UniProt.ws: A package for retrieving data from the UniProt web service

Marc Carlson

April 26, 2022

1 Configuring `uniport.ws`

The *UniProt.ws* package provides a `select` interface to the UniProt web service.

```
suppressPackageStartupMessages({
    library(UniProt.ws)
})
up <- UniProt.ws(taxId=9606)
```

If you already know about the select interface, you can immediately learn about the various methods for this object by just looking it's the help page.

```
help("UniProt.ws")
```

When you load the *UniProt.ws* package, it creates a `UniProt.ws` object. If you look at the object you will see some helpful information about it.

```
up
```

```
## "UniProt.ws" object:
## An interface object for UniProt web services
## Current Taxonomy ID:
## 9606
## Current Species name:
## Homo sapiens
## To change Species see: help('availableUniprotSpecies')
```

By default, you can see that the `UniProt.ws` object is set to retrieve records from Homo sapiens. But you can change that of course. In order to change it, you first need to look up the appropriate taxonomy ID for the species that you are interested in. Uniprot provides support for over 20 thousand species, so there are a few to choose from! In order to make this easier, we have provided the helper function `availableUniprotSpecies` which will list all the supported species along with their taxonomy ids. When you call the `availableUniprotSpecies` function, it's recommended that you make use of the pattern argument to limit your queries like this:

```
availableUniprotSpecies(pattern="musculus")
```

```
##    taxon ID               Species name
## 1    520121         Anthocoris musculus
## 2    208057        Anthoscopus musculus
## 3    238007             Apomys musculus
```

```
## 4    213557            Baiomys musculus
## 5      9771       Balaenoptera musculus
## 6    197864        Blepharisma musculus
## 7     10090                Mus musculus
## 8     35531     Mus musculus bactrianus
## 9     10091      Mus musculus castaneus
## 10    57486     Mus musculus molossinus
## 11  1891730 Mus musculus polyomavirus 1
```

Once you have learned the taxonomy ID for the species of interest, you can then change the taxonomy id for the `UniProt.ws` object using `taxId` setter or by calling the constructor for `UniProt.ws`

```
mouseUp <- UniProt.ws(10090)
mouseUp
```

```
## "UniProt.ws" object:
## An interface object for UniProt web services
## Current Taxonomy ID:
## 10090
## Current Species name:
## Mus musculus
## To change Species see: help('availableUniprotSpecies')
```

As you can see the species is different for the `mouseUp` new object.

2 Using `UniProt.ws`

Once you are safisfied that you have an `uniport.ws` that is using the appropriate organsims, you can make use of the standard set of methods in a `select` interface. Specifically: `columns`, `keytypes`, `keys` and `select`.

You will probably notice that there are a large number of columns that can be retrieved.

```
head(keytypes(up))
```

```
## [1] "AARHUS/GHENT-2DPAGE" "AGD"                 "ALLERGOME"
## [4] "ARACHNOSERVER"       "BIOCYC"              "CGD"
```

And most (but not all) of these fields can also be used as keytypes.

```
head(columns(up))
```

```
## [1] "3D"                  "AARHUS/GHENT-2DPAGE" "AGD"
## [4] "ALLERGOME"           "ARACHNOSERVER"       "BIOCYC"
```

If necessary you can also look up the keys of a given type. But please be warned that the web service is slow at this particular kind of lookup. So if you really want to do this kind of operation you are probably going to want to save the result to your R session.

```
egs = keys(up, "ENTREZ_GENE")
```

Finally, you can loop up whatever combinations of columns, keytypes and keys that you need when using `select`.

```
keys <- c("1","2")
columns <- c("PDB","HGNC","SEQUENCE")
kt <- "ENTREZ_GENE"
res <- select(up, keys, columns, kt)
```

```
## Getting mapping data for 1 ...  and ACC
```

```
## Getting mapping data for P04217 ...  and PDB_ID
```

```
## Getting mapping data for P04217 ...  and HGNC_ID
```

```
## Getting extra data for P04217, V9HWD8, P01023
```

```
## 'select()' returned 1:many mapping between keys and columns
```

```
res
```

```
##   ENTREZ_GENE  PDB   HGNC
## 1           1 <NA> HGNC:5
## 2           1 <NA>   <NA>
## 3           2 1BV8 HGNC:7
## 4           2 2P9R HGNC:7
## 5           2 4ACQ HGNC:7
## 6           2 6TAV HGNC:7
##
## 1
## 2
## 3 MGKNKLLHPSLVLLLLVLLPTDASVSGKPQYMVLVPSLLHTETTEKGCVLLSYLNETVTVSASLESVRGNRSLFTDLEAENDVLHCVAFAVPKSSSNEEVMFL
## 4 MGKNKLLHPSLVLLLLVLLPTDASVSGKPQYMVLVPSLLHTETTEKGCVLLSYLNETVTVSASLESVRGNRSLFTDLEAENDVLHCVAFAVPKSSSNEEVMFL
## 5 MGKNKLLHPSLVLLLLVLLPTDASVSGKPQYMVLVPSLLHTETTEKGCVLLSYLNETVTVSASLESVRGNRSLFTDLEAENDVLHCVAFAVPKSSSNEEVMFL
## 6 MGKNKLLHPSLVLLLLVLLPTDASVSGKPQYMVLVPSLLHTETTEKGCVLLSYLNETVTVSASLESVRGNRSLFTDLEAENDVLHCVAFAVPKSSSNEEVMFL
```

sessionInfo()

```
sessionInfo()
```

```
## R version 4.2.0 RC (2022-04-19 r82224)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 20.04.4 LTS
##
## Matrix products: default
## BLAS:   /home/biocbuild/bbs-3.15-bioc/R/lib/libRblas.so
## LAPACK: /home/biocbuild/bbs-3.15-bioc/R/lib/libRlapack.so
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_GB              LC_COLLATE=C
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
```

```
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] UniProt.ws_2.36.0   BiocGenerics_0.42.0 RSQLite_2.2.12
##
## loaded via a namespace (and not attached):
##  [1] KEGGREST_1.36.0        tidyselect_1.1.2       xfun_0.30
##  [4] purrr_0.3.4            vctrs_0.4.1            generics_0.1.2
##  [7] htmltools_0.5.2        stats4_4.2.0           BiocFileCache_2.4.0
## [10] yaml_2.3.5             utf8_1.2.2             blob_1.2.3
## [13] rlang_1.0.2            pillar_1.7.0           withr_2.5.0
## [16] glue_1.6.2             DBI_1.1.2              rappdirs_0.3.3
## [19] bit64_4.0.5            dbplyr_2.1.1           GenomeInfoDbData_1.2.8
## [22] lifecycle_1.0.1        stringr_1.4.0          zlibbioc_1.42.0
## [25] Biostrings_2.64.0      memoise_2.0.1          evaluate_0.15
## [28] Biobase_2.56.0         knitr_1.38             IRanges_2.30.0
## [31] fastmap_1.1.0          GenomeInfoDb_1.32.0    curl_4.3.2
## [34] AnnotationDbi_1.58.0   fansi_1.0.3            highr_0.9
## [37] Rcpp_1.0.8.3           filelock_1.0.2         BiocManager_1.30.17
## [40] cachem_1.0.6           S4Vectors_0.34.0       XVector_0.36.0
## [43] bit_4.0.4              BiocStyle_2.24.0       png_0.1-7
## [46] digest_0.6.29          stringi_1.7.6          dplyr_1.0.8
## [49] cli_3.3.0              tools_4.2.0            bitops_1.0-7
## [52] magrittr_2.0.3         RCurl_1.98-1.6         tibble_3.1.6
## [55] crayon_1.5.1           pkgconfig_2.0.3        ellipsis_0.3.2
## [58] assertthat_0.2.1       rmarkdown_2.14         httr_1.4.2
## [61] R6_2.5.1               compiler_4.2.0
```