Deciphering cis–acting regulatory elements in plant and drosophila promoter sequences

Magali Lescoat1, Gert Thijs2, Stéphane Rombauts3, Patrice Déhais3, David Martin1, Denis Thieffry1, Bernard Jacq1, Yves Moreau3, Pierre Rouzé4 and Jacques van Helden5

1 Laboratoire de Génétique et Physiologie du Développement, Parc Scientifique de LUMINY, CNRS Case 907, F–13288 Marseille cedex 9, France.
3 Laboratorium voor Genetica, Vlaams Interuniversitair Instituut voor Biotechnologie (VIB), Universiteit Gent, K.L. Ledeganckstraat 35, B–9000 Gent, Belgium.
4 Laboratoire Associé de l’Institut National de la Recherche Agronomique, France. Universiteit Gent, B–9000 Gent, Belgium.
5 Unité de Conformation des Macromolécules Biologiques, Université Libre de Bruxelles, CP 160/16, 50 av. F. D. Roosevelt, B–1050 Bruxelles, Belgium
*e–mail: lescot@lgpd.univ–mrs.fr

Keywords: pattern discovery, regulon, cis–acting elements, regulatory sequence analysis, program evaluation, Arabidopsis thaliana, Drosophila melanogaster

Pattern discovery algorithms aim at detecting motifs shared by a set of functionally related sequences. Different approaches have been followed to tackle the problem of pattern discovery and they can be divided in two groups: string–based methods (detection of over–represented words or of spaced dyads) on the one hand, and methods based on a matrix representation of motifs on the other hand.

Several pattern discovery programs have been developed for detecting cis–acting regulatory elements from upstream regions of co–regulated genes. These programs were generally developed, evaluated and optimized on the basis of microbial sequences (principally Escherichia coli and Saccharomyces cerevisiae). The specificity of transcriptional regulation in the model organism (motif size, degeneracy, position–specificity, strand–dependency of the motifs...) plays an important role in the algorithmic choices, and the rate of success of each program may depend on the organism considered. The extension of existing approaches to higher organisms is thus not trivial, because regulatory elements are dispersed far away from the transcription start, and can be found upstream, downstream, and within introns.

We used different programs to detect cis–acting regulatory elements in several families of co–regulated genes (regulons). A regulon is defined as a set of genes regulated by a same transcription factor. We collected information on experimentally proven regulons from the database PlantCARE [4].

Eleven regulons were built based on promoter sequences from Arabidopsis thaliana and other plant species, and ten from Drosophila melanogaster promoters. Nine of the plant regulons were deduced from PlantCARE. Two regulons were built based on literature studies. The ten Drosophila datasets were extracted from literature. For each gene, we extracted the upstream sequences of 1 to 1.6 kb from the transcription start (or, in the case of lack of data about the transcription start, the translation start).

In our datasets, the consensus sequence of the motifs were 6 bp long and more or less conserved depending on the regulon. For this reason, in the word–counting approaches of pattern discovery, we have looked for hexanucleotides and used Pattern–assembly [8] to get a better motif description (longer and degenerated sites). The program Oligo–analysis [7] was used with different options to compute the expected frequencies: predefined frequency tables based on the whole set of intergenic sequences [7], Markov chains [8], or non–overlapping segmentation (unpublished). The statistical significance was calculated for all oligonucleotides, and a threshold established according to the Bonferroni rule. Dyad–analysis [8] was used to detect over–represented spaced dyads (pairs of trinucleotides separated by a spacer of 0 to 20 unspecified nucleotides). We also used CoreSearch, another word counting method [9], and two methods of matrix–based pattern discovery in unaligned sequences: Motif Sampler [6] based on Gibbs sampler [5] and Consensus [1, 2].

Preliminary results suggest that the programs Oligo–analysis, Dyad–analysis and Motif sampler work efficiently on conserved motifs. A difficult task is, however, to detect degenerated motifs. In this case, the matrix–based method, Motif sampler, is adapted. The string–based methods such as Oligo–analysis and Dyad–analysis also allow to tackle this problem, because each different motif was assembled using Pattern–assembly [8]. In comparison with the approaches computing the word expected frequencies using Oligo–analysis, the
calibration on non-coding sequences was efficient, indeed the motifs were found with a higher significance index. The Motif sampler was efficient in most of the regulons when using a motif length slightly wider than the motif itself. The analysis of Consensus results is in progress.

In this survey, motif search results show that conserved motifs can be found in the upstream sequences, despite the fact that the regions used for this analysis are relatively short (1.6 kb) for an organism such as Drosophila. The next step of our analysis will consist in evaluating how the sequence retrieval options (size of sequences in the dataset, addition of intronic sequences) affect the pattern discovery accuracy.

In summary, experimentally documented regulons provide useful data for calibration and comparison of the motif finding methods. After this calibration stage, these methods can be applied to clusters of co-regulated genes for which regulatory elements are unknown. We further work on the definition of reliable statistical criteria to compare results between string-based and matrix-based approaches.

Availability

References