Core promoter prediction for genome annotation

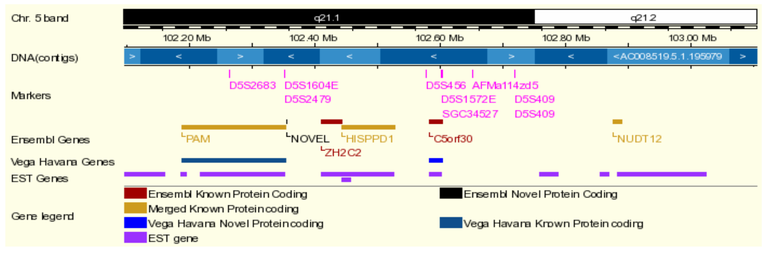
Thomas Abeel, Yvan Saeys and Yves Van de Peer

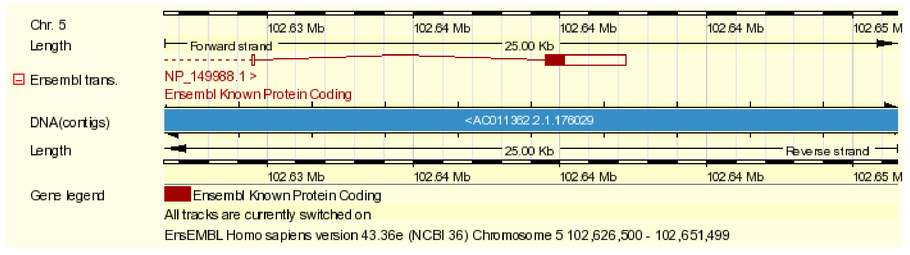
Department of Plant Systems Biology, VIB, Ghent University, Technologiepark 927, 9052 Gent, Belgium

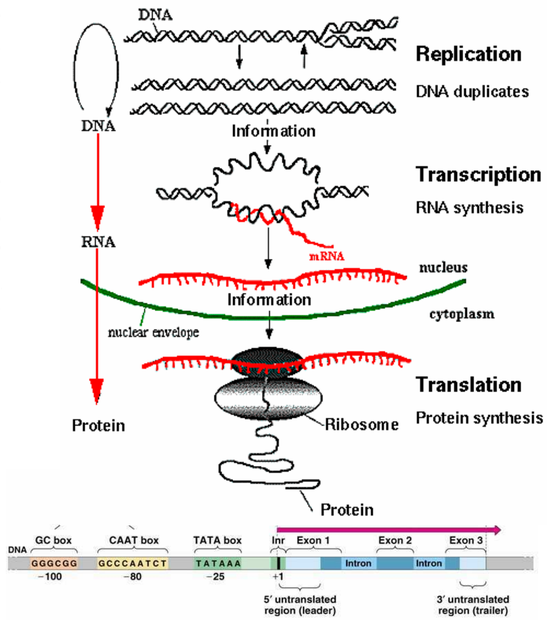
The identification and delineation of promoter regions is important for improving genome annotation and devising experiments to study and understand transcriptional regulation. Current methods to identify promoters have serious drawbacks (species specific, training, behave like black box, etc.). Here, by 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.

Background

Genomes are being sequenced at an ever increasing pace. Up to now a few dozens of eukaryotes are sequenced and at least the same amount are in the process of being sequenced. This ever rising amount of genomic data is only meaningful when one can pinpoint the regions that are most important to the organisms that host that piece of DNA. More specific it is important in biotechnology to identify the regions that code for genes and can thus produce proteins. The process of identifying genes and other features in the genome is called annotation. Besides annotating genes, it is even more interesting to identify the regions that are responsible for the regulation of genes. These regions are called promoters and are key elements in the regulation of the transcription of genes. They define when genes are expressed, how many times the gene is transcribed, etc. Transcription is the first step in the chain from DNA to protein; first the DNA is transcribed into pre-mRNA that is spliced to mRNA which is then translated into proteins. The core promoter is responsible for the initiation of transcription and is generally located in front of the gene it controls. It is easy to identify the part of the gene that codes for a protein, but it is much more difficult to identify the core promoter region. Therefore, computational techniques to identify core promoter regions are still in their infancy.

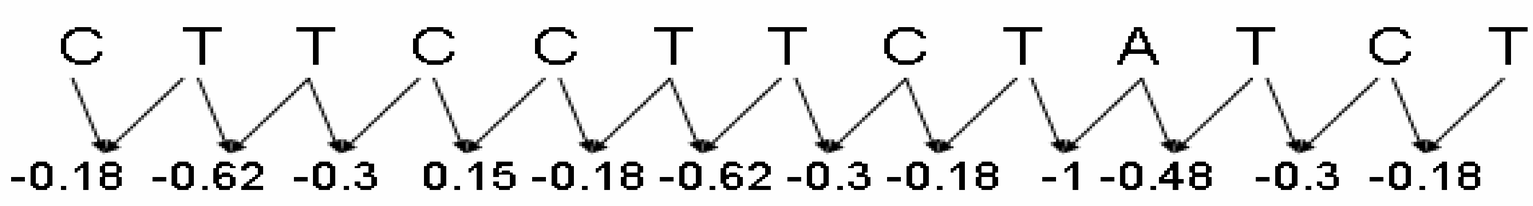






Converting DNA to physico-chemical properties

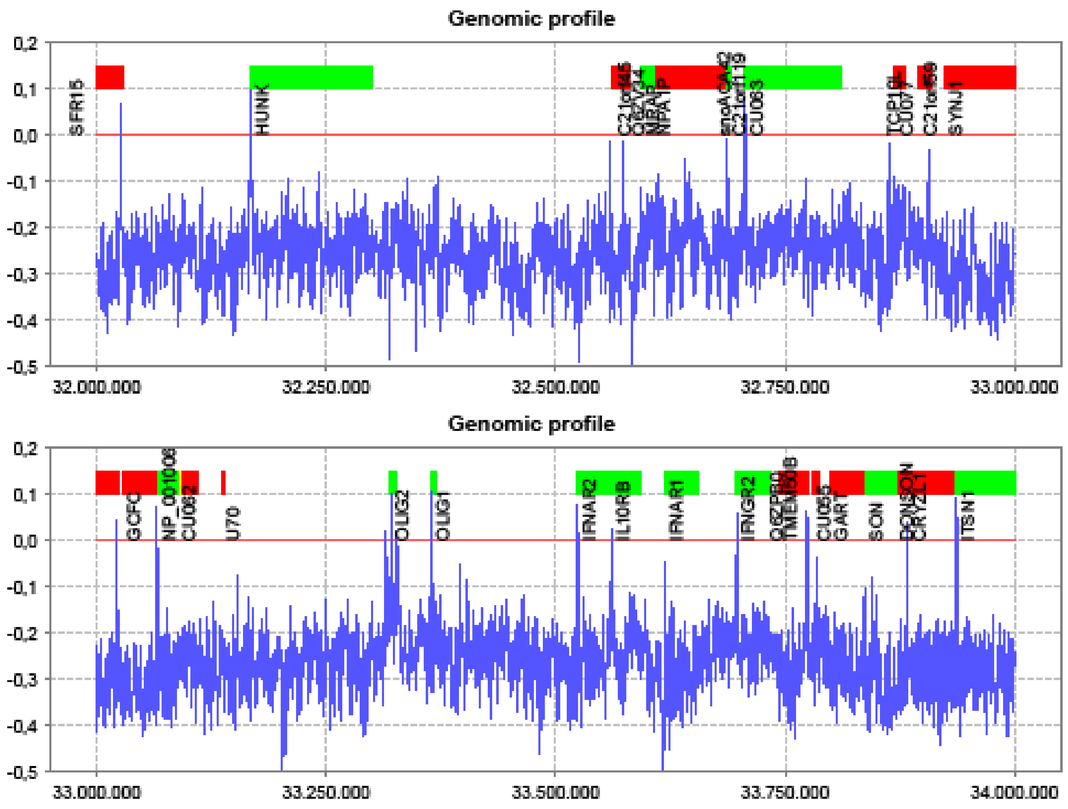
- DNA has physical and chemical properties which depend on the distribution of nucleotides A, T, G, C.
- Some of the structural properties are: Stacking energy, Propeller twist, DNA denaturation value (the one we used here), ...
- Experimentally calculated parameters allow the computation of a structural 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 the corresponding value. (see figure below)
  - 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**	64.93
ct	85.12	**tt**	66.51

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

Results and conclusion



By using different conversions, it is possible to model different physical and chemical properties of the DNA. One of these properties models the energy needed to melt the DNA. When the DNA melts, the two strands become separate and the nucleotides become available for interaction with proteins, including the proteins required for the initiation of transcription. The melting of the DNA in the core promoter is thus a prerequisite for transcription initiation. Using the melting model of a genome it is thus possible to identify regions that are suitable for transcription initiation.

We implemented this approach in a prediction tool to identify core promoter regions in a whole genome based on the melting model. This approach outperforms all existing state-of-the-art promoter prediction programs when tested on the complete human genome and referenced with a genome wide dataset of core promoters. Besides working well on the human genome the program is also one of the only programs that is applicable to more than one species because it is more robust for small genomic variations and it requires no training so that it can be applied to newly sequenced genomes for which little experimental data is available.

While other programs are usually restricted to a single species due to the different composition and structure of different genomes, our approach is clearly able to capture a more universal property of promoters defined by the chemo-physical properties of DNA. As a result, it is capable of performing very well on different species without the need for any retraining, which is a unique achievement among promoter prediction software, allowing true '*ab initio*' promoter prediction.

References

- Bajic, V. B., Brent, M. R., Brown, R. H., Frankish, A., Harrow, J., Ohler, U., Solovyev, V. V., & Tan, S. L. (2006) *Genome Biol* **7 Suppl 1,** S3.1--S313.
- Florquin, K., Saeys, Y., Degroeve, S., Rouze, P., & Van de Peer, Y. (2005) *Nucl. Acids Res.* **33,** 4255-4264.
- Sonnenburg, S. o., Zien, A., & Rätsch, G. (2006) *Bioinformatics* **22,** e472--e480.