Package 'geneHummus'

April 5, 2019

Title A Pipeline to Define Gene Families in Legumes and Beyond

Version 1.0.11

Description A pipeline with high specificity and sensitivity in extracting proteins from the RefSeq database (National Center for Biotechnology Information). Manual identification of gene families is highly time-consuming and laborious, requiring an iterative process of manual and computational analysis to identify members of a given family. The pipelines implements an automatic approach for the identification of gene families based on the conserved domains that specifically define that family. See Die et al. (2018) <doi:10.1101/436659> for more information and examples.

Depends R (>= 3.4.0)

License MIT + file LICENSE

Encoding UTF-8

LazyData true

Imports rentrez (>= 1.2.1), stringr (>= 1.4.0), dplyr (>= 0.8.0.1), httr (>= 1.4.0), utils, curl (>= 3.3)

Suggests knitr, rmarkdown

URL https://github.com/NCBI-Hackathons/GeneHummus

BugReports https://github.com/NCBI-Hackathons/GeneHummus/issues

RoxygenNote 6.1.1

NeedsCompilation no

Author Jose V. Die [aut, cre] (<https://orcid.org/0000-0002-7506-8590>), Moamen M. Elmassry [ctb], Kimberly H. LeBlanc [ctb], Olaitan I. Awe [ctb], Allissa Dillman [ctb], Ben Busby [aut]

Maintainer Jose V. Die <jose.die@uco.es>

Repository CRAN

Date/Publication 2019-04-04 22:30:03 UTC

R topics documented:

`accessions_by_spp` *Compute the total number of accession proteins per species*

Description

Summarizes a dataframe of protein ids and return the total number of accessions per organism.

Usage

```
accessions_by_spp(my_accessions)
```

Arguments

`my_accessions`	A data frame with accession protein ids and organisms

Value

A `data.frame` of summarized results including columns:

- organism, taxonomic species
- N.seqs, total number of sequences

Author(s)

Jose V. Die

See Also

`getAccessions` to create the data frame with acession id and organism for each protein identifier.

Examples

```
my_prots = data.frame(accession = c("XP_014620925", "XP_003546066",
   "XP_025640041", "XP_019453956", "XP_006584791", "XP_020212415",
   "XP_017436622", "XP_004503803", "XP_019463844"),
   organism =  c("Glycine max", "Glycine max", "Arachis hypogaea",
   "Lupinus angustifolius", "Glycine max", "Cajanus cajan",
   "Vigna angularis", "Cicer arietinum", "Lupinus angustifolius"))

accessions_by_spp(my_prots)
```

`accessions_from_spp` *Extract the accession ids (XP accession) for a given organism*

Description

Filter a dataframe of protein ids and return the accessions for a given species or organism.

Usage

```
accessions_from_spp(my_accessions, spp)
```

Arguments

`my_accessions`	A data frame with accession protein ids and organisms
`spp`	A string with the scientific name of the species or organism.

Value

A string vector with protein accession (XP accession, RefSeq database)

Author(s)

Jose V. Die

See Also

`getAccessions` to create the data frame with accession id and organism for each protein identifier.

Examples

```
my_prots = data.frame(accession = c("XP_014620925", "XP_003546066",
   "XP_025640041", "XP_019453956", "XP_006584791", "XP_020212415",
   "XP_017436622", "XP_004503803", "XP_019463844"),
   organism =  c("Glycine max", "Glycine max", "Arachis hypogaea",
   "Lupinus angustifolius", "Glycine max", "Cajanus cajan",
   "Vigna angularis", "Cicer arietinum", "Lupinus angustifolius"))

accessions_from_spp(my_prots, "Glycine max")
```

`accessions_warning` *Get acessions and organism for each protein identifier*

Description

Core function used by `getAccessions`.

Usage

```
accessions_warning(protein_ids)
```

Arguments

`protein_ids`	A string vector containing protein identifiers.

Author(s)

Jose V. Die

`archids_warning` *Get architecture identifiers for the conserved domains*

Description

Parses SPARCLE database (NCBI) and extract electronic identifiers for each conserved domain.

Usage

```
archids_warning(gene_family)
```

Arguments

`gene_family`	A string with conserved domain(s).

Author(s)

Jose V. Die

extract_proteins *Get the protein identifiers*

Description

Extract the protein identifier for the given taxonomic species, which are hosted by the RefSeq database (NCBI).

Usage

```
extract_proteins(targets, taxonIds)
```

Arguments

targets	A string with the electronic links for the SPARCLE architecture.
taxonIds	A string with the taxonomic species identifiers; legume species (by default).

Details

First, get the protein ids from RefSeq database. Then, extract only the ids for the selected taxonomic species (by default, legume species).

Author(s)

Jose V. Die

filterarchids_warning *Filter protein architectures based on conserved domains*

Description

Parse the architecture identifiers and extract those that contain, at least, the conserved domaind selected as filter.

Usage

```
filterarchids_warning(archs_ids, filter)
```

Arguments

archs_ids	A string with the architecture identifiers.
filter	A string with the domains as filter.

Author(s)

Jose V. Die

`filterArch_ids` *Filter the protein architectures based on conserved domains*

Description

Parse the architecture identifiers and extract those that contain, at least, those selected in the filter.

Usage

```
filterArch_ids(archs_ids, filter)
```

Arguments

`archs_ids`	A string with the architecture identifiers that contain, at least, one of the conserved domains defining the gene family.
`filter`	A string with the domains (and order) that are required (at least) for the proteins to have.

Value

the architecture identifiers from all the potential protein architectures defined by `getArch_ids` that contain, at least, the conserved domains explicitily show by the filter.

Author(s)

Jose V. Die

See Also

`getArch_ids`

Examples

```
## Not run:
archs_ids <- getArch_ids("pfam02362")
my_filter <- c("B3_DNA", "Auxin_resp")

filterArch_ids(archs_ids, my_filter)

## End(Not run)
```

geneHummus *genehummus: A pipeline to define gene families in Legumes and beyond*

Description

genehummus is a pipeline with high specificity and sensitivity in extracting proteins from the RefSeq database (NCBI).

Author(s)

Jose V. Die <jose.die@uco.es>, Moamen M. Elmassry, Kimberly H. LeBlanc, Olaitan I. Awe, Allissa Dillman, Ben Busby

See Also

see the preprint in BioRvix

getAccessions *Get the acessions ids and the organism for each protein identifier*

Description

The `getAccessions` function parses the protein page for each identifier and extracts the accession id (usually referred as XP accession in the RefSeq database) and the organism given by the scientific name.

The `accessions_by_spp` and `accessions_from_spp` functions are convenient filters for further cleaning of `getAccessions` by giving the total number of XP accessions per species or extracting the XP accessions for a given species, respectively.

Usage

```
getAccessions(protein_ids)
```

Arguments

`protein_ids`	A string vector containing protein identifiers.

Value

A `data.frame` of protein ids including columns:

- accession
- organism

Author(s)

Jose V. Die

See Also

accessions_by_spp to summarize the total number of accession proteins per species.

accessions_from_spp to filter the accession ids for a given species

Examples

```
prot_ids <- c("593705262", "1379669790", "357520645",  "1150156484")
getAccessions(prot_ids)
```

---

getArch_ids *Get the potential architecture identifiers for the conserved domains*

Description

Parses the SPARCLE database (NCBI) and extract the electronic identifiers for each conserved domain.

Usage

```
getArch_ids(gene_family)
```

Arguments

gene_family	A string with the conserved domain(s) defining the gene family. The domains have to be shown in the same order appearing in the sequences.

Value

the architectures identifiers for each of the conserved domains.

Author(s)

Jose V. Die

Examples

```
arf <- c("pfam06507")
getArch_ids(arf)
```

getArch_labels *Get the description label for a protein architecture identifier*

Description

Parses the architecture identifiers and extract their corresponding labels.

Usage

```
getArch_labels(arch_ids)
```

Arguments

arch_ids	A string with the architecture electronic identifiers.

Value

print out the description label for the candidate architectures that contain the proteins we are looking for.

Author(s)

Jose V. Die

See Also

filterArch_ids

Examples

```
filtered_archids <- c("12034188", "12034184")
getArch_labels(filtered_archids)
```

getProteins_from_tax_ids

Get the RefSeq protein identifiers for the given taxonomic species

Description

Parse the RefSeq database using protein architecture identifiers (SPARCLE dabatse) and extract the protein ids. for the selected taxonomic species.

Usage

```
getProteins_from_tax_ids(arch_ids, taxonIds)
```

Arguments

arch_ids	A string with the electronic links for the SPARCLE.
taxonIds	A vector string with taxonomy ids; Legume species available in RefSeq, by default.

Value

RefSeq protein identifiers for selected species.

Author(s)

Jose V. Die

Examples

```
filtered_archids <- c("12034184")
medicago <- c(3880)
getProteins_from_tax_ids(filtered_archids, medicago)
```

getProtlinks

Get the protein identifiers for a given architecture

Description

Parse the RefSeq database and extract all the protein identifiers that have a given architecture.

Usage

```
getProtlinks(archs_ids)
```

Arguments

archs_ids	A string with the architecture identifiers (SPARCLE database, NCBI)

Author(s)

Jose V. Die

---

getSparcleArchs *Get the electronic architecture for a conserved domain*

Description

Parses the SPARCLE database (NCBI) and extract the electronic links for a given conserved domain.

Usage

```
getSparcleArchs(CD)
```

Arguments

CD	A string with the conserved domain(s)

Author(s)

Jose V. Die

---

get_spp *Get the species name from the description sequence*

Description

Parse a string vector with sequence descriptions (title and species) and extract the species name.

Usage

```
get_spp(description)
```

Arguments

description	A string vector with the sequence description (title and species).

Value

for each sequence description, extract the species name.

Author(s)

Jose V. Die

labels_warning *Get description label for a protein architecture identifier*

Description

Parses the architecture identifier and extract the corresponding labels.

Usage

```
labels_warning(arch_ids)
```

Arguments

arch_ids	A string with the architecture electronic identifiers.

Author(s)

Jose V. Die

legumesIds *NCBI taxonomy ids for the legume family*

Description

Taxonomy identifier for about 10,000 legume species

Usage

```
data(legumesIds)
```

Format

a numeric vector with 10.009 elements

Source

Taxonomy identifiers for Fabaceae species (Taxonomy databse, NCBI).

my_legumes *ARF proteins per legume specie*

Description

XP accessions for the Auxin Response Factor gene family in the Legume Legume taxonomy (NCBI RefSeq database, as of SEP 2018).

Usage

```
data(my_legumes)
```

Format

a list of length 10 with 563 elements.

- 1. chickpea
- 2. medicago
- 3. soybean
- 4. arachis_duranensis
- 5. arachis_ipaensis
- 6. cajanus
- 7. vigna_angulata
- 8. vigna_radiata
- 9. phaseolus
- 10. lupinus

Source

Protein identifiers for Fabaceae species (RefSeq databse, NCBI).

proteins_warning *Get RefSeq protein identifiers for the given taxonomic species*

Description

Parse the RefSeq database using protein architecture identifiers and extract protein ids. for selected taxonomic species. Core function used by `getProteins_from_tax_ids`.

Usage

```
proteins_warning(arch_ids, taxonIds)
```

Arguments

arch_ids	A string with the electronic links for the SPARCLE.
taxonIds	A vector string with taxonomy ids.

Author(s)

Jose V. Die

---

sizeIds *Build a list containing N elements per element list*

Description

Split a vector into N elements, so that each element contains a given length.

Usage

```
sizeIds
```

Format

An object of class `numeric` of length 1.

Author(s)

Jose V. Die

Index