論 文 の 内 容 の 要 旨

論文題目　　Learning the Promoter Architecture of Tissue-Expressed Genes
（組織特異的発現遺伝子におけるプロモーター構造の学習）

氏　　名　　ヨスバニ　ロペス　アルバレス

Transcription is one of the most important biological processes in the cell. As the first level in the cascade of gene expression, the comprehensive understanding of the transcriptional mechanism is still a great challenge for life science researchers. For a gene to be expressed, the genomic (promoter) region surrounding its transcription start site (TSS) has to be bound by specific regulatory proteins known as transcription factors. A great body of studies have hypothesized that those genes (or a part of them) expressed in the same tissue, cell type or physiological condition might be regulated by similar mechanisms and accordingly share common regulatory structures. Therefore, the finding of structural binding patterns in promoter regions could contribute to better explain the regulatory mechanism of such genes and search for co-expressed genes with unknown biological functions. This thesis presents three studies conducted under the above-mentioned assumption.

Although several studies have focused on the analysis of promoter regions of expressed genes in distinct metazoan tissues, little research has been carried out in plants. The plant *Arabidopsis thaliana* offers a valuable opportunity for modeling promoters because of its small genome and short intergenic regions. By taking advantage of these characteristics and the availability of microarray data from *A. thaliana* structures, one method intended to uncover motif-combination patterns in promoters of genes expressed in structures such as flower, root, seed and shoot, and in the whole plant was developed. Initially, *de novo* motifs were predicted in five different sets (each comprising the promoters of genes expressed in the previous plant structures) and eight of them appeared to be novel. The average of the binding distances of identified motifs in both strands from the translation start site was subsequently computed and input into a support vector machine. The correctly classified promoter regions per plant structure were then taken for creating specific patterns of sets of motifs able to describe the promoter architecture of co-expressed genes. These five patterns were used to scan the entire *A. thaliana* promoter set and detect genes with unknown biological functions. Significant percentages of genes expressed in petal differentiation, root hair, synergid cells, trichome and housekeeping genes were found.

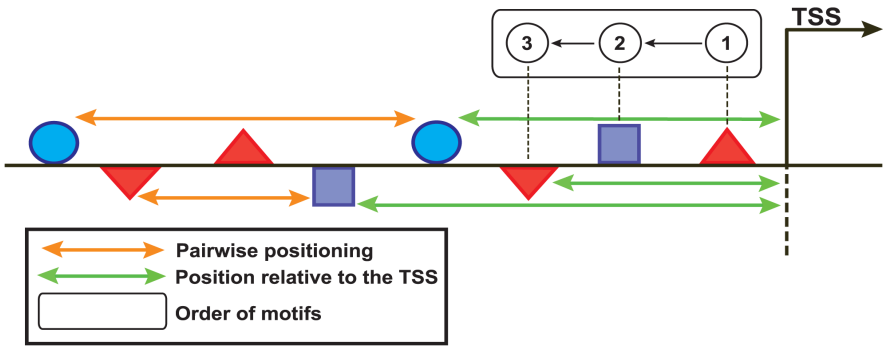

Figure 1. Scanning of the regulatory regions. Geometrical forms above and under the black line represent the binding sites on plus and minus strands. Orange and green arrows along with the rectangle indicate the computed features.

Given the huge amount of genomic data from well-studied organisms as *Drosophila melanogaster*, the second computational method was designed and validated in *cis*-regulatory modules of antenna-expressed genes in *D. melanogaster*. This approach simultaneously combined diverse structural features such as relative positioning to the TSS, pairwise positioning, binding order and strand orientation of regulatory motifs (Figure 1). Predictions of *de novo* motifs in the regulatory regions of the genes uncovered six potentially interesting antenna-related motifs from which three turned out to be novel. The regions were then scanned in search for the aforementioned features and a correlation-based filter was introduced to remove irrelevant characteristics. Afterwards a genetic algorithm was designed to reach those highly informative features common to the regions. As a result, eight structural features (Table 1) were obtained and used to score the entire set of *D. melanogaster* regions for unknown antenna-expressed genes with a similar regulatory architecture. Validations were conducted with two independent RNA-sequencing datasets from eye-antenna disc-derived and antenna disc-derived cell lines in the third instar larval stage from the Model Organism Encyclopedia of DNA Elements database (modENCODE). Expressed genes were compared to those with highly scoring regions predicted by the method, resulting in roughly 76.7% of overlapped genes. Conservation signals of the structural features were also found in regions of orthologs in eleven *D. melanogaster* sibling species. This approach showed comparable results to a former study while uncovered relevant features related to binding order and strand orientation of regulatory motifs.

Table 1. Eight features discovered in *cis*-regulatory modules of antenna-expressed genes.

	Description
Feature 1	DME-3 positioned ~0-100 bp from DME-3 on + strand
Feature 2	DME-5 positioned ~100-200 bp from TSS on - strand
Feature 3	DME-4 positioned ~200-300 bp from TSS on +/- strand
Feature 4	DME-5 positioned ~0-100 bp from TSS on + strand
Feature 5	DME-5 positioned ~600-700 bp from TSS on +/- strand
Feature 6	DME-5 positioned ~500-600 bp from TSS on - strand
Feature 7	DME-2 positioned ~1100-1200 bp from TSS on + strand
Feature 8	DME-6 positioned ~300-400 bp from TSS on +/- strand

The previous computational method has been extended to model the *cis*-regulatory modules of genes expressed in twenty-two different developmental stages of *D. melanogaster*. RNA-sequencing data profiling the whole developmental cycle were downloaded from the modENCODE to build and validate the models. Two additional structural features, namely relative distance of motif pairs to the TSS and presence of motifs anywhere in the regulatory region were included. As a result, 13 (59%) out of 22 models showed statistical significance (p-value $< .01$). Table 2 depicts the seven models with highest performance (F-score ≥ 0.7).

Table 2. Seven models with the highest performances.

Model	**F-score**	***p*-value**
Embryo 0-2h	0.729	0.0
Embryo 12-14h	0.937	0.0
Embryo 14-16h	0.701	2.8e-05
Embryo 22-24h	0.857	0.0
L3 stage larvae	0.861	0.0
Adult male eclosion + 30 days	0.975	0.0
Adult female eclosion + 5 days	0.736	3.64e-03

These studies evidence the reliability of measures as positioning and orientation of regulatory motifs at specific distances from the translation start site for differentiating the promoters of genes expressed in *A. thaliana* structures. The integration of diverse structural features including binding order and strand orientation of motifs into a single approach has proved to better describe the promoter regions of tissue-expressed genes. The combination of correlation-based filter and genetic algorithm has also contributed to reach highly informative features hidden in the promoter architecture. Despite the developed approach could be generalized for modeling the promoters of genes expressed in other biological conditions, its effectiveness is still comparable to that of former studies conducted under the same premise.