Automatic annotation tools

Coral del Val, Mechthild Falkenhahn, Karl-Heinz Glatting, Agnes Hotz-Wagenblatt, Barbara Pardon, and Sándor Suhai
Deutsches Krebsforschungszentrum (DKFZ), Abt. Mol. Biophysik, Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany

Keywords. Genomics, Annotation

The aim of a high-quality annotation is to identify the key genome features - in particular, the genes and their products in order to bridge the considerable gap between large-scale data-collection and its interpretation. The tools and resources for annotation are developing rapidly, however, high throughput data analysis requires the correct combination of these applications and databases.
In this context we have developed different analysis tools for semi-automatical annotation of sequences. In collaboration with other NGFN/DHGP research groups we have designed tools supporting cDNA mapping and characterization of full length cDNAs (cDNA2Genome), semi-automatical analysis of EST sequences supporting the search of functional annotations of novel transcript sequences (ESTAnnotator), classification of protein sequences and inference of function, combining different classification and search methods of the commonly used protein family databases (Prosite, Prints, Pfam, Blocks, Smart, Prodom, Tigrfams), as well as Interpro and Gene Ontology (DomainSweep), protein secondary structure prediction (2Dsweep), etc... cDNA2Genome gives information about the chromosomal and contig location of the cDNA. It extracts the genomic sequence where the cDNA is located and predicts the exons and introns in this region in both strands using tools likeGenScan, HMMgene, GeneID, GeneWise and Sim4. Additionally the genomic sequence is used as input for different homology searches with the aim of getting more evidences for the annotation process. This includes comparisons against the mouse genomic database, against EST databases (our EST databases contains all the unified expressed sequences contained in EMBL, Genbank, and their updates), against mRNA databases, UniGene, Swissprot and cDNA databases.
The final result is an XML file which contains all relevant information obtained by the task. This XML output can be easily used in successive analysis (i.e. in pipelines or for integrating the data into databases). At the same time, the web user can easily inspect the xml output through the graphical display using standard web browsers. The graphical representation provides an interactive view of the annotations.

All tools have been implemented under the W3H-Task-System (P.Ernst et al., Bioinformatics (2003)). Availability: http://genome-dkfz-heidelberg.de.

References

[1] Ernst, P., Glatting, K., Suhai, S. A task framework for the web interface W2H. Bioinformatics **19**, 278-282, 2003

[2] Hotz-Wagenblatt, A., Hankeln, T., Ernst, P., Glatting, K., Schmidt, E.R., Suhai, S. ESTAnnotator: a tool for high throughput EST annotation. Nucleic Acids Res. (in press)

[3] del Val, C. Bräuning, R., Glatting, K. and Sándor Suhai PATH: a task for the inference of phylogenies. Bioinformatics. **18**, 646-647. 2002.