Strategies and Tools for EST Data Mining

A.Guffanti, S.Banfi, G.Borsani and G.Simon

TIGEM - Telethon Institute of Genetics and Medicine - San Raffaele
Biomedical Science Park -Via Olgettina, 58, 20132 Milano, Italy.
email: guffanti@tigem.it Tel: 0221560212 Fax: 0221560220

Expressed Sequence Tags (ESTs) constitute an important source of information for laboratories interested in the identification of novel gene sequences. We developed bioinformatic strategies and tools which rely on dbEST sequence data mining in order to support the effort of disease gene identification both at TIGEM and worldwide. One example of systematic human EST analysis is the DRES (Drosophila Related Expressed Sequences) project. As a starting strategy, we applied the power of Drosophila genetics to identify novel human genes of high biological interest. Sixty-six human cDNAs (called DRES clones) showing significant homology to Drosophila mutant genes were identified by screening dbEST with keywords, and their map position was determined experimentally. Based on this approach we developed the "DRES Search Engine", a tool for the systematic identification of human cDNAs homologous to Drosophila genes through an automated sequence database searching procedure. The homepage of the DRES project is at the WWW address http://www.tigem.it/LOCAL/drosophila/dros.html. Other tools of interest to the researchers interested in maximizing the information associated with a single cDNA sequence are freely available at the WWW address http://www.tigem.it/LOCAL/sequtils.html : - the "In Situ Blast" server performs a library-specific (and consequently tissue-specific) Blast search against one or more given cDNA libraries belonging to the UniGene EST cluster database; - the "UniBlast" server performs a local Blast search against the UniGene database or against UniNewGene, a locally generated version of UniGene devoid of all the clusters containing an already known mRNA or coding sequence; - the "EST Assembly Machine" and the "EST Extractor" will build sequence contigs (corresponding to "virtual transcripts") from the UniGene EST cluster database or from dbEST respectively, starting from a sequence Accession Number or a plain DNA/Protein sequence. This procedure extends a given cDNA sequence information through repeated cycles of sequence comparison, ideally providing the sequence of a full-length transcript starting from a single query sequence.

References

S.Banfi, G.Borsani, E.Rossi, L.Bernard, A.Guffanti, F.Rubboli, A.Marchitiello, S.Giglio, E.Coluccia, M.Zollo, O.Zuffardi & A.Ballabio: Identification and mapping of human cDNAs homologous to Drosophila mutant genes through EST database searching. Nature Genetics, Vol. 13, p 167-174, 1996.

S.Banfi, A.Guffanti & G.Borsani. 'How to get the best of dbEST'. Trends in Genetics, Vol. 14 No. 2, pagg. 80-81, February 1998.

A.Guffanti & G.Simon. 'UniBlast and the Est Extractor: new WWW resources for EST Data Mining' . Trends in Genetics, Vol. 14 No. 7, pag. 293, July 1998.