Sugarcane Genome Annotation: Prediction Analysis and Plant Protein Mapping

ALMEIDA, A.S.S.A[1,4] ;CAMPOS, R.A.[1]; KASHIWABARA, A. Y.4; NISHIYAMA, M.Y.JR.[1]; FARIA, P.I.G.4; HOTTA, C.T.[1]; SETTA, N.[2]; CRUZ, G.M.Q.[2]; GAIARSA, J.W.[2]; LEMBKE, C.G.[1]; SATO, P.M.[1]; ANDRADE, R.F. [1]; NUNES, S.L.[1]; KITAJIMA, J.P.[3]; VAN SLUYS, M.A.[2] ; DURHAM, A. M.[2]; SOUZA, G.M.[1]

[1]Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, Brasil; [2]Departamento de Bioinformática, Instituto de Matemática e Estatística, Universidade de São Paulo, Brasil; [3]Departamento. de Botânica, Instituto de Biociências, Universidade de São Paulo, Brasil; [4]Departamento. de Ciência da Computação, Instituto de Matemática e Estatística, Universidade de São Paulo, Brasil. [5]Centro de Pesquisa Experimental, Instituto Israelita de Ensino e Pesquisa Albert Einstein, Brasil;

Sugarcane (Saccharum spp.) is a grass with a high commercial value due its ability to accumulate high amounts of sucrose. It is a first-generation biofuel crop used for ethanol and sugar production. The sugarcane genome has an estimated size of 10 Gb and contains a large number of repetitive sequences. It is polyploid and aneuploid. The aim of this work is to annotate the sugarcane genome from BACs assembled into one-contig identifying genes and its UTR regions. The BACs were constructed from the hybrid cultivar R570. After sequencing, the BAC reads were assembled in one-contig with the Phrap/Phred/Consed software. The sugarcane genome annotation pipeline was based on mapping of ESTs, Fl-cDNA and plant proteins into the BACs. For prediction analysis, we used MYOP and AUGUSTUS Predictor Pipelines for the identification of genes, splice sites and UTR regions. We mapped the BACs on the sorghum genome in order to have a reference and subsequently identified homologies. We have observed that both are mostly collinear in the genic regions. We are currently using 18 BACs to establish the parameters for the annotation of genes on additional BACs. To date, we have identified 320 genes in 18 BACs. Each BAC presents an average of 10 genes. We are not considering genes coding for transposases or other proteins related with insertion of repetitive sequences. The studies include the development of a gene promoter database where putative regulatory regions are associated to sugarcane transcripts.