Package 'refseqR'

October 30, 2024

Type Package

Title Common Computational Operations Working with RefSeq Entries (GenBank)

Version 1.1.5

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

Description Fetches NCBI data (RefSeq <https://www.ncbi.nlm.nih.gov/refseq/> database) and provides an environment to extract information at the level of gene, mRNA or protein accessions.

License MIT + file LICENSE

URL https://github.com/jdieramon/refseqR

BugReports https://github.com/jdieramon/refseqR/issues

Encoding UTF-8

Imports IRanges, rentrez, tibble, Biostrings

RoxygenNote 7.2.3

Suggests knitr, rmarkdown

VignetteBuilder knitr

NeedsCompilation no

Author Jose V. Die [aut, cre] (<https://orcid.org/0000-0002-7506-8590>), Lluís Revilla Sancho [ctb] (<https://orcid.org/0000-0001-9747-2570>)

Repository CRAN

Date/Publication 2024-10-30 00:50:01 UTC

Contents

refseqR 2
refseq_AAlen 2
refseq_AAlen_action 3
refseq_AAmol_wt 4
refseq_AAseq 5
refseq_CDScoords 6

refseq_CDSseq 7
refseq_description 8
refseq_fromGene 9
refseq_fromGene_action 10
refseq_GeneID 10
refseq_geneSymbol 11
refseq_geneSymbol_action 12
refseq_mRNAfeat 13
refseq_protein2RNA 14
refseq_RNA2protein 15

Index **16**

`refseqR` *refseqR: Common computational operations working with RefSeq*

Description

refseqR is a framework of common computational operations working with RefSeq entries (GenBank)

Author(s)

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

See Also

Useful links:

- https://github.com/jdieramon/refseqR
- Report bugs at https://github.com/jdieramon/refseqR/issues

`refseq_AAlen` *Get the amino acid length from a protein accession*

Description

`refseq_AAlen()` Returns the amino acid length from a single protein accession.

Depending on the function, available accessions in `refseqR` include RefSeq models with the prefixes XM_ (mRNA), XR_ (non-coding RNA), and XP_ (protein), as well as subsequently curated RefSeq records with NM_, NR_, or NP_ accession prefixes.

Usage

```
refseq_AAlen(protein)
```

Arguments

protein	A character string of the protein id.

Value

A numeric value representing the aa length of the protein.

Author(s)

Jose V. Die

See Also

refseq_RNA2protein to obtain the protein ids encoded by a set of transcript ids.

Examples

```
# Get the amino acid lengths from a set of protein accessions
protein = c("XP_004487758", "XP_004488550")
sapply(protein, function(x) refseq_AAlen(x), USE.NAMES = FALSE)
```

---

refseq_AAlen_action *Get the amino acid length from a protein accession*

Description

refseq_AA_len_action() Returns the amino acid length from a single protein accession.

Depending on the function, available accessions in refseqR include RefSeq models with the prefixes XM_ (mRNA), XR_ (non-coding RNA), and XP_ (protein), as well as subsequently curated RefSeq records with NM_, NR_, or NP_ accession prefixes.

Usage

```
refseq_AAlen_action(protein, retries)
```

Arguments

protein	A character string of the protein id.
retries	A numeric value to control the number of retry attempts to handle internet errors.

Value

A numeric value representing the aa length of the protein.

Author(s)

Jose V. Die

See Also

refseq_RNA2protein to obtain the protein ids encoded by a set of transcript ids.

---

refseq_AAmol_wt *Extract the molecular weight from a protein accession*

Description

refseq_AAmol_wt() Parses a protein accession output (RefSeq format) and extract the molecular weight (in Daltons).

Depending on the function, available accessions in refseqR include RefSeq models with the prefixes XM_ (mRNA), XR_ (non-coding RNA), and XP_ (protein), as well as subsequently curated RefSeq records with NM_, NR_, or NP_ accession prefixes.

Usage

```
refseq_AAmol_wt(protein)
```

Arguments

protein	A character string of the protein id.

Details

First, get the character vector containing the fetched record. Then, this function parses the fetched record and returns the molecular weight.

Value

A numeric vector representing the molecular weight of the protein.

Author(s)

Jose V. Die

See Also

refseq_RNA2protein to obtain the protein ids encoded by a set of transcript ids.

Examples

```
# Get the molecular weight from a single protein accession
protein <- "XP_020244413"
refseq_AAmol_wt(protein)


# Get the molecular weight from from a set of protein accessions
protein = c("XP_004487758", "XP_004488550")
```

```
sapply(protein, function(x) refseq_AAmol_wt(x), USE.NAMES = TRUE)
```

---

refseq_AAseq *Extract the amino acid sequence into a Biostrings object*

Description

refseq_AAseq() Parses a single/multiple protein accessions (RefSeq format) and extract the amino acid sequence(s) into a AAStringSet object.

Depending on the function, available accessions in refseqR include RefSeq models with the prefixes XM_ (mRNA), XR_ (non-coding RNA), and XP_ (protein), as well as subsequently curated RefSeq records with NM_, NR_, or NP_ accession prefixes.

Usage

```
refseq_AAseq(accession)
```

Arguments

accession	A character string containing a single/multiple accession ids.

Value

An object of AAStringSet class.

Author(s)

Jose V. Die

Examples

```
accession = c("XP_004487758", "XP_004488550", "XP_004501961")
my_aa <- refseq_AAseq(accession)
# Now, the `AAStringSet`can be easily used to make a fasta file :
# writeXStringSet(x= my_aa, filepath = "aa_result")
```

`refseq_CDScoords` *Extract the coding sequences (CDS) coordinates from a transcript accession*

Description

`refseq_CDScoords()` Parses a transcript accession (RefSeq format) and extract the CDS coordinates. The CDS coordinates refer to the mRNA molecule.

Depending on the function, available accessions in `refseqR` include RefSeq models with the prefixes XM_ (mRNA), XR_ (non-coding RNA), and XP_ (protein), as well as subsequently curated RefSeq records with NM_, NR_, or NP_ accession prefixes.

Usage

```
refseq_CDScoords(transcript)
```

Arguments

`transcript`	A character string of the single/multiple transcript id.

Value

An `IRanges` object with the start and end position of the CDS of the putative mRNAs.

Author(s)

Jose V. Die

See Also

`refseq_CDSseq`

Examples

```
transcript = c("XM_004487701")
refseq_CDScoords(transcript)


transcript = c("XM_004487701", "XM_004488493")
refseq_CDScoords(transcript)
```

refseq_CDSseq *Extract the CDS nucleotide sequence into a Biostrings object*

Description

refseq_CDSseq() Parses a single/multiple transcript accessions (RefSeq format) and extract the CDS nucleotide sequences into a DNAStringSet object.

Depending on the function, available accessions in refseqR include RefSeq models with the prefixes XM_ (mRNA), XR_ (non-coding RNA), and XP_ (protein), as well as subsequently curated RefSeq records with NM_, NR_, or NP_ accession prefixes.

Usage

```
refseq_CDSseq(transcript)
```

Arguments

transcript	A character string of the single/multiple transcript id.

Value

An object of DNAStringSet class.

Author(s)

Jose V. Die

See Also

refseq_CDScoords

Examples

```
transcript <-  c("XM_004487701", "XM_004488493", "XM_004501904")
my_cds <- refseq_CDSseq(transcript)
# Now, the `DNAStringSet` can easily used to make a fasta file :
# writeXStringSet(x= my_cds, filepath = "cds_result")
```

`refseq_description` *Get the sequence Description*

Description

`refseq_description()` Returns the sequence description from a single transcript, protein, or GeneID accession.

Depending on the function, available accessions in `refseqR` include RefSeq models with the prefixes transcript_ (mRNA), XR_ (non-coding RNA), and XP_ (protein), as well as subsequently curated RefSeq records with NM_, NR_, or NP_ accession prefixes.

Usage

```
refseq_description(id)
```

Arguments

`id`	A character string of the transcript, protein, or GeneID accession.

Value

A character vector containing the sequence description corresponding to the specified sequence as `id`.

Author(s)

Jose V. Die

See Also

`refseq_protein2RNA` to obtain the transcript ids that encode a set of protein ids.

`refseq_RNA2protein` to obtain the protein ids encoded by a set of transcript ids.

Examples

```
## Not run:
 # Get the sequence descriptions from a set of transcript accessions
 transcript = c("XM_004487701")
 sapply(transcript, function(x) refseq_description(x), USE.NAMES = FALSE)

 # Get the sequence descriptions from a set of protein accessions
 protein = c("XP_004487758")
 sapply(protein, function(x) refseq_description(x), USE.NAMES = FALSE)

#' # Get the sequence descriptions from a set of Gene accessions
locs <- c("LOC101512347", "LOC101506901")
sapply(locs, function(x) refseq_description(x), USE.NAMES = FALSE)
## End(Not run)
```

refseq_fromGene *Get the mRNA or protein accession*

Description

refseq_fromGene() Returns the mRNA or protein accession from a single GeneID.

Usage

```
refseq_fromGene(GeneID, sequence)
```

Arguments

GeneID	A character string of the GeneID.
sequence	A character string of the mRNA or protein accession to fetch data from mRNA or protein databases, respectively.

Value

A character vector containing the mRNA or protein accession corresponding to the especified GeneID.

Author(s)

Jose V. Die

See Also

refseq_protein2RNA to obtain the transcript accessions that encode a set of protein accessions.

refseq_RNA2protein to obtain the protein accessions encoded by a set of transcript accessions.

Examples

```
# Get the XM accessions from a set of gene ids
locs <- c("LOC101512347")
sapply(locs, function(x) refseq_fromGene (x, sequence = "transcript"), USE.NAMES = FALSE)

# Get the XP accessions from a set of gene ids
locs <- c("LOC101512347")
sapply(locs, function(x) refseq_fromGene (x, sequence = "protein"), USE.NAMES = FALSE)
```

refseq_fromGene_action *Get the mRNA or protein accession*

Description

`refseq_fromGene_action()` Returns the mRNA or protein accession from a single GeneID.

Usage

```
refseq_fromGene_action(GeneID, sequence, retries)
```

Arguments

`GeneID`	A character string of the GeneID.
`sequence`	A character string of the mRNA or protein accession to fetch data from mRNA or protein databases, respectively.
`retries`	A numeric value to control the number of retry attempts to handle 502 errors.

Value

A character vector containing the mRNA or protein accession corresponding to the especified `GeneID`.

Author(s)

Jose V. Die

refseq_GeneID *Get the GeneID*

Description

`refseq_GeneID()` Returns the GeneID from a single transcript or protein accession.

Depending on the function, available accessions in `refseqR` include RefSeq models with the prefixes XM_ (mRNA), XR_ (non-coding RNA), and XP_ (protein), as well as subsequently curated RefSeq records with NM_, NR_, or NP_ accession prefixes.

Usage

```
refseq_GeneID (accession, db, retries)
```

Arguments

accession	A character string of the transcript or protein accession.
db	A character string of the "nuccore" or "protein" database.
retries	A numeric value to control the number of retry attempts to handle internet errors.

Value

A character vector containing the GeneID corresponding to the specified accession as accession.

Author(s)

Jose V. Die

See Also

refseq_protein2RNA to obtain the transcript accessions that encode a set of protein accessions.

refseq_RNA2protein to obtain the protein accessions encoded by a set of transcript accessions.

Examples

```
## Not run:
# Get the gene symbol from a set of transcript accessions
transcript = c("XM_004487701")
sapply(transcript, function(x) refseq_GeneID (x, db = "nuccore", retries = 4), USE.NAMES = FALSE)

# Get the gene symbol from a set of protein accessions
protein = c("XP_004487758")
sapply(protein, function(x) refseq_GeneID (x, db = "protein", retries = 4), USE.NAMES = FALSE)
## End(Not run)
```

---

refseq_geneSymbol *Get the gene symbol*

Description

refseq_geneSymbol() Returns the gene symbol from a single Gene id. accession.

Usage

```
refseq_geneSymbol (id, db)
```

Arguments

id	A character string of the transcript or protein id.
db	A character string of the "nuccore" or "protein" database.

Value

A character vector containing the gene symbol corresponding to the especified accession as `id`.

Author(s)

Jose V. Die

See Also

`refseq_protein2RNA` to obtain the transcript ids that encode a set of protein ids.

`refseq_RNA2protein` to obtain the protein ids encoded by a set of transcript ids.

Examples

```
# Get the gene symbol from a set of transcript accessions
id = c("XM_004487701", "XM_004488493")
sapply(id, function(x) refseq_geneSymbol (x, db = "nuccore"), USE.NAMES = FALSE)

# Get the gene symbol from a set of XP accessions
id = c("XP_004487758")
sapply(id, function(x) refseq_geneSymbol (x, db = "protein"), USE.NAMES = FALSE)
```

---

refseq_geneSymbol_action

Get the gene symbol

Description

`refseq_geneSymbol_action()` Returns the gene symbol from a single Gene id. accession.

Usage

```
refseq_geneSymbol_action (id, db, retries)
```

Arguments

`id`	A character string of the transcript or protein id.
`db`	A character string of the "nuccore" or "protein" database.
`retries`	A numeric value to control the number of retry attempts to handle internet errors.

Value

A character vector containing the gene symbol corresponding to the especified accession as `id`.

Author(s)

Jose V. Die

See Also

refseq_protein2RNA to obtain the XM ids that encode a set of XP ids.

refseq_RNA2protein to obtain the XP ids encoded by a set of XM ids.

---

refseq_mRNAfeat *Get mRNA features*

Description

refseq_mRNAfeat() Returns a number of features from a single/multiple mRNA accession(s).

Depending on the function, available accessions in refseqR include RefSeq models with the prefixes XM_ (mRNA), XR_ (non-coding RNA), and XP_ (protein), as well as subsequently curated RefSeq records with NM_, NR_, or NP_ accession prefixes.

Usage

```
refseq_mRNAfeat(transcript , feat)
```

Arguments

transcript	A character string of the transcript id.
feat	A character string of the selected features. Allowed features: 'caption', 'moltype', 'sourcedb', 'updatedate', 'slen', 'organism', 'title'.

Value

A tibble of summarized results including columns:

- caption, mRNA accession
- moltype, type of molecule
- sourcedb, database (GenBank)
- updatedate, date of updated record
- slen, molecule length (in bp)
- organism
- title, sequence description

Author(s)

Jose V. Die

See Also

refseq_fromGene to obtain the transcript or protein accession from a single GeneID accession.

refseq_RNA2protein to obtain the protein accessions encoded by a set of transcript ids.

Examples

```
# Get several molecular features from a set of mRNA accessions
transcript = c("XM_004487701", "XM_004488493", "XM_004501904")
feat = c("caption", "moltype", "sourcedb", "slen")
refseq_mRNAfeat(transcript ,feat)
```

`refseq_protein2RNA` *Get the transcript accession from the protein accession*

Description

`refseq_protein2RNA()` Returns the transcript accession from a single protein accession.

Depending on the function, available accessions in `refseqR` include RefSeq models with the prefixes XM_ (mRNA), XR_ (non-coding RNA), and XP_ (protein), as well as subsequently curated RefSeq records with NM_, NR_, or NP_ accession prefixes.

Usage

```
refseq_protein2RNA(protein)
```

Arguments

`protein`	A character string of the protein id.

Value

A character vector containing the transcript ids that encode the `protein`.

Author(s)

Jose V. Die

See Also

`refseq_RNA2protein` to obtain the protein ids encoded by a set of transcript ids.

Examples

```
## Not run:
 # Get the transcript id from a single protein accession
 protein <- "XP_020244413"
 refseq_protein2RNA(protein)

 # Get the transcript ids from a set of protein accessions
 protein = c("XP_004487758", "XP_004488550")
 sapply(protein, function(x) refseq_protein2RNA(x), USE.NAMES = FALSE)
## End(Not run)
```

`refseq_RNA2protein` *Get the protein accession from the transcript accession*

Description

`refseq_RNA2protein()` Returns the protein accession from a single transcript accession.

Depending on the function, available accessions in `refseqR` include RefSeq models with the prefixes XM_ (mRNA), XR_ (non-coding RNA), and XP_ (protein), as well as subsequently curated RefSeq records with NM_, NR_, or NP_ accession prefixes.

Usage

```
refseq_RNA2protein(transcript)
```

Arguments

`transcript`	A character string of the protein accession.

Value

A character vector containing the protein id encoded by the mRNA especified as `transcript`.

Author(s)

Jose V. Die

See Also

`refseq_protein2RNA` to obtain the transcript ids that encode a set of proteins ids.

Examples

```
## Not run:
# Get the protein id from a single transcript accession
transcript <- "XM_004487701"
refseq_RNA2protein(transcript)

# Get the protein ids from a set of transcript accessions
transcript = c("XM_004487701", "XM_004488493")
sapply(transcript, function(x) refseq_RNA2protein(x), USE.NAMES = FALSE)
## End(Not run)
```

Index

refseq_AAlen, 2
refseq_AAlen_action, 3
refseq_AAmol_wt, 4
refseq_AAseq, 5
refseq_CDScoords, 6, *7*
refseq_CDSseq, *6*, 7
refseq_description, 8
refseq_fromGene, 9, *13*
refseq_fromGene_action, 10
refseq_GeneID, 10
refseq_geneSymbol, 11
refseq_geneSymbol_action, 12
refseq_mRNAfeat, 13
refseq_protein2RNA, *8, 9*, *11–13*, 14, *15*
refseq_RNA2protein, *3, 4*, *8, 9*, *11–14*, 15
refseqR, 2
refseqR-package *(*refseqR*)*, 2