ProSOM: Core promoter identification in the human genome.

Thomas Abeel THOMAS.ABEEL@PSB.UGENT.BE
Yvan Saeys YVAN.SAEYS@PSB.UGENT.BE
Yves Van de Peer YVES.VANDEPEER@PSB.UGENT.BE
Department of Plant Systems Biology, VIB, Technologiepark 927, 9052 Gent, Belgium,
Department of Molecular Genetics, Ghent University, Technologiepark 927, 9052 Gent, Belgium

Abstract

More and more genomes are being sequenced, and to keep up with the pace of sequencing projects, automated annotation techniques are required. One of the most challenging problems in genome annotation is the identification of the core promoter. Better core promoter prediction can improve genome annotation and can be used to guide experimental work.

Comparing the average structural profile of transcribed, promoter and intergenic sequences demonstrates that the core promoter has unique features that cannot be found in other sequences. We show that unsupervised clustering by using self-organizing maps can clearly distinguish between the structural profiles of promoter sequences and other genomic sequences. An implementation of this promoter prediction program, called ProSOM, is available and has been compared with the state-of-the-art.

1. Introduction

Currently, the genomic sequence of over 50 eukaryotic organisms is available. So it is important to automate the identification of genes and regulatory sequences.

The core promoter is the region immediately upstream of the TSS, where the transcription initiation complex assembles.

Core promoters have distinct features that can be used to distinguish them from other sequences. One such property models the local base-stacking energy. High values denote regions that destack or melt easily. Two regions that seem to melt easily are located around -30 from the TSS and on the TSS, and are embedded in a large-scale region that is significantly more stable. We used this large-scale feature in earlier work to predict promoter regions in a wide range of species.

We present a novel promoter prediction technique, called ProSOM, that uses an unsupervised self-organizing map (SOM) to distinguish core promoter regions from the rest of the genome.

2. Material and methods

2.1. Data

We used the human genome assembly (hg17, May 2004). The cap analysis gene expression (CAGE) dataset was retrieved from the Fantom3 project. It contains 123,400 unique TSSs for human. The Ensembl gene annotation has been retrieved using the BioMart tool for Ensembl release 37. Sequences and annotation were retrieved from the ENCODE project. For the training of the SOM we retrieved promoter, transcribed and intergenic sequences from DBTSS and Ensembl.

2.2. Structural profiles

The nucleotide sequence is converted into a sequence of numbers (i.e., a numerical profile). This is done by replacing each dinucleotide with its energy value, which is obtained from experimentally validated conversion tables. We have used the conversion tables for base-stacking energy from Florquin et al. 2005.

2.3. Clustering and promoter prediction

The clustering technique we used is the self-organizing map (SOM), a special type of artificial neural network that can be used both for clustering and class prediction. A SOM consists of a rectangular grid of clusters, each of which has a weighted connection to every input node. In our case, the input nodes represent the different values of a structural profile associated to a potential promoter region. The SOM provides a

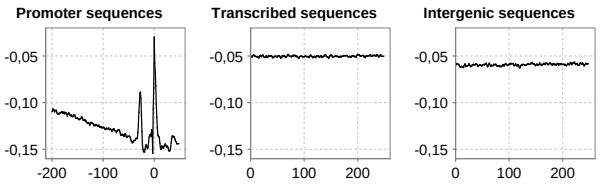

Figure 1. The structural profile of promoter (left), transcribed (center) and intergenic (right) human sequences. The profiles are the averages over all sequences in the respective training sets. We used the base-stacking energy as physical property. The left panel shows the region [-200,50] around the TSS, while for the other two panels there is no reference point and the location are numbered from 0 to 250.

mapping from a higher-dimensional feature space (the structural profile) to a lower-dimensional cluster space.

2.4. Validation

The validation of our technique was done on two sequence sets; first on the entire human genome assembly (hg17, May 2004) and secondly on the ENCODE regions. For both sets we retrieved a set of experimentally characterized TSSs and a gene annotation.

An aggregate measure for the performance of a classifier that is often used in the machine learning field is the F-measure. This is the harmonic mean of the recall (sensitivity) and the precision (specificity).

We have proposed a more objective way to assess the performance of a PPP based on the genome-wide screening for TSSs. This technique is based on the CAGE datasets that have been described earlier. The dataset contains locations where transcription starts. A TP is a known site that has a prediction within 50 bp of a true TSS, a FN is a TSS without a prediction and a FP is a prediction that has no associated TSS in the reference set within 50 bp.

3. Results

Figure 1 shows the average structural profile of base-stacking energy of the three datasets we used for training the SOM. The promoter sequences show a very striking profile with overall lower values than the other two graphs. It has two clear peaks at position -30 (TATA binding protein) and position 0 (TSS).

We used the trained SOM to predict promoter regions. To each cluster we attached a probability that a given sequence assigned to that cluster is a promoter. If the structural profile of a sequence maps to a cluster that has a probability equal to or above the threshold, we predict it as a promoter region.

Table 1. Evaluation of promoter prediction programs using the CAGE dataset with a maximum allowed distance of 50 bp.

program	recall	prec.	**F**
ProSOM	0.17	0.30	**0.22**
Eponine	0.14	0.35	**0.20**
EP3	0.11	0.27	**0.16**
ARTS	0.11	0.27	**0.15**
FirstEF	0.13	0.15	**0.14**

To validate our predictions we use the dataset of CAGE-tags from and a set of genes from Ensembl. To compare with the state-of-the-art, we used a maximum allowed distance from the TSS of 50 bp. Table 1 shows the performance of ProSOM versus a number of other PPPs.

We also analyzed the ENCODE regions of the human genome in more detail. The ENCODE project tries to annotate one percent of the human genome in great detail. ProSOM gets an F-measure of 0.28 on this validation set.

4. Discussion and conclusion

Self-organizing maps provide an intuitive way to cluster DNA sequences. They are unique among unsupervised clustering techniques in their ability to distinguish core promoters from other sequences. We packaged this technique as a full-fledged promoter prediction tool, called ProSOM, that performs as well as the best existing software packages.

Acknowledgments

T.A. is funded by a grant from the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen). Y.S. is funded by a post-doctoral grant from the Research Foundation Flanders (FWO-Vlaanderen).

References

Abeel, T., Saeys, Y., Bonnet, E., Rouzé, P., & Van de Peer, Y. (2008a). Generic eukaryotic core promoter prediction using structural features of DNA. *Genome Res*, *18*, 310–323.

Abeel, T., Saeys, Y., Rouzé, P., & Van de Peer, Y. (2008b). ProSOM: Core promoter prediction based on unsupervised clustering of DNA physical profiles. *Bioinformatics*, *(in press)*, –.