Fast and accurate core promoter prediction in human.

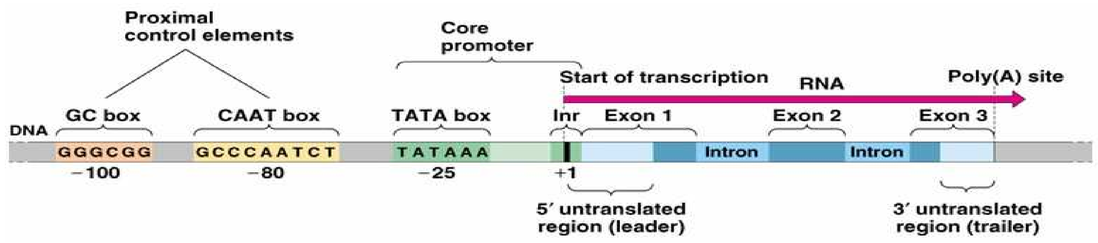
Thomas Abeel, Yvan Saeys and Yves Van de Peer

Bioinformatics and Evolutionary Genomics, Department of Plant Systems Biology, VIB, B-9052, Ghent, Belgium
Department of Molecular Genetics, Ghent University, B-9052 Ghent, Belgium

In silico identification of promoter regions is still in its infancy. However, the 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. Current methods to identify promoters have serious drawbacks because they are difficult to train, require large amounts of high quality training data, and often behave like black box models that output predictions that are difficult to interpret. Here, using the different chemical and physical properties of DNA in promoter and non-promoter regions we present a novel method for predicting promoters in whole genome sequences that is extremely fast, simple in design, and easy interpretable.

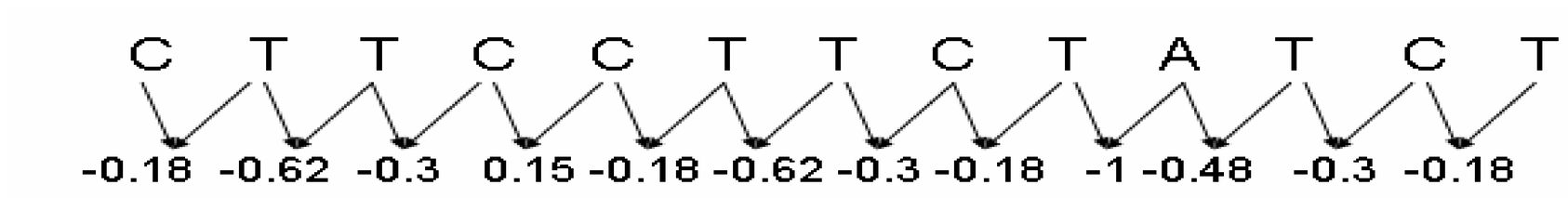
What is the core promoter?

- A promoter is a DNA sequence that enables a given gene to be transcribed from DNA into RNA. It is usually (but not always) located upstream of the gene to be transcribed.
- It tells when transcription shall occur, how much RNA should be produced, and where exactly on the genome the transcription should begin (Transcription Start Site, TSS).
- The core promoter is typically 50 bp upstream of the TSS, the proximal promoter is often said to be 200 bp upstream and the distal promoter up to a few thousand bp.



Converting the DNA sequence to the physico-chemical structure

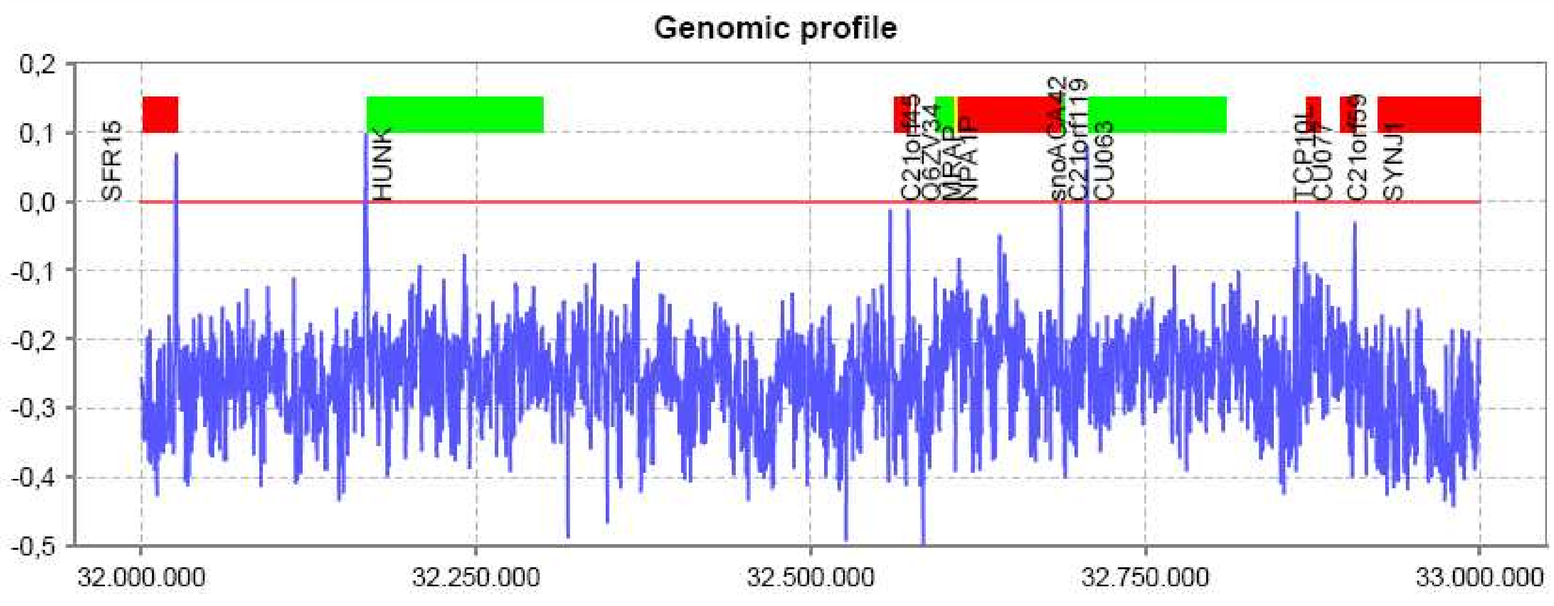
- DNA has physical and chemical properties which depend on the distribution of nucleotides A, T, G, C.
- Some structural properties are: Stacking energy, Propeller twist, DNA denaturation value (used here), ...
- Experimentally calculated parameters allow the computation of a structural profile for any given DNA sequence
  - Computed on di- or trinucleotide scales
  - Using experimental conversion tables (see table on the right)
  - Replace every di- trinucleotide with corresponding number. (see figure blow)
  - Convert the DNA sequence into a numeric sequence



aa	66.51	**ga**	80.03
ac	108.8	**gc**	135.83
ag	85.12	**gg**	99.31
at	72.29	**gt**	108.8
ca	64.92	**ta**	50.11
cc	99.31	**tc**	80.03
cg	88.84	**tg**	64.92
ct	85.12	**tt**	66.51

The conversion table for DNA denaturation value. Each dinucleotide corresponds with a numeric value.

Plotting the structure of the human genome with its gene annotation



The structural profile (DNA denaturation value) of a region of chromosome 21 of the human genome. The structural profile is shown in blue. On the X-axis is the position on chromosome 21, on the Y-axis is the normalized DNA denaturation value. The red and green boxes are respectively positive and reverse strand genes. The prediction threshold is depicted as a horizontal red line on 0 (zero).

- Calculate the structural profile of the whole human genome.
  - Hg17 genome assembly
- Plot this profile in a graph: see figure on the left, the blue line
  - Here we only show 1 million base pairs from chromosome 21, but the technique was applied to the whole genome.
- Put the gene annotation on the graph
  - Annotation was retrieved from Ensembl.
  - Positive strand genes in green
  - Negative strand genes in red
  - Names are added for reference
- Core promoters can be predicted where the profile peaks. This is determined with a cut-off (horizontal red line).

Comparing this technique to the current state-of-the art

- Make predictions for the whole human genome using the technique described above.
- Compare these predictions to a database of known transcription start sites.
- Use other state-of-the-art programs to make predictions on the same dataset and compare their predictions to the same database of known transcription start sites.
- The results of these analysis is shown in the table to the right. It is clear that our very simple approach is performing better than the current software
- There is still quite some room for improvements as the precisions and recall are still not 100%.

Name (Parameters)	TP	FP	FN	Precision	Recall	F-measure
SP4 (-0.125)	**173,254**	**78,001**	**239,057**	**0.69**	**0.42**	**0.52**
SP4 (-0.15)	**197,870**	**147,176**	**214,441**	**0.57**	**0.48**	**0.52**
FirstEF	152,866	37,253	259,445	0.80	0.37	0.51
McPromoter(-0.05)	299,214	477,046	113,097	0.39	0.73	0.50
CpGProD	139,192	37,313	273,119	0.79	0.34	0.47
DragonPF	226,671	351,840	185,640	0.39	0.55	0.46
Promoter2.0 (HLP)	262,073	498,345	150,238	0.34	0.64	0.45
DragonGSF	102,364	4,814	309,947	0.96	0.25	0.39
Promoter2.0 (MP)	396,581	1,300,624	15,730	0.23	0.96	0.38
ARTS	290,328	863,377	121,983	0.25	0.70	0.37
Eponine	84,192	4,073	328,119	0.95	0.20	0.34
McPromoter(0.0)	65,303	5,362	347,008	0.92	0.16	0.27

1. Bajic, V.B., et al. (2006) Performance assessment of promoter predictions on ENCODE regions in the EGASP experiment, *Genome Biol*, **7 Suppl 1**, S3.1--S313.
2. Bajic, V.B., Tan, S.L., Suzuki, Y. and Sugano, S. (2004) Promoter prediction analysis on the whole human genome, *Nature Biotechnology*, **22**, 1467--1473.
3. Florquin, K., Saeys, Y., Degroeve, S., Rouze, P. and Van de Peer, Y. (2005) Large-scale structural analysis of the core promoter in mammalian and plant genomes, *Nucl. Acids Res.*, **33**, 4255-4264.
4. Sonnenburg, S.o., Zien, A. and Rätsch, G. (2006) ARTS: accurate recognition of transcription starts in human, *Bioinformatics*, **22**, e472--e480.