Proceedings of the 50th Italian Society of Agricultural Genetics Annual Congress
Ischia, Italy – 10/14 September, 2006
ISBN **88-900622-7-4**

Poster Abstract – D.52

A COMPUTATIONAL SERVICE FOR THE ANALYSIS AND THE MANAGEMENT OF EXPRESSED SEQUENCE COLLECTIONS

N. D'AGOSTINO*, M. AVERSANO*, L. FRUSCIANTE**, M. L. CHIUSANO*

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

computational EST analysis, EST database, gene indices

Large Expressed Sequence Tag (EST) datasets are daily released to the scientific community. Despite their intrinsic shortcomings due to contaminations and limited sequence quality, ESTs represent a consistent resource for gene discovery, for genome annotation, for comparative genomics and for expression studies. The 'tag' nature and the vast quantity of ESTs require suitable and efficient approaches to harvesting the full potential from this data source.

We present here our effort for the construction of information frameworks, where the fragmented and error-prone EST data are processed to provide high quality and biologically meaningful collections. We implemented the ParPEST (Parallel Processing of ESTs) pipeline using public software integrated by in-house developed Perl scripts. Input EST sequence data are screened i) for trimming low quality sequences ii) for cleaning vector contaminations and iii) for filtering and masking low complexity sub-sequences and interspersed repeats. Than, sequences are clustered and assembled to detect sequence redundancy and to generate gene indices. Automated annotation is obtained from BLAST similarity searches against protein databases. For the description of sequence function, Gene Ontology (GO) terms and Enzyme Commission (EC) numbers are assigned to directly classify gene products according to international standards and to map *on the fly* the expressed sequences onto KEGG metabolic pathways.

All the data resulting from each single step of the pipeline are collected into a MySQL relational database. We implemented also a web application, based on PHP language, with a pre-defined query system to support the interactive browsing of the results by non expert users. The described methodology has already been applied to the analysis of i) a collection of ~ 200.000 ESTs from different tomato species (http://biosrv.cab.unina.it/tomatestdb) and ii) a collection of over 200.000 ESTs sequenced from different libraries of *Solanum tuberosum* (http://biosrv.cab.unina.it/potatestdb) in the frame of our effort for the International Solanaceae Genomics Network; iii) a collection of ~ 10.000 ESTs from a saffron stigma cDNA library in a collaborative effort with the ENEA (D'Agostino et al., SIGA 2006) .

Acknowledgements: This work is supported by the Agronanotech Project (Ministry of Agriculture, Italy) and by the EU-SOL Project (VI Frame Programme, European Union) .