"Young Researchers and Life Science" Meeting
Institut Curie in Paris, France
from the 23rd to the 25th of April 2008

Virginie BERNARD,
Véronique BRUNAUD, Alain LECHARNY

Unité de Recherche en Génomique Végétale (URGV),
UMR INRA – CNRS - UEVE 2, rue Gaston Crémieux,
CP5708, 91057 Evry cedex, France

Arabidopsis thaliana promoter architecture

The knowledge of promoter architecture is crucial to understand the regulation of gene expression. The core and the proximal promoters are regions playing a central role due to the presence of many Transcriptional Factor Binding Sites (TFBS) close to the Transcription Start Site (TSS). Several regulatory elements are observed at a preferential distance from the TSS. 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 a set of promoters by extracting 1000 bases upstream the TSS and all the 5'UTRs of the genes. Second, we carried out the search of 2 to 8 bases long Preferentially Located Motifs (PLMs) in respect to the TSS, i.e. distributions of motifs with a significant peak using statistical approach.

Among the motifs analysed, we identified 5110 PLMs with specific topological constraints. Roughly 90% of the PLM distributions showed a wide peak and reflected the presence of motifs on a large area. Upstream of the TSS, PLMs are present in the proximal promoter of small sets of genes and they are explained by either known TFBSs or potential novel TFBSs. Downstream of the TSS, PLMs are mainly microsatellites present in almost all promoters. We assume these sequences are involved in the DNA structure. The 10% of remaining PLMs whose distributions showed a sharp peak are located in two strict areas of the core promoter. One base upstream the TSS, our method has identified the CA, TG and TA di-nucleotides in 50% of promoters. Roughly 32 bases upstream of the TSS, several PLMs are due to the presence of a TATA-box in 17% of *A. thaliana* genes and of putative functional variants of this TFBS that are present in at least 23% of the promoters. An in-depth analysis of the promoters containing neither a TATA-box nor a variant of the box established that T and C rich PLMs observed at the expected place of the TATA-box characterized roughly 5% of *A. thaliana* promoters. This result highlights the efficiency of our approach to search for novel TFBS in sub-sets of promoters.

The presence of the TATA-box and of the CA di-nucleotide at the –1 position is positively correlated. A preferential distance between the TFBSs has confirmed their link. Moreover, we provided evidences of specific functional annotations of genes with both TFBSs. Thus a link with gene expression is suggested. We plan to search for correlation between promoter architecture and gene expression taking advantage of transcriptome data available on *Arabidopsis thaliana.*