trome, trEST and trGEN: databases of predicted protein sequences

Peter Sperisen1,*, Christian Iseli1,3, Marco Pagni1,3, Brian J. Stevenson1,3, Philipp Bucher1,2 and C. Victor Jongeneel1,3

1Swiss Institute of Bioinformatics, 2ISREC and 3Office of Information Technology, Ludwig Institute for Cancer Research, Chemin des Boveresses 155, 1066 Epalinges s/Lausanne, Switzerland

Received September 15, 2003; Revised and Accepted September 29, 2003

ABSTRACT

We previously introduced two new protein databases (trEST and trGEN) of hypothetical protein sequences predicted from EST and HTG sequences, respectively. Here, we present the updates made on these two databases plus a new database (trome), which uses alignments of EST data to HTG or full genomes to generate virtual transcripts and coding sequences. This new database is of higher quality and since it contains the information in a much denser format it is of much smaller size. These new databases are in a Swiss-Prot-like format and are updated on a weekly basis (trEST and trGEN) or every 3 months (trome). They can be downloaded by anonymous ftp from ftp://ftp.isrec.isb-sib.ch/pub/
databases.

DESCRIPTION OF DATABASES

High-throughput genome (HTG) and expressed sequence tag (EST) sequences are currently the most abundant nucleotide sequence classes in the public databases. The large volume, high degree of fragmentation and lack of gene structure annotations prevent efficient searches of HTG and EST data for protein sequence homologies by standard search methods. We have compiled three databases of predicted and annotated protein sequences to facilitate the use of proteomics tools. All databases are in a Swiss-Prot-like format and are updated on a weekly basis (trEST and trGEN) or every 3 months (trome). They can be downloaded by anonymous ftp from ftp://ftp.isrec.isb-sib.ch/pub/
databases.

trome

trome is an attempt to map transcribed RNA from different sources to the NCBI RefSeq genome sequence (1,2). As an example, for Homo sapiens the transcribed RNA sources are: the human EST section of the EMBL database (3), the human HTC section of the EMBL database, human mRNA documented in the EMBL database, ORESTES sequences from the LICR/FAESP Human Cancer Genome project (4,5), human mRNA documented in the NCBI-curated RefSeq database [http://www.ncbi.nih.gov/RefSeq (6)], published HR21 gene list and SEREX sequences [http://www2.licr.org/CancerImmunomeDB (7)]. For other species, similar sources are used. Currently four species are represented: H.sapiens, Mus musculus, Arabidopsis thaliana and Caenorhabditis elegans (Table 1). The mapping of the transcribed RNA sources to the genome is a three-step process (1,2):

(i) The program Megablast (8) is used to identify pairwise similarities between all known transcript sequences and the genomic data.
(ii) For each pair of matching RNA and genomic sequences, local alignments were generated using a modified version of sim4 (9).
(iii) The output of sim4 was filtered to eliminate all alignments that did not contain at least one region (exon) matching with at least 95% identity over their high-quality part and 88% over the remainder.

The output of sim4 is then used to generate directed acyclic graphs using the program tromer (locally developed program to automate the reconstruction of transcripts from transcript to genome mapping). These graphs (the edges and nodes represent exons/introns or splice donors/acceptors, respectively) represent transcribed loci of the genome and contain in a condensed form the information about all possible alternative splice variants that are experimentally documented. They can be used to reconstruct virtual transcripts from the underlying genomic sequence following a path from 3’ tags along experimentally verified exon boundaries. Transcript generation is a three-step process: (i) a seed edge is selected; (ii) this edge is extended toward the 5’ end and (iii) toward the 3’ end. The seed edge is first selected among unused 5’-most exons, then among any unused edge. The extension process always attempts to include unused edges, which were derived from the same RNA elements as the seed edge. The resulting virtual transcripts are translated into protein sequences using the program ESTScan (10). These protein sequences are the basis of the trome database. ESTScan detects the coding frame and corrects most frameshift errors introduced by sequencing errors, but predicts their position within a range of a few amino acids. Simulation experiments have shown that in 95% of the cases the range is seven or fewer amino acids. To visualize this uncertainty, the FT key UNSURE was used, indicating the range within which the predicted sequence is more likely to contain errors. However, due to the mapping of transcribed RNA data onto the genome, this is a rare event in contrast to
improvements were introduced: the program Genscan (13). The following sequences are searched for putative genes and their coding from HTG sequences from the EMBL database. The predicted from genomic sequences from the NCBI database or database for the ESTs that do not belong to UniGene clusters.

Entries that are based on UniGene clusters and to the EMBL compared to the database trome. As well as internal stop codons are found more frequently as UniGene clusters as well as the ESTs, since frameshift errors ESTScan are adapted to the error prone contigs produced from correction of internal stop codons. The parameters used with frameshift error correction by the program ESTScan plus the UNSURE was used to reflect the uncertainty range in UniGene cluster were also introduced into trEST. Protein sequences of coding ESTs that are not present in any produced from UniGene clusters (12) using ESTScan. That were generated through the translation of contigs corrections found in the database trEST (see below). The new FT key EXON was introduced to indicate the positions of the exon boundaries with respect to the NT contig.

(iii) The new FT Key GENSCAN was introduced. It contains the predictions made by the program Genscan. These are: FIRST EXON, INTERNAL EXON, LAST EXON and SINGLE EXON together with their associated p-values (sum over all parses containing exon) calculated by GENSCAN, which serve as an indication about the degree of certainty that should be ascribed to exons predicted by the program

(iv) The ID is composed of either the EMBL or NCBI accession number of the contig on which the protein was predicted, plus a number (_#) that enumerates the proteins as they are found on the contig.

(v) trGEN is cross-referenced to either the EMBL or the NCBI RefSeq database, with a cross-link to the underlying contig.

UPDATE TO THE DATABASES

The trEST and the trGEN databases are updated on a weekly basis. The trome database is updated roughly every 3 months.

ACCESS

FTP


WWW

Several web pages offer services that include the trome, trEST and trGEN databases. http://www.ch.embnet.org/software/fetch.html allows one to retrieve individual entries of trome, trEST and trGEN. http://www.ch.embnet.org/software/ aBLAST.html allows the three databases of hypothetical proteins to be searched using BLAST.

ACKNOWLEDGEMENTS

This work was supported by a grant from the Swiss Federal office for education and health (OFES): 01.0101 which is part of the TEMBLOR project (QLRI-CT-2001-00015) of the Quality of Life and Management of Living Resources Programme and by the Ludwig Institute for Cancer Research.

REFERENCES


