Nucleic Acids Research, 2004, Vol. 32, Database issue D509–D511
DOI: 10.1093/nar/gkh067

trome, trEST and trGEN: databases of predicted protein sequences

Peter Sperisen[1,*], Christian Iseli[1,3], Marco Pagni[1,3], Brian J. Stevenson[1,3], Philipp Bucher[1,2] and C. Victor Jongeneel[1,3]

[1]Swiss Institute of Bioinformatics, [2]ISREC and [3]Office 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 distributed in a Swiss-Prot-like format, with features that are specific to the databases presented below.

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/FAPESP Human Cancer Genome project (4,5), human mRNA documented in the NCBI-curated RefSeq database [http://www.ncbi.nih.gov/RefSeq (6)], published CHR21 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*, *Drosophila melanogaster*, *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

[*]To whom correspondence should be addressed. Tel: +41 21 6925956; Fax: +41 21 6925945; Email: peter.sperisen@isb-sib.ch

Table 1. Number of entries in each database for each species represented (established September 5, 2003)

	trGEN	trEST	trome
Homo sapiens	196110	1709305	121072
Mus musculus	225759	1018677	95754
Rattus norvegicus	293151	274511	n.a.
Drosophila melanogaster	128071	28254	20717
Arabidopsis thaliana	31424	39924	n.a.
Oryza sativa	111723	57269	n.a.
Bos taurus	n.a.	88378	n.a.
Danio rerio	n.a.	125453	n.a.
Hordeum vulgare	n.a.	79315	n.a.
Triticum aestivum	n.a.	146462	n.a.
Xenopus laevis	n.a.	116084	n.a.
Zea mays	n.a.	121846	n.a.
Caenorhabditis elegans	n.a.	n.a.	25841
Total	986238	3805478	263384

Due to the limited amount of data, not all species are represented in all databases. Missing data are indicated by 'n.a.'.

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.

```
FT EXON 1 173 Exon E0;NT_026943[46041..46562].
FT EXON 174 174 AA on splice site: tt/g -> L.
FT EXON 175 405 Exon E1;NT_026943[56062..56757].
FT EXON 406 406 AA on splice site: a/tg -> M.
FT EXON 407 514 Exon E2;NT_026943[63614..64087]
FT UNSURE 507 514 Frameshift error at pos.: 514;
base inserted:
```

trEST and trGEN

trEST is an attempt to produce contigs from clusters of ESTs and to translate them into proteins (11). In the past 2 years the following improvements have been introduced:

(i) Initially trEST was composed only of protein sequences that were generated through the translation of contigs produced from UniGene clusters (12) using ESTScan. Protein sequences of coding ESTs that are not present in any UniGene cluster were also introduced into trEST.

(ii) The species list has been increased (see Table 1).

(iii) Exactly as described for the trome database, the FT key UNSURE was used to reflect the uncertainty range in frameshift error correction by the program ESTScan plus the correction of internal stop codons. The parameters used with ESTScan are adapted to the error prone contigs produced from UniGene clusters as well as the ESTs, since frameshift errors as well as internal stop codons are found more frequently as compared to the database trome.

trEST is cross-referenced to the UniGene database for the entries that are based on UniGene clusters and to the EMBL database for the ESTs that do not belong to UniGene clusters.

The amino acid sequences of the trGEN database are predicted from genomic sequences from the NCBI database or from HTG sequences from the EMBL database. The sequences are searched for putative genes and their coding regions using the program Genscan (13). The following improvements were introduced:

(i) The species *Rattus norvegicus* was added.

(ii) The predictions for *H.sapiens*, *M.musculus*, *R.norvegicus* and *A.thaliana* are now made on the basis of the NCBI reference genome sequences (NT contigs).

(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

```
FT GENSCAN 1 226 FIRST EXON; p-value: 0.159.
FT GENSCAN 227 262 INTERNAL EXON; p-value: 0.093.
FT GENSCAN 263 269 INTERNAL EXON; p-value: 0.065.
FT GENSCAN 270 320 LAST EXON; p-value: 0.074.
```

(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

The files for the three databases are available by anonymous ftp from the directories: ftp://ftp.isrec.isb-sib.ch/pub/databases/trest, ftp://ftp.isrec.isb-sib.ch/pub/databases/trgen and ftp://ftp.isrec.isb-sib.ch/pub/databases/trome. In addition user manuals which provide more details about the format of each database are found in the individual directories.

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

1. Iseli,C., Stevenson,B.J., de Souza,S.J., Samaia,H.B., Camargo,A.A., Buetow,K.H., Strausberg,R.L., Simpson,A.J., Bucher,P. and Jongeneel,C.V. (2002) Long-range heterogeneity at the 3′ ends of human mRNAs. *Genome Res.*, **12**, 1068–1074.

2. Stevenson,B.J., Iseli,C., Beutler,B. and Jongeneel,C.V. (2003) Use of transcriptome data to unravel the fine structure of genes involved in sepsis. *J. Infect. Dis.*, **187** (Suppl. 2), S308–S314.
3. Stoesser,G., Sterk,P., Tuli,M.A., Stoehr,P.J. and Cameron,G.N. (1997) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.*, **25**, 7–14.
4. Dias,N.E., Correa,R.G., Verjovski-Almeida,S., Briones,M.R., Nagai,M.A., da Silva,,W.,Jr, Zago,M.A., Bordin,S., Costa,F.F., Goldman,G.H. *et al.* (2000) Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proc. Natl Acad. Sci. USA*, **97**, 3491–3496.
5. Camargo,A.A., Samaia,H.P., Dias-Neto,E., Simao,D.F., Migotto,I.A., Briones,M.R., Costa,F.F., Nagai,M.A., Verjovski-Almeida,S., Zago,M.A. *et al.* (2001) The contribution of 700 000 ORF sequence tags to the definition of the human transcriptome. *Proc. Natl Acad. Sci. USA*, **98**, 12103–12108.
6. Pruitt,K.D., Katz,K.S., Sicotte,H. and Maglott,D.R. (2000) Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet.*, **16**, 44–47.
7. Tureci,O., Sahin,U. and Pfreundschuh,M. (1997) Serological analysis of human tumor antigens: molecular definition and implications. *Mol. Med. Today*, **3**, 342–349.
8. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.
9. Florea,L., Hartzell,G., Zhang,Z., Rubin,G.M. and Miller,W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.*, **8**, 967–974.
10. Iseli,C., Jongeneel,V. and Bucher,P. (1999) ESTScan: A program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proceedings of the Seventh ISMB*, pp. 138–148.
11. Pagni,M., Iseli,C., Junier,T., Falquet,L., Jongeneel,V. and Bucher,P. (2001) trEST, trGEN and Hits: access to databases of predicted protein sequences. *Nucleic Acids Res.*, **29**, 148–151.
12. Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, **75**, 694–698.
13. Burge,C.B. and Karlin,S. (1998) Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.*, **8**, 346–354.