Deciphering *cis*-acting regulatory elements in plant and drosophila promoter sequences

Magali Lescot[1*], Stephane Rombauts[2], Patrice Déhais[2], Gert Thijs[3], David Martin[1], Denis Thieffry[1], Bernard Jacq[1], Yves Moreau[3], Pierre Rouzé[4] and Jacques Van Helden[5]

[1] Laboratoire de Génétique et Physiologie du Développement, Parc Scientifique de LUMINY, CNRS Case 907, F-13288 Marseille cedex 9, France.

[2] Laboratorium voor Genetica, Vlaams Interuniversitair Instituut voor Biotechnologie (VIB), Universiteit Gent, K.L. Ledeganckstraat 35, B-9000 Gent, Belgium.

[3] ESAT-SISTA/COSIC, KULeuven, Kasteelpark Arenberg 10, B-3001 Leuven-Heverlee, 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

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-based 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, especially because In these organisms, 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 (Lescot *et al.,* 2002).

Ten regulons were built based on promoter sequences from *Arabidopsis thaliana* and other plant species, and four from *Drosophila melanogaster* promoters. Eight of the plant regulons were deduced from PlantCARE. Two regulons were built based on literature studies. The four *Drosophila* datasets were extracted from literature. For each gene, we extracted the upstream sequences of 1kb 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 restricted our analysis to hexanucleotides. The program *oligo-analysis* (van Helden, 1998) was used with different options to compute the expected frequencies: predefined frequency tables based on the whole set of intergenic sequences (van Helden, 1998), Markov chains (van Helden, 2000), or non-overlapping segmentation (*unpublished*). The statistical significance was calculated for all the possible oligonucleotides, and a threshold established according to the Bonferoni rule. *Dyad-analysis* (van Helden *et al.*, 2000) 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, developed by Wolfertstetter (1996), and two methods of matrix-based pattern discovery in unaligned sequences: *Gibbs motif sampler* (Neuwald *et al*., 1995; Thijs *et al.,* 2001) and *Consensus* (Hertz *et al.*, 1990; Hertz & Stormo, 1999).

Preliminary results shown that most of the methods evaluated worked efficiently on conserved motifs. A difficult task is however to detect degenerated motifs. The string-based methods such as *oligo-analysis* and *dyad-analysis* allow to tackle this problem, because each different motifs are assembled using *patter-assembly* (van Helden, 2000). The calibration on non-coding sequences using *oligo-analys*is was efficient if we compare to the other approaches computing the word expected frequencies, because the motifs were found with a higher significance index. *Gibbs motif sampler* and *Consensus* perform particularly well on the G/C rich motifs.

In this survey, the motif search results shown that most of the motifs are retrieved in the upstream sequences, despite the fact that the regulatory region upstream of the coding sequences used for this analysis are relatively short (1kb) for an organism such as *Drosophila*. The next step of our analysis will consist in evaluating how the sequence retrieval options (size in the dataset, addition of intronic sequences) affect the pattern discovery accuracy.

In summary, experimentally documented regulons provide ideal data for calibration and comparison of the motif finding methods. After this calibration stage, these methods can then be applied to clusters of co-regulated genes for which regulatory elements are unknown. Further work is in progress in order to define reliable statistical criteria to compare results between string-based and matrix-based approaches.

Availability

The Regulatory Sequence Analysis Tools are available on the web at http://www.ulb.ac.be/bioinformatics/rsa-tools/

References

Hertz, G. Z. & Stormo, G. D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15**(7-8), 563-77.

Hertz, G. Z., Hartzell, G. W. d. & Stormo, G. D. (1990). Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput Appl Biosci* **6**(2), 81-92.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. & Wootton, J. C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science 262(5131), 208-14.

Lescot, M., Déhais, P., Thijs, G., Marchal, K., Moreau, Y., Van de Peer, Y., Rouzé P. and Rombauts, S. (2002). PlantCARE, a database of plant *cis*-acting regulatory elements and a portal to tools for *in silico* analysis promoter sequences. *Nucleic Acids Res,* database issue, 30(1):325-327*.*

Neuwald, A. F., Liu, J. S. & Lawrence, C. E. (1995). Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci* 4(8), 1618-32.

Thijs, G., Lescot, M., Rombauts, S., Marchal, K., De Moor, B., Moreau, Y. and Rouzé, P. (2001). A higher order background model improves the detection by Gibbs sampling of potential promoter regulatory elements in DNA sequences. *Bioinformatics,* 17(12):1113-1122*.*

van Helden, J., Andre, B. & Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol* 281(5), 827-42.

van Helden, J., Rios, A. F. & Collado-Vides, J. (2000). Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res* 28(8), 1808-18.

Wolfertstetter, F., Frech, K., Herrmann, G. & Werner, T. (1996). Identification of functional elements in unaligned nucleic acid sequences by a novel tuple search algorithm. *Comput Appl Biosci* 12(1), 71-80.