FASTA headers

The following is a description of FASTA headers for UniProtKB (including alternative isoforms), UniRef, UniParc (including proteomes) and archived UniProtKB versions. NCBI's program formatdb (in particular its -o option) is compatible with the UniProtKB FASTA headers.

Note that in the document below square brackets indicate optional fields. [ ]

Table of contents

UniProtKB

```
>db|UniqueIdentifier|EntryName ProteinName OS=OrganismName OX=OrganismIdentifi
```

Where:

- *db* is 'sp' for UniProtKB/Swiss-Prot and 'tr' for UniProtKB/TrEMBL.
- *UniqueIdentifier* is the primary **accession number** of the UniProtKB entry.
- *EntryName* is the **entry name** of the UniProtKB entry.
- *ProteinName* is the **recommended name** of the UniProtKB entry as annotated in the field. For UniProtKB/TrEMBL entries without a field, the field is used. In case of multiple , the first one is used. The 'precursor' attribute is excluded, 'Fragment' is included with the name if applicable.`RecNameRecNameSubNameSubNames`
- *OrganismName* is the **scientific name of the source organism** of the UniProtKB entry.
- *OrganismIdentifier* is the **unique identifier of the source organism, assigned by the NCBI**.
- *GeneName* is the first **gene name** of the UniProtKB entry. If there is no gene name, or , the field is not listed.`OrderedLocusNameORFnameGN`
- *ProteinExistence* is the numerical value describing the **evidence for the existence** of the protein.
- *SequenceVersion* is the **version number** of the sequence.

Examples:

```
>sp|Q8I6R7|ACN2_ACAGO Acanthoscurrin-2 (Fragment) OS=Acanthoscurria gomesiana
>sp|P27748|ACOX_CUPNH Acetoin catabolism protein X OS=Cupriavidus necator (str
>sp|P04224|HA22_MOUSE H-2 class II histocompatibility antigen, E-K alpha chain

>tr|Q3SA23|Q3SA23_9HIV1 Protein Nef (Fragment) OS=Human immunodeficiency virus
>tr|Q8N2H2|Q8N2H2_HUMAN cDNA FLJ90785 fis, clone THYRO1001457, moderately simi
```

UniProtKB alternative isoforms

This format only applies to UniProtKB/Swiss-Prot alternative isoform sequences.

```
>sp|IsoID|EntryName Isoform IsoformName of ProteinName OS=OrganismName OX=Orga
```

Where:

- *IsoID* is the isoform identifier as assigned in the **ALTERNATIVE PRODUCTS** section of the UniProtKB entry.
- *IsoformName* is the isoform name as annotated in the ALTERNATIVE PRODUCTS field of the UniProtKB entry.Name

ProteinExistence and *SequenceVersion* do not apply to alternative isoforms (*ProteinExistence* is dependent on the number of cDNA sequences, which is not known for individual isoforms).

Example:

```
>sp|Q4R572-2|1433B_MACFA Isoform Short of 14-3-3 protein beta/alpha OS=Macaca
```

UniRef

```
>UniqueIdentifier ClusterName n=Members Tax=TaxonName TaxID=TaxonIdentifier Re
```

Where:

- *UniqueIdentifier* is the unique identifier of the UniRef cluster.
- *ClusterName* is the name of the UniRef cluster.
- *Members* is the number of UniRef cluster members.
- *TaxonName* is the scientific name of the lowest common taxon shared by all UniRef cluster members.
- *TaxonIdentifier* is the NCBI taxonomy identifier of the lowest common taxon shared by all UniRef cluster members.
- *RepresentativeMember* is the entry name of the representative member of the UniRef cluster.

Example:

```
>UniRef50_Q9K794 Putative AgrB-like protein n=2 Tax=Bacillus TaxID=1386 RepID=
```

UniParc

```
>UniqueIdentifier status=Status
```

Where:

- *UniqueIdentifier* is the unique identifier of the UniParc entry.
- *Status* is 'active' if the UniParc entry has at least one active cross-reference, and 'inactive' if it does not have any active cross-references.

Example:

```
>UPI0000000005 status=active
```

UniParc proteomes

This format is only available for UniParc entry sets that correspond to the sequences of a proteome. It contains biological information from the UniParc source database entries that are associated with the requested proteome.

```
>UniqueIdentifier[ ProteinNameList] OS=OrganismName OX=OrganismIdentifier[
GN=GeneNameList][ AC=UniProtKBAcList][ SS=SourceIdList]
PC=Proteome:ComponentList
```

Where:

- *UniqueIdentifier* is the unique identifier of the UniParc entry.
- *ProteinNameList* is an optional list of protein names.
- *OrganismName* is the scientific name of the proteome's source organism.
- *OrganismIdentifier* is the **unique identifier of the proteome's source organism, assigned by the NCBI**.
- *GeneNameList* is an optional list of gene names.
- *UniProtKBAcList* is an optional list of the UniProtKB primary accession numbers that are linked to the proteome.
- *SourceIdList* is an optional list of the 'Source Sequence' identifiers that are linked to the proteome. These are the identifiers of the proteins that are annotated on the associated genome records from EMBL, Ensembl or RefSeq.
- *Proteome:ComponentList* is the identifier of the proteome, separated by ':' from a list of the components where the genes are encoded.

Lists are allowed in several fields to represent paraloguous genes that encode the same sequence. List elements are separated by the pipe symbol '|'.

The protein and gene names come from the corresponding UniProtKB entries, where these exist, and otherwise from the 'Source Sequence' database entries.

Examples:

Sequence that exists in UniProtKB: protein and gene name come from the UniProtKB entry (AC=P37636):

```
>UPI000013B286 Multidrug resistance protein MdtE OS=Escherichia coli (strain K
```

Sequence that does not exist in UniProtKB: protein and gene name come from the 'Source Sequence' entry (SS=EMBL:CQR80801):

```
>UPI000000E135 Oligopeptide transporter subunit OS=Escherichia coli (strain K1
◄ ▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬ ▶
```

Sequence that corresponds to 3 paraloguous genes that are encoded on 3 different chromosomes:

```
>UPI00000000C1 Calmodulin-1|Calmodulin-2|Calmodulin-3 OS=Homo sapiens OX=9606
◄ ▬▬▬▬▬▬▬▬ ▶
```

Archived UniProtKB sequence versions

```
>db|UniqueIdentifier archived from Release ReleaseNumber ReleaseDate SV=Sequen
◄ ▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬ ▶
```

Where:

- *db* is 'sp' for UniProtKB/Swiss-Prot and 'tr' for UniProtKB/TrEMBL.
- *UniqueIdentifier* is the primary accession number of the UniProtKB entry.
- *ReleaseNumber* refers to the release from which the sequence was archived (Swiss-Prot or TrEMBL release numbers for releases prior to the first UniProt release, and both UniProt and Swiss-Prot or TrEMBL release numbers for releases after the first UniProt release).
- *ReleaseDate* is the date of the release form which the sequence was archived.
- *SequenceVersion* is the version number of the sequence.

Examples:

'pre-UniProt':

```
>sp|P05067 archived from Release 18.0 01-MAY-1991 SV=3
>tr|Q55167 archived from Release 17.0 01-JUN-2001 SV=1
```

'post-UniProt':

```
>sp|P05067 archived from Release 9.2/51.2 28-NOV-2006 SV=3
>tr|A0RTJ8 archived from Release 11.0/36.0 29-MAY-2007 SV=1
```

Related terms: FASTA header, FASTA format, FASTA comment

Page last modified: Thu Jun 12 2025

Related articles

What is the canonical sequence? Are all isoforms described in one entry?

How to retrieve sets of protein sequences?

UniProtKB advanced search options

Programmatic access - Format conversion

Programmatic access - Retrieving entries via queries

Load more articles

UniProt

EMBL-EBI · PIR · SIB

Get in touch ✉

UniProt is an ELIXIR core data resource

UniProt is a GBC global core biodata resource

GLOBAL CORE BIODATA RESOURCE

Core data	Supporting data	Tools	Information
Proteins (UniProtKB)	Literature citations	BLAST	Cite UniProt"
Species (Proteomes)	Taxonomy	Align	About & Help
Protein clusters (UniRef)	Keywords	Retrieve/ID mapping	UniProtKB manual
Sequence archive (UniParc)	Subcellular locations	Peptide search	Technical corner
	Cross-referenced databases	Tool results	Expert biocuration
	Diseases		Statistics

Main **funding** by: **National Institutes of Health** EMBL-EBI · State Secretariat for Education, Research and Innovation SERI