Sequence analysis

A pattern-based nearest neighbor search approach for promoter prediction using DNA structural profiles

Yanglian Gan1, Jihong Guan1,∗, and Shuigeng Zhou2,3,∗

1Department of Computer Science and Technology, Tongji University, Shanghai 201804,
2School of Computer Science and 3Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai 200433, China

Received on December 19, 2008; revised on June 5, 2009; accepted on June 6, 2009
Advance Access publication June 10, 2009
Associate Editor: Martin Bishop

ABSTRACT

Motivation: Identification of core promoters is a key clue in understanding gene regulations. However, due to the diverse nature of promoter sequences, the accuracy of existing prediction approaches for non-CpG island (simply CGI)-related promoters is not as high as that for CGI-related promoters. This consequently leads to a low genome-wide promoter prediction accuracy.

Results: In this article, we first systematically analyze the similarities and differences between the two types of promoters (CGI- and non-CGI-related) from a novel structural perspective, and then devise a unified framework, called PNNP (Pattern-based Nearest Neighbor search for Promoter), to predict both CGI- and non-CGI-related promoters based on their structural features. Our comparative analysis on the structural characteristics of promoters reveals two interesting facts: (i) the structural values of CGI- and non-CGI-related promoters are quite different, but they exhibit nearly similar structural patterns; (ii) the structural patterns of promoters are obviously different from that of non-promoter sequences though the sequences have almost similar structural values. Extensive experiments demonstrate that the proposed PNNP approach is effective in capturing the structural patterns of promoters, and can significantly improve genome-wide performance of promoters prediction, especially non-CGI-related promoters prediction.

Availability: The implementation of the program PNNP is available at http://admis.tongji.edu.cn/Projects/pnnp.aspx.

Contact: jhguan@tongji.edu.cn; sgzhou@fudan.edu.cn

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

One of major challenges coming with the genome sequencing wave is to functionally annotate genomes and analyze gene regulatory networks. Promoters are functional regions surrounding the transcription start sites (TSSs) and are responsible for the initiation and regulation of DNA transcription (Pedersen et al., 1999). Gene transcription is controlled by RNA polymerase and various transcription factors binding to the promoters. Consequently, accurate recognition of promoters is an important issue in genomics (Carninci et al., 2006), which may reveal clues about how, where and when the transcription takes place.

Recently, a variety of computational promoter models have been proposed. These approaches to delineating promoter regions are based on a common fact. A primary observation is that the characteristics of promoter sequences are different from that of non-promoters. Current methods depend mostly on local DNA sequence information of core promoter regions (Thomas and Chiang, 2006), such as CpG islands (CGIs) (Bajic and Seah, 2003), TATA boxes (Ohler, 2002), CAAT boxes (Knudsen, 1999), specific transcription factor binding sites (TFBSs) (Solovyev and Shahmuradov, 2003), pentamer matrix (Ohler, 2002) and oligonucleotides (Schierf et al., 2000). Although much progress has been made, recent studies intriguingly indicate that available promoter prediction solutions share two major drawbacks (Solovyev et al., 2006). On one hand, the number of false positives is high at the whole genome level. Due to the complexity and heterogeneity of promoter architectures, not all cis-elements are consistently shared by promoters. Thus, local sequence composition signals alone cannot accurately discriminate promoters from non-promoters. On the other hand, the prediction accuracy of non-CGI-related promoters is quite low. Statistical analysis of promoters has revealed that non-CGI-related promoters account for a large proportion of all promoters in the whole genome, but no model performs satisfactorily on this type of promoters (Bajic et al., 2004). Thus, there is still much space to improve promoter prediction performance.

Up to now, the selection of right biological signals to predict promoters remains an open issue. Current findings indicate that protein-DNA binding specificity is also modulated by energetics (Baldis et al., 1998; Florquin et al., 2005; Goni et al., 2007; Abeel et al., 2008). In the eukaryotic nucleus, over two meters of DNA are packaged in the form of chromatin by folding nucleosome arrays. The higher order structures are stabilized by interactions with other nuclear proteins and associated with the magnitude of gene regulation in response to different signals. Although the structural features are ultimately determined by the nucleotide sequence itself, studies have shown that promoters indeed have distinct structural profiles when compared with coding or non-regulatory sequences. These findings imply that the widely used local signals are insufficient to define promoters. Thus, in this work we try to develop a more sophisticated prediction model by integrating structural and physical information.

∗To whom correspondence should be addressed.
We show that CGI- and non-CGI-related promoters exhibit almost similar changes in their structural profiles. Our comprehensive analysis shows that CGI- and non-CGI-related promoters exhibit similar changes in their structural profiles. These changes are consistent with the observation that CGI-related promoters have higher CpG content. This suggests that the presence of CGI islands is an important feature in promoter prediction.

In this article, we analyze the structural and physical characteristics of promoter sequences from a novel structural viewpoint. We show that CGI- and non-CGI-related promoters have similar changes in their structural profiles. These changes are consistent with the observation that CGI-related promoters have higher CpG content. This suggests that the presence of CGI islands is an important feature in promoter prediction.

2 MATERIALS AND METHODS

2.1 Core promoter dataset

To build our promoter prediction model, we use a collection of promoter sequences as positive training dataset, the transcribed and intergenic sequences as negative training dataset. The promoter sequences are obtained from ENCODE data. Our comprehensive analysis shows that CGI- and non-CGI-related promoters exhibit almost similar changes in their structural profiles. These changes are consistent with the observation that CGI-related promoters have higher CpG content. This suggests that the presence of CGI islands is an important feature in promoter prediction.
The widely used similarity measures include the Euclidean distance metric like the efficient quantify the similarity between two objects in a meaningful way. Therefore, the classification of an unknown instance can be realized more likely they fall into the same class, according to an appropriate distance function. Thus, a smaller distance indicates a larger similarity. The process to predict promoters. The classification process consists of three steps: (i) computing pattern distance of each training instance to the unknown instance; (ii) searching the NNs in the instance space; and (iii) assigning the unknown instance the majority class of the NNs as follows:

\[
q = \sum_{j=1}^{d} \frac{C(S_j)}{PDict(S_j, X)}
\]

where \(C(S_j)\) is 

\[
\begin{align*}
1, & \quad \text{if } S_j \text{ is the positive class} \\
-1, & \quad \text{otherwise}
\end{align*}
\]

In formula (3), if the number of NNs \(k\) is larger than 1, this approach is usually referred to as \(k\)NN classifier. A higher value of \(k\) can avoid susceptibility to noise in the training data, yielding a more robust classifier. As for the weighted scheme, an instance further from the tested instance has less influence on the classification result compared with the nearer instances. Distance-weighted voting is a good solution as far as local sensitivity is concerned. The tested instance \(X\) is labeled as the positive class if \(q\) is positive, otherwise it is assigned to the negative class.

In PNNP there are two parameters, one is the sequence position \(p\) and another is the number of NNs \(k\). Both parameters are determined empirically. We test different values from the start position to the end position on a sequence and record the corresponding results. As Theorem 1 shows, the value of \(d\) has little impact on the similarity measure. The proof was given in [35,45].

\[
\begin{align*}
\text{Performance measures} & \quad \text{We use Sensitivity (Se), specificity (Sp) and F-measure to evaluate the performance of the proposed method. Se and Sp are two metrics widely used for promoter prediction evaluation. F-measure is a unified score of prediction performance. They are defined as follows:} \\
Se &= \frac{TP}{TP + FN} \quad \text{(5)} \\
Sp &= \frac{TP}{TP + FP} \quad \text{(6)} \\
F-measure &= \frac{2(Se \cdot Sp)}{Se + Sp} \quad \text{(7)}
\end{align*}
\]

where \(TP\), \(FP\) and \(FN\) represent the numbers of predicted true positives, false positives and false negatives, respectively. If a prediction falls within \([-L, L]\)bp region relative to the TSS location [Bajic et al., 2004], it is regarded as a true positive; otherwise the prediction is denoted as a false positive. If a known promoter is missed by the approach, it means a false negative. Se is the proportion of correct predictions of TSSs with regard to all experimental TSSs. Sp is the proportion of correct prediction of TSSs out of all positive predictions. Usually, the larger the Se is, the smaller the Sp is. So there is a tradeoff between Se and Sp. As a unified measure, the advantage of F-measure is that it can be used to compare different techniques with different sensitivity.
This page contains text discussing promoter prediction models and their performance. It highlights the importance of having clear peaks near the TSS in promoter sequences. The text explains how different promoter prediction models are used to identify promoters, focusing on the thermodynamic properties of DNA, particularly the Duplex Free Energy. The text elaborates on the differentiation between CGI- and non-CGI-related promoters, emphasizing the role of CpG islands in these differences. It also touches on the impact of TATA boxes in transcription initiation and the importance of understanding these structures for accurate prediction.
Fig. 3. Comparison among different structural profiles of six features: (a) Duplex disrupt energy, (b) DNA denaturation, (c) Bendability, (d) Z-DNA, (e) Protein DNA-twist, (f) B-DNA twist. The other results are given in the Supplementary Material.

Fig. 4. Se, Sp and F-measure of each structural features. The results of PNNP are based on 5-fold cross-validation.

while the profiles in Figure 3d-f are more ambiguous and noisy, which implies their different discriminative abilities.

We then go ahead to test how well these features can identify promoters. Each feature is utilized to construct a PNNP classifier. Figure 4 shows the results of a 5-fold cross-validation of all the 13 features. The experimental results reveal that the structural features indeed differentiate in discriminative capability for promoter prediction. Protein deformation, DNA-bending stiffness and protein–DNA twist have fairly high Se values, but their Sp values are low. On the contrary, duplex free energy, DNA denaturation and Z-DNA manifest a balance between Se and Sp. So F-measures of these features are much higher than that of other features. Generally, duplex free energy outperforms other features. Thus, we choose duplex free energy for our further analysis.

3.3 Performance on different types of promoters

Recent analysis of promoters revealed that a surprisingly large number of promoters are non-CGI-related. However, previous promoter prediction models focus mainly on CGI-related promoters and their performance on non-CGI-related promoters is very poor, which constitutes the main reason of existing models’ low accuracy at the whole genome level.

We collect four datasets of promoters of four species from the latest promoter annotations in DBTSS and PlantProm. The species are human, mouse, C.elegans and plant. First, all promoters without splitting are treated as a group. Then, we utilize the CpG island detection program (Yong et al., 2004) to classify these promoters into CGI- and non-CGI-related promoters. The program detects 71.93%, 64.46%, 40.32%, 53.78% of CGI-related promoters on the four promoter datasets above. As a result, there are three groups of promoters for each species, i.e. all promoters, CGI- and non-CGI-related promoters. Then, we test the performance of our PNNP method on the three promoter groups of each promoter dataset (corresponding to one species) separately.

Figure 5 shows the prediction results of PNNP on CGI-, non-CGI-related promoters and the whole promoters sets of human and mouse. The results of plant and C.elegans are given in the Supplementary Material. We can see that PNNP performs better on CGI-related promoters than on non-CGI-related promoters in terms of Sp, and vice versa in terms of Se. Se values on non-CGI-related promoters are >70% and considerably higher than that on CGI-related promoters. By using structural features, PNNP achieves fairly good Se and Sp for the whole set of promoters. In short, the experiments on four different species validate that
the unique patterns of structural profiles can efficiently capture the inherent characteristics of promoters, which makes our approach to be effective in predicting both types of promoters, especially the non-CGI-related promoters.

3.4 Comparison with existing promoter prediction methods over the whole human genome

Here, we compare PNNP with six existing promoter prediction models, including DragonGsf (Bajic and Seah, 2003), FirstEF (Davuturi et al., 2001), McPromoter (Ohler, 2002), Eponine (Down and Hubbard, 2002), CoreBoost (Zhao et al., 2007) and EP3 (Abeel et al., 2008). Based on several types of compositional features, such as CpG island and the frequencies of fixed-length motifs, FirstEF adopts a rule-based solution with two quadratic discriminant functions to predict TSS. DragonGsf is implemented as an artificial neural networks (ANN), utilizing CpG islands as a global landmark in prediction process. McPromoter uses a sequence of six Markov chain models for different subregions and elements, including TATA box, spacer and initiator regions. Eponine is also based on information of GC content and TATA box, implemented as a relevance vector machine with an optimal set of positioned weight matrices. The four methods above are mainly based on the compositional features of sequences, and they perform well in a comparative review (Bajic et al., 2004). CoreBoost is based on simple LogitBoosting with stumps. It mainly uses compositional features and mechanical features, and also needs some prior information such as Chi-chip data, expressed sequence tags to enhance the recognition model. EP3 predicts promoters based on structural features. It directly uses the deviation from the average structural values to locate the promoter regions.

The comparison is done on human genome, carried out in the following steps. First, we apply PNNP and other methods to predict promoters on human genome. For PNNP, the DNA sequence of each chromosome is divided into a series of 250bp segments by shifting a window over the sequence with a step size of 50bp. If a segment is classified as a promoter, a possible TSS candidate is marked. With similar criterion proposed in Bajic and Seah (2003), the possible TSS candidates within 1000bp are merged into a cluster. Then a new prediction is output by averaging all the candidates in the cluster. Second, prediction results are evaluated based on the reference TSS set—CAGE dataset. As mentioned above, we extract only tag clusters identified by two or more tags as reliable promoters and subsume these promoters into CGI- and non-CGI-related promoters. By using CpG island detection tool (Yong et al., 2004), we find that 68.9% promoters in CAGE database are CGI-related. We evaluate the methods on the three groups of promoters, i.e., CGI-, non-CGI-related and all promoters, separately. All the predictions are subject to similar evaluation criteria. If a prediction is within 2 bp of an annotated TSS, we call it a true positive hit. To achieve unbiased results, we conduct experiments at three levels of resolution, low (1000bp), medium (200bp) and high (50bp) resolution. All existing methods are tested with their default settings, which are provided as the optimal parameters by their developers.

For PNNP, we select the trained model with best performance on DBTSS to predict promoters on the whole human genome. The Se and Sp results are shown in Figure 6 and the Supplementary Material. Figure 6a and b are Se and Sp values of seven promoter prediction methods at medium resolution (200bp), when there is no overlap between the training and testing datasets. Corresponding results for a lower (1000bp) and a higher (50bp) resolutions are shown in Figure 4 of the Supplementary Material. We also present results for the case of about 10% overlap between the training and testing datasets in Figure 5 of the Supplementary Material.

For CGI-related promoter prediction, we can see that FirstEF, Eponine and McPromoter have high Sp, but middle Se. EP3 has relatively high Se and Sp. While PNNP has the highest Se, and its Sp is among the highest. What is more, PNNP has more balanced Se and Sp than the others. For all methods, as the resolution gets higher, the performance turns worse. At high resolution, PNNP substantially outperforms the others.

For non-CGI-related promoters, all methods perform worse than over CGI-related promoters. Specifically, at high resolution, McPromoter has virtually no prediction. Due to adopting structural features to predict promoters, EP3 and CoreBoost achieve relatively good performance. However, EP3 uses similar deviation from average to identify promoters, which disregards the difference between the structural values of CGI- and non-CGI-related promoters and thus leads to still unsatisfactory performance. The dependance on CGI information also causes performance degradation of CoreBoost over non-CGI-related promoters. On the contrary, PNNP performs best among the seven methods, and its performance on non-CGI-related promoters is as well as that on CGI-related promoters. The reason is that the structural profiles of
these two types of promoters exhibit similar patterns, which are different from that of non-promoters. Considering the prediction of all promoters, PNNP outperforms the others in all resolution cases. In summary, PNNP can effectively discriminate promoters and non-promoters. Compared with the existing prediction methods, PNNP obtains a significant improvement on Se and a moderate increase in Sp.

4 CONCLUSION

From the structural perspective, our analysis on promoters and non-promoters sequences provide interesting findings. Although the absolute structural values of CGI-related sequence are different from that of non-CGI-related ones, all promoter sequences exhibit coherent structural patterns, rising and falling nearly synchronously. The more important finding is that promoters and non-promoters have distinctive structural patterns. It conforms to the fact: promoters bear a specific structure that is essential to the assembly of the transcription machinery, regardless of whether they are CGI- or non-CGI-related. These inherent structures may be specific to the interaction between proteins and DNA, and essential for transcription initiation. Based on the findings above, we have further presented a new promoter prediction model—PNNP that can predict promoters effectively based on structural features. In contrast to existing promoter prediction methods that have unsatisfactory performance on non-CpG-related promoters, PNNP is not only well suitable for CGI-related promoters, but also for non-CGI-related promoters. So our method provides a uniform model to predict both types of promoters, and subsequently improves the prediction accuracy at the level of whole genome.

We have applied our approach to four species and compared it with six existing promoter prediction solutions on the whole human genome. The results show that PNNP is an effective promoter prediction model that outperforms the other compared methods. The boosted performance is attributed to two major factors, biologically relevant features and an efficient classification algorithm.

In summary, PNNP is well suitable for vertebrate promoters, and shows high effectiveness in eukaryotic promoters. Furthermore, our approach can be easily extended to other genomic site detection problems. As for the future work, we plan to construct a classifier that takes into account all features and aggregates the results of different features by ensemble learning.

ACKNOWLEDGEMENTS

We would like to thank Abeel, T. et al. for letting us share the conversion models.

Funding: Natural Science Foundation of China (60873040 and 60873070); Shanghai Leading Academic Discipline Project No. B114; the Program for New Century Excellent Talents in University of China (NCET-06-0376).

Conflict of Interest: none declared.

REFERENCES


