survey_uniref

Daniel Vogel

2022-10-12

R Markdown

This document checks the contents of the uniref download and reports the number of organisms, proteins, lines, etc.

```
# UniProtKB Fasta headers
#
# Example:
# >db|UniqueIdentifier|EntryName ProteinName OS=OrganismName OX=OrganismIdentifier [GN=GeneName ]PE=Pro
#
# Actual Examples from a UniProt Fasta file.  Note that there are | delimiters, positional, and named c
# >sp|O65039|CYSEP_RICCO Vignain OS=Ricinus communis OX=3988 GN=CYSEP PE=1 SV=1
# >sp|B9RK42|GPC1_RICCO Glycerophosphocholine acyltransferase 1 OS=Ricinus communis OX=3988 GN=GPC1 PE=
# >sp|B9RU15|ATXR5_RICCO Probable Histone-lysine N-methyltransferase ATXR5 OS=Ricinus communis OX=3988
#
# Where:
#
# db is 'sp' for UniProtKB/Swiss-Prot and 'tr' for UniProtKB/TrEMBL.
# | separator
# UniqueIdentifier is the primary accession number of the UniProtKB entry.
# | separator
# The rest of the values are all together space separated
# EntryName is the entry name of the UniProtKB entry.
# ProteinName is the recommended name of the UniProtKB entry as annotated in the RecName field.
#   For UniProtKB/TrEMBL entries without a RecName field, the SubName field is used. In case of multipl
#   the first one is used. The 'precursor' attribute is excluded, 'Fragment' is included with the name
# OrganismName is the scientific name of the 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, OrderedLocusName or
# ProteinExistence is the numerical value describing the evidence for the existence of the protein.
# SequenceVersion is the version number of the sequence.



# read a variable with all of the headers from uniref50 so we can determine how many organisms
uniref <-readRDS( "/nbacc/uniref50_names.rds")
head( uniref )
```

```
## [1] "UniRef50_A0A5A9P0L4 Peptidylprolyl isomerase n=1 Tax=Triplophysa tibetana TaxID=1572043 RepID=A
## [2] "UniRef50_A0A410P257 Glycogen synthase n=2 Tax=Candidatus Velamenicoccus archaeovorus TaxID=19305
## [3] "UniRef50_A0A8J3NBY6 Uncharacterized protein n=2 Tax=Actinocatenispora rupis TaxID=519421 RepID=A
```

```
## [4] "UniRef50_Q8WZ42 Titin n=2871 Tax=Vertebrata TaxID=7742 RepID=TITIN_HUMAN"
## [5] "UniRef50_A0A401TRQ8 Uncharacterized protein (Fragment) n=2 Tax=Chiloscyllium TaxID=34767 RepID=A
## [6] "UniRef50_A0A6J2WDG0 titin n=196 Tax=cellular organisms TaxID=131567 RepID=A0A6J2WDG0_CHACN"
```

```
#
# UniRef fasta fields
# >UniqueIdentifier ClusterName n=Members Tax=TaxonName TaxID=TaxonIdentifier RepID=RepresentativeMembe
# Where:
#
# UniqueIdentifier is the primary accession number 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 clust
# RepresentativeMember is the entry name of the representative member of the UniRef cluster.
# e.g.
#"UniRef50_A0A5A9P0L4 Peptidylprolyl isomerase n=1 Tax=Triplophysa tibetana TaxID=1572043 RepID=A0A5A9P
#"UniRef50_A0A410P257 Glycogen synthase n=2 Tax=Candidatus Velamenicoccus archaeovorus TaxID=1930593 Re
#"UniRef50_A0A8J3NBY6 Uncharacterized protein n=2 Tax=Actinocatenispora rupis TaxID=519421 RepID=A0A8J3
#"UniRef50_Q8WZ42 Titin n=2871 Tax=Vertebrata TaxID=7742 RepID=TITIN_HUMAN"
#"UniRef50_A0A401TRQ8 Uncharacterized protein (Fragment) n=2 Tax=Chiloscyllium TaxID=34767 RepID=A0A401
#"UniRef50_A0A6J2WDG0 titin n=196 Tax=cellular organisms TaxID=131567 RepID=A0A6J2WDG0_CHACN"
# ....
# Sanity check to verify that all of these fields are contained in the uniref database fasta for each e
# Tax=      found 53625855 times in uniref50
# TaxID=    found 53625855 times in uniref50
# RepID=    found 53625855 times in uniref50
# UniRef50_ found 53625855 times in uniref50
# n=        found 53625855 times in uniref50
#
# let's try to pull the TaxID out to determine how many different organism

#Split twice?
#TaxID <-strsplit( uniref,"TaxID=")
#sapply(TaxID,"[[",2)
# Better option is to use RegEx

# Determine how many indivual organisms in uniref50
# TaxID <- stringr::str_match( uniref, "TaxID=([0-9][0-9]+)")
# TaxID_table <- table( TaxID[,2])
# nrow( TaxID_table )
# 118887 unique organisms in uniref50
# saveRDS( TaxID_table, "/nbacc/uniprot/taxid_table.RDS")
```