Package 'UniProt.ws'

October 13, 2020

Type Package

Title R Interface to UniProt Web Services

Version 2.28.0

Depends methods, utils, RSQLite, RCurl, BiocGenerics (>= 0.13.8)

Imports AnnotationDbi, BiocFileCache, rappdirs

Suggests RUnit, BiocStyle, knitr

Description A collection of functions for retrieving, processing and repackaging the UniProt web services.

Collate AllGenerics.R AllClasses.R getFunctions.R methods-select.R utilities.R

License Artistic License 2.0

biocViews Annotation, Infrastructure, GO, KEGG, BioCarta

VignetteBuilder knitr

LazyLoad yes

git_url https://git.bioconductor.org/packages/UniProt.ws

git_branch RELEASE_3_11

git_last_commit 499a943

git_last_commit_date 2020-04-27

Date/Publication 2020-10-12

Author Marc Carlson [aut],
Csaba Ortutay [ctb],
Bioconductor Package Maintainer [aut, cre]

Maintainer Bioconductor Package Maintainer <maintainer@bioconductor.org>

R topics documented:

UniProt.ws-objects *UniProt.ws objects and their related methods and functions*

Description

UniProt.ws is the base class for interacting with the Uniprot web services from Bioconductor.

In much the same way as an AnnotationDb object allows acces to select for many other annotation packages, UniProt.ws is meant to allow usage of select methods and other supporting methods to enable the easy extraction of data from the Uniprot web services.

select, columns and keys are used together to extract data via an UniProt.ws object.

columns shows which kinds of data can be returned for the UniProt.ws object.

keytypes allows the user to discover which keytypes can be passed in to select or keys via the keytype argument.

keys returns keys for the database contained in the UniProt.ws object . By default it will return the primary keys for the database, which are UNIPROTKB keys, but if used with the keytype argument, it will return the keys from that keytype.

select will retrieve the data as a data.frame based on parameters for selected keys and columns and keytype arguments.

The UniProt.ws will be loaded whenever you load the UniProt.ws package. This object will be set up to retrieve information from Homo sapiens by default, but this value can be changed to any of the species supported by Uniprot. The species and taxId methods allow users to see what species is currently being accessed, and taxId<- allows them to change this value.

species shows the genus and species label currently attached to the UniProt.ws objects database.

taxId shows the NCBI taxonomy ID currently attached to the AnnotationDb objects database. Using the equivalently names replace method (taxId<-) allows the user to change the taxon ID, and the species represented along with it.

availableUniprotSpecies is a helper function to list out the available Species along with their official taxonomy IDs that are available by Uniprot. Because there are so many species represented at UniProt, there is also a pattern argument that can be used to restrict the range of things returned to be only those whose species names match the searth term. Please remember when using this argument that the Genus is always capitalized and the species never is.

lookupUniprotSpeciesFromTaxId is another helper that will look up the species of any tax ID that is supported by Uniprot.

Usage

```
columns(x)
keytypes(x)
select(x, keys, columns, keytype, ...)
species(object)
taxId(x)

availableUniprotSpecies(pattern, n=Inf)
lookupUniprotSpeciesFromTaxId(taxId)
UniProt.ws(taxId, ...)
```

Arguments

x	the UniProt.ws object.
object	the UniProt.ws object.
keys	the keys to select records for from the database. All possible keys are returned by using the keys method.
columns	the columns or kinds of things that can be retrieved from the database. As with keys, all possible columns are returned by using the columns method.
keytype	the keytype that matches the keys used. For the select methods, this is used to indicate the kind of ID being used with the keys argument. For the keys method this is used to indicate which kind of keys are desired from keys
pattern	A string passed in to limit the results
n	the maximim number of results to return.
taxId	a taxonomy id
...	other arguments

Value

keys,columns,keytypes, species and lookupUniprotSpeciesFromTaxId each return a character vector of possible values.

taxId returns a numeric value that corresponds to the taxonomy ID.

select and availableUniprotSpecies each return a data.frame.

Author(s)

Marc Carlson

See Also

select

Examples

```
## Make a UniProt.ws object
up <- UniProt.ws(taxId=9606)

## look at the object
up

## get the current species
species(up)

## look up available species with their tax ids
availableUniprotSpecies("musculus")

## get the current taxId
taxId(up)

## look up the species that goes with a tax id
lookupUniprotSpeciesFromTaxId(9606)

## set the taxId to something else
```

```
taxId(up) <- 10090
up

## list the possible key types
head(keytypes(up))

## list the columns that can be retreived
head(columns(up))

## list all possible keys of type entrez gene ID.
## (this process is not instantaneous)
if(interactive()){
  egs = keys(up, "ENTREZ_GENE")
}

## use select to extract some data
res <- select(up,
              keys = c("22627","22629"),
              columns = c("PDB","HGNC","SEQUENCE"),
              keytype = "ENTREZ_GENE")
head(res)
```

UNIPROTKB *Descriptions of available values for* `columns` *and* `keytypes`.

Description

This manual page enumerates the kinds of data represented by the values returned when the user calls `columns` or `keytypes`

Details

All the possible values for `columns` and `keytypes` are listed below. Users will have to actually use these methods to learn which of the following possible values actually apply in their case.

:UNIPROTKB The central ID for UniProt and swissprot

:UNIPARC UniParc

:UNIREF50 UniRef50

:UNIREF90 UniRef90

:UNIREF100 UniRef100

:EMBL/GENBANK/DDBJ EMBL/GenBank/DDBJ

:EMBL/GENBANK/DDBJ_CDS EMBL/GenBank/DDBJ CDS

:PIR PIR

:ENTREZ_GENE Entrez Gene (GeneID)

:GI_NUMBER* GI number

:IPI IPI

:REFSEQ_PROTEIN RefSeq Protein

:REFSEQ_NUCLEOTIDE RefSeq Nucleotide

:PDB PDB

:DISPROT DisProt
:HSSP HSSP
:DIP DIP
:MINT MINT
:ALLERGOME Allergome
:MEROPS MEROPS
:PEROXIBASE PeroxiBase
:PPTASEDB PptaseDB
:REBASE REBASE
:TCDB TCDB
:PHOSSITE PhosSite
:DMDM DMDM
:AARHUS/GHENT-2DPAGE Aarhus/Ghent-2DPAGE
:ECO2DBASE ECO2DBASE
:WORLD-2DPAGE World-2DPAGE
:DNASU DNASU
:ENSEMBL Ensembl
:ENSEMBL_PROTEIN Ensembl Protein
:ENSEMBL_TRANSCRIPT Ensembl Transcript
:ENSEMBL_GENOMES Ensembl Genomes
:ENSEMBL_GENOMES PROTEIN Ensembl Genomes Protein
:ENSEMBL_GENOMES TRANSCRIPT Ensembl Genomes Transcript
:KEGG KEGG
:PATRIC PATRIC
:TIGR TIGR
:UCSC UCSC
:VECTORBASE VectorBase
:AGD AGD
:ARACHNOSERVER ArachnoServer
:CGD CGD
:CONOSERVER ConoServer
:CYGD CYGD
:DICTYBASE dictyBase
:ECHOBASE EchoBASE
:ECOGENE EcoGene
:EUHCVDB euHCVdb
:EUPATHDB EuPathDB
:FLYBASE FlyBase
:GENECARDS GeneCards
:GENEFARM GeneFarm

:GENOLIST GenoList
:H-INVDB H-InvDB
:HGNC HGNC
:HPA HPA
:LEGIOLIST LegioList
:LEPROMA Leproma
:MAIZEGDB MaizeGDB
:MIM MIM
:MGI MGI
:NEXTPROT neXtProt
:ORPHANET Orphanet
:PHARMGKB PharmGKB
:POMBASE PomBase
:PSEUDOCAP PseudoCAP
:RGD RGD
:SGD SGD
:TAIR TAIR
:TUBERCULIST TubercuList
:WORMBASE WormBase
:WORMBASE_TRANSCRIPT WormBase Transcript
:WORMBASE_PROTEIN WormBase Protein
:XENBASE Xenbase
:ZFIN ZFIN
:EGGNOG eggNOG
:GENETREE GeneTree
:HOVERGEN HOVERGEN
:KO KO
:OMA OMA
:ORTHODB OrthoDB
:PROTCLUSTDB ProtClustDB
:BIOCYC BioCyc
:REACTOME Reactome
:UNIPATHWAY UniPathWay
:CLEANEX CleanEx
:GERMONLINE GermOnline
:DRUGBANK DrugBank
:GENOMERNAI GenomeRNAi
:NEXTBIO NextBio
:CITATION citations
:CLUSTERS clusters

:COMMENTS comments

:DOMAINS domains

:DOMAIN domain

:EC ec ID

:ID ID

:EXISTENCE existence

:FAMILIES families

:FEATURES features

:GENES genes

:GO go term

:GO-ID go id

:INTERPRO interpro

:INTERACTOR interactor

:KEYWORDS keywords

:KEYWORD-ID keyword-id

:LAST-MODIFIED last-modified

:LENGTH length

:ORGANISM organism

:ORGANISM-ID organism-id

:PATHWAY pathway

:PROTEIN NAMES protein names

:REVIEWED reviewed

:SCORE score

:SEQUENCE sequence

:3D 3d

:TAXON taxon

:TOOLS tools

:VERSION version

:DATABASE(PFAM) PFAM ids

:DATABASE(PDB) PDB ids

:

Author(s)

Marc Carlson

Examples

```
up <- UniProt.ws(taxId=9606)
## List the possible values for columns
columns(up)
## List the possible values for keytypes
keytypes(up)
## get some values back
## list all possible keys of type entrez gene ID.
## (this process is not instantaneous)
if(interactive()){
  keys <- head(keys(up, keytype="UNIPROTKB"))
  keys
}
select(up, keys=c("P31946","P62258"), columns=c("PDB","SEQUENCE"),
keytype="UNIPROTKB")
```

`utilities` *Utility functions*

Description

UniProt uses custom coding of organism names from which protein sequences they store. These taxon names are used also in the protein names (not in the UniProt IDs!). These functions help to translate those names to standard scientific (Latin) taxon names and other useful identifiers.

- taxname2species(): converts UniProt taxonomy names to scientific species names
- taxname2taxid(): converts UniProt taxonomy names to NCBI Taxonomy IDs
- taxname2domain(): converts UniProt taxonomy names to the following taxonomical domains: 'A' for archaea (=archaebacteria)\ 'B' for bacteria (=prokaryota or eubacteria)\ 'E' for eukaryota (=eukarya)\ 'V' for viruses and phages (=viridae)\ 'O' for others (such as artificial sequences)\
- updatespecfile(): The updatespecfile helper function attempts to download the current version of the controlled vocabulary of species table from UniProt controlled vocabulary of species. If it fails to download, an archived version of the table in (in extdata/) will be used.

Usage

```
taxname2species(taxname, specfile)
taxname2taxid(taxname, specfile)
taxname2domain(taxname, specfile)
updatespecfile()
```

Arguments

`taxname`	Character string up to 6 uppercase characters, like HUMAN, MOUSE, or AERPX. Also works for a vector of such taxon names.
`specfile`	An optional local file where speclist.RData is saved from UniProt.org. When `specfile` is missing, a cached file from the extdata/ package directory is used.

Value

Function `taxname2species` returns a character vector of scientific taxon names matching to the UniProt taxon names supplied as `taxname`.

Function `taxname2taxid` returns a numeric vector of Taxonomy IDs matching to the UniProt taxon names supplied as `taxname`.

Function `taxname2domain` returns a character vector of one letter domain symbols matching to the UniProt taxon names supplied as `taxname`.

Function `updatespecfile` returns a file location where specfile.txt is downloaded. If the download fails, path is the location of the archived version in the package.

Author(s)

Csaba Ortutay

See Also

UniProt controlled vocabulary of species, which defines the taxon names.

Examples

```
taxname2species("PIG")
taxname2species(c("PIG","HUMAN","TRIHA"))

taxname2taxid("PIG")
taxname2taxid(c("PIG","HUMAN","TRIHA"))

taxname2domain("PIG")
taxname2domain(c("PIG","HUMAN","TRIHA"))

newspecfile <- updatespecfile()
taxname2domain("PIG", specfile = newspecfile)
taxname2domain(c("PIG","HUMAN","TRIHA"), specfile = newspecfile)
```

Index