UniSave: the UniProtKB Sequence/Annotation Version database

Rasko Leinonen, Francesco Nardone, Weimin Zhu and Rolf Apweiler*

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

Received on February 10, 2006; revised and accepted on March 17, 2006
Advance Access publication March 21, 2006
Associate Editor: Dmitrij Frishman

ABSTRACT
Summary: The UniProtKB Sequence/Annotation Version database (UniSave) is a comprehensive archive of UniProtKB/Swiss-Prot and UniProtKB/TrEMBL entry versions. All changed Swiss-Prot and TrEMBL entries are loaded into the UniSave as part of the public bi-weekly UniProtKB releases. Unlike the UniProtKB, which contains only the latest Swiss-Prot and TrEMBL entry versions, the UniSave provides access to previous versions of these entries.
Availability: http://www.ebi.ac.uk/uniprot/unisave
Contact: rolf.apweiler@ebi.ac.uk

INTRODUCTION
The Universal Protein Resource (UniProt) combines the activities of Swiss-Prot, TrEMBL, and Protein Information Resource (PIR) databases (Bairoch et al., 2005). The UniProt Knowledgebase (UniProtKB), the central part of UniProt, consists of the UniProtKB/Swiss-Prot and UniProtKB/TrEMBL databases. Swiss-Prot entries are manually curated to the highest standard, and the TrEMBL entries are annotated using powerful automated annotation, classification and cross-referencing algorithms.

The Swiss-Prot and TrEMBL entries are subject to changes, but only the most recent versions are preserved in the UniProtKB. However, access to previous entry versions may be highly desirable, especially when references to entries are made from journal articles. Entries in UniProtKB/Swiss-Prot and UniProtKB/TrEMBL go through numerous annotation changes, become secondary to other entries (are replaced) or are removed from UniProtKB without replacement (are deleted). Because of constant annotation improvements, the original annotation may only be accessible by having access to earlier entry versions.

In this article we will describe UniSave, a new publicly available service, which provides interactive and programmatic access to all versions of Swiss-Prot and TrEMBL entries. It is similar to the EMBL sequence version archive (Leinonen et al., 2003) and complements the UniProt Archive (Leinonen et al., 2004), which is the world’s most comprehensive protein sequence repository.

CONTENT OF UNISAVE
All new and updated UniProtKB/Swiss-Prot and UniProtKB/TrEMBL entries are distributed to the public in bi-weekly releases. These entries are made accessible from UniSave shortly after they are made public as part of the UniProtKB releases. All obtainable entry versions, starting from the ninth Swiss-Prot release in November 1988, and from the first TrEMBL release in November 1996, are available through UniSave. By the UniProtKB release 7.0 in February 2006, there were 27 539 591 and 5 071 382, respectively, different entry versions for TrEMBL and Swiss-Prot, respectively.

ENTRY STORAGE
The entry versions are stored in an Oracle database. To minimize space consumption changed entry versions are not stored as a whole, but are compared against previous entry versions using the Hunt–Szymanski algorithm (Hunt and Szymanski, 1977). If the entry differential is smaller in size then the original entry, only the differential is stored in the database. When the entry is being unloaded from the archive, the entry version is reconstructed by applying the differential to the original entry. If the entry differential is larger than the original entry, then the entry is stored as a whole, and subsequent versions will be compared against the new version.

As a result, entry differentials are only ever applied to one whole entry. To further reduce storage requirements, the entries and their differentials are compressed in 16 kb blocks using zlib (http://www.zlib.org). This increases compressibility of the entries by introducing more redundancy in each compressed unit.

PROGRAMMATIC ACCESS
UniProtKB/Swiss-Prot and UniProtKB/TrEMBL entries and Fasta formatted sequences can be retrieved programmatically using dbfetch (HTTP GET protocol) at http://www.ebi.ac.uk/cgi-bin/dbfetch, using UniSave/Batch (HTTP POST protocol) at http://www.ebi.ac.uk/uniprot/unisave?&do_batch=1 or SOAP at http://www.ebi.ac.uk/umiprot/unisave/unisave.wsdl. Up to 200 and 10 000 entries can be downloaded using dbfetch and UniSave/Batch, respectively, by providing a list of primary accession numbers or entry names. As an example, the following URL returns all UniProtKB entry versions with accession number Q00001 using

*To whom correspondence should be addressed.
dbfetch: (http://www.ebi.ac.uk/cgi-bin/dbfetch?db=UniSave&id=Q00001&format=default&style=raw). The \( n \)-th entry version is returned by id=Q00001, and the latest entry version by id=Q00001. A more fine-grained access is provided through the SOAP service, which is designed to support rich interactive clients.

INTERACTIVE ACCESS

UniProtKB entries and Fasta formatted sequences can be viewed and downloaded interactively at http://www.ebi.ac.uk/uniprot/unisave. Entries can be retrieved using primary accession numbers or entry names. The first result of a query is a list of matching entry versions together with the UniProtKB database name, entry status, primary accession number, entry name, entry version, sequence version, release and the release date (Fig. 1). The matches are ordered by the release date, the latest version first. If a snapshot date is provided then only the version of the entry that was current at that date is displayed. The entry version status is either incorporated, active, changed, replaced or deleted. An incorporated entry version is the first entry version added into UniProtKB, an active entry version is part of the latest public release, a changed entry version has been superseded by a newer entry version, a replaced entry has become secondary to another entry and a deleted entry has been removed from the UniProtKB without becoming secondary to any other entry. For replaced entry versions, the status ‘Replaced’ can be clicked to return all entries, which have the given entry as a secondary entry. Comparison between entry versions is straightforward by selecting two entries and clicking the ‘Compare Selected’ button. Whenever comparisons are made a Smith–Waterman sequence alignment is computed using SSEARCH (Pearson and Lipman, 1988), and displayed at the bottom of the entry.

ACCESS FROM SRS AT EBI

The interactive web client at http://www.ebi.ac.uk/uniprot/unisave is also accessible from SRS at http://srs.ebi.ac.uk by following links provided with UniProtKB query results.

ACKNOWLEDGEMENTS

The authors thank Allyson Williams and Daniel Barrell for help with old Swiss-Prot and TrEMBL releases, Maria-Jesus Martin, Claire O’Donovan, Elisabeth Gasteiger, Nicole Redaschi, Isabelle Phan, Raja Mazumder, Baris Suzek, Darren Natale and Eric Jain, for their suggestions for the web client, Quan Lin and Andrey Stinov for their contribution to the bi-weekly UniSave production, Mike Donnelly for database support, Alberto Labarga for web support and Mikael Andersson for dbfetch integration. Funding to pay the Open Access publication charges for this article was provided by the authors.

Conflict of Interest: none declared.

REFERENCES


