Arabidopsis thaliana core-promoter architecture

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
{bernard, brunaud, lecharny} @evry.inra.fr
[2]Université Paris-Sud, Institut de Biotechnologie des Plantes (IBP), UMR CNRS 8618 -UPS,
Bâtiment 630, 91405 Orsay Cedex, France

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, Transcription start site, Regulatory element, TATA-box

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]. 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'UTRs of the genes. Second, we analysed the distribution of each motif from 2 to 8 bases long to identify potential regulatory elements in the TSS overlapping region. The method to detect local over-representation of motifs described previously [4] has been improved to get a better sensibility. Using a simple linear regression, the distribution model is learned allowing defining for each motif a score reflecting its relevance. By adapting the sliding window size to each distribution profile, we obtained a significant improvement in the definition of the motif position.

Among the 87376 motifs analysed, we identified 5110 Preferentially Located Motifs (PLM) with specific topological constraints. Almost 90% of them showed a wide distribution. We focused our analyses on the 10% of remaining PLMs whose distribution gave a sharp peak and that were thus located in a strict area into the core promoter.

The first area, 35 bases upstream the TSS, is T and A rich and corresponds to the TATA-box [5] 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 other PLMs with sequences and topological constraints similar to the TATA-box are detected. These 34 PLMs could be putative functional variants of the TATA-box and are present in 23% of the promoters. An in-depth analysis of the promoters containing neither a TATA-box nor a variant of this box established for the first time that T and C rich PLMs are located at the TATA-box expected place. These T and C rich PLMs are observed in 5% of *A. thaliana* promoters.

In the second region, overlapping the TSS, we identified the CA, TG and TA di-nucleotides, all preferentially located at the -1 and at the TSS positions. They are present in 50% of the promoters.

CA was the most frequent and its distribution is extended downstream of the TSS. This CA rich region may well represent alternative CA that would be involved in the regulatory process.

We used this characterization of the *A. thaliana* core promoter architecture to re-examine the nature of the potential initiating element in this plant. We demonstrated that only CA, TA and TG are initiating di-nucleotides constituting the core of the potential Initiator element (Inr). Then we developed a promoter classification based on the PLMs of both the TATA-box and the initiating di-nucleotide regions. We used this classification to evaluate the accurate of the TSS prediction. Finally, we provided evidences for links between specific GO categories and promoter architecture.

Finally, our extensive description of the TSS region provides the necessary material to further analysed the question of the functional link between promoter architecture and gene expression by data-mining of transcriptome data ([6-8]).

References

[1] FitzGerald PC, Shlyakhtenko A, Mir AA, Vinson C, Clustering of DNA sequences in human promoters. *Genome Res*, 14(8):1562-1574, 2004

[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*, 9(1):89, 2008.

[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*, 8:67, 2007.

[4] Bernard V, Brunaud V, Serizet C, Martin-Magniette ML, Caboche M, Aubourg S, Lecharny A, Sélection de motifs candidats pour la régulation des gènes chez *Arabidopsis thaliana* sur des critères topologiques. In: *Journée Ouvertes de la Bioinformatique et des Mathématiques: july 2006; Bordeaux*; 17-28, 2006.

[5] 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*, 13(24):3217-3230, 1999.

[6] 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*, 33(Database issue):D562-566, 2005.

[7] 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*, 31(1):68-71, 2003.

[8] 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, 36(Database issue):D986-990, 2008.