Alignment-free genome-wide comparative motif discovery in 4 Monocot species

Dieter De Witte1, Michiel Van Bel2,3, Piet Demeester1, Bart Dhoedt1,

Klaas Vandepoele2,3 and Jan Fostier1

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.

Phylogenetic footprinting approaches to motif discovery usually rely on whole genome alignments (WGAs) for the detection of conserved regulatory sites. Due to complex genome rearrangements WGAs are not possible in plants. We therefore analyze the promoter sequences of a large set of orthologous gene families. We do not rely on the alignment of the promoters since it has been shown that regulatory sites are often not aligned correctly.

In this work we study 4 species of the Monocotyledon family. Our dataset contains 17724 gene families each consisting of 4 orthologous promoter sequences and on average one paralogous sequence. We use the Branch Length Score (BLS) to assess the degree of conservation of a motif inside a gene family.

We developed an exhaustive alignment-free algorithm based on generalized suffix trees to discover the conserved motifs in a gene family. We use a 5-character alphabet, including the Any character (N) to represent the motifs. After combining the results from the different gene families we are able to calculate the confidence of each motif-BLS pair.

A parallel version of the algorithm has been developed and implemented using the Message Passing Interface (MPI) which significantly reduces runtimes and makes it possible to keep the vast amount of candidate motifs in memory.

We investigated the algorithm's ability to recover known rice motifs from Transfac.. The results are compared with the results obtained by aligning the promoter sequences with Dialign-TX. Since the alignment quality might depend on the length of the sequences, we compare the results for both the 500b promoters and the 2kb upstream promoter regions.