Does UniProtKB contain all protein sequences?

The two sections of UniProtKB - UniProtKB/Swiss-Prot and UniProtKB/TrEMBL - give access to **most protein sequences which are available to the public**. However, UniProtKB excludes the following protein sequences:

1. Most non-germline **immunoglobulins** and T-cell receptors
2. Synthetic sequences
3. Most patent application sequences
4. Small fragments encoded from nucleotide sequence (<8 amino acids)
5. Pseudogenes
6. Sequences from **redundant proteomes**
7. Sequences from proteomes that NCBI genomes/RefSeq considers to be **low quality assemblies** , i.e. **excluded proteomes**
8. Fusion/truncated proteins
9. Not real proteins

The first 5 are identified automatically by the UniProtKB/TrEMBL creation program and never enter UniProtKB. However some proteins belonging to these classes are also identified during the UniProtKB/Swiss-Prot annotation process by the curators and then removed from UniProtKB.

Protein sequences originating from proteomes that are considered **redundant** or **low quality** are identified automatically at every release, and are either removed from UniProtKB/TrEMBL, or never enter UniProtKB but stay in **UniParc**.

Fusion/truncated proteins and those classified as not real proteins are only manually identified by the curators and removed from UniProtKB/TrEMBL or UniProtKB/Swiss-Prot. All these excluded sequences are available at **UniParc**. The corresponding UniParc entries have been flagged with the reason for the absence of that sequence from UniProtKB.

See also

- **Where do the UniProtKB protein sequences come from?**
- **What are UniProtKB's criteria for defining a CDS as a protein?**
- **Why have some UniProtKB accession numbers been deleted? How can I track them?**
- **Why do we keep dubious sequences in UniProtKB? How to discard them from a protein set?**
- **Sequences**

Page last modified: Wed Apr 27 2022

Related article

How redundant are the UniProt databases?

Do you use data resources managed by EMBL-EBI? Please spare a few minutes to tell us about your experience.

Bayer EX1075
Bayer v. Corteva - IPR2023-01037

Do you use data resources managed by EMBL-EBI? Please spare a few minutes to tell us about your experience.

https://www.surveymonkey.com/r/HJKYKTT?channel=uniprot

Next

© 2002 – 2024 UniProt consortium

Bayer EX1075
Bayer v. Corteva - IPR2023-01037