An integrated computational workbench for structural and functional analysis in the solanaceae genomics network

ID - 172

Chiusano Maria Luisa[1], D'agostino Nunzio[1], Traini Alessandra[1], Licciardello Concetta[2], Raimondo Enrico[1], Aversano Mario[1], Frusciante Luigi[1]

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

[2]Istituto Sperimentale per l'Agrumicoltura, CRA, Acireale (CT), Italy

Motivation

In the frame of the International Tomato Genome Sequencing Project (http://www.sgn.cornell.edu/), the Bioinformatics Committee coordinates the management and the analysis of the genomics data generated within the Solanaceae Genomics Network. The aim is to offer, in a distributed platform, integration of data sources from large scale analysis in order to support the research of the Solanaceae family which includes many species of relevant agricultural interest such as tomato and potato. As participants to the Bioinformatics of the Network, we set up a workbench to support i) the annotation of the Solanum lycopersicum (tomato) genome, ii) the characterization of the tomato genome and of its functional properties, iii) the comparative analysis of sequence collections from other Solanaceae species. The actual organization of the workbench and its novel features are presented.

Methods

The workbench is built and updated to improve the current available methodologies, to expand the data according to their increase, and to enhance integration of different data sources. To date the main resources of the workbench are the genome and the transcriptome data.

Transcriptome analysis. Within the Eu-SOL (EU VI Frame Programme), we are committed to maintain EST sequence collections from Solanaceae species retrieved from dbEST.

ESTs are processed to remove redundant sequences and then are released to the community through our ftp site.

The ParPEST pipeline (DAgostino et al., 2005) is performed at each increase of 5000 sequences, to cluster/assemble ESTs and to generate tentative consensus sequences (TCs) per putative transcripts. The resulting data are organized into relational databases (DAgostino et al., 2007) and are released as a complete set of annotated transcript indices. The computational annotation of the expressed sequence data is based on BLASTX analysis versus the UniProtKB/Swiss-Prot database and on the use of controlled vocabularies such as the Gene Ontologies and the Enzyme Commission numbers. We implemented the on the fly mapping of the transcripts onto the KEGG metabolic pathways (Kaneisha et al., 2004) to support expression pattern analysis. We included in the annotation procedure the analysis of non-coding RNAs and correspondences between the EST dataset and the probe-sets from both the tomato reference arrays TOM1,

a cDNA microarray (http://ted.bti.cornell.edu/), and the actual Affymetrix chip.

Genome analysis. The preliminary annotation of the BAC sequences released by the SOL Genomics Network was based on the spliced-alignments of Solanaceae ESTs and on the tentative consensus sequences (TCs) as generated by our group and as retrieved from the SGN (Mueller et al., 2005) and the TIGR repositories (Lee et al., 2005).

Arabidopsis thaliana coding sequences and the TIGR collection of Solanaceae repeats are mapped onto the genomic sequences too. Gene models are defined by the software GeneModelEST (DAgostino et al., 2007) which evaluates the quality and reliability of the tentative consensus sequences from tomato and other Solanaceae species.

The annotated BACs can be accessed through the Generic Genome Browser (GBrowse) at http://biosrv.cab.unina.it/GBrowse/.

Results

The computational workbench here presented was built to provide an Italian resource for the genomics of the Solanaceae and for the International Tomato Genome Sequencing Project. The workbench can be accessed via web based interfaces. The workbench is useful to support the experimental annotation of the genomic data; indeed, EST data are a quick route for discovering new genes and for confirming coding regions in genomic sequences. Moreover, it supports the analysis of coding and non-coding gene families. However, the most relevant aim of the workbench design is to provide methodologies to allow genome scale structural and functional analyses. The idea is to allow investigations on specific expression patterns as they can be derived from a holistic view of the transcriptome data integrated with arrays results contributed by the community. Our effort aims to provide novel approaches for investigations on the organization and functionalities of genomes to meet the bioinformatic challenges of comprehensive analyses within the Sol Genomics Network based on a systems biology view.

Availability: http://cab.unina.it

Email: chiusano@unina.it