts and pseudocounts at the ith site, respectively. If there are relatively few real counts, many k-mer variations may not be presented because of the small sample of sequences. The goal of adding pseudocounts is to obtain an improved estimate of the probability p_{xi} of k-mer oligonucleotide x at the ith site. A relatively few pseudocounts should be added when there is a good sampling of sequences, and more pseudocounts should be added when the data is sparser. One simple formula that has worked well in some studies is to make B_i equal to $\sqrt{N_i}$ and b_{xi} equal to $p_0\sqrt{N_i}$ (p_0 is the average background frequency) in Eq. 2 (Wasserman and Sandelin 2004; Kielbasa et al. 2005), respectively. As N_i increase, the influence of pseudocounts decrease because $\sqrt{N_i}$ increase more slowly. Due to the existence of pseudocounts, the estimated probabilities are strictly positive (Kielbasa et al. 2005). Based on the probabilities p_{xi}, the PCSF of an arbitrary sequence can be defined as (Li and Lin 2006):

$$F = \sum_i \ln(p_{xi}/p_0) \tag{3}$$

where p_0 is average background probability of k-mer. The score F shows the degree of sequence closed to matrix resource.

ID

According to the concept of diversity (Laxton 1978), if a sequence X can be denoted as a d-dimension vector $X : \{n_1, n_2, \ldots, n_i, \ldots, n_d\}$, the diversity of this sequence can be defined as (Zhang and Luo 2003):

$$D(X) = D(n_1, n_2, \ldots n_d) = N \ln N - \sum\nolimits_i^d n_i \ln n_i \tag{4}$$

here $N = \sum_i^d n_i$, n_i indicates the absolute frequency of ith k-mer oligonucleotide. If n_i equals zero, $n_i \ln n_i = 0$. It is easily deduced that the diversity equals N-fold of information entropy.

In general, for two sequences X and Y with the same space of d dimension, $X : \{n_1, n_2, \ldots, n_i, \ldots, n_d\}$ and $Y : \{m_1, m_2, \ldots, m_i, \ldots, m_d\}$, the increment of diversity is defined as the following formula:

$$ID(X, Y) = D(X + Y) - D(X) - D(Y) \tag{5}$$

here $D(X)$ and $D(Y)$ are the diversity of X and Y, respectively. $D(X + Y)$ denotes the diversity of the mixed source $X + Y$. Therefore, ID is essentially a measure of incremental information of mixed system. It is easily proved that the smaller the ID, the higher the similarity of two sequences.

Modified MD

Based on the theory of multivariate normal probability density functions, the modified MD (Goni et al. 2007; Levitsky and Katokhin 2003) can be defined as:

$$\mathrm{MD}(s, \mu) = (s - \mu)^\mathrm{T} \Sigma^{-1} (s - \mu) + \log|\Sigma| \tag{6}$$

here μ and Σ are the group mean and covariance matrix of training dataset, respectively. Σ^{-1} and $|\Sigma|$ are the inverse matrix and determinant, respectively. The symbol s denotes the feature vector of each test sequence. The symbol T denotes the transposition of vector $(s - \mu)$. It can be proved that, for the covariance matrix Σ, there are no negative eigenvalues and, in general, no zero eigenvalue either (Chou 1995). Here, the MD was used to measure the similarity level between test sequence and training data. Commonly, the prediction rule can be formulated by (Chou et al. 1998):

$$\mathrm{MD}(s,\mu_{\varsigma}) = \mathrm{Min}\left\{\mathrm{MD}\left(s,\mu_{\mathrm{promoter}}\right),\mathrm{MD}\left(s,\mu_{\mathrm{coding}}\right),\right.\\ \left.\mathrm{MD}\left(s,\mu_{non-coding}\right)\right\} \tag{7}$$

According to the quadratic discriminant formulation for the classification of multi-groups (Zhang 1997; Feng and Luo 2008), we shall generalize the decision function as:

$$\xi = \mathrm{MD}\left(s,\mu_{promoter}\right)\\ - \mathrm{Min}\left\{\mathrm{MD}\left(s,\mu_{\mathrm{coding}}\right),\mathrm{MD}\left(s,\mu_{\mathrm{non-coding}}\right)\right\} \tag{8}$$

where the operator Min means taking the minimum value among those in the parentheses. The score ξ represents the degree to which class the test sequence belongs to. Setting a threshold value ξ_0 will determine a predictive result. If ξ value of the test sequence is lower than the threshold ξ_0, the test sequence is predicted as promoter; otherwise it is predicted as non-promoter. Then Eq. 8 provides a method to select possible optimal models and to discard suboptimal ones independently from the cost context or the class distribution.

Performance evaluate

The following five parameters: sensitivity (S_n), specificity (S_p), false positive rate (FPR), precision, correlation coefficient (CC), and overall accuracy (Ac) are used to evaluate the predictive performance of our method.

$$S_n = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}), \tag{9}$$

$$S_p = \mathrm{TN}/(\mathrm{TN}+\mathrm{FP}), \tag{10}$$

$$\mathrm{FPR} = \mathrm{FP}/(\mathrm{TN}+\mathrm{FP}), \tag{11}$$

$$\mathrm{Precision} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP}) \tag{12}$$

$$\mathrm{CC} = \frac{(\mathrm{TP}\times\mathrm{TN})-(\mathrm{FP}\times\mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FN})\times(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TN}+\mathrm{FP})}} \tag{13}$$

$$\mathrm{Ac} = (\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{FN}+\mathrm{TN}+\mathrm{FP}) \tag{14}$$

where TP denotes the numbers of the correctly recognized promoters, FN denotes the numbers of the promoters recognized as non-promoters, FP denotes the numbers of the non-promoters recognized as promoters, TN denotes the numbers of correctly recognized non-promoters.

Results and discussions

Benchmark data selection

Five species promoters are used to evaluate the performance of the method. An important issue is how to select benchmark data to train proposed methods. Some works have used promoters and coding sequences, and promoters and non-coding sequences, respectively, to examine their approaches (Gordon et al. 2003). Other researches have proposed a more realistic method that used promoters and non-promoters (coding sequences and non-coding sequences) to train their approaches (Ohler 2006; Yang et al. 2008; Rani et al. 2007). However, these works have not considered the difference between coding sequences and non-coding sequences. Because coding sequences are distinctly different from non-coding sequences, one may argue that the combination of coding sequences and non-coding sequences together as one negative set confuses real properties of both coding sequences and non-coding sequences, which can result in poor performance of proposed method. Experiments in "prediction accuracy" prove this hypothesis.

Feature selection

The PCSF algorithm is used to estimate the occurrence of k-mer sequences at specific position. Here, conservation of trimer oligonucleotide is measured by Eq. 1. The results are exhibited in Fig. 1. As it can be seen from Fig. 1, the most conserved region of eukaryotic promoters usually occurs in Inr. The TATA box is less clear, which is consistent with the results of Aerts et al. (2004). For prokaryotic promoters, the −35 box, −10 box, and initiator are more conserved than other regions. On the contrary, there are not distinctly conserved regions in coding and non-coding sequences. By conservation analysis and accuracy evaluation of sites, the conserved sites of five species promoters can be selected. It initially evaluates each site and selects the one with the maximum accuracy. Subsequently, the site with second maximum accuracy is selected out and merged into conserved site subset. This process is repeated until increasing the size of the current conserved site subset leads to a lower prediction rate. By a great number of examinations, the conserved sites of five species promoters are selected and listed in Table 1. Therefore, the trimer oligonucleotide probability matrix with 64 rows (one row for each trimer oligonucleotide) and a number of columns of conserved sites is constructed according to Eq. 2. Based on this probability matrix, Eq. 3 can be used to calculate the weight score (F_{promoter}) of sequences. Correspondingly, another two weight scores (F_{coding} and $F_{\text{non-coding}}$) can be calculated according to the probability matrices of coding sequences and non-coding sequences.

Mahdi and Rouchka (2009) have suggested dividing sequences into some sub-regions. According to this, we divide eukaryotic and prokaryotic promoter sequences into three and two sub-regions. The three sub-regions of eukaryotic promoters are non-transcribed region, transcribed region, and core promoter region. The two sub-regions of prokaryotic promoters are initiator combined

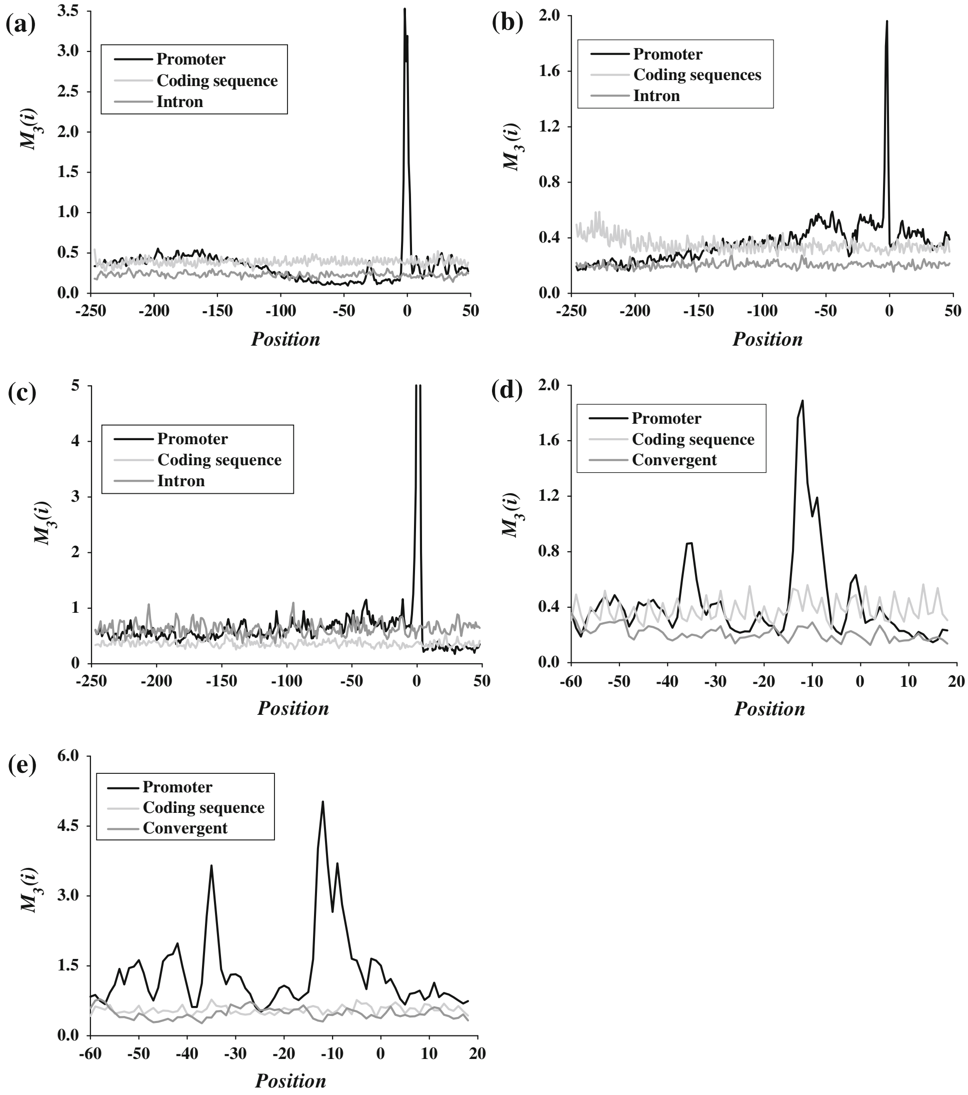

Fig. 1 The conservation of promoters, coding sequences and non-coding sequences for *D. melanogaster* (**a**), *H. Sapiens* (**b**), *C. elegans* (**c**), *E. coli* (**d**), and *B. subtilis* (**e**)

with −10 box and −35 box combined with upstream sequences, respectively. The detailed divisions are shown in Table 1. Hexamer oligonucleotide frequencies are important parameters which have been widely used to identify promoters or *cis*-regulatory motifs (Hutchinson 1996; Chan and Kibler 2005; Down and Hubbard 2002), so the hexamer oligonucleotide frequencies in each sub-region are measured by ID algorithm (Eq. 5) for evaluating the similarity of each sub-region between test sequences and training sequences. For eukaryotic promoters with three sub-regions and three datasets, nine IDs can be obtained. Correspondingly, six IDs (two sub-regions multiply three data sets) can be obtained for prokaryotic promoters.

Based on above two kinds of algorithms, sequences of eukaryotes and prokaryotes can be described as twelve (three PCSFs and nine IDs) and nine dimension vectors (three PCSFs and six IDs), respectively.

Prediction accuracy

Different statistical strategies are used to analyze and estimate the performance of our approach. First, all sequences are randomly split into two sets: one is training set, the other is test set. We use five ratios (1:9, 2:8, 3:7, 4:6, and 5:5) of test set and training set to examine our approach. For each examination, the predictive results are objective as the test set is completely independent from training set. The results in Table 2 show that the overall accuracy of ~95% for *D. melanogaster*, ~91% for *H. sapiens*, ~89% for *C. elegans*, ~88% for *E. coli*, ~85% for *B. subtilis* are achieved, suggesting that IPMD method is steady and robust.

The *n*-fold cross-validation is more rigorous and objective method for evaluating the predictive performance of the proposed method. For *n*-fold cross-validation, the

Table 1 The conserved sites used in PCSF and sub-regions used in ID for five species

Species	Conserved sites of promoters used in PCSF	Sub-regions used in ID
Eukaryotes		
D. melanogaster	−35, −34, −33, −32, −31, −30, −29, −28, −27, −5, −4, −3, −2, −1, + 1, + 25, +26, + 27, + 28	−249 to −1 bp +1 to +50 bp −40 to +30 bp
H. sapiens	−31, −30, −29, −28, −27, −26, −25, −24, −23, −22, −21, −4, −3, −2, −1, 0, + 1, + 2, + 25, + 26, + 27, + 28, + 29	−249 to −1 bp +1 to +50 bp −40 to +10 bp
C. elegans	−108, −87, −54, −49, −42, −41, −40, −39, −37, −11, −3, −2, −1, 0, + 1, + 2, + 3	−249 to −1 bp +1 to +50 bp −20 to +10 bp
Prokaryotes		
E. coli	−51, −37, −36, −35, −34, −16, −15, −14, −13, −12, −11, −10, −9, −8, −7, −2, −1	−60 to −25 bp −25 to +20 bp
B. subtilis	−42, −36, −35, −34, −16, −15, −14, −13, −12, −11, −10, −9, −8, −7, −2, −1, 0	−60 to −20 bp −20 to +20 bp

Table 2 The prediction results of IPMD method using different ratio of test set to training set

Ratio (test set:training set)	1:9	2:8	3:7	4:6	5:5
D. melanogaster					
S_n (%)	93.6	91.0	91.7	90.0	90.2
S_p (%)	96.1	97.0	96.7	96.7	96.6
Ac (%)	95.4	95.3	95.3	94.8	94.8
CC	0.889	0.884	0.884	0.872	0.872
H. sapiens					
S_n (%)	86.0	86.6	88.6	87.6	87.5
S_p (%)	94.4	92.5	92.2	93.4	93.9
Ac (%)	91.7	90.5	91.0	91.5	91.8
CC	0.811	0.788	0.800	0.808	0.815
C. elegans					
S_n (%)	88.1	90.0	89.4	85.8	89.3
S_p (%)	91.7	88.8	89.7	90.4	88.5
Ac (%)	90.5	89.2	89.6	88.9	88.8
CC	0.788	0.767	0.774	0.754	0.758
E. coli					
S_n (%)	85.1	82.4	82.4	80.7	77.8
S_p (%)	89.3	90.7	92.4	91.8	94.3
Ac (%)	87.9	87.9	88.9	88.0	88.6
CC	0.735	0.731	0.754	0.732	0.744
B. subtilis					
S_n (%)	74.1	72.2	75.3	84.3	80.7
S_p (%)	91.7	91.7	90.0	84.6	85.7
Ac (%)	86.2	85.6	85.4	84.4	84.1
CC	0.672	0.657	0.658	0.660	0.644

dataset is divided into n equal parts. Of these n parts, $n - 1$ parts are used for training and the nth is used for testing. This is done repeatedly n times for all n parts. Ten-fold cross-validation is used in this article. The receiver operating characteristic curve (ROC curve) and precision recall curve (PRC curve) can describe the performance of our model across the entire range of classification thresholds ξ (Eq. 8). We plot the ROC curves and PRC curve of five species in Fig. 2. Results show that the areas under the ROC and PRC curves are, respectively, 98.5 and 97.2% for *D. melanogaster*, 96.9 and 93.9% for *H. sapiens*, 94.6 and 92.3% for *C. elegans*, 95.3 and 92.0% for *E. coli*, and 93.3 and 84.7% for *B. subtilis*. By adjusting the threshold ξ to the optimal value ξ_0, the maximum overall accuracy (Ac) and correlation coefficient (CC) can be achieved. The experiments show that the optimal threshold ξ_0 are 0.20 for *D. melanogaster*, −0.70 for *H. sapiens*, 0.14 for *C. elegans*, −1.20 for *E. coli*, and 2.60 for *B. subtilis*. Table 3 shows that the sensitivities and specificities are, respectively, 90.6 and 97.4% for *D. melanogaster*, 88.1 and 94.1% for *H. sapiens*, 83.3 and 95.2% for *C. elegans*, 84.9 and 91.4% for *E. coli*, and 80.4 and 91.3% for *B. subtilis*, demonstrating that IPMD method is an excellent and powerful method for eukaryotic and prokaryotic promoter prediction.

Table 3 also records the performance of IPMD method for discriminating promoters from non-promoters (coding sequences and non-coding sequences) on five species. Ten-fold cross-validated accuracies are 89.1% for *D. melanogaster*, 87.7% for *H. sapiens*, 90.3% for *C. elegans*, 88.7% for *E. coli*, 87.7% for *B. subtilis*, which are lower than that

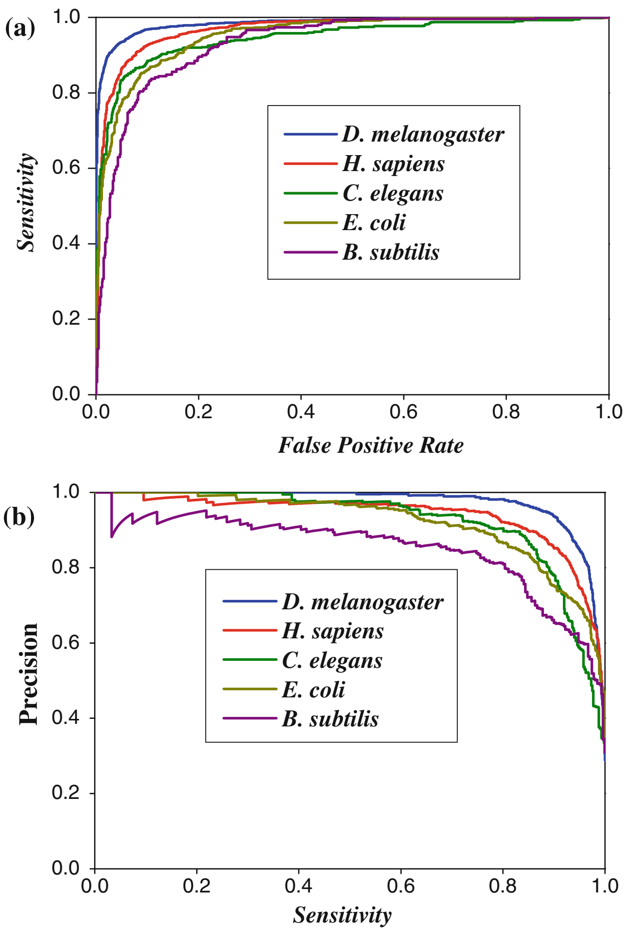

Fig. 2 ROC (**a**) and PRC curves (**b**) for five species

Table 3 The accuracy of IPMD method using 10-fold cross-validation

Species	Benchmark dataset	S_n (%)	S_p (%)	Ac (%)	CC
D. melanogaster	Three datasets	90.6	97.4	95.4	0.888
	Two datasets	89.2	89.1	89.1	0.751
H. sapiens	Three datasets	88.1	94.1	92.1	0.822
	Two datasets	85.7	88.8	87.7	0.731
C. elegans	Three datasets	83.3	95.2	91.2	0.800
	Two datasets	84.1	93.3	90.3	0.780
E. coli	Three datasets	84.9	91.4	89.2	0.761
	Two datasets	81.0	92.7	88.7	0.747
B. subtilis	Three datasets	80.4	91.3	87.9	0.718
	Two datasets	72.6	94.5	87.7	0.705

using three benchmark datasets. Results demonstrate that using three datasets (promoters, coding sequences, and non-coding sequences) to train the proposed method is more objective and realistic.

Comparison accuracies

It is necessary to investigate whether our method has a better performance than other existing approaches. Nevertheless, it is difficult to compare our results with other published results due to differences in data and experimental protocol. Currently, many on-line available tools have been developed for eukaryotic promoter prediction, which provides a chance to compare the performance of our approach with several methods.

Neural Network Promoter Prediction (NNPP version2.2) (Reese 2001) is a widely used on-line tool for the recognition of eukaryotic promoters. Recently, a successful on-line tool called *McPromoter* (Ohler et al. 2002; Ohler 2006) has been specially developed for *D. melanogaster* promoter prediction based on hidden Markov model (HMM). We compare our method with these two tools on *D. melanogaster* data. Total of 400 randomly selected sequences including 200 promoters, 100 coding sequences, and 100 introns are used to check the prediction accuracy. The results in Table 4 clearly show that the performance of *McPromoter* is better than NNPP2.2, but worse than our method.

We also compare our method with six kinds of promoter prediction tools, namely NNPP2.2 (Reese 2001), TSSW (Bajic et al. 2002), Promoter Scan version 1.7 (Prestridge 1995), Promoter 2.0 (Knudsen 1999), FirstEF (Davuluri et al. 2001) and Eponine (Down and Hubbard 2002). Total of 400 randomly selected *H. sapiens* sequences including 200 promoters, 100 coding sequences, and 100 introns are used to estimate the predictive accuracies. Table 5 reveals that the IPMD method has a better performance than other tools. It should be noted that the sensitivities of both Promoter 2.0 and FirstEF are 0.0%. The possible reason is that the default cutoff values of two tools are too deflective to detect true positives.

However, some methods are not currently available on-line. Thus, it is not feasible to compare our method with these methods using same sequences. We are only able to give a rough comparison between our method and other methods. Gangal and Sharma have presented a SVM method using non-linear time series descriptors to discriminate between *H. sapiens* promoters and non-promoters (Gangal and Sharma 2005). Ten-fold cross-validated accuracy of 87% was obtained. Another work has described a fisher discriminant algorithm to predict *H. sapiens* promoters (Yang et al. 2008). The overall accuracy of ~89% was achieved. Our 10-fold cross-validated accuracy is 92.1% for *H. sapiens*. Recently, by use of promoters and CDSs as benchmark data set, a SVM-based method was developed to discriminate promoter sequences from non-promoter sequences (Anwar et al. 2008). The overall accuracies of 94.82% for *D. melanogaster* and 91.25% for

Table 4 Comparison of *D. melanogaster* promoter prediction

Algorithm	NNPP2.2 (0.95)	NNPP2.2 (0.99)	McPromoter (0.8) (one model)	McPromoter (0.03) (five model)	IPMD
S_n (%)	35.0	23.0	59.0	65.5	78.5
S_p (%)	83.5	96.0	95.5	99.5	97.0
CC	0.21	0.28	0.59	0.69	0.77

Table 5 Comparison of *H. sapiens* promoter prediction

Promoter programs	S_n (%)	S_p (%)	CC
NNPP2.2 (threshold 0.8)	52.5	77.5	0.310
TSSW	67.0	93.5	0.627
Promoter Scan 1.7	43.5	97.0	0.479
Promoter 2.0	0	100	Infinite
FirstEF	0	100	Infinite
Eponine (threshold 0.999)	29.0	97.5	0.364
IPMD	88.5	95.0	0.837

H. sapiens were achieved using 7-fold cross-validation. It must be noted that intron was not considered here for prediction. It is known that it is more difficult for the discrimination between promoters and introns as they are both non-coding sequences. For the comparison, as was done by Anwar et al. (2008), promoters and coding sequences are used to examine IPMD method. Overall accuracies of 96.52% for *D. melanogaster* and 92.67% for *H. sapiens* are achieved, respectively, using 10-fold cross-validation. The comparison suggests that our approach yields superior performance.

Many methods have been applied in prokaryote promoter prediction. Gordon et al. (2003) have proposed a SVM-based method to recognize *E. coli* σ^{70}-promoters. The sensitivities and correlation coefficients are 82% and 0.67 for promoters versus coding sequences as well as 81% and 0.63 for promoters versus non-coding sequences, respectively. Huerta and Collado-Vides (2003) have used a two-stage PWM method to identify the promoters on 250-bp long regions upstream of gene starts. The cover function of 86% was achieved with the accuracy of 53%. Burden et al. (2005) have obtained the sensitivity of 64.1% with 3939 predicted sequences by combining the distance from TSS to translation initiation site (DGS) with NNPP2.2. Another article has developed a more successful approach which used DGS in conjunction with PWM and SVM to recognize *E. coli* promoters (Gordon et al. 2006). The area under the detection error trade-off (DET) curve was 0.047. Later, based on stress-induced DNA duplex destabilization (SIDD) properties and −10 motif scores, an accuracy of >82% was achieved using linear classification function (Wang and Benham 2006). Rani et al. (2007) have achieved an accuracy of 80% for *E. coli* promoter prediction using dinucleotide feature and neural network. Our former work has achieved the sensitivities and correlation coefficients of 91% and 0.68 for promoters versus coding sequences as well as 90% and 0.65 for promoters versus non-coding sequences by use of PCSF algorithm (Li and Lin 2006). Recently, based on relative stability of DNA, Rangannan and Bansal (2007, 2009) have obtained the sensitivities and precisions of 99 and 58% for *E. coli* as well as 95 and 60% for *B. subtilis*, respectively.

For making a rough comparison, promoters versus coding sequences and promoters versus non-coding sequences are, respectively, used to train and test IPMD model. We achieve the sensitivities and correlation coefficients of 94.5% and 0.844 for promoters versus coding sequences as well as 82.7% and 0.728 for promoters versus non-coding sequences using 10-fold cross-validation. Comparisons demonstrate that IPMD method can correctly detect prokaryotic promoters.

Conclusion

In this article, we develop an efficient approach for eukaryotic and prokaryotic promoter predictions. Five species promoters are used to evaluate the performance of IPMD method. And high predictive accuracies are achieved. Although this method exhibits good performance on the aspect of promoter prediction, there is still large space to improve prediction accuracy. The current researches can be considered as the draft of promoter annotation. The future work will focus on DNA structural information and complete genome prediction. This approach also may play an important complementary role to other existing methods for predicting promoters and transcription start sites.

Acknowledgments The authors are grateful to the anonymous reviewers for their valuable suggestions and comments, which have led to the improvement of this article. This study was supported in part by the Fundamental Research Funds for the Central Universities (ZYGX2009J081) and the National Natural Science Foundation of China (61063016).

References

Abeel T, Saeys Y, Bonnet E, Rouzé P, Van de Peer Y (2008a) Generic eukaryotic core promoter prediction using structural features of DNA. Genome Res 18:310–323

Abeel T, Saeys Y, Rouzé P, van de Peer Y (2008b) ProSOM: core promoter prediction based on unsupervised clustering of DNA physical profiles. Bioinformatics 24:i24–i31

Aerts S, Thijs G, Dabrowski M, Moreau Y, Moor BD (2004) Comprehensive analysis of base composition around the transcription start site in Metazoa. BMC Genomics 5:34

Akan P, Deloukas P (2008) DNA sequence and structural properties as predictors of human and mouse promoters. Gene 410:165–176

Anwar F, Baker SM, Jabid T, Mehedi Hasan M, Shoyaib M, Khan H, Walshe R (2008) pol II promoter prediction using characteristic 4-mer motifs: a machine learning approach. BMC Bioinformatics 9:414

Bajic VB, Seah SH, Chong A, Zhang G, Koh JL, Brusic V (2002) Dragon promoter finder: recognition of vertebrate RNA polymerase II promoters. Bioinformatics 18:198–199

Bajic VB, Choudhary V, Hock CK (2004) Content analysis of the core promoter region of human genes. In Silico Biol 4:109–125

Burden S, Lin YX, Zhang R (2005) Improving promoter prediction for the NNPP2.2 algorithm: a case study using *E. Coli* DNA sequences. Bioinformatics 21:601–607

Chan B, Kibler D (2005) Using hexamers to predict cis-regulatory motifs in Drosophila. BMC Bioinformatics 6:262

Chou KC (1995) A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins 21:319–344

Chou KC, Liu WM, Maggiora GM, Zhang CT (1998) Prediction and classification of domain structural classes. Proteins 31:97–103

Davuluri RV, Grosse I, Zhang MQ (2001) Computational identification of promoters and first exons in the human genome. Nat Genet 29:412–417

Down TA, Hubbard TJ (2002) Computational detection and location of transcription start sites in mammalian genomic DNA. Genome Res 12:458–461

Feng Y, Luo L (2008) Use of tetrapeptide signals for protein secondary-structure prediction. Amino Acids 35:607–614

Gangal R, Sharma P (2005) Human pol II promoter prediction: time series descriptors and machine learning. Nucleic Acids Res 33:1332–1336

Goni JR, Pere A, Torrents D, Orozco M (2007) Determining promoter location based on DNA structure first-principles calculations. Genome Biol 8:R263

Gordon L, Chervonenkis AY, Gammerman AJ, Shahmuradov LA, Solovyev VV (2003) Sequence alignment kernel for recognition of promoter regions. Bioinformatics 19:1964–1971

Gordon JJ, Towsey MW, Hogan JM, Mathews SA, Timms P (2006) Improved prediction of bacterial transcription start sites. Bioinformatics 22:142–148

Grech B, Maetschke S, Mathews S, Timms P (2007) Genome-wide analysis of chlamydiae for promoters that phylogenetically footprint. Res Microbiol 158:685–693

Grech B, Mathews S, Timms P (2008) Phylogenetic comparison of the known Chlamydia trachomatis σ^{66} promoters across to *Chlamydia pneumoniae* and *Chlamydia caviae* identifies seven poorly conserved promoters. Res Microbiol 159:550–556

Hawley DK, McClure WR (1983) Compilation and analysis of *Escherichia coli* promoter DNA sequences. Nucleic Acids Res 11:2237–2255

Horton PB, Kanehisa M (1992) An assessment of neural network and statistical approaches for prediction of *E. coli* promoter sites. Nucleic Acids Res 20:4331–4338

Huerta AM, Collado–Vides J (2003) Sigma70 promoters in *Escherichia coli*: specific transcription in dense regions of overlapping promoter-like signals. J Mol Biol 333:261–278

Hutchinson G (1996) The prediction of vertebrate promoter regions using differential hexamer frequency analysis. Bioinformatics 12:391–398

Janky R, van Helden J (2008) Evaluation of phylogenetic footprint discovery for predicting bacterial cis-regulatory elements and revealing their evolution. BMC Bioinformatics 9:37

Kielbasa SM, Gonze D, Herzel H (2005) Measuring similarities between transcription factor binding sites. BMC Bioinformatics 6:237

Knudsen S (1999) Promoter2.0: for the recognition of pol II promoter sequences. Bioinformatics 15:356–361

Laxton RR (1978) The measure of diversity. J Theor Biol 70:51–67

Levitsky VG, Katokhin AV (2003) Recognition of eukaryotic promoters using a genetic algorithm based on iterative discriminant analysis. In Silico Biol 3:81–87

Li QZ, Lin H (2006) The recognition and prediction of σ^{70} promoters in *Escherichia coli* K–12. J Theor Biol 242:135–141

Mahdi RN, Rouchka EC (2009) RBF–TSS: identification of transcription start site in human using radial basis functions network and oligonucleotide positional frequencies. PLoS One 4:e4878

Makita Y, Nakao M, Ogasawara N, Nakai K (2004) DBTBS: database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics. Nucleic Acids Res 1:D75–D77

Ohler U (2006) Identification of core promoter modules in Drosophila and their application in accurate transcription start site prediction. Nucleic Acids Res 34:5943–5950

Ohler U, Harbeck S, Niemann H, Noth E, Reese MG (1999) Interpolated Markov chains for eukaryotic promoter recognition. Bioinformatics 15:363–369

Ohler U, Niemann H, Liao GC, Rubin GM (2001) Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. Bioinformatics 17:S199–S206

Ohler U, Liao GC, Niemann H, Rubin GM (2002) Computational analysis of core promoters in the Drosophila genome. Genome Biol 3:RESEARCH0087

Pedersen AG, Engelbrecht J (1995) Investigations of Escherichia coli promoter sequences with artificial neural networks: new signals discovered upstream of the transcriptional startpoint. Proc Int Conf Intell Syst Mol Biol 3:292–299

Pedersen AG, Baldi P, Brunak S, Chauvin Y (1996) Characterization of prokaryotic and eukaryotic promoters using Hidden Markov models. Proc Int Conf Intell Syst Mol Biol 4:182–191

Pedersen AG, Baldi P, Brunak S (1999) The biology of eukaryotic promoter prediction—a review. Comput Chem 23:191–207

Ponger L, Mouchiroud D (2002) CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. Bioinformatics 18:631–633

Prestridge DS (1995) Predicting pol II promoter sequences using transcription factor binding sites. J Mol Biol 249:923–932

Rangannan V, Bansal M (2007) Identification and annotation of promoter regions in microbial genome sequences on the basis of DNA stability. J Biosci 32:851–862

Rangannan V, Bansal M (2009) Relative stability of DNA as a generic criterion for promoter prediction: whole genome annotation of microbial genomes with varying nucleotide base composition. Mol Biosyst 5:1758–1769

Rani TS, Bhavani SD, Bapi RS (2007) Analysis of *E. coli* promoter recognition problem in dinucleotide feature space. Bioinformatics 23:582–588

Reese MG (2001) Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome. Comput Chem 26:51–56

Salgado H, Gama-Castro S, Martinez-Antonio A, Diaz-Peredo E, Sanchez-Solano F, Peralta-Gil M, Garcia-Alonso D, Jimenez-Jacinto V, Santos-Zavaleta A, Bonavides-Martinez C, Collado-Vides J (2004) RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K–12. Nucleic Acids Res 32:D303–D306

Satija R, Pachter L, Hein J (2008) Combining statistical alignment and phylogenetic footprinting to detect regulatory elements. Bioinformatics 24:1236–1242

Schmid CD, Perier R, Praz V, Bucher P (2006) EPD in its twentieth year: towards complete promoter coverage of selected model organisms. Nucleic Acids Res 34:D82–D85

Shahmuradov IA, Solovyev VV, Gammerman AJ (2005) Plant promoter prediction with confidence estimation. Nucleic Acids Res 33:1069–1076

Shepelev V, Fedorov A (2006) Advances in the exon–intron database (EID). Brief Bioinform 7:178–185

Solovyev VV, Shahmuradov IA (2003) PromH: promoters identification using orthologous genomic sequences. Nucleic Acids Res 31:3540–3545

Sonnenburg S, Zien A, Ratsch G (2006) ARTS: accurate recognition of transcription starts in human. Bioinformatics 22:e472–e480

Wang HQ, Benham CJ (2006) Promoter prediction and annotation of microbial genomes based on DNA sequence and structural responses to superhelical stress. BMC Bioinformatics 7:248

Wasserman WW, Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet 5:276–287

Yang JY, Zhou Y, Yu ZG, Anh V, Zhou LQ (2008) Human pol II promoter recognition based on primary sequences and free energy of dinucleotides. BMC Bioinformatics 9:113

Zhang MQ (1997) Identification of protein coding regions in the human genome by quadratic discriminant analysis. Proc Natl Acad Sci USA 94:565–568

Zhang MQ (2005) Using CorePromoter to find human core promoters. Curr Protoc Bioinformatics Chapter 2: Unit 2.9

Zhang LR, Luo LF (2003) Splice site prediction with quadratic discriminant analysis using diversity measure. Nucleic Acids Res 31:6214–6220

Zhang X, Kassim A, Bajic VB (2004) Digital signal processing for potential promoter. In: IEEE international workshop on biomedical circuit and systems, pp S2/7/INV–S2/16-19