A tool for storage, automated annotation and analysis of Conserved Sequence Tags (CSTs) - (session: Comparative Genomics and Molecular Evolution)

A. Boccia*, M. Petrillo^, D. Di Bernardo°, S. Banfi°, A. Guffanti#, G. Pesole+, G. Paolella*^§

*BIOGEM, Ariano Irpino (AV); ^CEINGE, Napoli; °TIGEM, Napoli; #IFOM, Milano; +Universita' di Milano; §Universita' del Molise, CB.

Comparative analysis of DNA sequences from multiple species at varying evolutionary distances is a powerful approach in the identification of coding and functional noncoding sequences. Recently TIGEM initiated a large project which now involves several other institutions including IFOM, CEINGE, BIOGEM, University of Milano and other italian institutions aimed to the identification, automatic annotation and characterization of conserved sequence tags (CST) from about 1000 genes known to be involved in genetically transmitted diseases, through a number of approaches ranging from bioinformatic to laboratory experiments.
Here we report on the development of a database system for collection, storage and automatic annotation of CSTs, which also includes facilities for interrogation and graphic display in the chromosomal context. DNA regions from orthologous human and mouse genes are identified using the BLASTZ program and resulting alignments processed with the Strong-hits program from the PipMaker package. Conserved regions are stored in the CST database, together with annotation data derived from automatic scanning of ENSEMBL, LocusLink, Gene Ontology and other databases. Information regarding CST mapping on the chromosome sequence, relationships to intron exon gene structure, conservation related to taxonomial distribution of genes, similarity to other sequences as found by alignment with various nucleic and proteic datasets using the blast program, expression data shown by matches in EST databases are all easily accessed and searched. A simple, but effective, tool for graphic visualization of CSTS within the gene context is also included, which allows fast browsing along the chromosome. Data from other analysis tools may easily be added; we are currently including information on coding potential, as well as information about general sequence conservation.
Preliminary results of statistical analysis of the CSTs contained in the DB will be reported.