PromSearch: A Hybrid Approach to Human Core-Promoter Prediction

Byoung-Hee Kim, Seong-Bae Park, and Byoung-Tak Zhang

Biointelligence Laboratory, School of Computer Science and Engineering
Seoul National University, Seoul 151-744, Korea
{bhkim,sbpark,btzhang}@bi.snu.ac.kr

Abstract. This paper presents an effective core-promoter prediction system on human DNA sequence. The system, named PromSearch, employs a hybrid approach which combines search-by-content method and search-by-signal method. Global statistics of promoter-specific contents are included to represent new significant information underlying the proximal and downstream region around transcription start site (TSS) of DNA sequence. Local signal features such as TATA box and CAAT box are encoded by the position weight matrix (PWM) method. In the experiment for the sequence set from the review by J.W.Fickett, PromSeach shows 47% positive predictive value which surpasses most of previously systems. On large genomic sequences, it shows reduced false positive rate while preserving true positive rate.

1 Introduction

A promoter is a part of DNA sequence that regulates gene expression, i.e. initiates and regulates the transcription of a gene. Gene regulation is one of the most important research topics in molecular biology and is closely related with several hot issues like regulatory network construction and target gene finding. It is therefore very important to identify regulatory regions exactly in DNA sequences, in that it makes it possible for one to examine the characteristics of the regions in more detail and to understand mechanism that control the expression of genes. Since huge amount of data out of DNA sequencing is being generated rapidly and requirement for gene/protein research emerges constantly, a reliable computational prediction of promoter regions becomes more indispensable to cellular and molecular biology fields than ever before.

There have been many algorithms and systems to predict promoter region computationally, especially promoter of higher eukaryotes, but few of them are applicable to real DNA sequence mainly owing to high false positive rates [4].

The approach to computational promoter prediction can be categorized roughly into three: search-by-signal approach, search-by-content approach and combination of the two [5]. Search-by-content algorithms identify regulatory regions by using measures based on the sequence composition of promoter and non-promoter regions. Search-by-signal algorithms make predictions based on the detection of core pro-

Z.R. Yang et al. (Eds.): IDEAL 2004, LNCS 3177, pp. 125–131, 2004.

moter elements such as the TATA box, the initiator, and/or transcription factor binding sites outside the core.

PromSearch takes a hybrid approach. From the search-by-signal point of view, we utilize well-known position weight matrices (PWMs) of four core-promoter elements [3]. As a search-by-content approach, we introduce new features extracted from wide range around core-promoter region. Searcy-by-content approach has a tendency to lose positional information. These shortcomings are supplemented by features from core element signal in our system and PromSearch includes a two-step criterion to determine transcription start site (TSS) position on DNA sequences of human more reasonably.

2 Methods and Materials

Fig. 1 shows the outline of the PromSearch system. PromSearch takes four steps to predict TSS position on the fragment within a window which slides every 10 base pairs (bp) from 5' to 3' direction on the input DNA sequence.

The first step is applying our model on the main target region of PromSearch, which leads from 250bp upstream from TSS to 50bp downstream including TSS, total 300bp region. We represent this region as [-250, 49][1]. Fig. 2 shows the model on this region. We take two sorts of features into account: locally over-represented signals and globally distinguishable content information.

There are four well-known core promoter elements; INR (initiator or cap signal), TATA-box, GC-box and CAAT-box [3]. These can be regarded as good local signals.

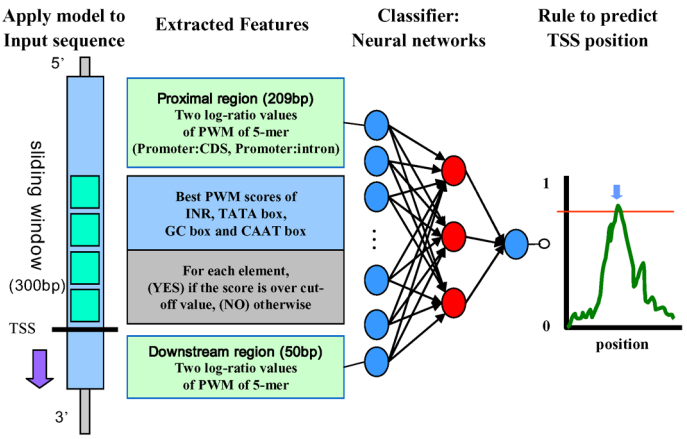

Fig. 1. Outline of the PromSearch system.

[1] Usually, the TSS position is marked by +1, where there is no 0 point and the coordinate immediately 1bp up from TSS is -1.In this paper, for the description purpose, we denote TSS position by 0, or the origin.

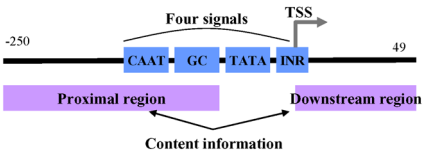

Fig. 2. Model on the 300bp region around TSS applied to PromSearch. The model combines locally overrepresented signals with globally distinguishable content information around TSS.

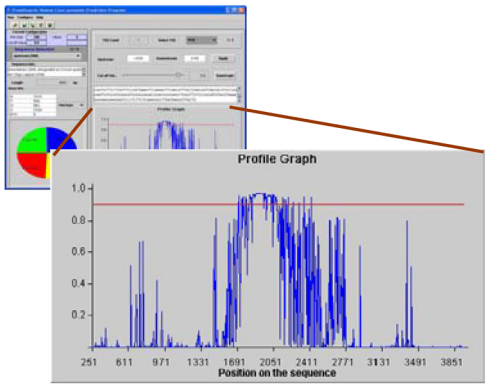

Fig. 3. (right) Software implemented by Java. Profile graph shows a 'plateau' around TSS.

But the combination and constitution of them are not well conserved among promoters and there are many fakes on non-promoter region that resemble these elements. So, these local signals are not sufficient to locate promoter accurately. Hence, we extract more global features out of proximal region and downstream region by search-by content approach.

Feature Extraction and Classification

We utilize position weight matrix (PWM) scores of core promoter elements as features to catch local signals. PWMs for each core element included in TRANSFAC [10] 6.0 are out-dated and built from only 502 eukaryotic RNA polymerase II promoter sequence. So, we have constructed new PWMs out of 2,538 vertebrate promoter sequences of eukaryotic promoter database (EPD) [8] release 76 with the aid of the pattern optimization (PatOp) algorithm of Bucher [3], which is served at the SSA web site [1]. Bucher's algorithm produces another two output values other than matrix: applying range of the matrix and cut-off value. We apply each PWM on its own applying range and searched the position where the score has maximum value and checked if the score is over cut-off value. These maximum score within given range and a binary value which represents if the score is over cut-off value, are used as representative features of corresponding promoter element. So, there are two features for each element, total 8 features.

We apply 5-mer pseudo-PWM score, as a measure for promoter and non-promoter. We introduce the definition that was used in dragon promoter finder (DPF) [2]. Two 5-mer PWMs are constructed from proximal region, [-250, -42], and downstream region, [1, 50], respectively. As non-promoter examples, we construct additional 5-mer PWMs from coding sequence (CDS) and intron sequences. Two are from 209 bp sequences of CDS and intron respectively as comparison examples of proximal PWM. Also for downstream PWM, two 5-mer matrices are built out of 50bp sequences of CDS and intron, respectively. So, there are three 5-mer PWMs for proximal and downstream region, respectively. Now, we construct two new features for each region. On the 209bp-long sequence of input window that corresponds to the

proximal region of the model in fig 2, we produce three scores (v_p, v_c, v_i) from each PWM and calculate log ratio values (F_1, F_2) as following: $F_1 = \ln(v_p / v_c)$, $F_2 = \ln(v_p / v_i)$. Same process goes on a 50bp long sequence. This results in four content features.

As described above, PromSearch extracts total 12 features from input window of which size is 300bp. The next step is to judge if the window contains core-promoter region using these features. This process results in a binary classification.

We use neural networks as a classifier. Training set is made up of positive samples composed of 1871 human promoter sequence from EPD release 76 and negative samples constituted with 890 CDS and 4345 intron sequences of human from 1998 GENIE set, which was extracted from GenBank version 105. On the input layer of the neural network, we apply 'standardization' to all input features, i.e., extract mean value and divide by standard deviation. Trained network was built with the aid of Weka 3.4 S/W. On the output node of the neural networks, we omit the final classification step and take the output value of sigmoid function as a profile for given sequence within a window. The final step is replaced by a step to decide a TSS position.

Deciding TSS Position and Evaluation

If current window is classified as containing core-promoter, we reflect signal information to determine TSS position. First, if INR PWM score is larger than cut-off value, we take the origin of INR PWM as TSS position. If INR_cut value is 0 and TATA_cut value is 1, 30bp downstream from the origin of TATA PWM is predicted as a TSS. Otherwise, 50bp upstream from the 3' end of the input window is selected. We define a 'core range' as the region from 5' end of TATA region to 3' end of INR region. In case of 300bp window size, the core range is 61bp. If several TSS' predicted within a core range, they are regarded as one. With this criterion, unreasonable duplicative check, or overlapping of several core ranges, is prevented.

Applying previous criterion, there are two problems to be solved. First, there is a trend that the profile score jumps up before TSS and forms a plateau (see fig. 3). We decide a series of TSS as a plateau if any adjacent two are within 250bp. On the basis of analysis on DBTSS [9], we select a point 152bp down from jump-up point of any plateau as a TSS. Second, we need to adjust the cut-off value on the final decision step of the classifier. On the usual output node of neural networks, the class is determined as positive if the probability is over 0.5 and negative if less than 0.5. But, the value 0.5 produces many meaningless false positives and we need to apply more strict condition to reflect the tendency of jumping-up near TSS. All the results on the following section are from the system with 0.9 as new cut-off value.

Fig. 3 shows the software implemented in Java. With this software, we analyzed the sequence set from the review by J.W.Fickett [4], long sequence set that was used to evaluate PromoterInspector [6] and human chromosome 22 release 3.1b. Fickett's data consist of 24 promoters covering a total of 18 sequences and 33,120 bp. Long sequence data from the paper of (Scherf et al.) [6] consist of 35 promoters covering a total 6 sequences and about 1.38 million bp. We follow the criterion of Fickett in his paper [4]: we consider our prediction is correct if predicted TSS is within 200 bp upstream, or 100 bp downstream of any experimentally mapped TSS.

3 Results

Results on the Fickett's data set are presented in table 1. PromSearch shows better positive predictive value (PPV), which represents the reliability of positive predictions of the induced classifier, except PromoterInspector. But, considering that PromoterInspector predicts promoter 'regions', PromSearch has an advantage as a TSS predictor submitting to slightly more false positive rate. It also shows better PPV than DPF.

Table 2 shows results from several systems on several long sequences. Result of DPF1.4 was not available because the service on WWW supports only the analysis of maximum 10,000 bp sequence. Results except PromSearch are from [6]. L44140 shows very different pattern. PromSearch shows 98 FPs on L44140 and PromoterInspector also shows 14 FPs on this sequence, resulting in 3 times more FPs. Excluding L44140, PromSearch shows 17.8% PPV and one false positive in every 25kbp. It is far better than TSSG which shows 3.9% of PPV in this case.

Table 3 shows the result of the analysis of human chromosome 22 release 3.1b, applying the same conditions and criteria with [7]. One out of about ten predictions are supported by current gene annotation and the total number of predictions (3,517) is far less than that of other programs (11,890 or more) [7], which means reduced FP rate.

Table 1. Results for the Fickett data set. It is based on the assumption of [6] : the gene orientation is known, i.e. if the experimental TSS is found on the sense strand of the sequence, only TSS predictions on the sense strand were considered. The result of PromoterInspector is from [6]. We set the sensitivity of DPF1.4 as 65% for the data below. TP: true positive, FP: false positive, PPV (positive predictive value) = TP/(TP+FP), Sensitivity = TP/(TP+FN).

Method	TP	FP	PPV	Sensitivity	Specificity(FP rate)
PromFind	11	24	0.31	0.46	1/1,380
TSSG	10	17	0.37	0.42	1/1,948
TSSW	14	33	0.30	0.58	1/1,004
PromoterInspector	7	3	0.7	0.29	1/11,040
DPF1.4 (65% Se)	10	19	0.34	0.42	1/1,743
PromSearch	8	9	0.47	0.33	1/3,680

Table 2. Results for the dataset consisting of long promoter sequence from [6]: AC002397, D87675, AF017257, AF146793, AC002368. We exclude L44140 which shows exceptional pattern. Total length is about 1.16million bp and the number of TSS is 24.

Method	TP	FP	PPV (%)	Sensitivity	Specificity(FP rate)
TSSG	11	274	3.9	0.46	1/4,234
TSSW	11	317	3.4	0.46	1/3,660
NNPP2.1	15	2,979	0.5	0.63	1/389
Promoter2.0	6	1,408	0.4	0.25	1/824
PromoterInspector	10	5	66.7	0.42	1/232,048
PromSearch	10	46	17.9	0.42	1/25,223

Table 3. The results on the analysis of human chromosome 22 release 3.1b.

	Number of annotated genes	Annotation-supported prediction	Sensitivity
All genes	936	300	0.32
Coding genes	393	143	0.36
Additional predictions	3,217	PPV = 8.53%	

4 Discussion

PromSearch is a hybrid system of search-by-signal and search-by-content approach. It combines features of four core-promoter elements with that of proximal and downstream region. Newly introduced features are from proximal and downstream regions and they help to discriminate promoter from non-promoter. We have set some rules to locate TSS more reasonably and precisely. We utilize the information that can be extracted from PWM of TATA box and INR. The profile graph built with the values from neural networks output node shows notable 'plateaus' around TSS in the middle of low and flat region on non-core-promoter part. Reflecting an analysis on all human sequences in DBTSS, we determine a position on a plateau as a TSS.

Experiments demonstrated that PromSearch predicts TSS position with much lower false positive rates than most of previous systems. One out of about five predictions can be expected to be true positives. PromoterInspector, which shows superior performance, focuses on the genetic context of promoters and predicts some promoter region, rather than their exact location [6]. Experiments show that PromSearch can be a good alternative for those who want exact TSS location submitting to the increase of FP. DPF [2] takes multiple model approach on vertebrate promoters and it predicts TSS position based on search-by-content approach only. We applied its 5-mer PWM concept but processed the score in different view and we tried to locate TSS position more robustly by search-by-signal approach as well as a two-step prediction rule.

The main advantage of PromSearch comes from the fact that it produces profile that constitute notable plateau around TSS with reduced FPs and it predicts a TSS more precisely based on the signal information and the result of analysis on DBTSS.

References

1. Ambrosini, G., Praz, V., Jagannathan, V., Bucher, P.: Signal search analysis server, *Nucleic Acids Res*., 31, 3618-3620, 2003.
2. Bajic, V.B., Chong, A., Seah, S.H., Krishnan, S.P.T., Koh, J.L.Y., Brusic, V.: Computer model for recognition of functional transcription start sites in polymerase II promoters of vertebrates, *Journal of Molecular Graphics & Modeling*, 21, 323-332, 2003.
3. Bucher, P.: Weight matrix description of four eukaryotic RNA polymerase II promotor elements derived from 502 unrelated promotor sequences. *J Mol Biol*., 212, 563-578, 1990.

4. Fickett, J. W., Hatzigeorgiou, A.G.: Eukaryotic Promoter Recognition. *Genome Research*, 7, 861-878, 1997.
5. Ohler, U.: Computational promoter recognition in eukaryotic genomic DNA, PhD thesis, Technische Fakultät Erlangen-Nürnbeg, 2001.
6. 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*., 297, 599-606, 2000.
7. Schef, M., et al.: First Pass Annotation of Promoters on Human Chromosome 22. *Genome Research*, 11, 333-340, 2001.
8. Schmid, C.D., Praz, V., Delorenzi, M., Périer, R., Bucher, P.: The Eukaryotic Promoter Database EPD: the impact of in silico primer extension. *Nucleic Acids Res*. 32, D82-5, 2004.
9. Suzuki, Y., Yamashita, R., Nakai, K., and Sugano, S.: DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Res*., 30, 328-331, 2002.
10. Wingender, E., et al.: The TRANSFAC system on gene expression regulation. *Nucleic Acids Res*., 29, 281-283, 2001.