A High Performance Computing Approach to the Discovery of Conserved Motifs

Dieter De Witte[1], Michiel Van Bel[2,3], Piet Demeester[1], Bart Dhoedt[1],

Klaas Vandepoele[2,3] and Jan Fostier[1]

[1] Department of Information Technology (INTEC), Gaston Crommenlaan 8, bus 201, Ghent University – IBBT, Ghent, Belgium.

[2] Department of Plant Systems Biology, Flemish Institute for Biotechnology (VIB), Ghent, Belgium.

[3] Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 927, Ghent, Belgium.

Poster Abstract

The inference of regulatory networks relies on the accurate identification of functional motifs in non-coding sequences. We use a comparative approach to detect the transcription factor binding sites in 4 monocot species. In plants complex genome rearrangements take place, which takes away the opportunity of aligning their genomes. Our method detects motifs that are highly conserved within the promoter sequences of orthologous genes. Aligning the promoters is a possibility, but it has been shown that multiple alignments do not necessarily align known regulatory elements correctly. We therefore turn to an alignment-free method based on generalized suffix trees.

In order to quantify the degree of conservation, we use the Branch Length Score (BLS), which represents the normalized evolutionary distance over which a motif is conserved in the phylogenetic tree of the gene family.

We use an exhaustive method to test all motifs-BLS pairs (mbps) up to a length k. Degenerate alphabets further increase the number of candidate motifs. Therefore a parallel version of the algorithm has been developed which significantly reduces runtimes.

We investigated the algorithm's ability to recover known rice motifs from Transfac. We used an exact alphabet with an Any (N) character to represent the motifs. The mbps with a confidence above 90% were retained and from this pool the most similar motifs to the Transfac rice motifs were selected. For almost every motif a highly similar motif was found, which proves the method is promising.