Genomics 97 (2011) 112–120

Contents lists available at ScienceDirect

Genomics

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

Identification of TATA and TATA-less promoters in plant genomes by integrating diversity measure, GC-Skew and DNA geometric flexibility

Yong-Chun Zuo, Qian-Zhong Li *

Laboratory of Theoretical Biophysics, School of Physical Science and Technology, Inner Mongolia University, Hohhot, 010021, China

ARTICLE INFO

Article history:
Received 9 March 2010
Accepted 12 November 2010
Available online 26 November 2010

Keywords:
Plant promoter
Support vector machine
GC-compositional bias
DNA structural pattern
Correlative conservation

ABSTRACT

Accurate identification of core promoters is important for gaining more insight about the understanding of the eukaryotic transcription regulation. In this study, the authors focused on the biologically realistic promoter prediction of plant genomes. By analyzing the correlative conservation, GC-compositional bias and specific structural patterns of TATA and TATA-less promoters in PlantPromDB, a hybrid multi-feature approach based on support vector machine (SVM) for predicting the two types of promoters were developed by integrating local word content, GC-Skew and DNA geometric flexibility. Compared with the TSSP-TCM program on the same test dataset, better prediction results were obtained. Especially for the TATA-less promoter, the accuracy is 10% higher than the result of TSSP-TCM program. The good performance of the hybrid promoters and the experimental data also indicate that our method has the ability to locate the promoter region of the plant genome.

1. Introduction

A promoter is usually defined as a function region located immediately upstream the transcription start site (TSS), which usually contains some specific DNA sequences and regulatory elements. The main function of a promoter is integrating information about the status of the cell, determining where the transcription of a gene should be initiated and on what condition. Therefore, promoter recognition can be considered as the fundamental problem in finding genes and elucidating the regulation of gene expression.

In the past decades, a number of methods for promoter prediction have been developed [1–4]. These methods are mainly based on DNA sequence properties, and they can be classified into three groups: (1) Search-by-signals algorithms make prediction by detecting the TATA-box, CAAT-box and the Initiator core promoter elements, or by finding transcription factor binding sites (TFBs). Promoter 2.0 [5] and NNPP 2.2 [6] belong to this category. However, the recent studies showed that the regulatory elements around the TSSs are more diverse than previously estimated [7–9]. Only applying some specific known signal motifs cannot exactly predict the promoter and often introduces many false positives. (2) Based on the difference in local word composition between the regulatory and non-regulatory DNA regions, search-by-content algorithms for identifying promoter region have been proposed, such as CorePromoter [10], Promoter–Inspector [11], Dragon Promoter Finder [12], Bayesian networks [13], Promoter Explorer [14] and Relative entropy [15]. Among these methods, the frequency of pentamer and hexamer were used mostly [14–16]. (3) The specific structure properties of DNA, such as DNA denaturation, DNA bending stiffness, propeller twist and duplex disrupt energy, are also important in terms of biological understanding of the transcription initiation mechanisms [17]. Therefore, the predicting methods for incorporating structure properties have been developed quickly in the past few years [18–21].

Although a large amount of genomic and full-length cDNA sequence data for plants have been now publicly available [22,23], the knowledge of the plant promoter is still limited. The TSSP-TCM program is the first program for recognizing the plant TATA and TATA-less promoters by combining the confidence estimations, and the prediction accuracies are 93.3% and 82.3%, respectively [24]. It has been reported that the plant promoters usually show more pronounced curvature and higher information content than the non-promoter regions [25]. In addition, there are many local distributions of short sequences (LDSSs) of hexamer and octamer segments in the promoter regions of *Arabidopsis thaliana* and rice [26,27]. Based on their localization patterns, these LDSSs can be classified into three groups: pyrimidine patch (Y Patch), TATA-box and regulatory element group (REG).

In fact, the sequence-dependent DNA flexibility originated from DNA 3D structure plays an important role in guiding transcription factors to the target site in transcription initiation [28–30], and compared with other regulatory regions of promoters, the positions of TATA and INR boxes contain more distinct flexible sequences. In order to reveal the biological meaning of effective features and improve the computational prediction accuracy, we analyzed the correlative

* Corresponding author. Fax: +86 471 4951761.
E-mail addresses: yczuo@imu.edu.cn (Y.-C. Zuo), qzli@imu.edu.cn (Q.-Z. Li).

0888-7543/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2010.11.002

conservation of specific sites, GC-compositional bias (GC-Skew) and the profiles of DNA geometric properties for TATA promoter and TATA-less promoter in this paper. A new approach combined with signal features, content features, and DNA flexibility features was developed to recognize TATA and TATA-less promoters. The best average accuracies (*Accs*) for TATA and TATA-less promoters are 97.3% and 91.4%, respectively, and the *Accs* for the hybrid promoters are 96.5%, 95.4% and 87.3% by using the coding, intron and intergenic regions as negative examples, respectively. For the available experimental data, our method also gave a very clear indication of where transcription of the gene begins.

2. Result and discussion

2.1. Conservation analysis and distribution of TATA-box segments

In order to find the positional conservation property of the different promoter types, we applied the $M_k(l)$ to calculate the conservation degree of the segment at different positions. The conservation of sequence segment with length k at the site l can be defined as follows [31]:

$$M_k(l) = \sum_k \left[P(i,l) - 1/4^k\right]^2 \Big/ \left(1/4^k\right) \quad (1)$$

where k denotes the segment length, and $P(i,l)$ denotes the probability of i-th segment with length k at the site l. For a random sequence, M_k (l) equals to zero. The larger the value of M_k (l) is, the stronger conservation of the l-th site there is. For example, if $k=1$, M_k (l) shows the conservation of four bases (A, T, G and C) at l-th position.

The correlative conservation of segment with different lengths in regulatory region for 175 TATA promoters and 130 TATA-less promoters, M_K (l) value, was calculated by Eq. (1). By analyzing the position conservation of different segments with k-bp length, we concluded that the positional conservation of trinucleotide segments at the TATA-box region is more conservative than at other sites (shown in Fig. 1A). The regular expression is CTATA[AT]A[TA]A[CG][CA]. Since the TATA-box localization is correlated with the tissue specificity of the downstream transcript, we analyzed the localization distribution. The results showed that the preferential positions of TATA-box are constrained to a narrow window [−36 bp, −28 bp], the −32 bp upstream the TSS is the optimal position for plant promoter (Fig. 1A). In addition, the TATA-box localizations of mammalian core promoter are constrained to more narrow window [−32 bp, −29 bp] [32,33]. The presence of TATA elements represent more frequently in EST collections and the correlation genes usually contain shorter 5′ untranslated regions (5′ UTRs) [33,34]. The wide localizations of the TATA-box allow the plant species to achieve higher tissue specificity of transcription. For the 130 TATA-less promoters, there is no significant conservation site in the whole transcription initiation regions. However, about 35% TATA-less genes contain the TC-element. As shown in Fig. 1B, the TC-element microsatellites are found preferentially located in the basal transcription initiation region (from −50 bp to +50 bp). There may be the reason why the conservative degree of correlation position calculated by $M_K(l)$ score is not as outstanding as the TATA promoter. The latest study reported that the TC-elements might constitute a class of novel regulatory elements participating towards the complex modulation of gene expression in plants [35].

2.2. The GC/AT-Skew in up-/downstream region of the TSSs for plant promoter

The GC-Skew of the proximal region around the transcription start sites (TSSs) of TATA promoter in Fig. 2A showed prominent GC-compositional strand bias and the peaked value at the TSSs. The GC-Skew decrease quickly downstream the TSS. Moreover, there are two GC-Skew peaks in the TATA promoter at −35 bp and +5 bp around the TSS, respectively. The GC-Skew values are larger in these two positions. The number of C residues is twice as much as the G residues.

These trends in core promoter regions can be explained by the transcription-coupled effects of the DNA strand asymmetry [36]. The DNA transcriptions enhance the Cytosine deamination. The DNA unwinds at the TSS, which produces a transcription bubble during binding of the transcription complex in the transcription factor active region (especially for TATA-box region). The bubble makes both strands prone to C–T mutations, but the RNA polymerase preferentially protects the nucleotides on the non-transcribed strand. Hence the C–T mutations occur more frequently on the transcribed strand of TATA-box gene. The AT-Skew shows no significant bias as the GC-Skew, only existing two small peaks at −25 bp and +5 bp. The results in Fig. 2B suggested that the TATA-less promoter also has GC-Skew bias, but the peak is not as prominent as the TATA promoter.

2.3. Structure profiles of DNA geometric flexibility in different genome regions

Promoter sequences usually possess some distinct structural ability to allow functional protein binding and nucleosome

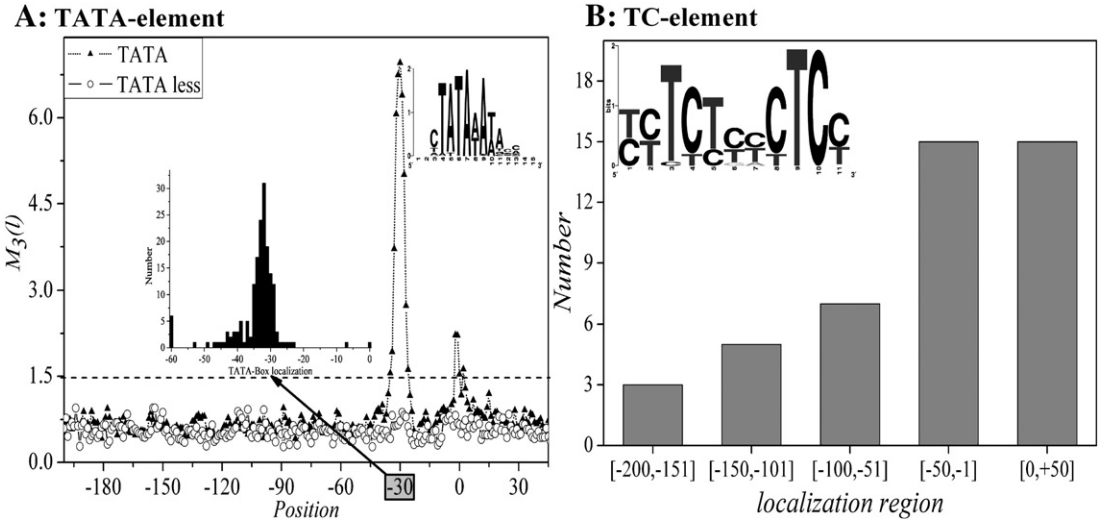

Fig. 1. The graph of M_3 (l) values and the localizations of conservative segments for plant promoters.

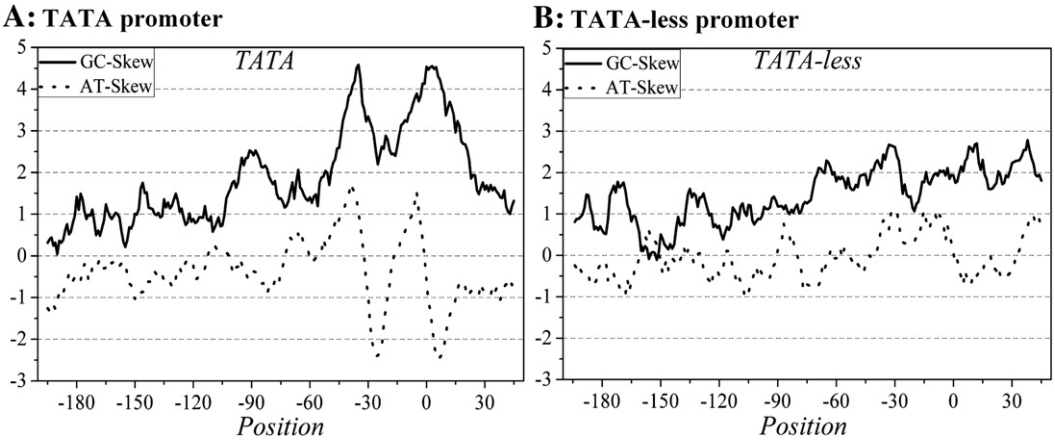

Fig. 2. The GC/AT-Skew in up-/downstream regions of the TSS for plant promoter.

positioning. Hence it is necessary to investigate promoter sequences in higher dimension. The DNA 3D structure can be characterized by three local angular parameters (Twist, Roll and Tilt) and three translational parameters (Shift, Slide and Rise) between two successive base-pair steps. According to the recent *ab initio* quantum mechanical calculations, the 16 DNA geometric flexibility values of dinucleotides had been calculated by Goñi et al. based on the long atomistic molecular dynamics (MD) simulations in water [29]. Using these experimental parameters of nearest-neighbor dinucleotides based on the structural models (Table 1), we converted the core promoter sequence into a string of numerical values for studying the structural flexibility. Here the sliding window approach was used to analyze the DNA geometric flexibility, and the flexibility profiles of the different DNA regions are shown in Fig. 3.

We can find that the Slide profiles of core promoter sequences show striking similarities, and both of them have the coherent shapes of profile, rising and falling nearly synchronously. The TATA-less promoter contains more valley regions, such as −160 bp, −130 bp, −100 bp, −15 bp and +10 bp. This may indicate that the transcription initiation of TATA-less promoter requires more complex structural mechanism for binding more transcription factors (TFs). The Rise profile of TATA promoter contains two significant valley regions [37], but this valley vanishes in the plant TATA-less promoter, which confirms that the valley is mainly due to the characteristic of TATA-box motif in this region. For the Shift profile, there is only one clear valley at the −35 bp site upstream of TSS in the TATA promoter. For the structural profiles of three rotation angles, we observed that the structure of Twist contains similar profiles with the Slide of the displacement structure (Fig. 3). The TATA-less promoter has more valley regions than the TATA promoter. For the Tilt profiles, both the TATA and TATA-less promoters have larger values than the non-promoter, except for the TATA valley region. Besides the TATA valley region, the structural patterns of Roll profile between the promoter and non-promoter sequences become more ambiguous and noisy than other structural features. For the different pattern profiles, we thought that the effect of the TATA-box signal plays an important role for the clear valley/peaks of plant TATA gene. The widespread motifs throughout the TATA-less genes cause the geometric flexibility more complicated.

Further, the intergenic sequences randomly abstracted from The *Arabidopsis* Information Resource (TAIR) [38] were used as another control group to parse out the specificity of the different regulatory regions. As shown in Fig. 3, we could clearly observe that the structure profiles of promoter region are specific to the intergenic region. Most of the profile shapes of the physical-geometry structure contain two clear valleys at −35 bp position (TATA-binding protein localization) and 0 bp position (TSS). The first region is crucial for allowing functional protein binding events and the assembly of the transcription machinery, and the second one is located around the initiator itself, which needs to denature for starting the transcription [39]. For the intron sequence, the profiles of Rise and Twist descriptors rise dips very substantially at the left hand side. It is well known that there are many regulatory elements around the splicing site. We thought the base structure preferred to Rise and Twist in the core splicing regions. The six variables of geometric structure demonstrate the 3D structure of DNA possess a degree of anisotropic flexibility. In conclusion, as the structural patterns depict in Fig. 3, the promoter region has the distinctive structural patterns along the sequences comprise with the non-promoter regions. These inherent structures may be specific to the interaction between proteins and DNA, and essential for transcription initiation. Moreover, these distinctive values of DNA geometric flexibility are also very useful for promoter identification.

Table 1
The 16 different dinucleotide values of DNA geometric flexibility.

Dinucleotides	The six geometric degrees of freedom					
	Twist	Tilt	Roll	Shift	Slide	Rise
AA/TT	0.026	0.038	0.02	1.69	2.26	7.65
AC/GT	0.036	0.038	0.023	1.32	3.03	8.93
AG/CT	0.031	0.037	0.019	1.46	2.03	7.08
AT	0.033	0.036	0.022	1.03	3.83	9.07
CA/TG	0.016	0.025	0.017	1.07	1.78	6.38
CC/GG	0.026	0.042	0.019	1.43	1.65	8.04
CG	0.014	0.026	0.016	1.08	2	6.23
GA/TC	0.025	0.038	0.02	1.32	1.93	8.56
GC	0.025	0.036	0.026	1.2	2.61	9.53
TA	0.017	0.018	0.016	0.72	1.2	6.23

Parameter values related to rotational parameters are in kcal/mol degree^2, while those related to translations are in kcal/mol Å^2.

2.4. The prediction results of TATA and TATA-less types evaluated on the large negative test datasets

Because the sizes of the test sets are small and the false positive is one of the greatest hurdles facing promoter identification, the prediction evaluation is needed to be assessed on the large negative datasets. For each validation, the complete dataset of 175 TATA promoter sequences was randomly divided into 140 training sequences (80%) and 35 testing sequences (20%). To verify the reliability of our method, the total negative datasets contain 10,000 CDSs, 10,000 intergenic sequences and 10,000 introns. Different 1000 non-promoters were randomly chosen from the negative dataset each time, and 140 sequences were chosen as the training negative set (the same as the training positive set) and the remaining 860 sequences were selected as the testing negative set (greater than the testing positive set). The number ratio of the testing positives to testing negatives is up to 1:25. Such 10 uncorrelated cross-validation sets were generated. For the TATA-less promoter prediction,

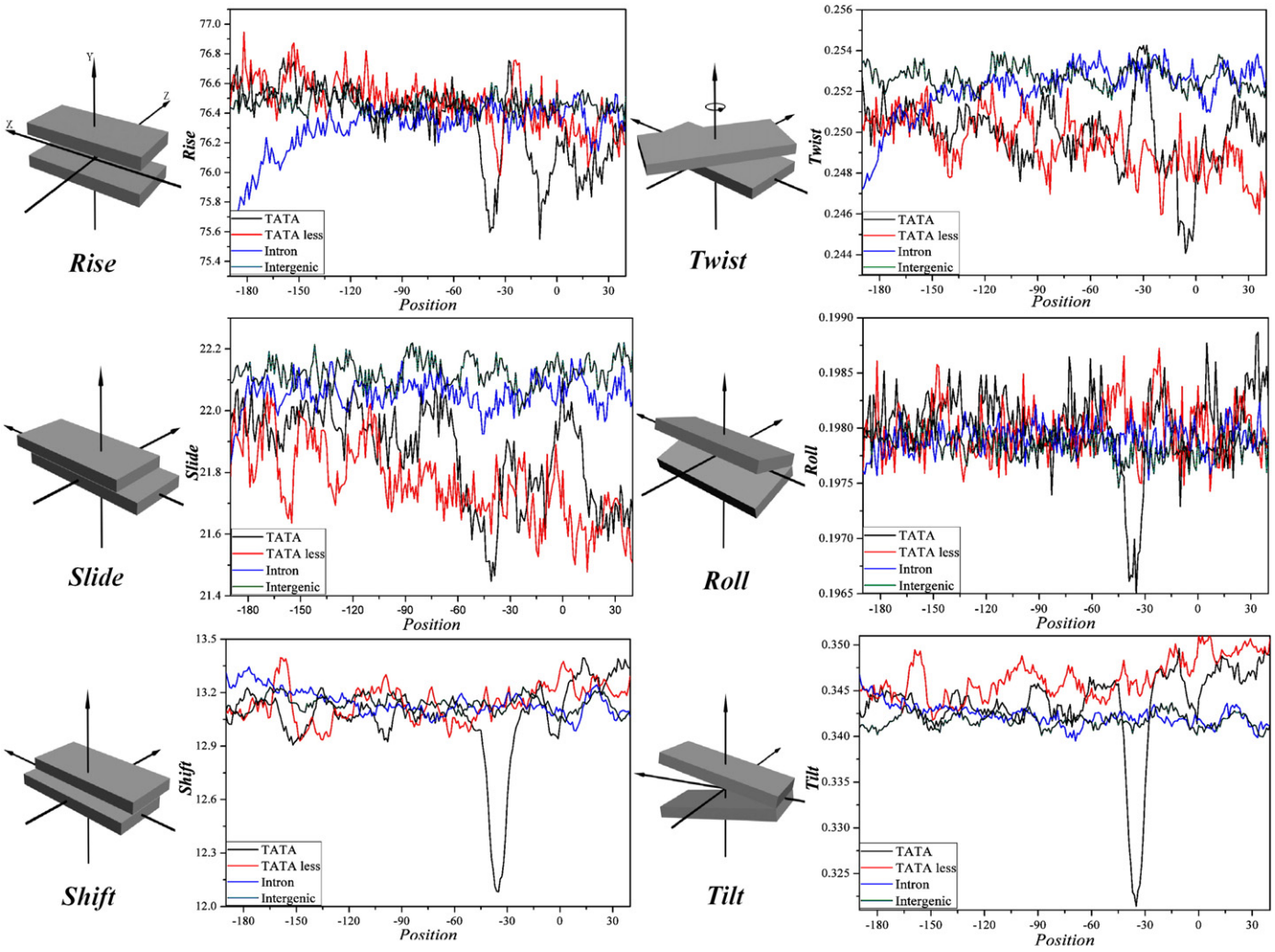

Fig. 3. The pattern profiles of physical-geometry structure for different sequence types based on structural model (with a step of 1 and a windows size of 10).

similar validations as TATA promoter sequences were generated and the number ratio of the testing positives to testing negatives is up to 1:36. The 24 values shown in Table 2 were selected as the input parameters of SVM, which contains signal feature, content feature, GC-Skew and DNA geometric flexibility feature. The flowchart of the experiment is described in Fig. 4 and the 10 cross-validation results are shown in Table 3.

To investigate the best type of local word for promoter prediction, the evaluation was done by using various values of *K*-mer (from 3 to 6). The prediction results based on combined multi-features are shown in Table 3. For different *K* values of subsequence, it is shown that the prediction ability is improving with *K* increase, up to the best performance when $K=5$. The reason may be that the increasing length of segment will also increase the dimension of parameter space, and most segments will not appear simultaneously in a training-unknown input sequence because the sequence length is limited. Moreover, computation becomes not only impractical but also susceptible to danger of over fitting. The best sensitivities for the TATA promoter are 96.3%, 93.4% and 84.7% by using CDSs, introns and intergenic sequences as the non-promoter training sets, respectively, and the specificities are 98.3%, 97.3% and 96.7%, respectively. It indicates that our method has the ability to reduce the rate of false positive effectively, and for TATA-less promoter, the best sensitivities are 90.5%, 81.3% and 70.8% by using CDSs, introns and intergenic as the negative training sets, respectively, and the specificities are 92.3%, 91.6% and 91.2%, respectively. For both TATA and TATA-less promoters, all of the specificities are more than 91.0% when experimented on large negative datasets.

2.5. Comparison with the TSSP-TCM program

Shahmuradov et al. firstly divided the plant promoter sequences into two categories: the TATA promoters and the TATA-less promoters. They predicted these two categories by using TSSP-TCM program [24]. In order to verify our method, we compared our method with TSSP-TCM methods based on the same dataset validation. The comparative results of the same cross validation for the sensitivity (*Sn*), specificity (*Sp*) and accuracy (*Acc*) for the TATA and TATA-less promoters are shown in Table 4.

For the TATA promoter, the results showed that the average prediction rates (*Sn*) are 94.2% and 96.3% by using CDS and introns as negative examples, respectively, and the specificities (*Sp*) are 98.3% and 95.7%. All of them are higher than the TSSP-TCM program. For the TATA-less promoter, the average prediction rates (*Sn*) are 92.8% and 87.5% by using CDS and introns as negative examples, respectively, and the specificities (*Sp*) are 93.1% and 92.7%. For recognition of TATA-less promoter, we could observe that the accuracy of our method is 10% better than the TSSP-TCM method by using both CDS and introns as negative examples.

2.6. The prediction performance of hybrid promoters among different types of feature extraction

To discuss how a particular type of selected features affects on the prediction algorithm, our method was used to evaluate on the whole hybrid promoters (175 TATA promoters and 130 TATA-less

Table 2
The input parameters of SVM for promoter prediction used in this study.

Feature type	Num	Promoter type	
		TATA	TATA-less
Signal feature	*KPCS* value	The *KPCS* value of 3-mer at 12 sites (−34,..., −26, −2, −1, and 0)	–
Content diversity feature	$ID1_{pos}$, $ID1_{neg}$	*ID1* value of *K*-mer from −200th nt to −118th nt between *M* and pos/neg set	*ID1* value of *K*-mer from −200th nt to −118th nt between *M* and pos/neg set
	$ID2_{pos}$, $ID2_{neg}$	*ID2* value of *K*-mer from −117th nt to −35th nt between *M* and pos/neg set	*ID2* value of *K*-mer from −117th nt to −35th nt between *M* and pos/neg set
	$ID3_{pos}$, $ID3_{neg}$	*ID3* value of *K*-mer from −34th nt to +50th nt between *M* and pos/neg set	*ID3* value of *K*-mer from −34th nt to +50th nt between *M* and pos/neg set
	$ID4_{pos}$, $ID4_{neg}$	*ID4* value of *K*-mer from −200th nt to −151th nt between *M* and pos/neg set	*ID4* value of *K*-mer from −200th nt to −151th nt between *M* and pos/neg set
	$ID5_{pos}$, $ID5_{neg}$	*ID5* value of *K*-mer from −150th nt to −101th nt between *M* and pos/neg set	*ID5* value of *K*-mer from −150th nt to −101th nt between *M* and pos/neg set
	$ID6_{pos}$, $ID6_{neg}$	*ID6* value of *K*-mer from −100th nt to −51th nt between *M* and pos/neg set	*ID6* value of *K*-mer from −100th nt to −51th nt between *M* and pos/neg set
	$ID7_{pos}$, $ID7_{neg}$	*ID7* value of *K*-mer from −50th nt to −1th nt between *M* and pos/neg set	*ID7* value of *K*-mer from −50th nt to −1th nt between *M* and pos/neg set
	$ID8_{pos}$, $ID8_{neg}$	*ID8* value of *K*-mer from −0th nt to +50th nt between *M* and pos/neg set	*ID8* value of *K*-mer from −0th nt to +50th nt between *M* and pos/neg set
GC Bias feature	*GC-Skew value*	GC-Skew score from −200th nt to 50th nt	GC-Skew score from −200th nt to 50th nt
DNA geometric flexibility	*Twist value*	Twist value from −150th nt to +30th nt	Twist value from −150th nt to +30th nt
	Tilt value	Tilt value from −50th nt to −40th nt	Tilt value from −160th nt to +30th nt
	Roll value	Roll value from −50th nt to −20th nt	Roll value from −160th nt to +30th nt
	Shift value	Shift value from −60th nt to −20th nt	Shift value from −160th nt to +30th nt
	Slide value	Slide value from −100th nt to +30th nt	Slide value from −200th nt to +30th nt
	Rise value	Rise value from −50th nt to +30th nt	Rise value from −200th nt to −100th nt

The *ID* values of content diversity feature indicate the increment of diversity of *M* with positive (promoter) or negative (non-promoter) set.

promoters). Details of the parameters are depicted in Table 2, and the prediction results of the 10-fold cross validation are shown in Table 5.

From the results shown in Table 5, we could find that the performance of DNA flexibility is better than the signal feature except for discriminating the intergenic sequences. The DNA structure features achieved the higher sensitivities. Among the six DNA flexibility descriptors, the Slide gives the best prediction performance among 3D geometric parameters. It is known that the local word contents have been widely used as the discriminating function in promoter recognition, and many regulatory motifs in promoter region are also constituted by various oligonucleotide fragments. The increment of diversity was applied to improve the efficiency of information extraction and resolve the problem of high-dimensional parameters resulted by the *K*-mer types in different regions. The results in our study also demonstrated that the content diversity gives the best performance in identifying the plant promoter. The content diversity of 5-mer local words in overlap regions get the sensitivity (*Sn*) 94.3% and specificity (*Sp*) 93.7% for discriminating the negatives and the accuracy (*Acc*) is 83.6% when using intergenic sequences as non-promoters. When combining all the 24 feature parameters, the *Accs* are 96.5%, 95.4% and 87.3% by using the coding, intron and intergenic regions as the negative examples, respectively. From the above results, we can be sure to conclude that our SVM program combined with signal feature, content information and distinct structural patterns of DNA geometric flexibility has the ability to locate the promoter region of the plant genome.

2.7. Validation with experimental annotation of Arabidopsis thaliana genome

Some independent identifications were evaluated for testing our method with data derived from the GenBank database. We also assessed the empirical usefulness of our prediction on the genome level. The TSSs annotated in literature and the wide upstream regions were selected as representative cases for this validation. Two examples are TATA-promoter genes (AC: U71080 and AB048395) and the other two examples are TATA-less genes (AC: AF039206 and AF319968). When evaluating on each actual transcription start region, the corresponding promoter would be excluded from the training model. The validation was experimented by using sliding window approach with window size 251 bp and step 10 bp. For each nucleotide in the selected regions, a prediction score was calculated and score profiles were shown as a plot along the genomic sequences (see Fig. 5).

For the TATA promoter, as shown in Fig. 5a, the upstream region of the *Arabidopsis thaliana* cinnamate-4-hydroxylase (atC4H) gene ATU71080 was characterized experimentally [40]. The ATU71080 gene with 5432 bp was located at chromosome 2. It includes several function segments, the upstream region of transcription start site (TSS) (1 bp...2865 bp), the 5′ untranslated region (5′ UTR) (2866 bp...2951 bp), three coding regions (2952 bp...3736 bp; 3822 bp...3955 bp; and 4176 bp...4774 bp) and 3′ untranslated region (3′ UTR) (4775 bp...4928 bp). In scanning process, the TSSs were mapped by using the combined features. The profile in Fig. 5a shows that the transcription initiates from more than 20 continuous scores. It spanned a 200 bp length region, from position 2590 bp to 2790 bp, and the position 2660 bp gave the maximum score (value = 0.98). Because the region searched by our sliding method was starting from −200 bp to +50 bp, the predictive annotation TSS of our method in fact is at 2861 bp site (2660 + 201 = 2861), and the score value of the 2861 bp site equals to 0.98. Thus, there are only five base-pairs that differ from the experimental TSSs (2866 bp). This may be that the window step is 10 bp gap in our sliding prediction. In this case, our method gave a very clear indication of where transcription of this gene begins. When using the cutoff score = 0.6, all of the four promoter regions of *Arabidopsis thaliana* gene could be identified correctly. This demonstrated that our plant promoter predictions strongly agreed with the available experimental data for well studied promoter regions.

3. Conclusion

Using bioinformatics techniques to identify the promoter from a given DNA sequence is one of the active areas in gene transcription regulation. However, false positives are still the main problem on promoter recognition, especially for the TATA-less promoters. The

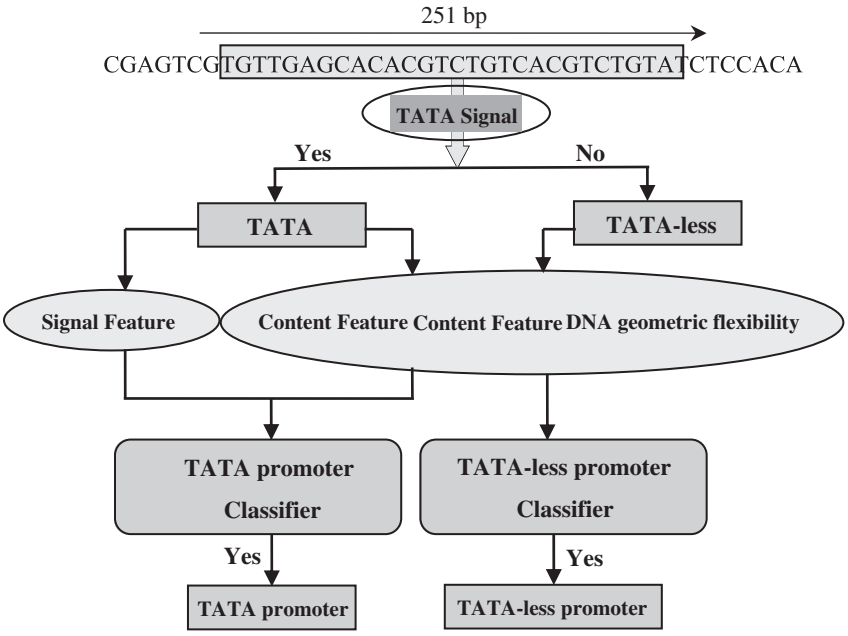

Fig. 4. The flowchart of our method for plant promoter recognition.

fundamental reason may be that all of the existent methods assume the whole genome is accessible for binding, without knowing which regions of chromatin are open or close [41]. It is necessary to incorporate more feature types for promoter prediction. In this paper, signal feature, content feature, GC-Skew and DNA geometric flexibility in the plant promoter regions were discussed, and all of them were combined to promoter recognition. The prediction results showed that our method is helpful for improving the promoter recognition in DNA sequence and useful for detecting the transcription start sites of plant genomes.

4. Material and theoretical algorithm

4.1. Datasets

The 175 TATA and 130 TATA-less plant promoters were downloaded from the PlantPromDB [42]. The database is the annotated non-redundant collection of proximal promoter sequences for RNA polymerase II with experimentally determined transcription start site (TSS) from various plant species. The sequences with 251 bp length from 200 bp upstream to 50 bp downstream of the transcription start sites (TSSs the 0th sites) were selected as the core promoters. The 20,000 sequences with 251 bp length from coding regions (CDS) and 20,000 sequences from the intron regions of plant genes annotated in the GenBank database were extracted as negative datasets (the same training negative sequences as the TSSP-TCM program [24]). The 10,000 intergenic sequences were obtained from The *Arabidopsis* Information Resource (TAIR). The TAIR database maintains the genetic and molecular biology data for the higher model plant genome of *Arabidopsis thaliana* [38].

Table 3
The average results of 10 cross validations for plant promoter.

K-mer	Negative type	TATA			TATA-less		
		Sn (%)	Sp (%)	Ac (%)	Sn (%)	Sp (%)	Ac (%)
3	Vs CDS	92.5	96.7	94.6	88.2	86.4	87.3
	Vs introns	92.2	94.8	93.5	82.7	85.8	84.2
	Vs intergenic	82.3	93.6	88.0	73.8	81.4	77.6
4	Vs CDS	95.6	97.9	96.7	89.4	90.3	89.9
	Vs introns	92.5	96.2	94.4	83.5	84.9	84.2
	Vs intergenic	82.1	92.4	87.3	74.3	82.6	78.5
5	**Vs CDS**	**96.3**	**98.3**	**97.3**	**90.5**	**92.3**	**91.4**
	Vs introns	**93.4**	**97.3**	**95.4**	**81.3**	**91.6**	**86.4**
	Vs intergenic	**84.7**	**96.7**	**90.7**	**70.8**	**91.2**	**81.0**
6	Vs CDS	95.8	96.9	96.4	85.0	94.5	89.8
	Vs introns	85.4	97.4	91.4	74.6	94.3	84.4
	Vs intergenic	73.6	97.9	85.8	45.2	98.7	72.0

The best prediction result was highlighted in bold.

4.2. Increment of diversity (ID)

The diversity measure is one kind of information description on state space and a measure of the whole uncertainty [43]. In order to compare the distribution of two diversities, the increment of diversity (*ID*) was defined by the difference between the total diversity measure of two systems and the diversity measure of the mixed system. The *ID* indexes have been successfully applied to the recognition of protein structural class [44,45], the exon–intron splice site prediction [46] and the subcellular localization of apoptosis protein [47]. A promoter sequence can be denoted by a source of diversity: $\{n_1, n_2, \ldots n_i, \ldots n_S\}$, where $n_1, n_2, \ldots n_S$ are the numbers of information parameters in the primary promoter sequences. Therefore, the diversity measure can be expressed as:

$$D(X) = N \log N - \sum_{i=1}^{S} n_i \log n_i \tag{2}$$

where $N = \sum_{i=1}^{S} n_i$. If n_l equals to zero, then $n_l \log n_l = 0$.

Table 4
The comparison results of our method with TSSP program in the independent dataset.

		Vs CDS		Vs introns	
		TSSP-TCM	Our method	TSSP-TCM	Our method
TATA	Sn (%)	92.6	**94.2**	96.5	**96.3**
	Sp (%)	94.0	**98.3**	91.3	**95.7**
	Ac (%)	93.3	**96.2**	93.9	**96.0**
TATA-less	Sn (%)	81.4	**92.8**	86.0	**87.5**
	Sp (%)	83.1	**93.1**	70.5	**92.7**
	Ac (%)	82.3	**93.0**	78.3	**90.1**

The best prediction result was highlighted in bold.

Table 5
The prediction results of our method for 305 PlanPromoters based on hybrid features.

Feature type	Num	Negative type	Sn	Sp	Ac
TATA-box GC-Skew	2	Vs CDS	71.3	90.3	80.8
		Vs introns	70.1	88.3	79.2
		Vs intergenic	67.9	88.3	78.1
DNA Flexibility	6	Vs CDS	86.9	84.6	85.8
		Vs introns	85.4	83.7	84.6
		Vs intergenic	76.6	73.5	75.1
Content diversity(K = 5)	16	Vs CDS	94.3	93.7	94.0
		Vs introns	92.7	92.7	92.7
		Vs intergenic	81.3	86.0	83.6
Content diversity, TATA-box, GC-Skew	18	Vs CDS	95.4	93.1	94.2
		Vs introns	94.7	91.7	93.2
		Vs Intergenic	83.4	87.1	85.3
All of the above features	24	Vs CDS	97.3	95.7	96.5
		Vs introns	96.1	94.7	95.4
		Vs intergenic	85.9	88.7	87.3

For two sources of diversity in the same space of S dimensions: X: $\{n_1, n_2, \ldots n_i, \ldots n_S\}$ and Y:$\{m_1, m_2, \ldots, m_i, \ldots m_S\}$, the increment of diversity is calculated by the following formula:

$$ID(X,Y) = D(X+Y) - D(X) - D(Y). \quad (3)$$

Here $D(X, Y)$, $D(Y)$ and $D(Y)$ denote the standard diversity measure of source: $X+Y$, X and Y, and they were calculated by Eq. (2), respectively. For an arbitrary promoter sequence M to be predicted, the two increments of diversity (ID_P, ID_N) between the sequence M and the two standard diversities of promoter and non-promoter in training sets were calculated. At last the two ID values were selected as two feature parameters for inputting into the SVM program.

4.3. The dinucleotide flexibility parameters used in this paper

Structure features originated from the DNA 3D structures can describe proximal promoter well. Based on the physical potentials derived from quantum chemical calculations of atomic molecular dynamics (MD) simulations, the Mahalanobis metrics combined with six geometric degrees of freedom have been developed for determining promoter localization by using a well-defined physical description of DNA deformability [29]. These helical stiffness parameters reveal the complexity of the DNA deformation pattern. The prediction results showed that these dinucleotide flexibility parameters can better define a promoter as a region of unique deformation properties, particularly, near TATA-box and TSS regions. The six physical structure parameters include three angular variables called Tilt, Roll, Twist and three distance variables called Shift, Slide and Rise. The 16 different nearest-neighbor interaction values for the six flexibility parameters are shown in Table 1.

4.4. K-mer position correlation score (KPCS)

Position weight matrices (PWM) have been successfully applied to represent the nucleotide sequence signal of the regulatory elements [48–50]. Besides the base composition, the evolution of base correlation deserves more attention. The DNA sequence has an

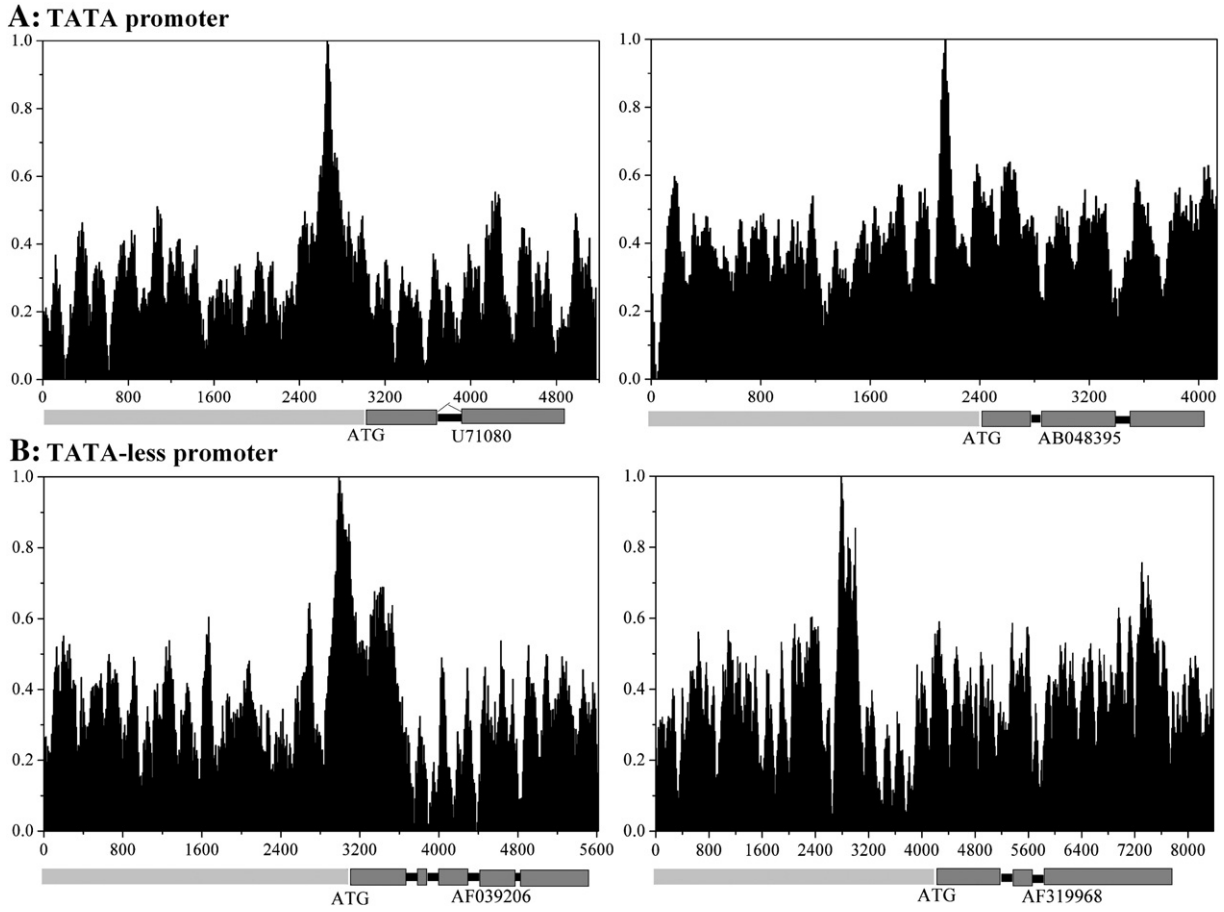

Fig. 5. The promoter identification in genome annotations with independent experimental data of *Arabidopsis thaliana* gene. (a) The two examples of TATA-box contained gene, U71080 and AB048395, respectively; (b) The two examples of TATA-less gene, AF039206 and AF319968, respectively. The profile was made using a window size of 251 bp and a step size of 10 bp. Experimentally annotated genes from GenBank are depicted at the bottom of each prediction profile. Peaks in the profile that exceed the classification threshold (score = 0.6) are predicted as promoter regions.

essential feature of short-range dominance of base correlations. For the plant genome, many local distributions of short sequences (LDSSs) have been identified by Yamamoto et al. firstly in the promoter regions of *Arabidopsis thaliana* and rice [26,27]. Thus the position-correlation weight matrix (PCWM) was established to describe the correlation characteristic of some specific positions. The PCWM method has been successfully applied to the recognition of prokaryotic promoter [31]. In our previous study, the PCWM scanning model showed better performance than the PWM model for discriminating the TATA and TATA-less promoters [51]. The position-correlation weight matrix of the segments with length k is defined as follows:

$$P_k(i,l) = \left(f_{i,l} + s_i\right) / \left(N_l + \sum_{k_i} s_i\right) \tag{4}$$

where l denotes the l-th site and i denotes the i-th segment. When $k=2$, the i is one of the 16 dinucleotides. $f_{i,l}$ is the real counts of the i-th segment in site l. N_l and S_i are the total numbers of real counts and pseudocount in site l, respectively. Here S_i equals to $\sqrt{N_l}/16$.

The K-mer position correlation score ($KPCS$) is defined as follows:

$$KPCS = \sum_{i}^{n} ln(P_K(i,l) / P_0) \tag{5}$$

where the P_0 is the average background frequency of each K-mer; the value of $KPCS$ shows the degree of sequence closing to the matrix resource. The larger the value, the closer it is to the matrix resource. In this paper, the trinucleotide position correlation score (TPCS) was calculated by using the 12 conversation sites for signal parameters of TATA promoter. They are −34th nt, −33th nt, −32th nt, −31th nt, −30th nt, −29th nt, −28th nt, −27th nt, −26th nt, −2th nt, −1th nt, and 0th nt. The M_3 (l) values in these sites are larger than 1.50 (shown in Fig. 1A).

4.5. K-mer increment of overlap content diversity (KIOCD)

For calculating the local features in the different regions of promoter, each primary sequence was split into several parts with different lengths. According to the "principle of resonance" [10], the true signals tend to remain at the same position while the noises tend to displace randomly when the window size is varied. If two profiles corresponding to different parameters are combined, the true signals will tend to enhance each other while the noises will tend to cancel each other. In this study, the subsequence compositions of K-mer in eight overlap regions were introduced to establish the content feature of promoter (5 windows of 50 bp, the length of the last window is 51 bp and 3 windows of 83 bp, the length of the last window is 85 bp) with the sample size of 251 bp. The increment of diversity (ID) was applied in order to avoid too large number variants. According to the description in the section "Increment of diversity", the subsequence composition in every region can be translated into two ID values (ID_{pos}, ID_{neg}). Therefore, the total 16 increments of overlap content diversity ($IOCD$) in the eight overlap regions were obtained by using Eq. (3). The length of subsequence segment ranged from 3-mer to 6-mer. The dimensions of diversity source S equal to 64, 256, 1024, and 4096, respectively. The longer segment was not chosen because it was not practical in real application. All of the 16 ID values were selected as the inputting parameters of the content feature for SVM program.

4.6. GC-Skew score (GCSS)

Recently, a prominent GC-compositional strand bias around the transcription start sites (TSS) in *Arabidopsis thaliana* gene was reported [36]; high C residue frequencies and low G residue frequencies are presented in the proximal promoters. The potential value of GC-Skew as an index is helpful for promoter prediction in plants [52]. Therefore, the GC-Skew score (GCSS) was defined and used as one of the input feature vectors in our recognition method. The GCSS is defined as follows:

$$GCSS = \sum_{i=-200}^{50} (C[i]-G[i]) / (C[i] + G[i]). \tag{6}$$

4.7. Support vector machine

The support vector machine (SVM) has been used widely in classification method based on statistical learning theory. It has been successfully used to deal with several bioinformatics problems, such as membrane protein classification [53], protein subcellular location prediction [47], operon prediction [54] and promoter prediction [55]. First, SVM method maps the input vectors into one feature space. Then, within this feature space, it constructs a hyperplane for separating two classes. The software toolbox used in this paper is the libsvm-2.88 developed by Chang and Lin [56]. More theory of SVM had been widely described in Chang and Li's literature. The software toolbox can be freely downloaded from the website http://www.csie.ntu.edu.tw/_cjlin/libsvm.

4.8. Assessment of performance

In order to evaluate the correct prediction rate and reliability of our prediction method, the sensitivity (Sn), specificity (Sp) and accuracy (Ac) are defined the same as the TSSP-TCM program [24]:

$$\text{Sensitivity}: S_n = TP / (TP+FN) \tag{7}$$

$$\text{Specificity}: S_p = TN / (TN+FP) \tag{8}$$

$$\text{Accuracy}: A_{cc} = \left(S_n+S_p\right) / 2 \tag{9}$$

where TP denotes the number of the correctly recognized positives, FN denotes the number of the positives recognized as negatives, FP denotes the number of the negatives recognized as positives, and TN denotes the number of correctly recognized negatives.

Acknowledgments

We wish to express our gratitude to the executive editor and two anonymous reviewers whose constructive comments were very important for this article. We would like to thank the administrators of PlanPromDB and Shahmuradov, I.A for access to their data. We thank Xing Huang at the University of Science and Technology Beijing for his help in graphs of the DNA geometric descriptors. This work was supported by the National Natural Science Foundation of China (No. 61063016), the Natural Science Foundation of Inner Mongolia Autonomous Region (No. 200607010101) and the "211 project" innovative talents training program of Inner Mongolia University.

References

[1] J.J. Gordon, M.W. Towsey, J.M. Hogan, S.A. Mathews, P. Timms, Improved prediction of bacterial transcription start sites, Bioinformatics 22 (2006) 142–148.
[2] M.W. Towsey, J.J. Gordon, J.M. Hogan, The prediction of bacterial transcription start sites using support vector machines, Int. J. Neural Syst. 16 (2006) 363–370.
[3] H. Wang, C.J. Benham, Promoter prediction and annotation of microbial genomes based on DNA sequence and structural responses to superhelical stress, BMC Bioinform. 7 (2006) 248.
[4] X.Y. Zhao, Z.Y. Xuan, M.Q. Zhang, Boosting with stumps for predicting transcription start sites, Genome Biol. 8 (2007) R17.
[5] S. Knudsen, Promoter 2.0: for the recognition of Pol II promoter sequences, Bioinformatics 15 (1999) 356–361.
[6] S. Burden, Y.X. Lin, R. Zhang, Improving promoter prediction for the NNPP2.2 algorithm: a case study using *Escherichia coli* DNA sequences, Bioinformatics 21 (2005) 601–607.

[7] N.I. Gershenzon, I.P. Ioshikhes, Synergy of human Pol II core promoter elements revealed by statistical sequence analysis, Bioinformatics 21 (2005) 1295–1300.
[8] U. Ohler, G.C. Liao, H. Niemann, G.M. Rubin, Computational analysis of core promoters in the *Drosophila* genome, Genome Biol. 3 (2002) 12.
[9] C. Yang, E. Bolotin, T. Jiang, F.M. Sladek, E. Martinez, Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters, Gene 389 (2007) 52–65.
[10] M.Q. Zhang, Identification of human gene core promoters in silico, Genome Res. 8 (1998) 319–326.
[11] M. Scherf, A. Klingenhoff, T. Werner, Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach, J. Mol. Biol. 297 (2000) 599–606.
[12] V.B. Bajic, et al., Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoters, Bioinformatics 18 (2002) 198–199.
[13] V. Narang, W.K. Sung, A. Mittal, Computational modeling of oligonucleotide positional densities for human promoter prediction, Artif. Intell. Med. 35 (2005) 107–119.
[14] X.D. Xie, S.H. Wu, K.M. Lam, H. Yan, PromoterExplorer: an effective promoter identification method based on the AdaBoost algorithm, Bioinformatics 22 (2006) 2722–2728.
[15] S.H. Wu, X.D. Xie, A.W. Liew, H. Yan, Eukaryotic promoter prediction based on relative entropy and positional information, Phys. Rev. E 75 (2007) 041908.
[16] H.Y. Jin, L.F. Luo, L.R. Zhang, Using estimative reaction free energy to predict splice sites and their flanking competitors, Gene 424 (2008) 115–120.
[17] P. Akan, P. Deloukas, DNA sequence and structural properties as predictors of human and mouse promoters, Gene 410 (2008) 165–176.
[18] T. Abeel, Y. Saeys, E. Bonnet, P. Rouzé, Y. Van de Peer, Generic eukaryotic core promoter prediction using structural features of DNA, Genome Res. 18 (2008) 310–323.
[19] K. Brick, J. Watanabe, E. Pizzi, Core promoters are predicted by their distinct physicochemical properties in the genome of *Plasmodium falciparum*, Genome Biol. 9 (2008) R178.
[20] K. Florquin, Y. Saeys, S. Degroeve, P. Rouzé, Y. Van de Peer, Large-scale structural analysis of the core promoter in mammalian and plant genomes, Nucleic Acids Res. 33 (2005) 4255–4264.
[21] P.J. Uren, M. Cameron-Jones, A.H.J. Sale, Promoter prediction using physico-chemical properties of DNA, in: Computational Life Sciences II. Lecture Notes in Bioinformatics (4216), Springer-Verlag: Lecture Notes in Computer Science, Berlin; New York, 2006, pp. 21–31.
[22] S. Rombauts, et al., Computational approaches to identify promoters and cis-regulatory elements in plant genomes, Plant Physiol. 132 (2003) 1162–1176.
[23] Y. Van de Peer, The future for plants and plants for the future, Genome Biol. 8 (2007) 308.
[24] I.A. Shahmuradov, V.V. Solovyev, A.J. Gammerman, Plant promoter prediction with confidence estimation, Nucleic Acids Res. 33 (2005) 1069–1076.
[25] S.P. Pandey, A. Krishnamachari, Computational analysis of plant RNA Pol-II promoters, BioSystems 83 (2006) 38–50.
[26] Y.Y. Yamamoto, et al., Differentiation of core promoter architecture between plants and mammals revealed by LDSS analysis, Nucleic Acids Res. 35 (2007) 6219–6226.
[27] Y.Y. Yamamoto, et al., Identification of plant promoter constituents by analysis of local distribution of short sequences, BMC Genomics 8 (2007) 67.
[28] W.K. Olson, A.A. Gorin, X.J. Lu, L.M. Hock, V.B. Zhurkin, DNA sequence-dependent deformability deduced from protein–DNA crystal complexes, Proc. Natl Acad. Sci. 95 (1998) 11163–11168.
[29] J. Goñi, A. Pérez, D. Torrents, M. Orozco, Determining promoter location based on DNA structure first-principles calculations, Genome Biol. 8 (2007) R263.
[30] A.V. Morozov, et al., Using DNA mechanics to predict in vitro nucleosome positions and formation energies, Nucleic Acids Res. 37 (2009) 4707–4722.
[31] Q.Z. Li, H. Lin, The recognition and prediction of sigma70 promoters in *Escherichia coli* K-12, J. Theor. Biol. 242 (2006) 135–141.
[32] J. Ponjavic, et al., Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters, Genome Biol. 7 (2006) R78.
[33] C. Molina, E. Grotewold, Genome wide analysis of *Arabidopsis* core promoters, BMC Genomics 6 (2005) 25.
[34] J. Lichtenberg, et al., The word landscape of the non-coding segments of the *Arabidopsis thaliana* genome, BMC Genomics 10 (2009) 463.
[35] V. Bernard, V. Brunaud, A. Lecharny, TC-motifs at the TATA-box expected position in plant genes: a novel class of motifs involved in the transcription regulation, BMC Genomics 11 (2010) 166.
[36] T. Tatarinova, V. Brover, M. Troukhan, N. Alexandrov, Skew in CG content near the transcription start site in *Arabidopsis thaliana*, Bioinformatics 19 (2003) i313–:i314.
[37] A. Berezhnoy, Y. Shckorbatov, Dependence of the *E. coli* promoter strength and physical parameters upon the nucleotide sequence, J. Zhejiang Univ. Sci. 6B 11 (2005) 1063–1068.
[38] D. Swarbreck, C. Wilks, P. Lamesch, T.Z. Berardini, M. Garcia-Hernandez, H. Foerster, D. Li, T. Meyer, R. Muller, L. Ploetz, A. Radenbaugh, S. Singh, V. Swing, C. Tissier, P. Zhang, E. Huala, The *Arabidopsis* Information Resource (TAIR): gene structure and function annotation, Nucleic Acids Res. 36 (2008) D1009–D1014.
[39] T. Abeel, Y. Saeys, P. Rouzé, Y. Van de Peer, ProSOM: core promoter prediction based on unsupervised clustering of DNA physical profiles, Bioinformatics 24 (2008) i24–i31.
[40] D.A. Bell-Lelong, J.C. Cusumano, K. Meyer, C. Chapple, Cinnamate-4-hydroxylase expression in *Arabidopsis*. Regulation in response to development and the environment, Plant Physiol. 113 (1997) 729–738.
[41] M.Q. Zhang, Computational analyses of eukaryotic promoters, BMC Bioinform. 8 (2007) S3.
[42] I.A. Shahmuradov, A.J. Gammerman, J.M. Hancock, P.M. Bramley, V.V. Solovyev, PlantProm: a database of plant promoter sequences, Nucleic Acids Res. 31 (2003) 114–117.
[43] R.R. Laxton, The measure of diversity, J. Theor. Biol. 71 (1978) 51–67.
[44] H. Lin, Q.Z. Li, Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components, J. Comput. Chem. 28 (2007) 1463–1466.
[45] Q.Z. Li, Z.Q. Lu, The prediction of the structural class of protein: application of the measure of diversity, J. Theor. Biol. 213 (2001) 493–502.
[46] L.R. Zhang, L.F. Luo, Splice site prediction with quadratic discriminant analysis using diversity measure, Nucleic Acids Res. 31 (2003) 6214–6220.
[47] Y.L. Chen, Q.Z. Li, Prediction of the subcellular location of apoptosis proteins, J. Theor. Biol. 245 (2007) 775–783.
[48] P. Bucher, Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoters, J. Mol. Biol. 212 (1990) 563–578.
[49] J.M. Claverie, S. Audic, The statistical significance of nucleotide position-weight matrix matches, CABIOS 12 (1996) 431–439.
[50] J.W. Fickett, Quantitative discrimination of MEF2 sites, Mol. Cell. Biol. 16 (1996) 437–441.
[51] Y.C. Zuo, Q.Z. Li, Analysis of plant TATA and TATA-less promoters by using sequence and structure features, Prog. Biochem. Biophys. 36 (2009) 863–871.
[52] S. Fujimori, T. Washio, M. Tomita, GC-compositional strand bias around transcription start sites in plants and fungi, BMC Genomics 6 (2005) 26.
[53] X. Pu, J. Guo, H. Leunga, Y.L. Lin, Prediction of membrane protein types from sequences and position-specific scoring matrices, J. Theor. Biol. 247 (2007) 259–265.
[54] G.Q. Zhang, Z.W. Cao, Q.M. Luo, Y.D. Cai, Y.X. Li, Operon prediction based on SVM, Comput. Biol. Chem. 30 (2006) 233–240.
[55] S. Sonnenburg, A. Zien, G. Rätsch, ARTS: accurate recognition of transcription starts in human, Bioinformatics 22 (2006) e472–e480.
[56] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machinesSoftware available at, http://www.csie.ntu.edu.tw/~cjlin/libsvm.