LARGE-SCALE GENE ONTOLOGY ANALYSIS OF PLANT GENOME AND TRANSCRIPTOME SEQUENCES RETRIEVED BY AFLP TECHNOLOGY

BOTTON A., GALLA G., RAMINA A., BARCACCIA G.

Department of Environmental Agronomy and Crop Science, Faculty of Agriculture, University of Padova, Viale dell’Università 16, Campus of Agripolis, 35020 Legnaro (Italy)

GO, annotation, genomic AFLP, cDNA-AFLP, ESTs

After 10-year-use of AFLP technology for DNA fingerprinting and mRNA profiling, large repertories of genome- and transcriptome-derived sequences are available in public databases for model, crop and tree species. AFLP marker systems have been and are being extensively exploited for genome scanning and gene mapping, as well as cDNA-AFLP for transcriptome profiling and differentially expressed gene cloning. The evaluation, annotation and classification of genomic markers and expressed transcripts would be of great utility for both functional genomics and systems biology research in plants.

We retrieved from both NCBI databases and private repertories a total of 7,806 cDNA-AFLP sequences related to roots, leaves, stems, flowers, fruits and seeds, along with the 285 publicly available genomic AFLP sequences. All these entries belong to 22 different species distributed among seven botanic families: Solanaceae, Fabaceae, Poaceae, Salicaceae, Rosaceae, Brassicaceae and Vitaceae. Redundant sequences were preliminarily clustered to select singlets and assemble contigs. BlastX analysis against non-redundant protein databases, GO terms mapping and annotation analysis were then performed using Blast2GO, a research tool designed with the main purpose of enabling GO based data mining on sequence sets for which no GO annotation is yet available. Descriptive statistics on the type, size and nature of chromosome regions and gene sequences mainly investigated using the AFLP technology were calculated. The gene sequences associated to mRNA transcripts and proteins were classified according to the GO vocabularies. In addition to cellular component, biological process and molecular function, other hierarchically structured GO terms were adopted to query sequences and to assign genes and gene products at different levels, depending on the depth of knowledge. A classification of all cDNA-AFLP records for the main GO vocabularies was also performed by splitting the sequence dataset in monocots and dicots and by comparing the two subgroups with all annotated ESTs of Arabidopsis and rice.

The examination and annotation of EST clones enabled basic inferences to be made on the potentials and drawbacks of AFLP technology for mRNA profiling and differential display gene cloning. Although the different number of sequences retrieved in gene banks for plant organisms and organs might have biased some of the descriptive statistics, the whole set of data emerged from gene ontologies is consistent with the existence of AFLP features exploitable across plant transcriptomes (e.g., ESTs associated to kinase activity can be assayed with very similar rates (11%) in each of the families analyzed in this study) and supports the reliability of expression patterns detection using very small amounts of messengers (i.e., DD-cDNA-AFLP applied to tissues where it is hard to isolate stage-specific mRNAs, such as flowers, fruits and seeds). On the whole, experimental steps and statistical parameters adopted for the in silico AFLP technology-derived sequence analysis proved to be critical for obtaining robust ontology data. Annotation results for the
whole sequence dataset and also for botanic families, single species and plant organs are presented
and the main features of genes and gene products detectable in plants by genomic AFLP
fingerprinting and cDNA-AFLP profiling discussed. To the best of our knowledge this is the first
large-scale survey of amplified fragment length polymorphism-derived sequences belonging to
plant angiosperms.