BIOINFORMATICS

Vol. 17 no. 6 2001
Pages 533–534

Technical comment to "Database verification studies of SWISS-PROT and GenBank" by Karp et al.

Rolf Apweiler[1,*], *Paul Kersey*[1], *Viv Junker*[1] *and Amos Bairoch*[2]

[1]*The EMBL Outstation—The European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK and* [2]*Swiss Institute of Bioinformatics, Centre Medical Universitaire, 1 rue Michel Servet, 1211 Geneva 4, Switzerland*

In their paper "Database verification studies of SWISS-PROT and GenBank" Karp *et al.* (2001) conclude: (1) "SWISS-PROT is more incomplete than we expected…"; (2) "Even if we combine SWISS-PROT and TrEMBL, some sequences from the full genomes are missing from the combined dataset"; (3) "In many cases, translated GenBank genes do not exactly match the corresponding SWISS-PROT sequences, …"; and (4) "…that SWISS-PROT does not identify a significant number of experimentally characterized proteins".

These results, and the approach used to arrive at these results, are in our opinion somewhat misleading. Herein, we only focus on four major points.

First, there has never been a claim that SWISS-PROT is comprehensive. Thus, it is surprising that Karp *et al.* found that "SWISS-PROT is more incomplete than we expected…". To make sequences available as quickly as possible without diluting the quality of SWISS-PROT, the supplemental database TrEMBL was introduced in 1996 and contains the translation of all coding sequences (CDS) in the DDBJ/EMBL/GenBank nucleotide sequence database, except those already included in SWISS-PROT. Snapshots of the SWISS-PROT, TrEMBL and TrEMBLnew databases are released weekly, synchronised with the DDBJ/EMBL/GenBank nucleotide sequence database and provide comprehensive coverage (ftp://ftp.ebi.ac.uk/pub/databases/sp_tr_nrdb/). The weekly comprehensive SWISS-PROT/TrEMBL nonredundant database (SPTR) has been widely publicised on the EBI and ExPASy web-servers and in various publications (e.g. Apweiler, 2000).

Second, the authors' assertions that "Even if we combine SWISS-PROT and TrEMBL, some sequences from the full genomes are missing from the combined dataset." and "SWISS-PROT curators apparently chose not to replace existing SWISS-PROT sequences with sequences from complete-genome projects" are rather inaccurate. Karp *et al.* tried to establish corresponding sets of SWISS-PROT/TrEMBL proteins and DDBJ/EMBL/GenBank coding sequence translations by sequence similarity searches between SWISS-PROT data from release 38, data from an unspecified TrEMBL release, and the data originally submitted to GenBank, which represents an outdated version of the genomic sequences.

This methodology is questionable, since changes to sequence, both in SWISS-PROT and in the nucleotide sequence databases, imply that sequence identity cannot be used for tracking entries between databases. For this reason, we use the 'Protein Sequence Identifier' to cross-reference with coding sequences in the nucleotide sequence databases. The specific format for cross-references from SWISS-PROT or TrEMBL to CDS in the DDBJ/EMBL/GenBank nucleotide sequence database is:

```
DR EMBL; ACCESSION_NR; PROTEIN_ID;
    STATUS_IDENTIFIER.
```

For example:

```
DR EMBL; AJ000012; CAA03857.1; -.
```

The secondary identifier is here the 'protein_id', which stands for the 'Protein Sequence Identifier'. It is a string, which is stored in a qualifier called '/protein_id' tagged to every CDS in the DDBJ/EMBL/GenBank nucleotide databases. For instance:

```
FT CDS 302..2674
FT /protein_id=''CAA03857.1''
FT /db_xref=''SWISS-PROT:P26345''
```

Use of these identifiers allows the identification of all proteins in SWISS-PROT and TrEMBL that correspond to coding sequences in a given completed genome sequence. In this way, up-to-date non-redundant protein sets are produced each week for each completed genome (Apweiler *et al.*, 2001; http://www.ebi.ac.uk/proteome/).

The reason these sets are produced weekly is that genome sequence data is frequently updated after the

*To whom correspondence should be addressed. Email: apweiler@ebi.ac.uk

original submission. SWISS-PROT and TrEMBL contain entries corresponding to every coding sequence in the current DDBJ/EMBL/GenBank genome sequence entries. For more details, see also http://www.ebi.ac.uk/proteome/CPhelp.html.

Third, while SWISS-PROT and TrEMBL are constantly updated to be synchronised with the nucleotide sequence databases, it seems that Karp *et al.* used for their comparisons the originally submitted data, which is now an outdated version of the genomic sequences. Apparently, these sequences have been updated a number of times since their first release. This implies the statements such as "Even if we combine SWISS-PROT and TrEMBL, some sequences from the full genomes are missing from the combined dataset." and "In many cases, translated GenBank genes do not exactly match the corresponding SWISS-PROT sequences, ..." refer to a comparison of different datasets.

Finally, our last remark refers to the statement "Contrary to claims by the SWISS-PROT authors, we conclude that SWISS-PROT does not identify a significant number of experimentally characterized proteins".

The assumption here is that the Description (DE) line in SWISS-PROT holds the information about the function of the protein. However, the function of a protein is described in the comment (CC) lines under the topic FUNCTION. For example:

```
CC -!- FUNCTION: PROFILIN PREVENTS THE
       POLYMERIZATION OF ACTIN.
```

SWISS-PROT states in its user manual the following as the definition of the DE line: "The DE (DEscription) lines contain general descriptive information about the sequence stored. This information is generally sufficient to identify the protein precisely".

Consequently, the DE line gives a name for a protein, which does not necessarily correspond to the protein function. For example, according to the SWISS-PROT user manual, the label hypothetical is used when "it is not known whether a sequence is actually translated into a protein"; this is a statement about existence, not function. Thus, using the DE line, it is not possible to accurately distinguish proteins whose functions were determined experimentally from entries with computationally determined functions. However, this information is stored elsewhere in the entry. Full details of how biochemical information is assigned to sequence entries is available at http://www.expasy.ch/cgi-bin/lists?annbioch.txt. This document is part of the extensive documentation we provide with SWISS-PROT (see http://www.expasy.ch/sprot/sp-docu.html).

We believe that we produce a high-quality database, although we are aware that the data in SWISS-PROT + TrEMBL are far from perfect. To further improve quality, we have introduced evidence tags to SWISS-PROT and TrEMBL. The aim of this is to allow users to identify the source of various data items and to enable SWISS-PROT staff to update data if the underlying evidence changes. This is an ongoing internal project (since July 2000) and we hope to provide a public version by the end of 2001. In TrEMBL release 16 (March 2001), 259 719 of 425 026 entries internally contain these tags. For more details, see ftp://ftp.ebi.ac.uk/pub/databases/trembl/evidenceDocumentation.html. We welcome feedback from the user community on the implementation of the evidence tags, as we appreciate the help of all scientists who are sending us update requests to help us to make the database even better.

REFERENCES

Apweiler,R. (2000) Protein sequence databases. In Richards,F.M., Eisenberg,D.S. and Kim,P.S. (eds), *Advances in Protein Chemistry.* Academic Press, New York, Vol. 54, pp. 31–71.

Apweiler,R., Biswas,M., Fleischmann,W., Kanapin,A., Karavidopoulou,Y., Kersey,P., Kriventseva,E.V., Mittard,V., Mulder,N., Phan,I. and Zdobnov,E. (2001) Proteome analysis database: online application of InterPro and CluSTr for the functional classification of proteins in whole genomes. *Nucleic Acids Res.*, **29**, 44–48.

Karp,P.D., Paley,S. and Zhu,J. (2001) Database verification studies of SWISS-PROT and GenBank. *Bioinformatics*, **17**, 526–532.