EXProt: a database for proteins with an experimentally verified function

Björn M. Ursing[1,*], Frank H. J. van Enckevort[1], Jack A. M. Leunissen[1] and Roland J. Siezen[1,2]

[1]Centre for Molecular and Biomolecular Informatics (CMBI), University of Nijmegen, PO Box 9010, 6500 GL Nijmegen, The Netherlands and [2]NIZO Food Research, PO Box 20, 6710 BA Ede, The Netherlands

Received August 16, 2001; Revised and Accepted October 30, 2001

ABSTRACT

EXProt is a non-redundant protein database containing a selection of entries from genome annotation projects and public databases, aimed at including only proteins with an experimentally verified function. In EXProt release 2.0 we have collected entries from the *Pseudomonas aeruginosa* community annotation project (PseudoCAP), the *Escherichia coli* genome and proteome database (GenProtEC) and the translated coding sequences from the Prokaryotes division of EMBL nucleotide sequence database, which are described as having an experimentally verified function. Each entry in EXProt has a unique ID number and contains information about the species, amino acid sequence, functional annotation and, in most cases, links to references in MEDLINE/PubMed and to the entry in the original database. EXProt is indexed in SRS at CMBI (http://www.cmbi.kun.nl/srs/) and can be searched with BLAST and FASTA through the EXProt web page (http://www.cmbi.kun.nl/EXProt/).

INTRODUCTION

One of the main bottlenecks in genome projects is currently the annotation of the open reading frames (ORFs). Most functional annotation is based on similarity searches to other already annotated sequences. Experimental verification of the predicted functions is predominantly performed after publication of the genome sequence. They are most often first put on web pages of the genome projects and much later the EMBL/DDBJ/GenBank files are updated. Subsequently, updates are made in TrEMBL and SWISS-PROT.

In annotating ORFs with similarity searches, potential errors in assigning the correct function can lead to inherited mistakes from annotation to annotation (1). The information about which sequences have an experimentally verified function exists in some databases. For instance, in the EMBL nucleotide sequence database (2) there is a qualifier in the feature table (FT) '/evidence=EXPERIMENTAL' which indicates that a feature is experimentally verified. This feature could be anything from ribosomal binding site or splicing site to coding sequence. However, this piece of information is presently not preserved when the data is transferred to TrEMBL and SWISS-PROT (3). Domain databases as PROSITE (4), Pfam (5) and SMART (6) build their alignments on sequences from SWISS-PROT and TrEMBL. Subsequently, they all lack the information of the protein function being experimentally verified. In the *Pseudomonas aeruginosa* community annotation project (PseudoCAP) database (http://www.pseudomonas.com) (7), the basis for functional annotation of ORFs is put in confidence levels one to four, where confidence level one is 'Function experimentally demonstrated in *P.aeruginosa*' and confidence level four is 'Homologs of previously reported genes of unknown function, or no homology to any previously reported sequences'. In EXProt we combine protein sequences, which are stated to have an experimentally verified function, from different databases in order to provide a non-redundant database with proteins having an experimentally verified function (8).

DATABASE STRUCTURE

In a given protein database the information about how the function has been determined/predicted for a specific ORF is not always present. In the database of PseudoCAP (7) it is described which 375 genes have an experimentally verified function, all of these being included in EXProt. The *Escherichia coli* genome and proteome database (GenProtEC) (http://genprotec.mbl.edu) (9) is a database for genes in *E.coli* K-12 from which we have added 2031 selected amino acid sequences (6) to EXProt 2.0.

In the EMBL nucleotide sequence database, the translated amino acid sequence for the coding sequence (CDS) is in the FT. Coding sequences having the qualifier '/evidence=EXPERIMENTAL' in the FT were selected as entry in EXProt. In each EMBL entry there can be more than one coding sequence with the qualifier, and in these cases we made separate entries for each of the coding sequences. Any sequence from EMBL sequence database from the same organism and having an identical amino acid sequence with an entry from any of the other databases was considered redundant and was excluded. From EMBL nucleotide sequence database (Release 67, June 2001) 6405 sequences were included in EXProt 2.0.

*To whom correspondence should be addressed. Tel: +31 24 365 3379; Fax: +31 24 365 2977; Email: ursing@cmbi.kun.nl

Table 1. Number of entries from the different databases included in EXProt, where it could be read from the abstracts of referred papers that the protein function was experimentally verified (in abstracts)

Source database	Checked	In abstracts	Ratio (%)
Pseudomonas genome database (7)	38 / 375	26	69
GenProtEC (8)	60 / 2031	37	62
EMBL nucleotide sequence database (1)	102 / 3999	69	67
Total for EXProt	200 / 6405	132	66

QUALITY CONTROL

In EXProt we only collect data from other sources and we do not evaluate the entries ourselves. We trust that the experimental verification claimed by the authors of the databases or database entries is correct. In an attempt to estimate the reliability of the entries in EXProt we read the abstracts of the referred articles from 200 entries in EXProt. In 66% of the entries we could read in the abstracts that the function of the protein was indeed experimentally verified. This figure ranged between 62 and 69% in the underlying databases (see Table 1 for separate values). The ratio of experimentally verified entries is probably higher, but the figures give an indication that to fully trust an entry one has to check the literature.

DATABASE ACCESS

EXProt is indexed in SRS at CMBI (http://www.cmbi.kun.nl/srs/) with, for example, unique ID numbers, accession numbers from original database, gene name, gene description, EC number and sequence, and entries can be searched with keywords. The EXProt database can also be searched with BLAST (http://www.cmbi.kun.nl/bioinf/tools/blast.shtml) and FASTA (http://www.cmbi.kun.nl/bioinf/tools/fasta.shtml). The data files of EXProt can be downloaded from our FTP site at ftp://ftp.cmbi.kun.nl/pub/EXProt.

FUTURE PERSPECTIVES

We have started collaboration with more genome and topic specific databases in order to increase the number of entries. In addition, new methods of selecting only those proteins that have an experimentally verified function have to be developed. We will also follow the initiative from European Bioinformatics Institute (EBI) in entering tags on TrEMBL and SWISS-PROT entries indicating source of annotation (10).

ACKNOWLEDGEMENTS

Thanks to Dr Margrethe Hauge Serres from Marine Biology Laboratory in Woods Hole, MA, USA, for help in retrieving the entries from GenProtEC.

REFERENCES

1. Karp,P.D. (1998) What we do not know about sequence analysis and sequence databases. *Bioinformatics*, **17**, 753–754.
2. Stoesser,G., Baker,W., van den Broek,A., Camon,E., Garcia-Pastor,M., Kanz,C., Kulikova,T., Lombard,V., Lopez,R., Parkinson,H., Redaschi,N., Sterk,P., Stoehr,P. and Tuli,M.A. (2001) The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **29**, 17–21. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 21–26.
3. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
4. Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1991) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 235–238.
5. 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. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 276–280.
6. Schultz,J., Copley,R.R., Doerks,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. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 242–244.
7. Stover,C.K., Pham,X.Q., Erwin,A.L., Mizoguchi,S.D., Warrener,P., Hickey,M.J., Brinkman,F.S.L., Hufnagle,W.O., Kowalik,D.J., Lagrou,M. *et al.* (2000) Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature*, **406**, 959–964.
8. Ursing,B.M., van Enckevort,F.H.J., Leunissen,J.A.M. and Siezen,R.J. (2001) EXProt – a database for EXPerimentally verified Protein functions. *In Silico Biol.*, **2**, 0001. (http://www.bioinfo.de/isb/2001/02/0001/).
9. Riley,M. (1998) Genes and proteins of *Escherichia coli* K-12. *Nucleic Acids Res.*, **26**, 54.
10. Apweiler,R., Kersey,P., Kunker,V. and Bairoch,A. (2001) Technical comments to 'Database verification studies of SWISS-PROT and GenBank' by Karp *et al. Bioinformatics*, **17**, 533–534.