A BIOINFORMATIC APPROACH FOR THE DETECTION OF PUTATIVE MYB ORTHOLOGUES IN LARGE PLANT EST DATASETS

CATALANO D., CALAMITA G., FINETTI M., DE VIRGILIO M., BLANCO E., PIGNONE D., SONNANTE GA.

Istituto di Genetica Vegetale CNR – Bari, Via Amendola 165/A, Cap 70122

Myb domain, bioinformatic, hidden Markov models

Myb proteins make up one of the largest families of transcription factors in plants. These proteins are characterised by having a conserved DNA-binding domain composed from one up to four repeat motifs of about 50 amino acids length, called R0R1R2R3.

MYB proteins play an important role in the regulation of various metabolisms including morphogenesis, meristem formation, cell cycle and secondary metabolism. Here we developed a bioinformatic pipeline to classify putative MYB transcript genes using a wide set of plant EST sequences. As a case study, the Asteraceae ESTs were considered.

First, we downloaded the complete dataset of the EST sequences stored in the Genbank database, then, Emboss packages were used to trim the polyA tail and clean the vector sequence possibly present in the dataset. The cleaned ESTs were clustered/assembled with the Cap3 program and we used the obtained contigs and singletons for the detection of the putative open reading frames. The obtained ORFs were analyzed with Hmmer program, a bioinformatic tool that processes the sequences with hidden Markov models. Hmmer is an implementation of profile hidden Markov models for biological sequence analysis (Krog et. al.)

The sequences containing at least two myb domains in the same open reading frame were considered the most relevant ones. In order to find the putative cluster of the orthologues, we used the Clustalw program to align the MYB transcription factors belonging to Arabidopsis and Oryza and the ORF obtained as described above. Afterwards, we analyzed the result of alignment by means the Dendroscope program, in order to find some putative orthologous clusters.