148–151 Nucleic Acids Research, 2001, Vol. 29, No. 1 

# trEST, trGEN and Hits: access to databases of predicted protein sequences

**Marco Pagni[1], Christian Iseli[1,3], Thomas Junier[1,2], Laurent Falquet[1,2], Victor Jongeneel[1,3] and Philipp Bucher[1,2,*]**

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

Received September 1, 2000; Revised and Accepted November 1, 2000

## ABSTRACT

**High throughput genome (HTG) and expressed sequence tag (EST) sequences are currently the most abundant nucleotide sequence classes in the public database. The large volume, high degree of fragmentation and lack of gene structure annotations prevent efficient and effective searches of HTG and EST data for protein sequence homologies by standard search methods. Here, we briefly describe three newly developed resources that should make discovery of interesting genes in these sequence classes easier in the future, especially to biologists not having access to a powerful local bioinformatics environment. trEST and trGEN are regularly regenerated databases of hypothetical protein sequences predicted from EST and HTG sequences, respectively. Hits is a web-based data retrieval and analysis system providing access to precomputed matches between protein sequences (including sequences from trEST and trGEN) and patterns and profiles from Prosite and Pfam. The three resources can be accessed via the Hits home page (http://hits.isb-sib.ch).**

## DESCRIPTION OF THE DATABASES

The three databases presented here are intended to help biologists to rapidly retrieve or discover proteins in expressed sequence tag (EST) and genomic sequences. trGEN and trEST are automatically generated collections of hypothetical proteins (see below). Although error-prone, they constitute a rich source of as yet undocumented proteins (1). Hits is an accompanying database that gathers lists of matches of protein-domain predictors (see below) against the databases of hypothetical proteins. These pre-compiled lists allow one to quickly query the protein databases for protein domains predicted by the most powerful tools available to date.

## trEST

trEST is an attempt to produce contigs from clusters of ESTs and to translate them into proteins. This is a three-step process:

(i) The ESTs are grouped into clusters that correspond to a single transcript. When available, the Unigene clusters (2) are used for that purpose, otherwise the clustering is performed using in-house software.

(ii) The ESTs of one cluster are assembled into one or several contigs with a script that makes use of the contig assembly programs Phrap and CAP (3). More than one contig is often produced from a single cluster and these contigs can be either disjoint or overlapping. In the latter case, they can either describe splice variants or reflect ambiguities in the contig assembly process.

(iii) Detection of the coding regions in the assembled contigs and translation of these regions into protein is perfomed by the program ESTscan (4) which corrects most frame shift errors and predicts their position with an error of a few amino acids. Benchmark experiments have indicated that ~95% of true coding regions longer than 30 amino acids are detected.

The trEST collection currently covers the following species: human, mouse, rat, *Drosophila melanogaster*, *Brachydanio rerio* and *Arabidopsis thaliana*.

## trGEN

The amino acid sequences of the trGEN database are predicted from High Throughput Genome (HTG) sequences and from genomic sequences of the non-HTG sections (HUM, ROD, INV, PLN) of the EMBL database. Entries under 10 000 bp are discarded. HTG sequences consisting of multiple unordered fragments are decomposed into individual sequences. Vectors and bacterial contaminants are then removed. The sequences are searched for putative genes and their coding regions with Genscan (5).

Although Genscan is one of the best gene prediction programs available, it is not foolproof, and it wrongly predicts a non-negligible fraction of all exons. While the majority of trGEN entries contain missing or extra exons, they usually also contain the correct predictions of a few contiguous exons. This often suffices for a particular protein domain to be recognized, if present. In this way trGEN entries provide links to genomic data from which a manual reconstruction of the gene can be undertaken.

trGEN is a highly redundant database that reflects the rapidly evolving situation prevalent in the HTG section of the EMBL

*To whom correspondence should be addressed at: Swiss Institute of Bioinformatics, Instititue of Experimental Cancer Research, Chemin des Boveresses 155, CH-1066 Epalinges, Switzerland. Tel: +41 21 692 59 91; Fax: +41 21 692 59 45; Email: philipp.bucher@isb-sib.ch

**Table 1.** The Hits database at the end of October 2000

| | SWISS-PROT (88 166) | TrEMBL (301 497) | TrEMBLnew (102 633) | trEST (165 758) | trGEN (501 714) |
|---|---|---|---|---|---|
| Prosite patterns | **36.8**% | **20.4**% | **27.4**% | **13.5**% | **8.7**% |
| (1304) | 74 465 | 130 516 | 46 913 | 39 834 | 95 558 |
| Prosite profiles | **28.0**% | **19.7**% | **24.3**% | **18.6**% | **13.9**% |
| (330) | 54 716 | 155 292 | 51 573 | 64 822 | 211 439 |
| Pfam HMMs | **56.1**% | **34.4**% | **51.3**% | **21.9**% | **13.9**% |
| (2216) | 94 973 | 190 537 | 88 541 | 64 928 | 161 337 |

The columns are the five collections of proteins, the rows are the three collections of protein-domain predictors, the figures in parentheses are the number of entries in each collection. In each cell, the percentage indicates the fraction of the protein sequences with at least one match by a predictor, the count below is the total number of significant matches. It is quite common that a protein is hit by more than one predictor as the three collections of predictors are partially redundant.

database. The trGEN collection currently covers the following species: human, *D.melanogaster*, mouse, rice and *A.thaliana*.

## HITS

Profile-based methods (6) and hidden Markov models (HMMs) (7) are currently the best techniques for detecting domains and other signatures, or motifs, in protein sequences. It is very expensive, in CPU cycles, to search a database for all proteins that match a given motif. To provide biologists with access to such a resource, a solution is to compute the matches of all predictors once and to make a database from these matches. Access to the data amounts to a simple lookup, which is very quick. This strategy is used, for example, by Pfam (8) and SMART (9); as well as by InterPro, a European project of a unified resource of protein domains and functional sites (10). Hits is an attempt to provide a comparable service for the two databases of hypothetical proteins presented here. Indeed, the updates of the Hits database require intensive computation and are mostly realized on dedicated hardware (GeneMatcher, Paracel). Hits currently includes a heterogenous collection of predictors, the Prosite collection of patterns and profiles (11) and the Pfam collection of HMMs (8).

The content of Hits at the end of October 2000 is summarized in Table 1. About half of the proteins of SWISS-PROT have a match by at least one predictor. The percentage of matched proteins decreases in TrEMBL and further in trEST and trGEN. This diminution does not equally affect the three collections of predictors: the Prosite patterns are selective predictors that primarily cover SWISS-PROT proteins from which they were designed. The collection of Pfam HMMs (2216 entries) is far larger than the collection of Prosite profiles (330 entries) and covers about twice the number of proteins in SWISS-PROT. But the performances of the two collections are comparable when considering trEST and trGEN. The decrease in coverage is partly due to the fragmentation of the protein sequences that happens to some extent in these databases. Indeed, if a long sequence with a single match is split into chunks, it is highly probable that only one chunk will retain the match, and that all the others will contribute to lessen the coverage. Another explanation for the diminution of the coverage concerns more specifically the Pfam collection of HMMs that includes many relatively long descriptors that are designed for automated annotation of full-length sequences and thus perform poorly on incomplete sequences.

## EXAMPLE

Figure 1 presents a diagonal plot of a sample protein (the neuropilin-2 precursor) versus two entries of trEST and trGEN that actually correspond to it. This summarizes the kinds of problems one has to deal with when using hypothetical proteins.

- The trEST prediction is globally correct but the boundaries of the coding region are only approximate: ESTscan is currently not capable of detecting translation start sites. The reconstructed sequence has an insertion near the C-terminus that corresponds to a known splice variant according to the SWISS-PROT entry.
- The prediction of the protein sequence in the trGEN entry contains several superfluous exons and the exon introduced at the C-terminus is completely wrong. Despite these errors, most of the protein sequence is retrieved and the two domains that form tandem repeats in the protein are clearly distinguishable in the reconstructed sequence. Indeed, these domains were correctly identified by the corresponding protein predictors and the sequence is easily retrieved using Hits.

One of the repeated motifs that occurs in the above example is a DS domain, which resembles the coagulation factor 5/8 type C repeat (12). At the end of October 2000, this domain was found in 75 entries in trGEN. The comparison of the sequence of these entries with those of the protein databases (see example in Fig. 2) indicated that at least six new proteins with a DS domain exist in the human genome.

## UPDATE TO THE DATABASES

The trEST and trGEN databases are updated weekly. The content of the databases appeared to evolve quite rapidly over the last months. This was primarily due to the rapid growth of the EMBL database, but also to improvements we made to the

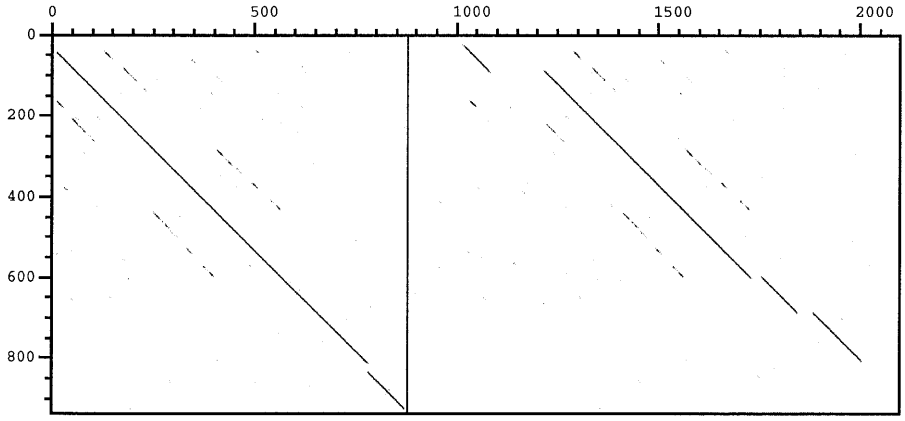


**Figure 1.** Diagonal plot of the neuropilin-2 precursor (O60462, vertical) versus one entry from trEST (Hs_17778_1, left horizontal) and another from trGEN (AC007362_1, right horizontal) that correspond to it. The two tandem repeats of the CUB and DS domains are clearly visible despite the errors in the reconstruction of the sequence (see text).

```
Query= tg:Z85999_1 Chromosome 6; Map 6q21; Clone RP1-94G16;
     (832 letters)
...
>tr|Z99572|043737 FACTOR V.[Homo sapiens]
            Length = 2224

 Score = 122 bits (302), Expect = 2e-26
 Identities = 71/168 (42%), Positives = 101/168 (59%)

GCSRSLSFE----PDGQIRASS---SWQSVNESGDQVHWSPGQARLQDQGPSWASGDSSN
GCS  L  E     + QI ASS   SW      GD  +W P +ARL  QG    A   +N
GCSTPLGMENGKIENKQITASSFKKSWW-----GD--YWEPFRARLNAQGRVNAWQAKAN

NHKPREWLEIDLGEKKKITGIRTTGSTQSNFNFYVKSFVMNFKNNNSKWKTYKGIVNNEE
N+K  +WLEIDL + KKIT I T G    +   YVKS+ +++     +WK Y+   +  +
NNK--QWLEIDLLKIKKITAIITQGCKSLSSEMYVKSYTIHYSEQGVEWKPYRLKSSMVD

KVFQGNSNFRDPVQNNFIPPIVARYVRVVPQTWHQRIALKVELIGCQI
K+F+GN+N +  V+N F PPI++R++RV+P+TW+Q IAL++EL GC I
KIFEGNTNTKGHVKNFFNPPIISRFIRVIPKTWNQSIALRLELFGCDI
```

**Figure 2.** Alignment of a newly discovered DS domain with closest relative in the protein databases.

algorithm used to produce the databases. We intend to pursue this effort and plan to add new species as soon as sufficient amounts of sequence are available.

The Hits database is updated on a monthly basis. The current development of the databases of protein domains is another factor that contributes to making the picture change very rapidly.

## ACCESS

### FTP

The files for the trEST, trGEN and Hits 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/hits.

### World Wide Web

Several web pages offer services that include the trEST, trGEN and Hits databases.

http://www.ch.embnet.org/software/fetch.html allows one to retrieve individual entries of trEST and trGEN.

http://www.ch.embnet.org/software/aBLAST.html allows the two databases of hypothetical proteins to be searched using BLAST.

http://hits.isb-sib.ch is the entry point of the web interface to the Hits database. Various integrated services are offered, which include several types of query forms, data-mining tools like SEView (13) and dotlet (14), links to other databases and online documentation.

## ACKNOWLEDGEMENTS

This work was partly supported by grant 3100-49669.96 of the Swiss National Science Foundation.

## REFERENCES

1. Jongeneel,C.V. (2000) Searching the expressed sequence tag (EST) databases: panning for genes. *Briefings in Bioinformatics*, **1**, 76–92.
2. Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, **75**, 694–698.
3. Huang,X. and Madan,A. (1999) CAP3: A DNA sequence assembly program. *Genome Res.*, **9**, 868–877.
4. Iseli,C., Jongeneel,C.V. and Bucher,P. (1999) ESTScan: A program for detecting, evaluating, and reconstructing potential coding regions in EST sequences, *Proc. 7th ISMB*, 138–148.
5. Burge,C.B. and Karlin,S. (1998) Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.*, **8**, 346–354.
6. Bucher,P., Karplus,K., Moeri,N. and Hofmann,K. (1996) A flexible motif search technique based on generalized profiles. *Comput. Chem.*, **20**, 3–24.
7. Krogh,A., Brown,M., Mian,I.S., Sjolander,K. and Haussler,D. (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.*, **235**, 1501–1531.
8. Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.
9. Schultz,J., Copley,R.R., Doerk,T., Ponting,C.P. and Bork,P. (2000) SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids. Res.*, **28**, 231–234.
10. Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti L., Corpet,F., Croning,M.D.R., Durbin,R., Falquet,L., Fleischmann,W., Gouzy,J., Hermjakob,H., Hulo,N., Jonassen,I., Kahn,D., Kanapin,A., Karavidopoulou,Y., Lopez,R., Marx,B., Mulder,N.J., Oinn,T.M., Pagni,M., Servant,F., Sigrist,C.J.A. and

Zdobnov,E.M. (2001) InterPro – An integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.
11. Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219.
12. Baumgartner,S., Hofmann,K., Chiquet-Ehrismann,R. and Bucher,P. (1998) The discoidin domain family revisited: new members from prokaryotes and a homology-based fold prediction. *Protein Sci.*, **7**, 1626–1631.
13. Junier,T. and Bucher,P. (1998) SEView: a java applet for browsing molecular sequence data. *In Silico Biol.*, **1**, 13–20.
14. Junier,T. and Pagni,M. (2000) Dotlet: diagonal plots in a web browser. *Bioinformatics*, **16**, 178–179.