TFBS prediction in sugarcane using binding sites prediction from PlantTFDB server

Mauro de Medeiros, Danielle Izilda Rodrigues da Silva, Alan Durham, Glaucia Souza Mendes

Ph.D student in Bioinformatics IME/USP, Ph.D student in International Cellular and Molecular Plant BiologyESALQ/USP, Associate Professor IME/USP and Associate Professor IQ/USP.

Brazil is the world's largest sugarcane producer with productivity close to 70 or 80 tons per hectare [t/ha]. However, such productivity is low if compared with its maximum yield (around 380-472 t/ha). Gene regulation is considered the primary mechanism for activating all biological potential. Thus, the search for new varieties that express higher productivity can begin by understanding gene regulation processes. However, due to the complexity and diversity of the gene regulation mechanisms, research efforts have limited the study mainly on the promoter region. This region is composed of two sets. The first set is the core promoter - DNA sequence of about 100 nucleotides (nt) and the subsequences TATA box, Inr and TSS (transcription start site). The second set consists of DNA sequences that are upstream of the core promoter subsequences such as the CAAT box and TFBS (Transcription Factor Binding Site). The discovery of these regions can be achieved through *in vivo* experiments such as Chromatin Immunoprecipitation followed by Sequencing (ChIP-Seq), systematic evolution of ligands by exponential enrichment (SELEX) and DNAse I hypersensitive mapping. However, these tests can be too expensive and are not always suitable for non-model organisms, like sugarcane. To address this challenge some studies have been using the in silico approach. The aim of this work was to search TFBSs - described in the literature - in the promoter region of differentially expressed genes of sugarcane samples subjected to water stress. To solve this task we used the FIMO tools and the PlantTFDB database binding sites. The FIMO application performs the individual search for motifs through Position-Specific Scoring Matrix - PSSM. The PlantTFDB is a non-redundant collection of 674 TFBSs distributed in 156 plant species. To task discovery, we used 6 groups from 100 DNA sequences of the promoter region of differentially expressed genes. According to our analysis, in all groups 25% to 30% of the evaluated sequences were classified as TATA-box promoter region. However, when the assessment was related to the number of TFBS from Transcription Factors responsive to drought, Promoter groups from stressed leaf samples were at least twice as large as its counterpart. Moreover, this scenario is not the same when compared with dry and irrigated root. This difference between the two organs may indicate that each tissue has its own mechanism of regulation.