Organisation of regulatory elements in *Arabidopsis thaliana* core promoters: toward a functional annotation

Virginie Bernard[1], Véronique Brunaud[1] and Alain Lecharny[1,2]

[1]Unité de Recherche en Génomique Végétale (URGV), UMR INRA 1165 -CNRS 8114 -UEVE,
2 Rue Gaston Crémieux, 91057 Evry Cedex, France
[2]Université Paris-Sud, Institut de Biotechnologie des Plantes (IBP), UMR CNRS 8618 -UPS,
Bâtiment 630, 91405 Orsay Cedex, France

{bernard, brunaud, lecharny} @evry.inra.fr

Abstract: *Taking advantages of the preferential position of some regulatory elements relative to the TSS, we carried out a systematic search of preferentially located motifs in the core and proximal promoters of Arabidopsis thaliana. This work led to the knowledge of the plant promoter architecture giving prominence to two areas of the core promoters containing regulatory element highly conserved, linked together, and characterized by specific Gene Ontology categories.*

Keywords: Promoter, Regulatory element, TATA-box, Initiating di-nucleotide

The knowledge of promoter architecture *i.e.* the topological organisation of regulatory sequences is crucial in order to understand the regulation of gene expression. Several regulatory elements are preferentially located relative to the Transcription Start Site (TSS) [1] and *ab-initio* approaches are using this preferential location characteristic to identify biologically relevant regulatory elements [2, 3]. These elements are specifically organized in promoters and their cooperation allows the control of the gene expression [4]. We propose to analyse the regulatory elements organisation in *Arabidopsis thaliana* and to identify co-present elements expected to be involved in a same regulatory pathways. We assume that a specific biological function of genes sharing such co-present regulatory elements allow a first indication of the regulatory elements function annotation.

First, taking biological information of the full-length cDNA and other transcript sequences, we predicted the TSS position of 14927 *Arabidopsis thaliana* genes. We constituted our set of promoters by extracting 1000 bases upstream the TSS and all the 5'UTR of the genes. Second, we build the distribution of each motif from 2 to 8 bases long. We searched for not evenly distributed motifs in the TSS overlapping region, *i.e.* motifs showing a distribution exhibiting a peak. Third, we selected genes sharing co-present regulatory elements putatively involved in the same regulatory pathways. The Gene Ontology annotations [5, 6] of gene sets characterized specific biological functions.

Among the 87376 motifs analysed, we identified 5105 Preferentially Located Motifs (PLMs) with specific topological constraints. Almost 90% of them showed a distribution with a wide peak. We focused our analyses on the 10% of remaining PLMs whose distribution gave a sharp peak and that were located in a strict area into the core promoter.

The first area, roughly 35 bases upstream the TSS, is T and A rich and corresponds to the TATA-box [7] expected region. This box involved in the recruitment of the transcription initiation complex is characterized by the minimal consensus TATAWA, with W for A or T. The canonical TATA-box is present in 17% of *A. thaliana* genes. Moreover, in the same area, 34 PLMs with sequences and topological constraints similar to the TATA-box are detected. They are observed in 32% of *A. thaliana* promoters without the canonical TATA-box. These PLMs could be variants of the TATA-box with different levels of sequence degeneration and with putative function specification.

In the second region, overlapping the TSS, we identified the YR dinucleotide, i.e. the CA, TG, TA and CG di-nucleotides, all preferentially located one base upstream of the TSS. They are present in quite half of the promoters and are *A. thaliana* initiating di-nucleotide as proposed by Yamamoto *et al.* [3]. Interestingly, the distribution of all these di-nucleotides is extended downstream of the TSS. A rich region that may well represent alternative initiating di-nucleotides followed each di-nucleotide peak. This specific promoter organisation may be involved in the regulatory process.

We used this characterization of the *A. thaliana* core promoter architecture to examine the links between the regulatory elements in this plant. We developed a gene classification based on the presence in their promoters of PLMs of both the TATA-box and the initiating di-nucleotide regions. The presence of the canonical TATA-box with the CA initiating dinucleotide is positively biased. The genes containing these regulatory elements are more involved in specific than in basic biological processes. This result is a first step in the establishment of relationships between the promoter architecture and gene function.

Finally, our extensive description of the TSS region provides the necessary material to further analysis of the question of a functional link between promoter architecture and gene expression by data-mining of transcriptome data ([8-10]).

References

1. FitzGerald PC, Shlyakhtenko A, Mir AA, Vinson C: **Clustering of DNA sequences in human promoters**. *Genome Res* 2004, **14**(8):1562-1574.
2. Casimiro AC, Vinga S, Freitas AT, Oliveira AL: **An analysis of the positional distribution of DNA motifs in promoter regions and its biological relevance**. *BMC Bioinformatics* 2008, **9**(1):89.
3. Yamamoto YY, Ichida H, Matsui M, Obokata J, Sakurai T, Satou M, Seki M, Shinozaki K, Abe T: **Identification of plant promoter constituents by analysis of local distribution of short sequences**. *BMC Genomics* 2007, **8**:67.
4. Moshonov S, Elfakess R, Golan-Mashiach M, Sinvani H, Dikstein R: **Links between core promoter and basic gene features influence gene expression**. *BMC Genomics* 2008, **9**(1):92.
5. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.
6. Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M *et al*: **The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community**. *Nucleic Acids Res* 2003, **31**(1):224-228.
7. Patikoglou GA, Kim JL, Sun L, Yang SH, Kodadek T, Burley SK: **TATA element recognition by the TATA box-binding protein has been conserved throughout evolution**. *Genes Dev* 1999, **13**(24):3217-3230.
8. Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, Edgar R: **NCBI GEO: mining millions of expression profiles--database and tools**. *Nucleic Acids Res* 2005, **33**(Database issue):D562-566.
9. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG *et al*: **ArrayExpress--a public repository for microarray gene expression data at the EBI**. *Nucleic Acids Res* 2003, **31**(1):68-71.
10. Gagnot S, Tamby JP, Martin-Magniette ML, Bitton F, Taconnat L, Balzergue S, Aubourg S, Renou JP, Lecharny A, Brunaud V: **CATdb: a public access to Arabidopsis transcriptome data from the URGV-CATMA platform**. *Nucleic Acids Res* 2008, **36**(Database issue):D986-990.