ISSN 0582-9879 Acta Biochimica et Biophysica Sinica 2004, **36**(4): 250–258 **CN** 31-1300/Q

Predicting Polymerase II Core Promoters by Cooperating Transcription Factor Binding Sites in Eukaryotic Genes

Xiao-Tu MA, Min-Ping QIAN*, and Hai-Xu TANG[1]

School of Mathematical Sciences, Peking University and Center for Theoretical Biology, Peking University, Beijing 100871, China;
[1]Department of Computer Science and Engineering, University of California, San Diego, La Jolla CA 92093-0114, USA

Abstract Several discriminate functions for predicting core promoters that based on the potential cooperation between transcription factor binding sites (TFBSs) are discussed. It is demonstrated that the promoter predicting accuracy is improved when the cooperation among TFBSs is taken into consideration. The core promoter region of a newly discovered gene *CKLFSF1* is predicted to locate more than 1.5 kb far away from the 5′ end of the transcript and in the last intron of its upstream gene, which is experimentally confirmed later. The core promoters of 3402 human RefSeq sequences, obtained by extending the mRNAs in human genome sequences, are predicted by our algorithm, and there are about 60% of the predicted core promoters locating within the ± 500 bp region relative to the annotated transcription start site.

Key words promoter; transcription factor binding site; transcription start site; gene finding

In the transcription stage of genes, various proteins bind to promoters—the transcriptional regulatory regions in genome. The main part of a promoter consists of many short sequence elements, TFBSs, with positive or negative effects on the transcription initiation [1]. Within this region, there is a core promoter defined by a minimal DNA element that is necessary and sufficient for transcription initiated by RNA polymerase II [2]. It is generally understood that the core promoter is the flank region of the transcription start site (TSS) [3]. This is always true for prokaryotes. With regard to eukaryotes, such as the homo sapience, it has been noted that the transcription regulation is much more sophisticated [4]. It is suggested that the "transcription machine" may not locate near the TSS and the DNA tertiary structure may play a role in the transcription regulation [5]. Consequently, the problem of the promoter identification may be different from that of the identification of the TSS [6].

From the evolutionary point of view, the TFBSs in core promoters should be conservative. During evolution, recombination and mutation frequently occur on the chromosomes, and only those changes keeping the necessary elements for survival can be observed in the present organs. Thus most of the segments in promoters that are necessary for transcription initiation should have a statistically significant conservation. In fact, Fickett *et al.* [7] pointed out: "TFBSs stand out clearly against a non-conserved background". Prestridge [8] first characterized eukaryotic promoters in terms of the density of TFBSs. There are several other attempts in this approach (see [9]), but one drawback of these methods is the large number of false positives. As a good alternative at present, PromoterInspector [10] makes use of pairs of IUPAC (International Union of Pure and Applied Chemistry) words with some distance and predicts promoters at a high specificity of several thousand base pairs per false positive with sensitivity rate of about 50% [11]. FirstEF makes use of the 5-tuple and 6-tuple around TSS, the donor information and the CpG-island information, and predicts 86% of the first exons with 17% false positives [11]. But these algorithms focusing on TSS identification are of limited use for biologists who are interested in the transcriptional regulation but not the accurate TSS.

On the other hand, Wagner [12] initiated the study of the cooperation between TFBSs and pointed out that the

Received: December 8, 2003 Accepted: February 5, 2004
This work is supported by the grants from National Natural Science Foundation of China (No. 90208022), and the National High Technology Research and Development Program of China (No. 2002AA234011)
*Corresponding author: Tel, 86-10-62759666; E-mail, qianmp@math.pku.edu.cn

transcription of many eukaryotic genes is co-regulated by transcription factors. Fickett *et al.* [7] showed that the co-binding site for the transcription factors is conservative among species. Besides, the example of HNF1 [13] strongly supports the cooperation between TFBSs.

Here, we provide an improved approach in terms of TFBS pairs to raise the signal-noise ratio for identifying core promoters. Our result supports the cooperation between TFBSs in core promoter sequences and demonstrates that the predicting accuracy is significantly improved by using TFBS pairs. Moreover, using the distance between TFBS pairs also helps to improve the predicting accuracy. Finally, we also report an interesting experimental result that the promoter of a gene discovered by Xu [14] is predicted locating in the last intron of its upstream gene by our program, which is confirmed later by wet experiment.

Materials and Methods

Data

The datasets in our experiments include 1300 TFBSs (with length from 5 to 12 bp) from TRANSFAC3.5 [15] and 575 vertebrate promoter sequences (except retroviruses) extracted from the EPD50 [16]. For consistency, each promoter sequence is cut from –250 to 50 bp, where TSS is at position +1. The entries with less than 40 bp upstream or 5 bp downstream are discarded with 565 entries left. The non-promoter data are the 890 human coding sequences in the 1998 GENIE multiple exon gene dataset. The genomic data in Fickett *et al.* [9] and the genomic data HMR195 [17] with complete genes are used for testing the performance of our algorithm on genomic sequence. The dataset COMPEL2.4 [18] is used for explaining the biological significance of TFBS pairs selected in our dictionary. The complete genomic sequences of 3402 mRNAs in human RefSeq are obtained by extending them in the human genome sequences to perform our program.

Algorithm description

From a linguistic point of view, we take core promoter sequences as sentences written on a random text background composed of A, C, G, and T, while the words are the TFBSs. One difficulty here is that there are many words shared in promoters and other DNA segments such as introns due to random chance. A reasonable assumption is that many genes have similar TFBS modules and the statistical features of these modules will stand out from the random background. Our approach is as following. First we try the single TFBS scoring system (SSS). The *k*-tuple ($4 < k < 13$) over-represented in promoters while under-represented in coding regions are considered (we do not consider the introns, because introns and promoters are functionally interchangeable in some cases (see "Promoter identification in gene sequences with known ATG"), and it will be demonstrated that the over-represented tuples (which are called "keywords" following the linguistics) in promoter region are mostly TFBSs (see "Keyword analysis"). We pick up TFBSs to build a keyword dictionary (*WD*) from the database TRANSFAC3.5, by counting their appearing frequencies in both the promoter and coding sequences, keeping those TFBSs that have relatively high appearing frequencies in promoters. A score $s(w)$ is assigned to each keyword w in the dictionary. For a given sequence S, we enumerate all the keywords w appearing in it and take the sum [Formula (1)]

$$T = \sum_{w \in S} s(w) \tag{1}$$

as its promoter-like score. A suitable threshold is determined by the following statistical method. Let FN_y be the number of sample promoters in learning dataset that have promoter-like score lower than y, and FP_y be the number of sample coding sequences in learning dataset with score higher than y. The threshold x is taken as Formula (2)

$$x = \arg\inf_y \{\frac{FN_y}{\#\text{Prom}} + \frac{FP_y}{\#\text{CDS}}\} \tag{2}$$

where #Prom and #CDS are the numbers of learning sample promoters and coding sequences respectively. A testing sequence will be accepted as a promoter if its score is above the threshold and rejected if below. Secondly, considering the cooperation between transcription factors, we try the TFBS pair scoring system (PSS). The same procedure as that for single TFBS is applied: a dictionary of over-represented TFBS pairs, a scoring function, and a threshold for TFBS pairs are set up. The result shows that considering TFBS pairs do improve the predicting accuracy. Then we try TFBS pair scoring system with distance (PSSD). To analyze the distance between TFBSs, we take the minimum non-overlapping distance between TFBSs of each pair in sample promoter sequences for statistics. Since there are not enough experimentally verified sample promoter sequences and the transcription factors may bind to various TFBSs, we cluster the TFBS pairs first and assume that the distance between TFBS pairs of a cluster is drawn from an identical and independent distribution. We will show that the information of the distance between TFBSs is useful for improving the predicting accuracy. Finally we take the PSSD as the discri-

minate function for promoter identification. For a given eukaryotic gene sequence with translation start site ATG known, which could often be provided by EST (expressed sequence tags) or RefSeq, we can predict promoters in the upstream region of the translation start site by the scoring systems given above.

Results

Keyword analysis

We count the number of *k*-tuple appearing in the TFBSs dataset. A *k*-tuple is defined as TFBS-*k*-tuple if it is either a TFBS itself or part of a TFBS in TRANSFAC 3.5. For $k = 5, 6, 7$, and 8, the number of *k*-tuple (denoted by *N*) and of TFBS-*k*-tuple (denoted by *nt*) are listed in Table 1. Apparently, only 1/4 and 1/13, which is a small portion of 7-tuple and 8-tuple respectively, appear in the TFBS set. This observation suggests that it would be better to take TFBS-*k*-tuples as the candidates of keywords instead of all *k*-tuples.

Table 1 Number of all *k*-tuples (*N*) and TFBS-*k*-tuples (*nt*)

k	*N*	*nt*
5	1024	1011
6	4096	3076
7	16384	4895
8	65536	4983

A detailed analysis of the *k*-tuple appearance distribution also supports the above idea. Let $N(x)$ be the number of 7-tuple that appears in x of the 565 sample promoter sequences. The bar plot of $N(x)$ along x is shown in Fig. 1. We can see that the ratio of the number of TFBS-7-tuple to 7-tuple in promoter sequences increases as x increasing, while it keeps almost the same for different x in random sequences. The overall distributions of 7-tuple appearance frequencies are similar for both promoter and random sequences, except that the distribution for promoter sequences has longer tail and larger $N(x)$ for lower x relative to that for random sequences. This indicates that there are certain 7-tuples, which are mostly TFBSs, with very high appearing frequencies, while some 7-tuples with low appearing frequencies in promoter regions.

A simple model can be used to explain the behavior of *k*-tuple frequency distribution in promoter sequences. This distribution can be decomposed into three components: (1) the background distribution, which is the contribution of the *k*-tuple appearance in random sequences; (2) the appearances of over-represented *k*-tuple; and (3) the appearances of under-represented *k*-tuple. In Fig. 2, we use the mixed model of three normal distributions (f_{ran}, f_{or} and f_{ur} for random, over-represented and under-represented frequency respectively) to fit the overall appearance distribution of the sample promoter data. Obviously the mixed model F_{prom} is very well fit to the true overall distribution R_{prom}. Meanwhile, the distribution of f_{ran} is quite similar to the true background distribution R_{ran}, acquired from the statistics on random sequences.

The result of this model suggests that a promoter

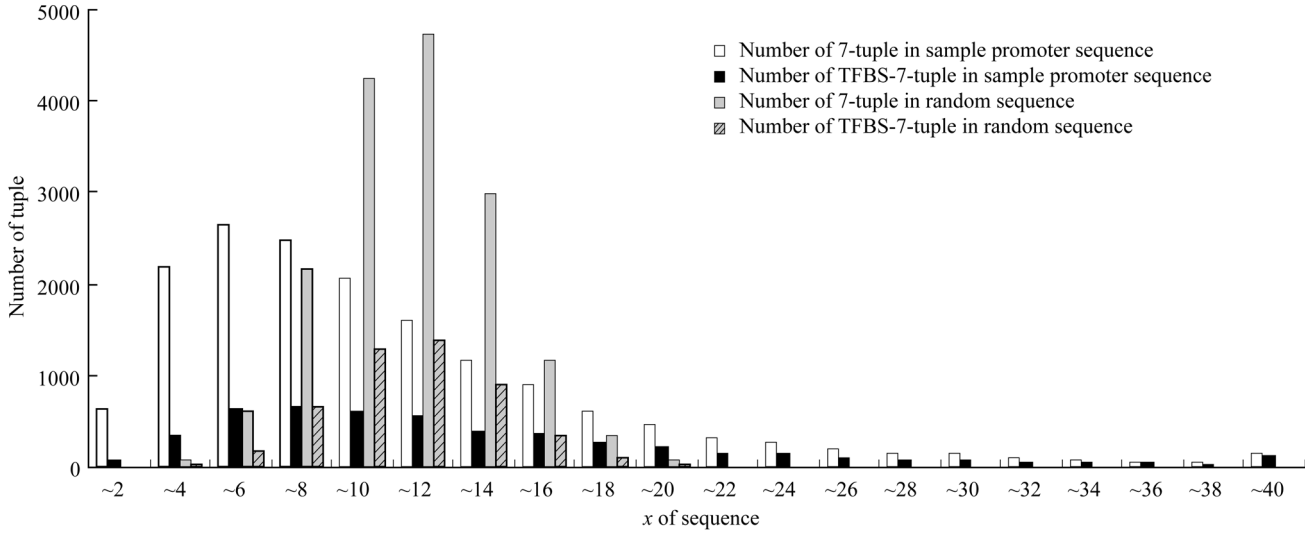

Fig. 1 The distributions of 7-tuple frequencies in promoter sequences and random sequences

The vertical axis is the number of the tuple that has corresponding appearance frequency shown in the horizontal axis. The 7-tuple and TFBS-7-tuple frequencies in 565 promoter sequences and 565 random sequences were analyzed.

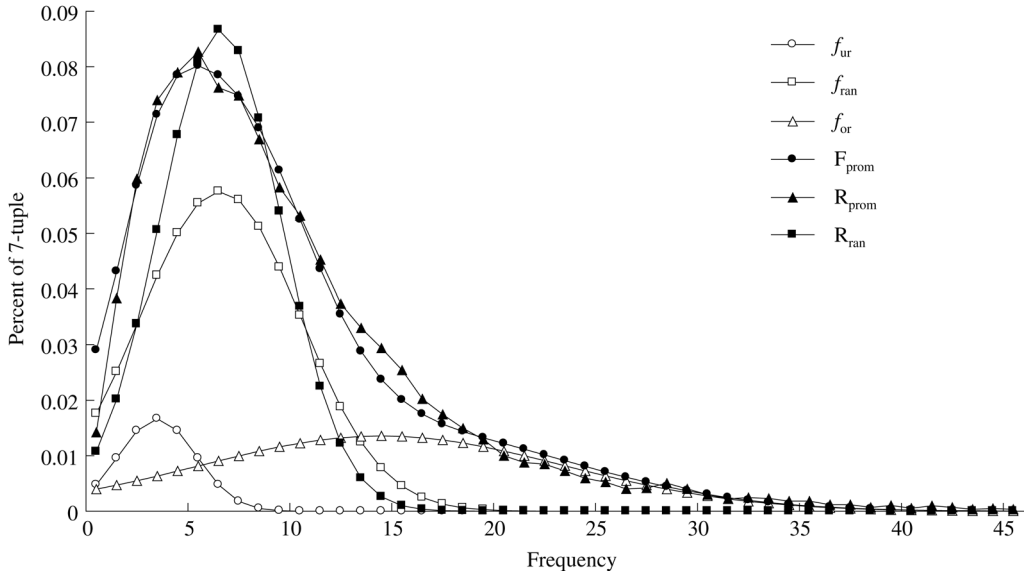

Fig. 2 Decomposition of frequency distribution [$N(x)/4^7$] of 7-tuple

The vertical axis is the percent of 7-tuple having corresponding appearing frequency x, indicated by the horizontal axis. R_{ran}, distribution of 7-tuple in random sequences; R_{prom}, 7-tuple distribution in promoters; F_{prom}, the sum of three component distributions f_{ur}, f_{or}, and f_{ran}; f_{ur}, weighted distribution of the under-represented 7-tuple; f_{or}, weighted distribution of the over-represented 7-tuple; f_{ran}, weighted distribution of the background 7-tuple.

sequence consists of some conservative (appearing more frequently than by chance) and biologically significant segments. These segments are a subset of keywords (TFBS), while the sequences between them form a random background. Therefore, we could conclude that the weight matrices computed from the mixture of random and conservative segments would bring in considerable amount of noise for promoter identification. It suggests that one way to amplify the signal carried by those keywords is to skip the random segments and make use of the TFBS. A keyword dictionary (*WD*) with 619 TFBS (from 5 bp to 12 bp) is built as described in section "Algorithm description".

Promoter identification by single TFBS scoring system (SSS)

By dividing the 565 promoter sequences and the 890 coding sequences into 60% and 40% as learning and testing dataset randomly, we apply cross-validation test for the single TFBS scoring system (SSS). Each keyword w in the *WD* is assigned a score [Formula (4)]

$$s(w) = \log\left[\frac{f_p(w)}{f_{np}(w)}\right] \quad (4)$$

where $f_p(w)$ and $f_{np}(w)$ are the frequencies of keyword w in promoter and non-promoter sequences in learning dataset respectively. As described in section "Algorithm description", we obtain a threshold by optimizing the *FP* (False Positive) and *FN* (False Negative) rate in learning dataset. With the threshold, the testing sequences are classified as promoters or non-promoter sequences according to its score. Then we obtain the true positive rate and true negative rate in test dataset. The cross-validation procedure is repeated 9000 times. The average of true positive rates and true negative rates in test data are 77.6% and 81.3% respectively.

Dependency of the TFBSs in over-represented pairs

In light of the knowledge of the cooperation among TFBSs in transcription initiation, it is natural to consider pair even triple of keywords. Since the currently available sample promoters experimentally verified are very limited, only pairs of keywords are considered here.

Totally there are 191,271 possible combinations from the 619 words in our *WD*. To build the pair dictionary (*PD*), we select a pair if it appears more than four times in promoter dataset and its appearance frequency in the promoter dataset is four times higher than that in the coding sequences as well. Then a *PD* including 3155 pairs is built. We take the following *D*-score [Formula (5)] to evaluate the dependency of the TFBSs $w1$ and $w2$ of each pair ($w1$, $w2$) in *PD*, where $f(w1,w2)$, $f(w1)$ and $f(w2)$ are the frequencies of pair ($w1$,$w2$), word $w1$ and $w2$ respectively.

$$D(w1,w2) = f(w1,w2)\times\log\left[\frac{f(w1,w2)}{f(w1)\,f(w2)}\right] + [1-f(w1,w2)]\times\log\left[\frac{1-f(w1,w2)}{1-f(w1)\,f(w2)}\right] \quad (5)$$

The distribution of the *D*-score of pairs in our dictionary is shown in Fig. 3. Two TFBSs in a pair are considered to be independent if the *D*-score is 0. One can see that the *D*-scores of a large part of pairs in *PD* are not close to 0, which indicates that the TFBSs in the pairs in *PD* are mostly dependent.

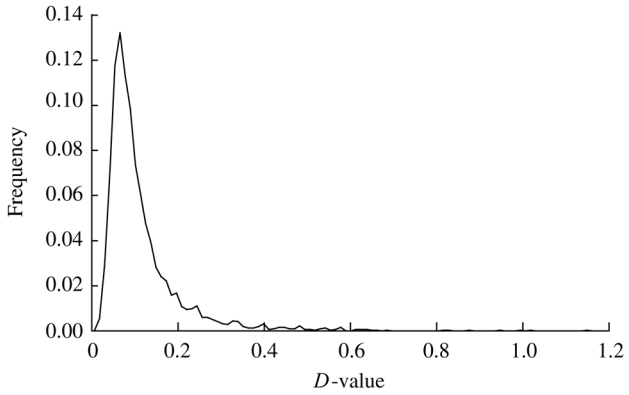

Fig. 3 *D*-value of the 3155 pairs we selected

The vertical axis is the percentage of TFBS pairs that have the corresponding *D*-value shown in the horizontal axis.

Clustering the TFBS pairs

One transcription factor may bind to various sites. Hence we group the TFBS pairs into clusters, which may be bound by the same transcription factor pairs. We use the Hamming distance of two TFBSs to group the 3155 pairs in the last section. Totally 235 clusters are obtained. Among them, 82 clusters (with 1551 pairs) are TATA-related and 16 clusters (with 297 pairs) have no TATA-box but are CAAT-related. The remaining 137 clusters contain 1307 pairs that are mostly GC rich TFBSs. In Table 2, some typical sequences of the clusters of TATA-CAAT pattern and the corresponding numbers of their members are listed. It can be seen that the complementary sequence of CAAT-box, GATTGG, is also statistically significant. We guess that it may be functionally similar to CAAT-box and we consider it as CAAT-box.

It is interesting that some TFBS pairs we selected statistically coincide with the experimentally selected cooperative pairs by comparing our pairs with the FACTOR table in TRANSFAC3.5. We find that 71% (2240 of the 3155) pairs in our *PD* have corresponding transcription factor pairs with known cooperation. We also compared

Table 2 TATA-CAAT pattern clusters

Center of cluster	Number of members
CCAAT-TATAA	42
CCAAT-GGGCGG	25
ATTGAA-TATAA	13
GATTGG-TATAA	14
ATTGC-TATAAA	12
ATTGG-GGGCGG	29
GCAAT-TATAAA	27

our TFBS pairs with the database COMPEL2.4. Among the 150 Composite regulatory Elements (CE) in COMPEL2.4, 24 are found in our dictionary. 10 of the 24 CEs match with our TFBS pair very well and are listed in Table 3. For example, the TFBS pair CTGGGTAAAAT has its counterpart C00051. The sequence of C00051 is CTGGGAAgat…aaATTAAATATTAAC, with the capital letters representing the binding sites of transcription factors IL-6 RE-BP and HNF-1 respectively.

We also find that there are some TFBS-triples over-represented in promoters relative to in coding sequences. Among them, several TATA-box related triples are listed in Table 4. But with the current limited promoter data, it is impossible to make stable statistics since most of the triples do not have appearing frequency high enough in such a small dataset.

Table 3 Some TFBS pairs in database COMPEL2.4 that also presented in our dictionary

TFBS pair	Counterpart in COMPEL24
CTGGG -TAAAAT	C00051
TAAAT-TGACG	C00096
CCGCCCCC-CGCGG	C00132
ATAAATA - MAMAG	C00041
GCCCC-TAAAT	C00046, C00045
TATTT-TAAAT	C00045
TAAAT-CTGGG	C00051
TAAAT-CCTGC	C00045
GAGGA-TATAAA	C00083

Table 4 Appear time of some TATA related TFBS triples over-represented in the 565 sample promoters while under-represented in the 890 sample CDS

TFBS triple	In prom	In cds
TATAAA-CCAAT-GGCGG	18	0
TATAAA-CATTT-CCAAT	19	1
TATAAA-CATTT-CAGAG	18	3
TATAAA-CCAAT-TCTCC	16	1
TATAAA-CATTT-AAGGAA	13	0
TATAAA-CCAAT-GGGCA	12	0

Promoter Identification by TFBS pair scoring system (PSS)

We also apply the same cross-validation procedure for PSS as that for SSS. Each keyword pair is assigned a score [Formula (6)]

$$S(w1, w2) = \log\left[\frac{f_{\mathrm{p}}(w1, w2)}{f_{\mathrm{np}}(w1, w2)}\right] \tag{6}$$

where $f_{\mathrm{p}}(w1,w2)$ and $f_{\mathrm{np}}(w1,w2)$ represent the frequency of TFBS pair $(w1,w2)$ in learning promoter and non-promoter sequences respectively. We perform the cross-validation test 1000 times (less than 9000 times for sake of saving time). The average true positive and true negative rates are 84.1% and 82.4% respectively. The result comparing with that of single TFBS is summarized in Table 5. We can see that considering the interaction between TFBSs does improve the discrimination rate. We then take the 565 promoter sequences and 890 coding sequences as learning data and make use of the dictionary with 3155 TFBS pairs obtained in "Dependency of the TFBSs in the over-represented pairs".

Table 5 TP rate and TN rate in cross-validation tests by SSS and PSS

	TP rate	TN rate
PSS	84.1%	82.4%
SSS	77.6%	81.3%

Promoter identification by TFBS Pair Scoring System with Distance (PSSD)

One natural question is whether the distance between the TFBSs in pairs is useful for promoter identification. Since the learning dataset is too limited, we use the TFBS pair clusters in "Clustering the TFBS pairs" to get the distance distribution. To evaluate how the distance works, we give pairs score with consideration of the distance between TFBSs in learning promoter dataset. Here only 70% of the 565 promoters with high TFBS density are taken into comparison, since when the total number of TFBSs in learning sample promoter is very small, it may not be a real core promoter. The score for each TFBS pair $(w1, w2)$ with minimum space d in promoter is defined as [Formula (7)]

$$s(w1, w2, d) = P_{C(w1,w2)} \times \log\left[\frac{f_p(w1, w2)}{f_{np}(w1, w2)}\right] \times P\left[d(w1, w2) = d\right] \tag{7}$$

where $C(w1,w2)$ is the cluster that contains the TFBS pair $(w1,w2)$ and $P_{C(w1,w2)}$ is the percentage of cluster $C(w1,w2)$ in our pair dictionary and $P[d(w1,w2)=d]$ is the probability of the minimum space between $w1$ and $w2$ being d. Shown in Table 6 is the change of the discriminate rates in the testing dataset with and without considering the TFBS distance information. Obviously the information of distance between the TFBSs does improve the discriminate rate significantly. In the following we will use the PSSD as our promoter identification algorithm.

Promoter identification in Gene sequences with known ATG

For gene sequences with start site ATG known, which is often available from such as RefSeq mRNA or EST sequences, the PSSD can be applied by scanning the upstream sequences of ATG with a 300 bp-sliding window and 10 bp step (for consistency with the learning data). The following is an interesting prediction result of the

Table 6 The change of the true negative rate when the true positive (TP) rate of PSS and PSSD change

TP rate	TN (PSS)	TN (PSSD)
0.99	0.298	0.368
0.95	0.600	0.876
0.90	0.754	0.897
0.85	0.856	0.901
0.80	0.895	0.918
0.75	0.913	0.927
0.70	0.945	0.935

The left column is the TP rates with different thresholds. The corresponding true negative rates of PSS and PSSD are represented by TN (PSS) and TN (PSSD) respectively.

human gene CKLFSF1, a member of CKLF gene family discovered by Xu [14]. The prediction was confirmed by wet experiment later. This is the first example that the promoter of a gene locates in the intron of another gene, which means that there is no absolute sequence difference between promoters and introns—a sequence could be both a part of promoter and a part of intron, and there is no way to distinguish them unless considering how they related to their coding sequences. The promoter predicting result by the PSSD is shown in Fig. 4. FirstEF [11] successfully predicts the first exon of CKLFSF1 but it fails to give the core promoter region. PromoterInspector has no predictions. This result also suggests that the prediction results should be carefully evaluated when a predicted promoter is not very close to TSS, or in the gene region (from the translation start site ATG to the stop site TAA, including introns).

Promoter identification in RefSeq database

We try to annotate the core promoters of human mRNA sequences in RefSeq by our algorithm. We first retrieve all human mRNAs in RefSeq from NCBI (ftp://ncbi.nlm.nih.gov) with restriction "Homo sapiens" and 17903 sequences are obtained. We align these sequences back to the human genome and 3402 sequences that have matching rate more than 90% in the coding regions are obtained. Then we extend [19] them from 5000 bp upstream to 500 bp downstream of the annotated translation start site ATG and build a dataset named RSPD. Shown in Fig. 5 is the length distribution of the 5′ UTR of the 3402 mRNAs in RSPD. It suggests that the 5′ UTR length of mRNA is rather conservative.

Then we scan promoter along the sequences in RSPD with the PSSD. The window with the maximum score is predicted to be core promoter if it is higher than the threshold determined in "Promoter identification by TFBS Pair Scoring System with Distance (PSSD)". No promoter will be predicted if the maximum score is lower than the threshold. Shown in Fig. 6 is the genomic distance distribution from the predicted core promoter region to the 5′ end of the mRNAs in RSPD. 1771 genes (52%) are predicted to have core promoter nearby the 5′ end region [–500, +500] while 1251 genes (37%) have core promoter in the region of [–3000, –500]. No significant core promoters are reported in the region [–3000, +500] for the remaining 379 genes (11%).

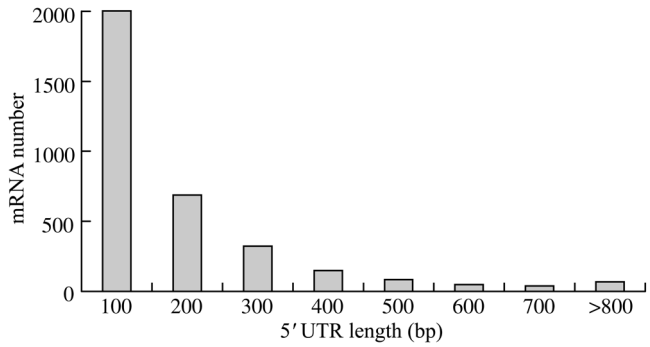

Fig. 5 Length distribution of the 5' UTR of the 3402 mRNAs

Shown in the vertical axis is the number of mRNAs that have corresponding 5' UTR length (in bps) shown in the horizontal axis.

Performance Evaluations

Comparison on the data by Fickett

Eighteen independent eukaryotic sequences with 20

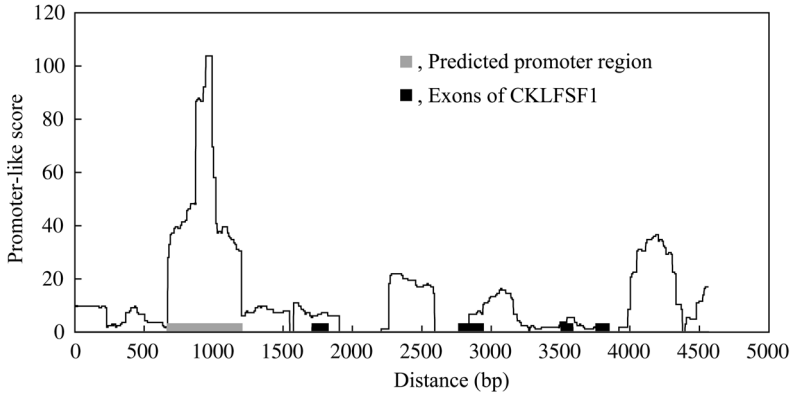

Fig. 4 Moving promoter-like score of gene CKLFSF1 by the PSSD

The gray bar is the predicted promoter region while the black bars above the horizontal axis are the exons of the newly discovered gene CKLFSF1 [14]. The predicted promoter region lies between the 3rd and 4th exon of the upstream gene CKLF1 (data not shown, see [14]).

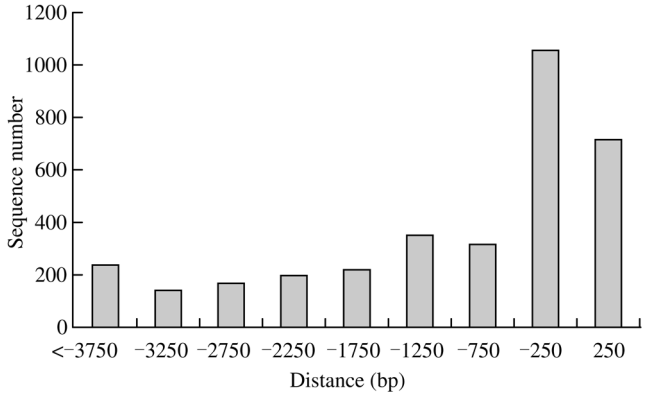

Fig. 6 Genomic distance distribution from the predicted core promoters to the 5′ end of the 3402 mRNAs

The horizontal axis is the distance (in bp) from the predicted promoter to the annotated TSS while the vertical axis is the corresponding numbers in the 3402 genes.

experimentally mapped TSS have been compiled by Fickett *et al*. [9] to evaluate the performance of the existing promoter predicting programs. We apply our algorithm on seventeen of these sequences data, since one anonymous sequence (Chu *et al.*) can not be found in either GenBank or EMBL. As what Fickett *et al*. had done, we also take predictions from 200 bp upstream to 100 bp downstream of TSS as correct predictions. The result is summarized in Fig. 7. It is clear that considering the distance between TFBSs in each pair can increase the specificity rate.

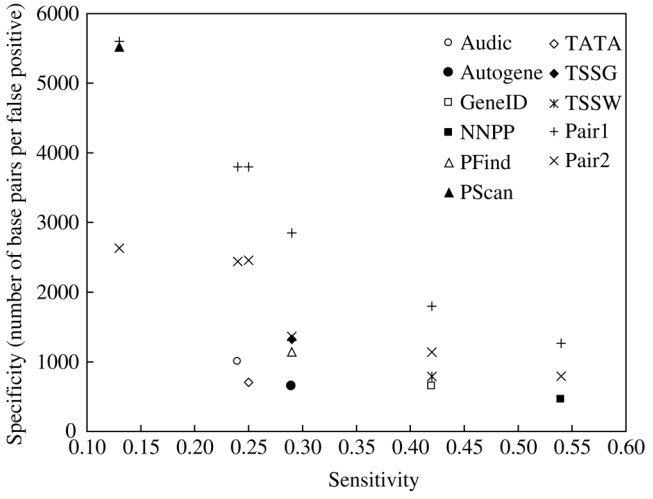

Fig. 7 Comparison with other programs on the data by Fickett *et al.*

The horizontal axis is the sensitivity and the vertical axis is the specificity (base pairs per prediction). The '+' (Pair1) and '×' (Pair2) are the result of the PSSD and PSS respectively. Multiple '+' and '×' represent the different specificity rates corresponding to different sensitivity rates accommodated to different programs in Fickett *et al.* [9].

Some predicted promoters that are not very close to the annotated TSS (12 of the 18 annotated TSSs have predicted promoter around them) might actually be the core promoters, but we still count them as false positives since there are not enough positive or negative evidences for core promoters from wet experiments. Hence the specificity for core promoters may be actually lower than it should be.

TFBS clusters and CpG-island

Although the PSSD reduces the false positive rate, it can be seen from the result in "Comparison on the data by Fickett" that the false positive rate is still too high to predict promoters in genomic sequences. Noticing that the TSS is strongly related to the CpG-island (see [11,20]), while the core promoters are closely related to TSS, we integrate the PSSD score and the CpG-island information together to reduce the false positive rate in terms of artificial neural networks [ANN (provided by Zhang Cheng-Fu, personal communication)]. 77 promoter sequences with length of 700 bps from dataset HMR195 compiled by Rodgic [17] are used as positive data points. The input for the ANN is the CpG feature and the PSSD feature, where the CpG features are defined in Zhang [20]. 687 segments with length of 700 bp extracted from translation start codon to stop codon in Rodgic [17] are used as negative data points. The data are processed by the ANN and plotted in Fig. 8. It can be seen that 60% core promoter region data points (46 of the 77 promoter regions) are correctly classified with 55 false positives. This result is consistent with that of FirstEF, which successfully predicts 60% (46 of the 77) TSS regions with 35 false positives. PromoterInspector predicts 30% (23 of the 77) TSS regions with 10 false positives.

Discussion

We have reported a new approach to analyze core promoters in eukaryotic genes. It is mainly based on the potential cooperation between transcription factors and their binding sites. It is biologically reasonable and is useful for the experimental biologists. Our preliminary study shows that this method has promising performance even when it is applied to genomic sequences. Moreover, the result strongly supports the basic assumption of this method: the promoter sequence should be interpreted as a number of keywords sitting in a random sequence background. The fact that 71% of the TFBS pairs in our dictionary are

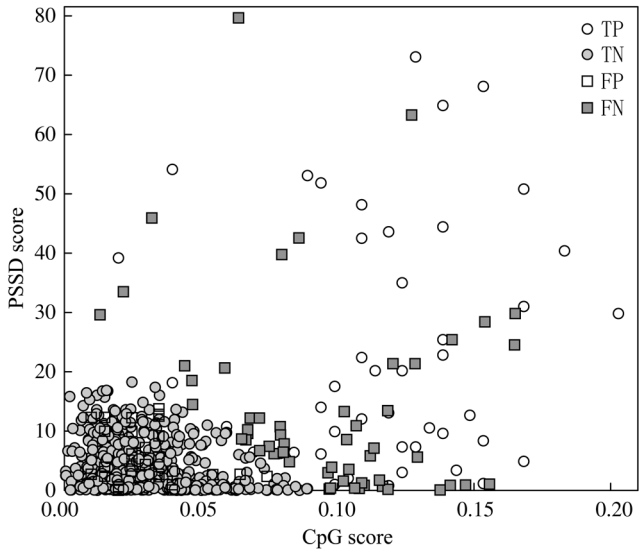

Fig. 8 Scatter plots of the CpG score and the PSSD score in promoter region and gene region

77 promoter data points are represented by 'white circle' (correct classification by the ANN) and 'white square' (wrong classification by the ANN). 'gray circle' (correct classification by the ANN) and 'gray square' (wrong classification by the ANN) signs are for 687 gene region data points.

associated by transcription factors gives the biologists light of finding cooperative TFBSs through statistics model to reduce the candidate numbers. Our result suggests that the cooperation between the TFBSs contributes much to the correctly transcriptional initiation of the genes. The result on RefSeq sequences may reflect two possibilities of the current RefSeq dataset: the lack of full 5′ UTR region of the genes or the existence of far upstream core promoters relative to the TSS. Our result on gene CKLFSF1 suggests the existence of the latter possibility—at least there are some genes with far upstream promoters. Maybe it is more appropriate to use the triples of TFBSs or more when there are enough data. This work is the first but main step of our efforts towards a more accurate algorithm for identifying the complete structure of a gene along the DNA sequences.

Acknowledgements

We thank Wen-Ling HAN in the Medical School of Peking University for her discussion on the biological background of promoters. We also thank Cheng-Fu ZHANG for providing advices and his program on Artificial Neuron Network (ANN).

References

1 Zhang MQ. Identification of human gene core promoters in silico. Genome Res, 1998, 8: 319–326
2 Roeder RG. The role of general initiation factors in transcription by RNA polymerase II. Trends Biochem Sci, 1996, 21: 327–335
3 Qiu P. Computational approaches for deciphering the transcriptional regulatory network by promoter analysis. BioSilico, 2003, 1: 125–133
4 Wagner A. Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. Bioinformatics, 1999, 15: 776–784
5 Ioshikhes I, Trifonov EN, Zhang MQ. Periodical distribution of transcription factor sites in promoter regions and connection with chromotin structure. Proc Natl Acad Sci USA, 1999, 96: 2891–2895
6 Pedersen AG, Baldi P, Chauvin Y, Brunak S. The biology of eukaryotic promoter prediction—a review. Computers & Chemistry, 1999, 23: 191–207
7 Fickett JW, Wasserman WW. Discovery and modeling of transcriptional regulatory regions. Current Opinion in Biotechnology, 2000, 11: 19–24
8 Prestridge DS. Prediction Pol II promoter sequences using transcription factor binding sites. J Mol Biol, 1995, 249: 923–932
9 Fickett JW, Hatzigeorgiou AG. Eukaryotic promoter recognition. Genome Res, 1997, 7: 861–878
10 Scherf M, Klingenhoff A, Werner T. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: A novel context analysis approach. J Mol Biol, 2000, 297: 599–606
11 Davuluri RV, Grosse I, Zhang MQ. Computational identification of promoters and first exons in the human genome. Nature Genetics, 2001, 29: 412–417
12 Wagner A. A computational genomics approach to the identification of gene networks. Nucleic Acids Res, 1997, 25: 3594–3604
13 Tronche F, Ringeisen F, Blumenfeld M, Yaniv M, Pontoglio M. Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome. J Mol Biol, 1997, 266: 231–245
14 Xu M, Han W, Qian M, Ma X, Ding P, Wang Y, Xia D *et al*. Last intron of the chemokine-like factor gene contains a putative promoter for the downstream CKLF super family member 1 gene. Biochem Biophy Res Comm, 2004, 313: 135–141
15 Wingender E, Chen X, Fricke E, Geffers R, Hehl R, Liebich I, Krull M *et al*. The TRANSFAC system on gene expression regulation. Nucleic Acids Res, 2001, 29: 281–283
16 Perier RC, Junier T, Bonnard C, Bucher P. The eukaryotic promoter database (EPD): Recent developments. Nucleic Acid Res, 1999, 27: 307–309
17 Rogic S, Mackworth AK, Quellette FBF. Evaluation of gene finding programs on mammalian sequences. Genome Res, 2001, 11: 817–832
18 Kel-Margoulis OV, Kel AE, Reuter I, Deineko IV, Wingender E, TRANSCompel®: A database on composite regulatory elements in eukaryotic genes. Nucleic Acid Res, 2002, 30: 332–334
19 Chen Z, Wang L, Sun H, Qian M. Error-tolerating searching and its application in translational signal extending. The 2nd Chinese Conference on Bioinfomatics, 2002, 42
20 Zhang MQ. Computational methods for promoter recognition. In: Jiang T, Xu Y, Zhang MQ eds. *Current Topics in Computation Molecular Biology*. Boston: MIT Press, 2002, 249–268

Edited by
Da-Fu DING