Towards accurate human promoter recognition: a review of currently used sequence features and classification methods

Jia Zeng, Shanfeng Zhu and Hong Yan
Submitted: 19th February 2009; Received (in revised form): 3rd May 2009

Abstract
This review describes important advances that have been made during the past decade for genome-wide human promoter recognition. Interest in promoter recognition algorithms on a genome-wide scale is worldwide and touches on a number of practical systems that are important in analysis of gene regulation and in genome annotation without experimental support of ESTs, cDNAs or mRNAs. The main focus of this review is on feature extraction and model selection for accurate human promoter recognition, with descriptions of what they are, what has been accomplished, and what remains to be done.

Keywords: human promoter recognition; genome annotation; feature extraction; model selection

INTRODUCTION
The past decade has seen a rapid development of promoter recognition algorithms since the first review paper [1] was published in 1997. Promoter recognition is a real problem that computationally identifies the transcription start site (TSS) or the 5′ end of the gene without time-consuming and expensive experimental methods that align ESTs, cDNAs or mRNAs against to the entire genome. In this article, we focus on humans because it is a representative species that has attracted much more attention in the past decade. In humans, the TSS is surrounded with a core-promoter region within around ±50 base pairs (bp). A proximal promoter region has several hundreds bp immediately upstream of the core-promoter. The capacity of transcription factors (TFs) to activate gene expression is encoded in both the core and proximal promoters, which are composed of short regulatory elements that function as transcription factor binding sites (TFBSs) for specific TFs to control and regulate the transcription initiation of the gene. Therefore, the rich information encoded in promoters is crucial to locate the accurate position of the TSS. Figure 1 shows a schematic representation of the locations of the promoter region, TFBSs, TSS, 5′ UTR, exons, introns and 3′ UTR. So far, high-resolution promoter recognition algorithms have at least two important motivations. First, they can improve the genome annotation when the experimental support of ESTs, cDNAs or mRNAs is not available. Second, they can efficiently narrow down the regions in transcriptional regulation for inspiring further experimental work because the computational approach is much cheaper. However, accurate human promoter recognition...
is still challenging since the state-of-the-art recognition algorithms reviewed and compared in recent reports [2–8] show unsatisfactory performance at the genome level due to the diverse nature of promoter sequences. This article aims to attract more interests from computer scientists in pattern recognition to solve the promoter recognition problem.

Extracting the most informative and discriminative features to differentiate the categories of promoters from nonpromoters is one of the key problems in promoter recognition. The properties of promoter regions are considerably different from those of other genomic regions, such as exons, introns, 3' UTRs and intergenic regions in Figure 1. Regarding the nature of promoters, three types of sequence features are now recognized, namely signal, context and structure features, to contribute essentially to locate core-promoter regions. Signal features are powerful biological signals including core-promoter elements [9, 10], some short modular TFBSs and CpG-islands [11], which play important roles in assembly of transcriptional machinery. If we view promoter and nonpromoter DNA sequences as two collections of documents, context features are the contents of the documents represented by the basic unit DNA words called n-mers (n-base-long nucleotide sequences) [12]. The direct readout mechanism suggests that some TFs recognize and bind specific n-mers through direct contact between amino acids and nucleotide bases [13]. Structure features originate from chemo-physical properties of the DNA three-dimensional structures, which have been intensively studied in biochemistry and biophysics fields. The indirect readout mechanism also suggests that some TFs recognize specific DNA sequences indirectly based on the sequence-dependent conformation and deformability of the DNA [13].

Figure 2A shows the ‘human promoter recognition’ topic trend from 1998 to 2008 in the PubMed MEDLINE database. We retrieve a total of 66 papers from 1998 to 2008 (two papers in early 2009) on ‘human promoter recognition’ according to the keywords including ‘human’, ‘promoter’, ‘TSS’, ‘prediction’ and ‘recognition’. Interested readers can find all important publications in the supplementary material. Based on our knowledge, we manually assign labels ‘signal’, ‘context’ and ‘structure’ to these papers if they use the signal, context, structure features or the combination of these features for promoter recognition. There are also ten review papers labelled as ‘review’. If a paper uses two different features such as ‘signal & context’, we shall add one when counting the number of papers on this feature. Note that the number of papers per year is fairly low for each topic, so the topic trends of different features can only reflect the relative preference of the corresponding features adopted by practical human promoter recognition systems.

Besides feature extraction, another important task is to select appropriate classifiers to differentiate categories of promoters from nonpromoters based on selected features. Generally, there are three major classification methods in human promoter recognition systems. The first is the discriminative model that finds the optimal thresholds or classification boundaries in the signal, context and structure feature space. Typical methods include artificial neural networks (ANNs), discriminant functions and support vector machines (SVMs). The second is the generative model that describes the generative process of signal, context and structure observations. Position weight matrix (PWM), nearest neighbour and hidden Markov models (HMMs) belong to generative models. The third lies in the classifier...
ensemble that combines multiple classifiers for multiple features in order to achieve a consensus and robust recognition result. Adaboost, boosting techniques with strumps and classifier combinations are typical examples. Because the performance of most current classifiers on the entire human genome remains at almost the same low level, we are in great need of powerful and efficient classifiers that can take full advantage of informative signal, context, structure as well as epigenetic features.

This article is organized as follows. In the next section, we introduce the DNA sequence features in detail. We also introduce the newly discovered epigenetic features for promoter recognition. Subsequently, we review the classification algorithms based on all these features. Finally, we discuss currently used datasets and performance measures, and envision a number of future research directions.

FEATURE EXTRACTION
Signal features are biologically functional regions including core-promoter elements, combinations of TFBSs and mammalian CpG islands. Figure 2B shows the number of papers published on signal features each year. We see that signal features were widely studied from 2000 to 2007. The core-promoter elements consist of the TATA box, initiator (Inr), downstream core-promoter element (DPE), TFIIB recognition element (BRE) and the motif 10 element (MTE) [9]. These core elements work in a synergetic way to start gene transcription.

Figure 2: The number of publications on promoter recognition in the PubMed MEDLINE database from 1998 to 2008. (A) After the first review [1], interests in promoter recognition was increasing until another review [5] evaluated the state-of-the-art promoter recognition algorithms on the entire human genome. Hereafter, more and more attention has been paid to genome-wide human promoter recognition. (B) The number of publications on signal features by year. The CpG-islands have become standard signal features since FirstEF [36] published in 2001. (C) The number of publications on context features by year. The n-mer representation has been widely adopted as a standard context feature since PromoterInspector [12] published in 2000. (D) The number of publications on structure features by year. More and more attention has been paid to structural properties of promoters in accurate core promoter recognition since 1998 [26].
Towards accurate human promoter recognition: a review

In addition, some significant position-specific TFBSs in proximal and core-promoters are also important signals that bind TFs to initiate transcription [14, 15]. Earlier promoter recognition algorithms searched by the TATA box, Inr element and consensus sequences for TFBSs [1]. However, it has been realized recently that these signals vary significantly and exist only in a small proportion of all human promoters [10, 16]. This is a major reason for the high false discovery rate when attempting to detect core-promoter elements and other specific TFBSs in promoters. Since 2001, the CpG-island signal has attracted more and more interests because it is a strong large-scale signal existing in the majority of the known human promoters (~72%) [17]. More specifically, the recent studies have shown that 5′ end promoters are much more likely to be located within CpG islands than 3′ end alternative promoters [18]. The CpG-island, which is a region of DNA longer than 200 bp enriched with phosphodiester-linked cytosine (C) and guanine (G) pairs. According to the cap analysis of gene expression, mammalian promoters can be broadly classified into TATA- and CpG-rich promoters, and different tissues and families of genes differentially use distinct types of promoters [19]. Here we show two important CpG-island features including the global frequency of GC content (GC_p) and the ratio of expected to observed CG dinucleotides:

\[
\frac{GC_p}{e} = \frac{p(G) + p(C)}{p(C)p(G)}
\]

where \(p(CG)\), \(p(C)\) and \(p(G)\) are percentages of CG, C and G in a DNA sequence, respectively. The CpG-island signal is so strong that if we mix CpG-rich and CpG-poor promoters together, the overall characteristics will be dominated by the majority of CpG-rich promoters. Recent studies [18, 20–22] show that it is necessary to analyze CpG-rich and CpG-poor promoters separately. In general, most existing classifiers can achieve a much higher recognition resolution regarding locations of TSSs in CpG-rich promoters than in CpG-poor promoters on a genome-wide scale [18]. Since CpG-rich and CpG-poor promoters have distinct properties, a practical choice in designing promoter recognition systems is to first examine whether the test sequence is CpG-rich or CpG-poor, and subsequently use two different classifiers focusing on CpG-rich and CpG-poor promoters, respectively.

Context features are extracted from the genomic content of promoters as a set of \(n\)-mers whose statistics are estimated from training samples. Indeed, the \(n\)-mers cover the signal features. For example, TATA box is an 8-mer TATAWAAR (W denotes A/T and R denotes A/G), Inr is a 7-mer YYANWYY (Y denotes C/T and N denotes A/G/C/T), and the CpG islands are GC-rich 5-mers like CGGCG. Intuitively, the \(n\)-mers encode contextual information of promoters with the following advantages. First, the distribution of \(n\)-mers may have biological significance, such as TATA box, Inr and CpG islands. Second, \(n\)-mers may reveal fine details of yet unknown promoter units. Therefore, the context features may help reduce the false positive rates while maintaining a relatively high sensitivity in promoter recognition [6]. Figure 2C shows the number of papers published on context features each year. We see that interests in context features continue to increase since PromoterInspector [12] was proposed in 2000. If \(n\) is fixed, the term frequency-inverse document frequency (TF-IDF) is an important feature applied to text mining, where each document is deemed as a ‘bag of words’ and represented by the TF-IDF [23]. PromoterInspector considers promoters and nonpromoters as two collections of documents, and uses ‘bag of \(n\)-mers’ as features. Besides \(n\)-mer frequency, inverse document frequency based on Kullback–Leibler (KL) divergence [24] is also used in promoter recognition. One serious problem of \(n\)-mer representation is that the search space \(4^n\) becomes overwhelmingly large when \(n\) increases. For example, the total number of 6-mer is \(4^6 = 4096\). After testing the discriminant power of 6-, 7- and 8-mers based on the KL divergence, it shows that 6-mers keep a good balance between the discriminative power and the computational complexity [24].

Structure features are derived from DNA three-dimensional structures, which play important roles in guiding DNA-binding proteins to target sites efficiently [13, 16]. Typical properties include DNA flexibility, DNA curvature, stabilizing energy of Z-DNA, DNA denaturation values, base stacking values, propeller twist and nucleosome positioning preference. Recent studies [2, 25] show that local changes of structure features around the classical TATA box position and the TSS are informative in differentiating promoters from nonpromoters. As an important structure feature, the sequence-dependent
DNA flexibility has been extensively examined in several organisms, and it has been suggested that it influences promoter activities [16], nucleosome positions [26], TF binding [27] and yeast DNA replication [28]. Figure 2D shows the number of papers published on structure features each year. We see that little interest has been paid to structure features until 2005. From 2006 to 2008, structure features become popular in promoter recognition systems. In 1998, DNA flexibility was first proposed to be an effective feature for promoter recognition [26]. The reasons why DNA flexibility has not been widely adopted in later promoter recognition methods lie in two aspects [29]. First, individual promoter flexibility profile is extremely noisy. Second, DNA flexibility is sequence-dependent (di-, tri- or tetra-nucleotide) so that it is unknown if it might already be implicitly encoded in the sequence context alone, or if it could lead to an improvement in promoter recognition. Using large-scale structure features is relatively new to the field of promoter recognition since previous studies have focused on 200 bp around the TSS. Currently, there are three types of structure features demonstrated to be effective for human promoter recognition. The first is the DNA flexibility parameters based on the tetranucleotide potential energy surface model [22, 30]. The second is the DNA flexibility parameters derived from long atomic molecular dynamics simulations in water [31]. The last is the DNA base-stacking property [2, 32]. Without loss of generality, we show how to calculate the flexibility profile based on the tetranucleotide flexibility parameters [33]. At each position of the nucleotide sequence, the flexibility value is calculated based on the \( n \)-mer (\( n \) can be empirically fixed) by averaging the flexibility parameters of \( n-3 \) consecutively overlapping tetranucleotides,

\[
f = \frac{\sum_{i=1}^{n-3} t_i}{n-3},
\]

where \( t_i \) denotes the flexibility parameter of the \( i \)th tetranucleotide in \( n \)-mer stored in the conversion table [33]. In this way, it is possible to calculate the flexibility profile for any given sequence based on 136 unique tetranucleotide parameters.

Signal, context and structure features are correlated DNA sequence features. Context features are basis of signal and structure features. Obviously, core-promoter elements, TFBSs and CpG islands are all position-specific functional \( n \)-mers. Also, most structure features are mapping from di-, tri- or tetra-nucleotides (\( n \)-mers, \( n = 2, 3, 4 \)) to physical parameters based on conversion tables. One may question if structure features can provide additional information compared with context features in order to improve promoter recognition. Several recent studies [2, 21, 22, 30, 31] have demonstrated that structure features work very well in genome-wide human promoter recognition. One possible reason is that different \( n \)-mers may correspond to similar physical properties, which implicitly group different \( n \)-mers together to perform similar functions. As far as the DNA flexibility is concerned, highly rigid DNAs are found in both TATA-rich and TATA-less promoters [16] at canonical TATA box position, which demonstrate that structure features may be the same when \( n \)-mers are quite different. In addition, signal features have close relationships with structure features. TFBSs are generally more flexible DNA sequences in order to wrap around proteins [27]. In particular, TATA box contains a highly rigid ‘AAAA’ tail that may be an important signal to guide protein to locate the target site accurately [27]. CpG-rich and CpG-poor promoters have similar flexibility patterns but at different flexibility levels. Note that replication origins of yeast DNA sequences have similar structure features with those of promoters [28]. This phenomenon warns us that we may find many false positives that may be replication origins rather than promoters.

Besides sequence features, epigenetic features, such as those encoding important information to predict chromatin opening and closed states have been suggested for promoter recognition for several years [3, 8]. Overall 75–90% eukaryotic genomes are packaged into chromatin structures. The basic repeat unit of chromatin is the nucleosome that prevents DNA from interacting with most DNA-binding proteins because of its stable structure. If we do not know which regions of chromatin are open or closed, we have to assume the whole genome is accessible for proteins, which is obviously wrong and will lead to more false positives. Epigenetic features may help determine the opening degree of chromatin so that we can identify those false promoters that are within the closed chromatin structure. The nucleosome is composed of an octamer of four core histones. The recent technology of chromatin immunoprecipitation followed by sequencing produces genome-wide profiles of histone modification patterns in mammalian cells [34]. These histone
modification profiles, including their shape and absolute intensity around the TSS, are important epigenetic features that may contribute positively to identify core-promoter regions. Further experiments show that CpG-rich and CpG-poor promoters have significantly different histone modification profiles around the TSS [21], which implies that epigenetic features correlate with signal features to some extent. Two features based on histone modification profiles have been used for promoter recognition [21]. The first captures the shape information using the Pearson correlation coefficient between the profile of local modification and the average profile of this modification for all promoters in the training set. The second encodes the weighted intensity of this modification using the dot product of these two profiles. Compared with currently used sequence features, epigenetic features based on histone modification can provide additional information in terms of high-resolution core-promoter recognition.

MODEL SELECTION
As far as signal features are concerned, NNPP 2.2 [35] combines the characteristics of TATA box and Inr using the time-delay ANNs and linear discriminant function. It specially focuses on the region between TSS and translation start site and scans a DNA sequence of length 51 bp and outputs a list of predictions with scores greater than the user defined threshold. FirstEF [36] first scans [–1500, +500] bp to detect CpG-islands, and then uses two different quadratic discriminant functions, one for CpG-rich and the other for CpG-poor first exons, to predict the TSS within a window [–500, +70] bp. It is one of the first classifiers to handle CpG-rich and -poor promoters separately. CpgProd [37] identifies CpG islands using a sliding window when the moving average GC frequency is above 0.5 and the ratio is above 0.6. Since CpG-islands are strong signal features associated with promoters, correct prediction of CpG-islands will lead to high accurate TSS prediction. All these methods are mainly discriminative models. Eponine [38] applies relevance vector machines to capture the most important sequence-based signals, such as the TATA box together with flanking regions of CG enrichment. It consists of a collection of positioned constraints, each represented by a PWM, which is a simple generative model having a probability distribution over consensus sequences. When the PWM scans the test sequence, motifs matching consensus sequences will receive high scores. The FProm [39] evaluates the occurrence of TSS using two linear discriminant functions (separate for TATA-rich and -poor promoters encoded by PWMs) with characteristics computed in the [–200, +50] bp region for each position on a given sequence.

For context features, PromoterInspector [12] builds two disjoint groups of n-mers with variable gaps or wildcards, one for promoter group and the other for nonpromoter group. It uses a nearest neighbourhood strategy in that if an unknown genomic sequence contains more n-mers in the promoter group, it will be classified into promoters. The KL divergence based classifier (KLC) utilizes the inverse document frequency in document classification [23]. KLC selects the most discriminative n-mers to characterize promoters against exons, introns and 3’UTRs based on the relative entropy criterion. Three groups of PWMs are built for these selected n-mers. However, PWMs have very high dimensionality, when n is very large. Therefore, principle component analysis (PCA) [40] is a choice to reduce the high dimensionality of PWMs. The eigenvectors of PWMs with the largest eigenvalues correspond to the dimensions that have the strongest correlation in the n-mer space. The reduced PWM has a smaller size so as to speed up the scanning process on a genome-wide scale. PSPA [20] uses the propensity of position specific n-mers (1 ≤ n ≤ 5) such as CCGTTT to differentiate promoters from nonpromoters. Because position and distance specific motif pairs around promoters are strongly associated with TSSs [14, 15], PSPA trains two propensity models of these position specific n-mers for CpG-rich and -poor promoters within [–100, +100] bp separately. As far as discriminative models are concerned, Promoter 2.0 [41] uses ANNs optimized by genetic algorithms to process a small window of 200 bp DNA sequence, which is represented as binary codes. Each ANN consists of only one hidden and one output neuron with input from 6-mers. The genetic optimization algorithm can make ANNs generalize well to unknown genomic sequences. Since DNA sequence can be viewed as a dynamic process, a recent study [42] shows that n-mer frequency along with nonlinear time series descriptors, such as the Lyapunov component stability and the Tsallis entropy, and supervised learning such as SVMs can improve promoter recognition accuracy. DPF [43] and DGSF [44] encode both CpG-island
and context features around the TSS (the region downstream of the TSS) by feedforward four-layer ANNs (input layer, two hidden layers with 10 and 15 neurons, and output layer). Normalized signal and context features are used to train ANNs by the standard back-propagation algorithm. Note that DPF [43] and DGSF [44] are suitable for analysis and discovery of CpG-rich promoters.

For structure features, McPromoter [29] analyzes [−250, +250] bp region and incorporates different structure features including DNA flexibility and GC content using HMMs to improve promoter recognition. Each state of HMMs is a Gaussian distribution to generate observed features. Furthermore, McPromoter uses ANNs to combines decisions from context and structure profiles. ProSOM [32] applies a clustering technique, self-organizing maps (SOM), to core-promoter recognition based on structural profiles of the transcribed base stacking energy. Based on 36 clusters of promoters and nonpromoters, ProSOM classifies unknown genomic sequence to the category of promoters if it is mapped to a cluster with a promoter probability higher or equal to the threshold. In contrast to above generative models, ARTS [45] identifies the TSS through designing kernels of SVMs to combine n-mers (n ≤ 20) and other structure features such as twisting angles and stacking energies of the DNA sequence. It specially designs the kernels for upstream of the TSS, downstream of the TSS as well as further downstream of introns and coding regions, and three-dimensional structure of DNA near the TSS. ProStar [31] classifies [−250, +250] bp sequences by the Mahalanobis distance based on dinucleotide flexibility parameters derived from atomic simulations. If a genomic sequence has a normalized distance score higher than the predefined threshold, it should be classified into promoters. EP3 [2] examines fifteen large-scale structure features of DNA and shows that the base-stacking property gives the best results. In particular, EP3 depends on the expert’s empirical knowledge to set the proper threshold. As a result, EP3 can handle many eukaryotic genomes with a small modification of that threshold. One advantage of ProStar and EP3 is that the speed and memory requirements are much less when scanning the entire human genome.

Recently classifier ensemble techniques become popular in human promoter recognition systems. PromoterExplorer [46] uses the AdaBoost algorithm to select most discriminative CpG-island and 5-mer features for promoter and nonpromoter classification within [−200, +50] bp. CoreBoost [22] applies a boosting technique with strumps to feature selection from core-promoter elements, TFBSs, DNA flexibility, Markovian score, and 1- and 2-mer frequencies related to nucleotide G or C. CoreBoost_HM [21] further combines epigenetic features such as histone modification profiles based on LogitBoost with strumps. All these methods can select informative features and use an assembly of weak classifiers to build a strong classifier. Note that CoreBoost and CoreBoost_HM focus only on core-promoter recognition. As a hierarchical recognition system, SCS [30] describes CpG-islands, 6-mer frequency and DNA flexibility profiles within [−250, +50] bp by Gaussian mixture models and naïve Bayes classifiers, and integrates output likelihoods from component classifiers by a decision tree. MetaProm [18] integrates predictions from six complementary promoter classifiers (PSPA, FirstEF, McPromoter, DPF, DGSF and FProm) using an ANN, which implicitly combines signal, context and structure features to discover alternative human promoters. Similarly, EnsemPro [47] also combines the classification results of the existing promoter predictors including NNPP, FirstEF, DPF, Eponine and Promoter 2.0. It applies three representative ensemble schemes such as the majority voting, the weighted voting and the Bayesian approach, for the TSS prediction of hundreds of human genomic sequences.

**DATASETS AND PERFORMANCE MEASURES**

It is important to build a benchmark dataset as well as reasonable performance measures to evaluate the state-of-the-art promoter recognition algorithms. Currently there are three publicly available promoter databases. The EPD [48] provides the manually annotated promoters, which is highly reliable but has relatively low coverage of the 1871 human promoters. The Ensembl [49] contains regions around gene starts, which ensures a high coverage of the 28 253 genes but provides mixed quality. It also contains the entire human genome of about three billion base pairs. Intergenic sequences can be also extracted from the Ensembl. The DBTSS [50] collects 30 964 sequences [−1000, +200] bp around the experimentally determined TSSs, which at
present provides the best combination of coverage and quality. Besides these three databases, the RefSeq database [51] includes annotations for coding regions, gene and protein product names. Recently, the CAGE technique [19] identifies 123,400 possible TSSs with high accuracy by considering only tag clusters with at least two mapped tags on the same genomic location. For nonpromoters, exons and introns can be obtained from the EID [52], and 3′UTRs can be retrieved from the UTRdb [53]. Current promoter recognition algorithms are usually trained by promoters from the DBTSS, and are validated on the entire human genome according to the Ensembl gene annotation or the CAGE annotation [2, 32]. Another validation dataset is based on the ENCODE project [4], which aims to carefully annotate all functional elements in a small portion about 1% of the human genome. Recently, promoter recognition algorithms are strongly encouraged to be assessed on both CAGE and ENCODE annotations [2, 4, 31, 32].

An evaluation method in terms of TSS prediction for different promoter recognition systems is shown in Figure 3. On the entire human genome, if one or more predictions fall in the region [−500, +500] bp relative to the annotated true TSS, they are counted as one true positive (TP). If the annotated TSS is missed by this count, it represents one false negative (FN). All predictions that fall outside the range [−500, +500] bp are counted as false positives (FPs). We may choose other resolutions such as [−2000, +2000] bp, [−1000, +1000] bp and [−50, +50] bp. The highest resolution [−50, +50] bp corresponds to the core-promoter recognition. Three performance measures, sensitivity (Se), positive predictive value (PPV) and F-measure (F), are defined as follows,

\[
\text{Se} = \frac{\text{TP}}{\text{TP} + \text{FN}},
\]

\[
\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}},
\]

\[
F = \frac{2 \times \text{Se} \times \text{PPV}}{\text{Se} + \text{PPV}},
\]

where F-measure reflects the balance between Se and PPV. Generally, the higher F-measure means the better recognition performance.

**DISCUSSION**

Although much progress has been made in promoter recognition and *cis*-regulatory motif discovery, we are still far away from the goal of accurate human promoter recognition because no current algorithms can achieve the level with both Se and PPV over 0.4 by the resolution [−50, +50] bp (see recent comprehensive comparisons [18, 21, 30, 32, 54]). It means that we cannot accurately identify half of the existing TSSs, and on the other hand, we can identify only one true TSS by the cost of more than one false TSS. Interestingly, we find that some complex classifiers such as SVMs [45] and Adaboost [46] do not outperform other simple classifiers [2]. In pattern recognition, Occam’s razor [55] interprets the phenomenon that one should not use algorithms that are more complicated than are necessary, where ‘necessary’ is determined by the quality of fit to the training data. In addition, SCS [30] shows that classifiers often overfit the training data and generalize to unknown genomic sequences with more than a 15% drop in the performance.

Figure 3: Counting false positive (FP) and true positive (TP) for genome-wide human promoter recognition with a resolution [−500, +500]. We may choose other resolutions such as [−2000, +2000], [−1000, +1000] and [−50, +50]. The black region denotes the first exon of a gene.
Accurate human promoter recognition relies largely on feature extraction and model selection. First, we need to investigate more informative and discriminative features. Current sequence and epigenetic features seem to be able to cover only about 60% of promoters. In particular, CpG-rich promoters can be accurately identified compared with CpG-poor promoters [21, 22, 36, 54]. Therefore, extracting effective features from CpG-poor promoters is one of the urgent tasks in the future. Second, we need to solve a series of computational problems. For example, how to model the combinatorial modules of TFBSs? How to effectively reduce the search space of n-mers? How to denoise the structural profiles of a long DNA sequence? How to take the DNA topology into account? These problems are important but remain largely unsolved.

We have to mention that most promoter recognition systems are limited to upstream TSSs around the 5′ end of the gene. High throughput 5′ cap analysis has shown that 58% of mammalian genes have alternative TSSs and promoters [19]. More specifically, the majority of genes contain at least two and even up to more than 20 alternative promoters. The same promoter may contain multiple different TSSs used in a tissue and pathway specific manner. So, it is difficult to evaluate promoter recognition systems in terms of TPs and FPs in Figure 3 because we still lack genome-wide true annotated TSSs currently. One possible evaluation method [32] is based on highly accurate CAGE [19] alternative TSS annotations. In addition, EPD [48] and DBTSS [50] TSS annotations are also very useful in evaluation. Many recent promoter recognition systems [2, 21, 32] use all CAGE, EPD and DBTSS annotations to evaluate their performance in core-promoter recognition. Also, MetaProm [18] predicts alternative TSSs including middle TSS and most downstream TSS.

It is promising to partition human promoters into different functional groups and use different features and classifiers for each group in recognition. The presence of alternative promoters for a single gene is related to tissue-specific gene expression. Similar to the broad classification of CpG-rich and CpG-poor promoters, these tissue-specific promoters may have quite different properties that can be handled separately. So, tissue-specific promoter recognition may shed more light on studying diverse nature of promoters. The TiProD [56] provides a collection of human tissue-specific promoters with high specificity for this potential task. On the other hand, ProSOM [32] achieves a relatively better performance in core-promoter recognition by clustering human promoters into different groups based on DNA physical properties. Whether these groups of promoters correlate with tissue-specific promoters requires future investigations.

**Key Points**
- Human promoter recognition
- Genome annotation
- Feature extraction
- Model selection

**SUPPLEMENTARY DATA**
Supplementary data are available online at http://bib.oxfordjournals.org/. online.

**FUNDING**
Hong Kong Research Grant Council (Project CityU 123408) and the City University of Hong Kong (Project 9610034).

**References**
Towards accurate human promoter recognition: a review


