Characterizing the tomato transcriptome

ID - 220

D'agostino Nunzio[1], Chiusano Maria Luisa[1]

[1]Department of Soil, Plant, Environmental and Animal Production Sciences, University of Naples Federico II, Portici (NA), Italy

Motivation

The tomato genome sequencing initiative aims to determine the functional gene space of Solanum lycopersicum by sequencing the gene-rich regions of the genome. A noteworthy and parallel contribution to the study of the tomato gene content involves the screening of the EST collections worldwide available which have dramatically increased after the kick off of the International Tomato Genome Sequencing Project.

In December 2006, there were over 250000 tomato sequences deposited in the dbEST.

This is not surprising because ESTs represent the most abundant resource for gene discovery, genome annotation, and comparative genomics. Herein we provide a large-scale sampling of the tomato transcriptome based on EST analysis.

Methods

250552 ESTs generated from 102 cDNA libraries have been retrieved from dbEST.

Over-represented ESTs are removed from the collection in order to clip the original dataset. This dataset was used to feed the ParPEST pipeline (D'Agostino et al., 2005) which removes vector contaminations, clusters and assembles ESTs into contigs and performs BLASTX searches against the UniProt database in order to annotate both ESTs and contigs. Transcripts are associated to Gene Ontology (GO) terms and to the Enzyme Commission (EC) numbers via the UniProt matching protein. All the GO terms related to each transcript were convertd to plant GO slim terms using the map2slim.pl script, distributed as part of the go-perl package. The plant GO Slim file was downloaded from the GO website. Transcripts that are associated to EC numbers are mapped onto KEGG metabolic pathways

Results

The 250552 ESTs have been generated from 102 distinct cDNA libraries representing 18 tissue types covering both the sexual reproductive and the vegetative parts of the plant. Our first purpose was to remove from the original collection over-represented ESTs in order to clip the original dataset where we distinguished container (22873), contained (59789) and stand-alone (167890) sequences. The next step of the analysis, involving the removal of vector contaminations, reduced the dataset from 190 763 to 190 593. The subsequent clustering/assembling procedure generated 44759 clusters.

Each cluster (i.e. gene index) should correspond to a unique gene. 28005 clusters are made up of a single EST and they are classified as singletons. The remaining 16754 clusters are assembled into 17629 contigs. 658 clusters are assembled into multiple contigs, ranging in size from 2 to 25 members, because of alternative transcription and putative paralogy. Two main types of contigs resulted from the analysis: contigs generated from ESTs which come from different libraries (14598) and library/tissue-specific contigs (3031). Of the 45624 transcripts, 33174 have at least one BLAST hit, while the remaining 12450 have no hit found. Considering this latter set, 8 955 are yet uninformative hits since they are hypothetical, unknown or expressed proteins. We focused on the full length protein encoding sequences, that can be specifically useful for gene identification, for gene model building and for the definition of a tomato 'virtual proteome'. 767 transcripts presented full length BLAST hits. Among these, 426 are from members of the Solanaceae family, 2 of the Rubiaceae family, 183 of the Brassicacea family, the remaining 156 are from different plant families.

10638 transcripts are mapped to GO hierarchy. Each transcript could present multiple GO term assignment. Thus, 440119 assignments were made to the class molecular function, 292286 were assigned to the biological process and 172517 to the cellular component class. In the first ontology class the 42% of the GO terms are members of the binding categories, the 18% are members of the catalytic activity category while the remaining categories are less represented. Considering the biological process class, the vast majority of the GO assignments correspond to the more generic metabolism category, the 13% of the GO terms are members of the transport category.

The 4% of the GO annotations describeS response to biotic or abiotic stimula, while the 3% concerns response to stress. This is not surprising since a division of the ESTs was derived from tissues/libraries responding to plant pathogen challenge or salt stressed. Finally, for the cellular component class the vast majority of the assignments were to the membrane (21%), mitochondrion (19%), plastid (17%), cytoplasm (12%) and nucleus category (11%).

2544 transcripts are associated to 519 distinct enzymes. 1986 transcripts encoded 395 enzymes had

mappings to 113 biochemical pathways, but only 224 enzymes act in a single phatway and are classified as pathway-specific.
In order to better refine the functional annotation of the transcript data a BLASTN analysis was performed to compare S. lycopersicum transcripts to the Rfam database.
170 out of 45624 transcripts have a match record in Rfam. In particular, 95 transcripts have matches with both UniProt and Rfam databases.
Availability: http://biosrv.cab.unina.it/tomatestdb
Email: fattakkie@libero.it