Linking the tomato and the potato genomes: the comparative season is now opened

Traini A[1],D'Agostino N[1], Aversano M[1], Frusciante L[1], Chiusano ML[1]

Motivation

The considerable ease in sequencing genome and transcript data leads to the need of appropriate tools for comparative genomics. An interesting case study is provided by the goal of the International Solanaceae (SOL) Genome Project. In particular, since Solanaceae share a high genome conservation, the long-term goal of the SOL Consortium is to exploit the information generated from the sequencing of solanaceae species, such as tomato (Solanum lycopersicum) and potato (Solanum tuberosum), to analyze, by comparative approaches, the genome organization, the functionality and the molecular evolution of the entire Solanaceae family. A centralized bioinformatics resource is an important prerequisite for the investigation of such an extensive data collection. Tools to integrate and compare genome annotations from multiple-species are still poor and therefore efforts are necessary to fully exploit such intriguing data collections. Currently available genome browsers by themselves, indeed, are not yet adequate to establish a comprehensive information system and to take full advantage from the integration of various collections from different species. This is why we designed ISOLA, an Italian SOLAnaceae bioinformatics resource, made up of different databases integrated into a unique platform, which was first focused on the tomato genome (BITS, 2008). The extended version of ISOLA is here presented, as it now includes data from different species and novel tools for comparative analyses within Solanaceae genomics.

Methods

BAC sequences from S. lycopersicum cv. Heinz 1706 and from S. tuberosum cv. RH89-039-16 were retrieved from GenBank. ESTs and the corresponding tentative consensus sequences (TCs) from different Solanaceae species, which are collected in a dedicated MySQL database known as SolEST (D'Agostino et al., 2009), embedded within ISOLA, were also aligned along tomato and potato BAC sequences. The identification of interspersed repeats was performed by the RepeatMasker tool (http://repeatmasker.org) using the Plant Repeat Database at Michigan State University (http://plantrepeats.plantbiology.msu.edu/), RepBase.13.06 (http://www.girinst.org/server/archive/) and the SGN tomato UniRepeats (ftp://ftp.sgn.cornell.edu /tomato_genome/repeats/) as filtering

[1] Dept. of Soil, Plant, Environmental and Animal Production Sciences, University of Naples Federico II, Naples

databases. Annotated genome data from both tomato and potato were independently organized and made accessible using the GBrowse package (version 1.69). Pre-computed MySQL tables from transcript data (SolEST) and the genome data (GBrowse databases) were created to exploit data integration and achieve faster query rensponses. Data are accessible through web-based tools, developed in PHP (version 5.2.6), where user-friendly query systems and graphical approaches have been developed to enhance inter genome data browsing.

Results

ISOLA was conceived as a multi-level computational environment, that could be accessed through two convenient gateways. The ‘genome’ gateways, which now includes data from both S. lycopersicum and S. tuberosum, and the 'transcriptome' gateway which provides an access point to explore ESTs and tentative transcripts (TC) from different species included in SolEST. The 'genome' gateway allows to independently query the gbrowse based annotations of the S. lycopersicum and S. tuberosum BAC sequences, which include several tracks represented by the spliced-alignments of the EST/TC collections, repeat content and similarity to known proteins. The main novelty of the platform, is the possibility to crosslink both genome and transcript data based on comparative approaches. Comparative queries of the transcript collections are based on a unified UNIPROT annotation of all the unique transcripts from all the species included in SolEST. Tracks representing annotation on both genomes are dynamically crosslinked, permitting multiple hits from multiple tracks investigations, and comparative analyses of genome sequences with common information from tomato and potato. We think that the upgraded version of ISOLA, which combines data from multiple species and is supported by novel user friendly query systems as well as graphical approaches such as the real-time crosslinked network rappresentation viewer, is recommendable in order to better perform large scale analyses on multiple species. The integrated design of ISOLA here proposed, can be an example of a bioinformatics platform useful for comparative genomics.

Contact e-mail

chiusano@unina.it