Version VI of the ESTree db: an improved tool for peach transcriptome analysis

ID - 147

Lazzari Barbara[1], Caprera Andrea[1], Vecchietti Alberto[1], Merelli Ivan[2], Barale Francesca[3], Milanesi Luciano[2], Stella Alessandra[1], Pozzi Carlo[1]

[1]Parco Tecnologico Padano - CERSA, Lodi

[2]Istituto Tecnologie Biomediche - CNR, Segrate (MI)

[3]Universit degli Studi di Milano, Facolt di Scienze Agrarie, Milano

Motivation

The ESTree database (db) (http://www.itb.cnr.it/estree) is a collection of Prunus persica and Prunus dulcis EST sequences that in its current version (version VI, as of March 2007) encompasses 75,404 sequences from 22 libraries. Among these, three libraries (for a total of 3,864 sequences) are from almond, while the others are from peach. Nine peach genotypes (Bolero, Loring, OroA, Red Haven, Suncrest, Yumeyong, O'Henry, Baby gold 5 and Fantasia) and four peach tissues (mesocarp, skin, shoot, and leaf) are represented, and mesocarp sequences are from four different fruit developmental stages (post-allegation, pit hardening, pre-climacteric, and post-climacteric).

Aim of this work was to implement the already existing ESTree db adding new sequences and new analysis programs. Particular care was given to the implementation of the web interface, that allows querying the database and extracting data related to every feature that is taken into account in data analysis.

Methods

A Perl pipeline is the backbone of sequence analysis in the ESTree db project. The pipeline is modular, and integrates a number of public programs with scripts that were prepared by the authors to perform further analyses and to allow automatic data flow. Relevant outputs obtained during the pipeline steps automatically fill the fields of a MySQL database. Apart from ordinary analyses (for example sequence clustering and standard annotation), that have already been described in a previous work (Lazzari et al., BMC Bioinformatics 2005, 6(Suppl 4):S16), version VI of the ESTree db encompasses new tools for tandem repeats identification, annotation with BLASTn against an in-house prepared database of genomic rosaceae sequences, and positioning on the database of oligomer sequences that were used in a peach microarray study. Furthermore, the most probable putative protein sequences were inferred from each EST with FrameFinder and compared to the PROSITE database to search for known protein patterns and motifs. The already existing annotation procedure against the UniProtKB database was modified and a script was prepared to track positions of homologous hits on the GO tree and build statistics on the ontologies distribution in GO functional categories (referred to as 'GO statistics' in the web interface). GO statistics are given for a number of sequence subsets, to allow comparisons among the different tissues, genotypes or fruit developmental stages that are represented in the database. Data from the ESTree mapping project were integrated in the database and links to the GDR (http://www.bioinfo.wsu.edu/gdr/) Map Viewer pages were added for all the mapped sequences that are present in the db. All these modifications, together with a number of minor changes that were brought to the database, caused a reorganization of the PHP-based web interface that was upgraded and extended. Aim of the authors was to give the possibility to query the database according to all the biological aspects that can be investigated from the analysis of data available in the ESTree db. This is achieved allowing multiple searches on logical subsets of sequences that represent different biological situations or features, and retrieving only data that are relevant for each user's purpose.

Results

The version VI of ESTree db offers a broad overview on peach gene expression.

Sequence analyses results contained in the database represent a huge amount of information that can be extensively queried via the tools offered in the web interface, while links to external related resources offer the possibility to easily explore existing knowledge on peach and related species. Flexibility and modularity of the ESTree analysis pipeline and of the web interface allowed the authors to set up similar structures for different datasets, with limited manual intervention.

Availability: http://www.itb.cnr.it/estree

Email: barbara.lazzari@tecnoparco.org