BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 17 2004, pages 3236–3237
doi:10.1093/bioinformatics/bth191

UniProt archive

Rasko Leinonen, Federico Garcia Diez, David Binns, Wolfgang Fleischmann, Rodrigo Lopez and Rolf Apweiler*

EMBL Outstation, The European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received on September 17, 2003; revised on January 28, 2004; accepted on February 19, 2004
Advance Access publication March 25, 2004

ABSTRACT

Summary: UniProt Archive (UniParc) is the most comprehensive, non-redundant protein sequence database available. Its protein sequences are retrieved from predominant, publicly accessible resources. All new and updated protein sequences are collected and loaded daily into UniParc for full coverage. To avoid redundancy, each unique sequence is stored only once with a stable protein identifier, which can be used later in UniParc to identify the same protein in all source databases. When proteins are loaded into the database, database cross-references are created to link them to the origins of the sequences. As a result, performing a sequence search against UniParc is equivalent to performing the same search against all databases cross-referenced by UniParc. UniParc contains only protein sequences and database cross-references; all other information must be retrieved from the source databases.
Availability: http://www.ebi.ac.uk/uniparc/
Contact: rolf.apweiler@ebi.ac.uk

INTRODUCTION

Universal Protein Knowledgebase (UniProt) is a central database of protein sequences and functions created by joining the information in SWISS-PROT (Boeckmann *et al*., 2003), TrEMBL (Boeckmann *et al*., 2003) and PIR-PSD (Wu *et al*., 2002). UniProt Archive (UniParc), as part of UniProt, is the most comprehensive, publicly accessible non-redundant protein sequence database available. New and updated protein sequences are loaded daily from public databases including SWISS-PROT, TrEMBL, PIR-PSD, EMBL (Stoesser *et al*., 2003), Ensembl (Hubbard *et al*., 2002), PDB (Westbrook *et al*., 2003), IPI (http://www.ebi.ac.uk/IPI), RefSeq (Pruitt and Maglott, 2001), FlyBase (The FlyBase Consortium, 2003), WormBase (Harris *et al*., 2003), European Patent Office proteins, United States Patent and Trademark Office proteins and Japan Patent Office proteins. A protein sequence may exist in multiple databases, and it is not unusual for the same sequence to be found several times in a single database. For example, the sequence in UniParc entry UPI0000000001 appeared four times in EMBL in August 2003. To avoid this redundancy, each unique sequence is stored only once and assigned a UniParc identifier. These identifiers are stable and, once created, are never deleted or reassigned. Consequently, UniParc identifiers can be used to uniquely identify protein sequences in any protein database. The format of UniParc identifiers is UPI followed by 10 hexadecimal numbers, e.g. UPI000000000A.

Proteins extracted from source databases are linked to their origins by using database cross-references. Each cross-reference links one protein in UniParc to an accession number in a source database. If the source database provides a sequence version for the protein, it is stored as a part of the cross-reference. Database cross-reference is active as long as the sequence identified by the accession number remains unchanged. When the sequence is modified or removed, the cross-reference retires. Active cross-references can be used to directly access the source databases, but retired cross-references can only be used to access sequence archives, such as the EMBL Sequence Version Archive (Leinonen *et al*., 2003). In practice, retired database cross-references would mostly be used to retrieve old sequence versions from UniParc. Because few databases have sequence versions, a UniParc sequence version number is made available as part of each database cross-reference. This number is incremented each time the sequence identified by accession number changes, and enables the use of sequence versions in all source databases.

UniParc contains only protein sequences, sequence versions and database cross-references. All other information concerning the sequences must be retrieved from the source databases using database cross-references. This is done, e.g. when displaying UniParc entries with SRS (Etzold *et al*., 1996) at http://srs.ebi.ac.uk/.

There were 5.6 million database cross-references and 1.8 million protein sequences in UniParc by October 2003. One quarter of these sequences had only one database cross-reference, whereas some sequences had several hundreds of them. For example, the sequence in UniParc entry UPI00000FE91D had 198 database cross-references to EMBL, TrEMBL, PIR and PDB. In this case, EMBL and TrEMBL contributed almost 90 database cross-references

*To whom correspondence should be addressed.

each, whereas PDB contributed 22 and PIR only two. In EMBL, as in many other databases, one protein sequence may appear multiple times in different contexts. TrEMBL inherits some of this sequence redundancy when it is derived from EMBL. In August 2003, EMBL had 1,535,482 protein sequences, but only 1,196,047 of them were unique. SWISS-PROT is an example of a database with a low level of sequence redundancy. It had 132,242 protein sequences in August 2003 and only 124 of them were redundant.

INTERACTIVE ACCESS

UniParc is available for text- and sequence-based searches. Performing a similarity search against UniParc is equivalent to performing the same search against all databases cross-referenced in UniParc, as UniParc contains all proteins from its source databases. Sequence similarity searches can be done using Fasta (Pearson and Lipman, 1988) at http://www.ebi.ac.uk/fasta/, Blast (Karlin and Altschul, 1990; Altschul *et al*., 1997) at http://www.ebi.ac.uk/blast/ or the Smith–Waterman algorithm (Smith and Waterman, 1981) at http://www.ebi.ac.uk/MPsrch/. Sequences, which are no longer part of any source database, are excluded from these searches.

Text-based SRS searches can be performed at http://srs.ebi.ac.uk/. UniParc identifiers can be used as search strings on the quick search page, when retrieving protein sequences. UniParc is available as a protein sequence database on the library page, where protein identifiers and sequence versions extracted from the source databases as well as UniParc identifiers, UniParc sequence versions, sequence lengths and sequence checksums can be used as search strings. The result of the searches is a list of matching UniParc identifiers, which are linked to UniParc entries. Each entry contains a protein sequence with its UniParc identifier, and one or more database cross-references. Database cross-references are hyperlinked to the source databases for easy access to the original resources. When displaying a UniParc entry, database cross-references are used to retrieve information from other databases, including the description and the name of the organism from SPTR (Boeckmann *et al*., 2003). SRS also provides an easy access to a large number of sequence analysis tools, including EMBOSS applications (Rice *et al*., 2000).

PROGRAMMATIC ACCESS

UniParc entries and Fasta formatted sequences can be retrieved programmatically using HTTP at http://www.ebi.ac.uk/cgi-bin/dbfetch. Either UniParc identifiers or protein identifiers extracted from the source databases can be used for retrieval. As an example, the following URL returns the UniParc entry having the UniParc identifier UPI0000000001:(http://www.ebi.ac.uk/cgi-bin/dbfetch?db=UNIPARC&id=UPI0000000001&format=default).

REFERENCES

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.

Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I., Pilbout,S. and Schneider,M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Etzold,T., Ulyanov,A. and Argos,P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.*, **266**, 114–128.

Harris,T.W., Lee,R., Schwarz,E., Bradnam,K., Lawson,D., Chen,W., Blasier,D., Kenny,E., Cunningham,F., Kishore,R., *et al*. (2003) WormBase: a cross-species database for comparative genomics. *Nucleic Acids Res.*, **31**, 133–137.

Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T., *et al*. (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.

Karlin,S. and Altschul,S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci., USA*, **87**, 2264–2268.

Leinonen,R., Nardone,F., Oyewole,O., Redaschi,R. and Stoehr,P. (2003) The EMBL sequence version archive. *Bioinformatics*, **19**, 1861–1862.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci., USA*, **85**, 2444–2448.

Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.

Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.

Smith,T.F. and Waterman,M.S. (1981) Identification of Common Molecular Subsequences. *J Mol Biol.*, **147**, 195–197.

Stoesser,G., Baker,W., van den Broek,A., Garcia-Pastor,M., Kanz,C., Kulikova,T., Leinonen,R., Lin,Q., Lombard,V., Lopez,R., *et al*. (2003) The EMBL Nucleotide Sequence Database: major new developments. *Nucleic Acids Res.*, **31**, 17–22.

The FlyBase Consortium (2003) The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Res.*, **31**, 172–175.

Westbrook,J., Feng,Z., Chen,L., Yang,H. and Berman,H.M. (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Res.*, **31**, 489–491.

Wu,C.H., Huang,H., Arminski,L., Castro-Alvear,J., Chen,Y., Hu,Z., Ledley,R.S., Lewis,K.C., Mewes,H., Orcutt,B.C., *et al*. (2002) The protein information resource: an integrated public resource of functional annotation of proteins. *Nucleic Acids Res.*, **30**, 35–37.