Eukaryotic core promoter prediction using structural features of DNA

Thomas Abeel, Yvan Saeys and Yves Van de Peer

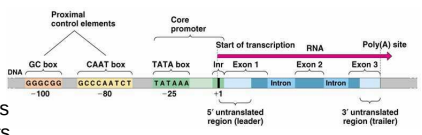
Contact: thomas.abeel@psb.ugent.be

Department of Plant Systems Biology, VIB, Ghent University, Technologiepark 927, 9052 Gent, Belgium

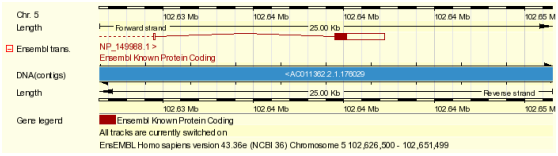
Despite many recent efforts, *in silico* identification of promoter regions is still in its infancy. Accurate identification and delineation of promoter regions is important for several reasons, such as improving genome annotation and devising experiments to study and understand transcriptional regulation. Here, we present a novel approach that requires no training for predicting promoters in whole genome sequences by using large-scale structural properties of DNA. We compared our approach to the state-of-the-art in promoter prediction.

Challenges

Promoter is small, only 0.05% of the genome
→ Many false positives



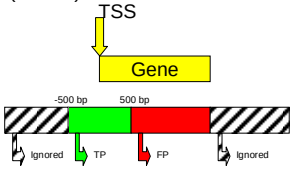
- Can be far from genes
- Non-coding transcripts
- In new genomes no experimental data



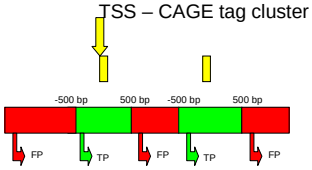
Data and validation

- DBTSS promoter sequences
- Human genome assembly
- CAGE TSS tags for human (~120,000 TSS)
- Ensembl gene annotation (~ 30,000 genes)

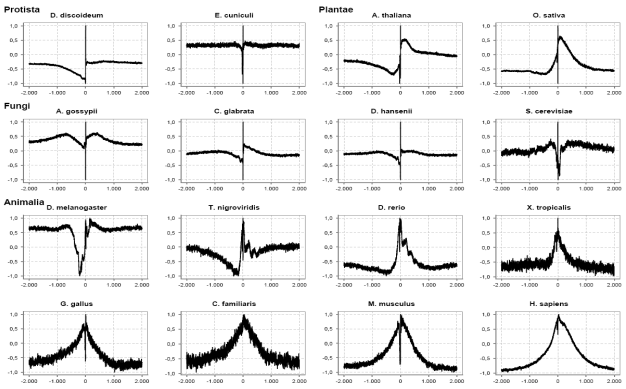
Gene annotation (classic)



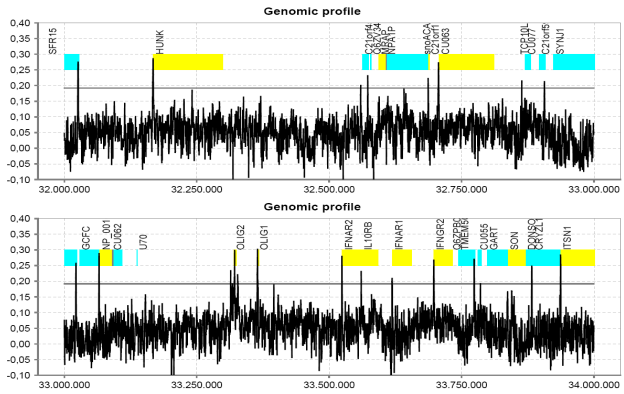
Cage tags (novel)



Structural features of the promoter



The average base stacking profiles of multiple promoter sequences for 16 organisms. The profile is the average over a large number of promoters with the TSS on position zero.



The base stacking profile of 2 Mbp of chromosome 21 taken with a window size of 400 bp. Positive strand genes are shown in yellow, negative strand genes are shown in blue.

Comparison with the state-of-the-art

Species	*F*
P. falciparum	0.17
O. pacifica	0.71
O. tauri	0.66
A. thaliana	0.37
O. sativa	0.53
P. trichocarpa	0.46
S. cerevisiae	0.42
S. pombe	0.31
C. elegans	0.26
D. melanogaster	0.19
T. nigroviridis	0.23
M. musculus	0.46
H. sapiens	0.44

→ **Right:** comparison of EP3 with state-state-of-the-art PPPs. The Ensembl column denotes the performance using the gene annotation and using the classic validation technique. The CAGE column shows the performance using the TSS data with the novel counting technique. The programs are ranked according to their F-measure on the CAGE data.

← **Left:** comparison of the performance of EP3 on different eukaryotic genomes. Performance between species differs, but in general the performance is in line with the performance on human. Only EP3 was tested on all genomes because the other programs require training to work on those genomes.

Program	Ensembl			CAGE		
	Recall	Prec.	F	Recall	Prec.	F
EP3	0.42	0.46	0.44	0.34	0.66	**0.45**
DragonGSF	0.45	0.63	0.53	0.31	0.75	**0.44**
PromoterInspector	0.38	0.7	0.49	0.29	0.81	**0.43**
FirstEF	0.58	0.34	0.43	0.41	0.42	**0.42**
Eponine	0.36	0.51	0.42	0.28	0.75	**0.41**
CpgProD	0.5	0.36	0.42	0.34	0.41	**0.37**
PromoterExplorer	0.55	0.24	0.33	0.39	0.3	**0.34**
McPromoter (0.0)	0.24	0.61	0.34	0.17	0.68	**0.28**
PromFD	0.55	0.14	0.22	0.44	0.16	**0.23**
DragonPF	0.65	0.11	0.19	0.51	0.11	**0.18**
ARTS	0.77	0.08	0.14	0.63	0.06	**0.12**
Promoter2.0	0.68	0.03	0.06	0.63	0.04	**0.08**
NNPP 2.2 (0.99)	0.03	0.02	0.02	0.03	0.03	**0.03**