ESTuber db: a tool for Tuber borchii EST sequence management

Caprera A (1,5), Cosentino C (2,6), Viotti A (2), Stella A (3), Milanesi L (4), Lazzari B (2,5)

(1) CISI, Via Fratelli Cervi 93, 20090 Segrate (MI), Italy
(2) Istituto di Biologia e Biotecnologia Agraria, via Bassini 15, 20133 Milan, Italy
(3) Parco Tecnologico Padano, Via Einstein - Località Cascina Codazza, 26900 Lodi, Italy
(4) Istituto Tecnologie Biomediche, Via Fratelli Cervi 93, 20090 Segrate (MI), Italy
(5) Current address: Parco Tecnologico Padano, Via Einstein - Località Cascina Codazza, 26900 Lodi, Italy
(6) Current address: Darmstadt University of Technology, Institute of Botany, Schnittspahnstrasse 3-5,
64287 Darmstadt, Germany

Motivation

The ESTuber database (http://www.itb.cnr.it/estuber/) represents a collection of 3,271 expressed sequenced tags (EST) of the white truffle Tuber borchii. The dataset consists of 2,389 sequences from an in-house prepared cDNA library obtained from vegetative mycelium, and 882 sequences downloaded from GenBank, representing four libraries obtained from vegetative hyphae and fruit bodies at different developmental stages. An automated pipeline was prepared to process EST sequences by using public software integrated with in-house developed Perl scripts. Data produced during EST processing were parsed and collected in a MySQL database. The database can be queried via a php-based web interface. The aim of this work was to create a public comprehensive resource of data and links related to truffle EST sequences.

Methods

A multifasta file containing the complete truffle EST dataset was used as input for the CAP3 program and 356 contigs were generated. Extensive sequence annotation was performed by blastx against the GenBank nr protein db and by blastn against an in-house prepared database of more than 42,000 genomic sequences from four filamentous fungi (Magnaporthe ssp, Aspergillus ssp, Fusarium ssp and Neurospora ssp) and a dimorphic fungus (Saccharomyces ssp). Blastx was also performed against the UniProtKB database, to annotate sequences according to the Gene Ontology (GO) project, and an algorithm was implemented to infer statistical classification from the ontologies occurrences. GO statistics are provided on the ESTuber db web site for the whole sequence set and for library-specific subsets. All the blast output pages are available and can be queried by text search. Detection of tandem repeats was performed on all the EST sequences, as well as on the contig consensus sequences, with the Tandem Repeats Finder software. Putative polypeptide sequences were deduced from nucleotide sequences with FrameFinder and compared to the PROSITE database of protein families and domains. Links to matching patterns are given in the database web interface. A local blast utility was set up to perform blast searches on the ESTuber db nucleotide dataset or on the inferred protein sequences. A text search utility allows querying all the database fields and query outputs can be downloaded in multiple formats. Logical sequence subsets (singlets/sequences participating to contig assembly, unigenes, repeats containing sequences, protein pattern containing sequences) were defined and can be searched as independent datasets. An exhaustive help page was included in the database to help users to surf the database and to interpret the program outputs.

Results

The resulting database represents, at present, the most comprehensive public resource for Tuber EST sequences, and can be helpful to all researchers interested in the molecular biology of truffle and of other filamentous fungi. The Perl pipeline automatically fills in all the database fields and very little manual participation is required for the complete assembly of new releases. The database structure is modular and new applications can easily be integrated. The in-house developed GO statistics tool was useful for comparison of inter-species ontologies occurrances as well as for database data analysis, and represents a powerful tool for investigating EST distribution in Gene Ontology categories. The ESTuber db is intended to be the main repository of information related to Tuber sequences, and will be updated and modified according to the needs of the Truffle Project.

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

Contact email: barbara.lazzari@tecnoparco.org