EST Analysis system for panning gene

Cheol-Goo Hur*, So-hyung Lim, Sung-Ho Goh, Min-Su Shin, Hwan-Gue Cho[1]
Genome Research Center, Korea Research Institute of Bioscience and Biotechnology,
Department of Computer Science, Pusan National University[1]

Abstract

Expressed sequence tags (ESTs) are the partial segments of cDNA produced from 5 or 3 single-pass sequencing of cDNA clones, error-prone and generated in highly redundant sets. Advancement and expansion of Genomics made biologists to generate huge amount of ESTs from variety of organisms-human, microorganisms as well as plants, and the cumulated number of ESTs is over 5.3 million. As the EST data being accumulate more rapidly, it becomes bigger that the needs of the EST analysis tools for extraction of biological meaning from EST data. Among the several needs of EST analyses, the extraction of protein sequence or functional motifs from ESTs are important for the identification of their function in vivo. To accomplish that purpose the precise and accurate identification of the region where the coding sequences (CDSs) is a crucial problem to solve primarily, and it will be helpful to extract and detect of genuine CDSs and protein motifs from EST collections. Although several public tools are available for EST analysis, there is not any one to accomplish the object. Furthermore, they are not targeted to the plant ESTs but human or microorganism. Thus, to correspond the urgent needs of collaborators deals with plant ESTs and to establish the analysis system to be used as general-purpose public software we constructed the pipelined-EST analysis system by integration of public software components. The software we used are as follows - Phred/Cross-match for the quality control and vector screening, NCBI Blast for the similarity searching, ICATools for the EST clustering, Phrap for EST contig assembly, and BLOCKS/Prosite for protein motif searching. The sample data set used for the construction and verification of this system was 1,386 ESTs from human intrathymic T-cells that verified using UniGene and Nr database of NCBI. The approach for the extraction of CDSs from sample data set was carried out by comparison between sample data and protein sequences/motif database, determining matched protein sequences/motifs that agree with our defined parameters, and extracting the regions that shows similarities. In recent future, in addition to these components, it is supposed to be also integrated into our system and served that the software for the peptide mass spectrometry fingerprint analysis, one of the proteomics fields. This pipelined-EST analysis system will extend our knowledge on the plant ESTs and proteins by identification of unknown-genes.

Curriculum Vitae

Name : Cheol-Goo Hur

Position : Senior Speciallist, Genome Research Center, Korea Research Institute of Bioscience and Biotechnology (KRIBB).

Address : 52 Oun-Dong, Yusung-Gu, DAEJEON, KRIBB, KOREA.

Education

1982-1990 Graduated, Department of Computer Science and Statistics, Chungnam National University

1992 - 1995 Master course, Department of Computer Science, Chungnam National Univesity.

2000 - Entered Ph.D Course, Department of Bioinformatics, Pusan National University.

Project Experience

1997.10 - 1998. 9 Construction of GenBank for Korea Genome Researcher.
1998.10 - 1998. 6 PDB Databases
1998.10 - 1999.7 Construction of Genome Information Processing system.
1999. 8 - 2000.7 Large-scale functional analysis of genes using computational technology

Current project

2000.7 - 2000.12 Construction of Gene Retrieve system using UniGene DB.
2000.7 - 2001. 2 Construction of GenBank and Coding sequences DB.
2000.7 - 2001. 6 Development of EST Analysis system using Public S/W and DB.
2000.9 - 2003. 7 (21c frontier program / Plant Diversity Research)
Plant EST Analysis and DNA Chip expression Analysis system.