TOMATEST DB: a database of expressed sequences to mine on Tomato functional genomics

D'Agostino N (1), Aversano M (1), Frusciante L (2), Chiusano ML (1)

(1) Department of Structural and Functional Biology, University 'Federico II', 80134 Naples, Italy
(2) Department of Soil, Plant and Environmental Sciences, University 'Federico II', 80055 Portici (NA), Italy

Motivation

TomatEST db is a secondary database with primary EST sequence information collected from Solanum lycopersicum (200,438), Solanum pennellii (8,346), Solanum habrochaites (8,000) and Solanum lycopersicum X Solanum pimpinellifolium (1,008) available at the dbEST in the release of November 2005. TomatEST can be considered a fundamental resource for the International Tomato Genome Sequencing Project (http://www.sgn.cornell.edu/help/about/tomato_sequencing.pl), since EST collections are a quick route for discovering new genes and for confirming coding regions in genomic sequences and a workbench for tomato functional genomics. One promising approach for the extraction of biologically meaningful information on genome functionalities is to find suitable way for the classification and the organization of the data.
For the description of sequence function, we choosethe Gene Ontologies (GO; The Gene Ontology Consortium 2000) and the Enzyme Commission(EC; Bairoch, 2000) numbers so to directly classify the annotated sequences according to their functionality and to link data to known pathways.

Methods

TomatEST db core structure is a MySQL relational database collecting all the results of the sequence analysis automatically produced by the execution of the software ParPEST (D'Agostino et al. 2005). TomatEST is integrated with two satellite databases: myGO and myKEGG. The first database is a mirror of the GO database. The second one is built from KEGG (Kaneisha et al., 2004) xml formatted files and related maps in GIFF format. The database provides a complete set of computationally defined transcript indices. The transcript indices represent unique sequences, singletons or tentative consensus (TC), derived from EST clustering analysis. The four different collections are clustered independently accounting for a total of 52,780 transcript indices, 18,813 TC and 33,967 singletons. We based the functional annotation, from both primary data (ESTs) and transcript indices, on BLAST similarity searches versus the UniProt database (Apweiler et al., 2004). In case of successful matches (e-value less equal than 0.01), the three best blast hits are stored into TomatEST db. When the UniProt identifier is in myGO db, the database is crosslinked to the gene ontologies and when the EC number occurs in the best blast hit description lines, the database is crosslinked with myKEGG. The TomatEST web site is a PHP-based interface that allows easy access to the data.

Results

TomatEST db can be queried through a pre-defined query system to support non expert users. All the results are displayed in detailed and friendly graphical views. The data can be queried via three different HTML forms. The first form produces an 'EST report page', the second form results in a 'Clusters report page', and the last form results in a 'Transcript Indices report page' which represents the core structure of the web interface. Here the data corresponding to user-selected criteria can be inspected considering two different classes of objects: the enzymes and the metabolic pathways. The results are reported as HTML based tree menus. This format allows immediate and informative data retrieval. The expressed sequences resulting from a query can be mapped 'on the fly' onto metabolic pathways that can be accessed as GIFF images.

Availability: http://cab.unina.it/

Contact email: chiusano@unina.it

Supplementary informations

Acknowledgements This work is supported by the Agronanotech Project (Ministry of Agriculture, Italy).