The structure of the promoter unraveled

Thomas Abeel, Yvan Saeys, Pierre Rouzé and Yves Van de Peer

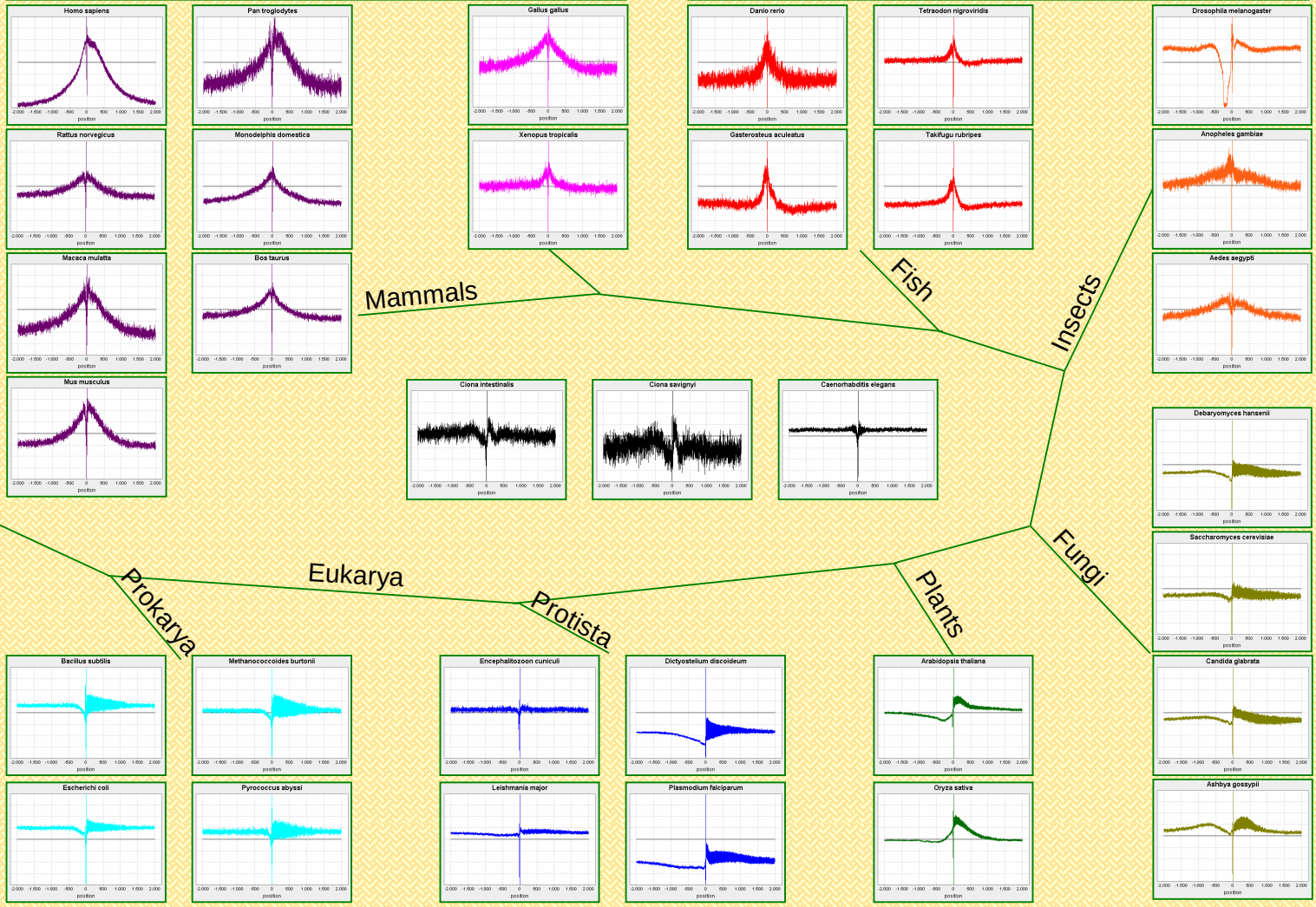
Bioinformatics and Evolutionary Genomics, Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University
Technologiepark 927, B-9052 Gent, BELGIUM
E-mail : thomas.abeel@psb.ugent.be

Introduction

Promoters play a crucial role in the regulation of the transcription of genes, yet their structure is largely unknown. While in prokaryotes the core promoter is thoroughly defined, the eukaryotic promoter is still under debate. Early studies indicated that eukaryotic promoters, like prokaryotic promoters, contained a TATA motif, but soon after, it became apparent there was more to it than only motifs.
Structural features are another way of representing DNA sequences (Baldi, 1998). This approach looks to secondary information contained within the sequence. Structural features are **physical and chemical properties** of the DNA like stacking energy, propeller twist, DNA denaturation value, etc.
Previous studies indicated that the promoters of **both prokaryotes and eukaryotes have distinct structural features** compared to the rest of the genome. However, at the time there was not sufficient data to compare the structure of multiple species. Thanks to the many genome project of the past few years, there is enough information to compare multiple genomes. Here we have compared a region of 2000 bp around the transcription start site (TSS). For this analysis we have used **4 prokaryotic and 29 eukaryotic genomes**.

Methods and material

The **datasets were downloaded** from various sources including Ensembl, JGI, Sanger institute, Genoscope, Flybase, EMBL-EBI and Genolevures. For each species we have extracted the region of 2000 base pairs around the transcription start site. This resulted in several thousand promoter sequences for each species. Unfortunately not all annotations are of very high quality which can explain some of the more noisy graphs, well studied species have in general much clearer graphs (human, mouse, Arabidopsis, rice, …).
Structural profiles are calculated as follows. First, the nucleotide sequence is converted into a sequence of numbers (i.e., a numerical profile). This is done by replacing each dinucleotide with its corresponding structural value. The structural values are obtained from experimentally validated conversion tables (Blake, 1998). Finally we take the average over all numerical profiles of a species, normalize such that all values are in the interval [-1,1] and plot it in a graph.
The graphs below are made using the **DNA denaturation value**, this is the energy needed to melt the DNA (cal/mol).



Results

We have discovered several features of the promoter, some of which are valid for all species, while others are valid for a large subset of species. In the picture above we show a phylogenetic tree with in the leaf nodes a group of species that are more or less *closely* related.
One **feature we see in all species** is the large peak and drop at position zero, the actual transcription start site. Besides this main theme we have **two categories**. The first one starts with prokaryotes and ends roughly after fungi. This category consists of species that have a lower value upstream of the TSS, a drop right before the TSS and a higher value downstream. These promoters have a **narrow region** with distinct values. The second category starts from insects and goes to mammals. Here we see a **very broad peak** (several hundreds up to several thousand bp).
This feature can be used to create a **promoter prediction program** that works well for a whole range of species with minimal parameter fitting. Preliminary data shows recall values up to 40% and precision of 99% in human. (research in progress) .

$$recall = \frac{TP}{TP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

References

- Baldi,P. et al., Computational Applications of DNA structural scales, *Proc Int Conf Intell Syst Mol Biol.* **6**, 35-42. (1998)
- Blake, R.D. & Delcourt, S.G. Thermal stability of DNA. *Nucl. Acids Res.* **26**, 3323--3332 (1998).
- Carninci, P. et al. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38**, 626--635 (2006).
- Florquin, K., Saeys, Y., Degroeve, S., Rouze, P. & Van de Peer, Y. Large-scale structural analysis of the core promoter in mammalian and plant genomes. *Nucl. Acids Res.* **33**, 4255-4264 (2005).
- Fukue, Y., Sumida, N., Tanase, J.-i. & Ohyama, T. A highly distinctive mechanical property found in the majority of human promoters and its transcriptional relevance. *Nucl. Acids Res.* **33**, 3821-3827 (2005).
- Fukue, Y., Sumida, N., Nishikawa, J.-i. & Ohyama, T. Core promoter elements of eukaryotic genes have a highly distinctive mechanical property. *Nucl. Acids Res.* **32**, 5834-5840 (2004).
- Pedersen, A. Gorm, Baldi, P., Chauvin, Y. and Brunak S. DNA Structure in Human RNA Polymerase II Promoters, *Journal of Molecular Biology* **281**, 663-673, (1998).