

# Package: refseqR (via r-universe)

August 17, 2024

**Type** Package

**Title** Common Computational Operations Working with RefSeq Entries  
(GenBank)

**Version** 1.1.2

**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

**Roxygen** list(markdown = TRUE)

**RoxygenNote** 7.2.3

**Suggests** knitr, rmarkdown

**VignetteBuilder** knitr

**Repository** <https://jdieramon.r-universe.dev>

**RemoteUrl** <https://github.com/jdieramon/refseqr>

**RemoteRef** HEAD

**RemoteSha** 3aa2898d21a97170046b1913978afaa286dab9ce

## Contents

extract_from_xm . . . . .	2
refseqR . . . . .	3
refseq_AAseq . . . . .	3
refseq_AA_length . . . . .	4

refseq_AA_mol_wt . . . . .	5
refseq_CDScoords . . . . .	6
refseq_CDSseq . . . . .	7
refseq_description . . . . .	8
refseq_fromGene . . . . .	9
refseq_GeneID . . . . .	10
refseq_geneSymbol . . . . .	11
refseq_mRNA2protein . . . . .	12
refseq_mRNAfeat . . . . .	13
refseq_protein2mRNA . . . . .	14

<b>Index</b>	<b>15</b>
--------------	-----------

---

extract_from_xm	<i>Extract some features from an XM accession</i>
-----------------	---

---

## Description

Parses an XM accession (Genbank format) and extract some features provided by the user.

## Usage

```
extract_from_xm(listName, feat = "tissue")
```

## Arguments

listName	a downloaded flat file from the nuccore NCBI database
feat	a feature to be extracted. Allowed features include "sex", "tissue" or "genotype"

## Author(s)

Jose V. Die

## Examples

```
xm <- "XM_020388824"
# First, get the character vector containing the fetched record
mrna_gb <- rentrez::entrez_fetch(db = "nuccore", id = xm, rettype = "gp")
extract_from_xm(mrna_gb, feat = "sex")
extract_from_xm(mrna_gb, feat = "genotype")
extract_from_xm(mrna_gb, feat = "tissue")
```

---

`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_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 sequences 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\_AA\_length      *Get the amino acid length from a protein accession*

---

## Description

refseq\_AA\_length() 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_AA_length(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\\_mRNA2protein](#) to obtain the protein ids encoded by a set of transcript ids.

## Examples

```
# Get the transcript ids from a set of protein accessions
protein = c("XP_004487758", "XP_004488550")
sapply(protein, function(x) refseq_AA_length(x, retries = 4), USE.NAMES = FALSE)
```

---

refseq_AA_mol_wt	<i>Extract the molecular weight from a protein accession</i>
------------------	--

---

### Description

refseq\_AA\_mol\_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_AA_mol_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

### Examples

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

# Get the molecular weight from from a set of protein accessions
protein = c("XP_004487758", "XP_004488550")
sapply(protein, function(x) refseq_AA_mol_wt(x), USE.NAMES = TRUE)
```

---

refseq_CDScoords	<i>Extract the coding sequences (CDS) coordinates from a transcript accession</i>
------------------	---

---

### 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	<i>Extract the CDS nucleotide sequence into a Biostrings object</i>
---------------	---

---

### 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\\_protein2mRNA](#) to obtain the transcript ids that encode a set of protein ids.

[refseq\\_mRNA2protein](#) to obtain the protein ids encoded by a set of transcript ids.

### Examples

```
# 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)
```

---

refseq_fromGene	<i>Get the mRNA or protein accession</i>
-----------------	--

---

**Description**

refseq\_fromGene() Returns the mRNA or protein accession from a single GeneID.

**Usage**

```
refseq_fromGene(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 specified GeneID.

**Author(s)**

Jose V. Die

**See Also**

[refseq\\_protein2mRNA](#) to obtain the transcript accessions that encode a set of protein accessions.

[refseq\\_mRNA2protein](#) to obtain the protein accessions encoded by a set of transcript accessions.

**Examples**

```
# Get the transcript accessions from a set of gene ids
locs <- c("LOC101512347")
sapply(locs, function(x) refseq_fromGene (x,sequence="transcript",retries=4),USE.NAMES=FALSE)

# Get the protein accessions from a set of gene ids
locs <- c("LOC101512347")
sapply(locs, function(x) refseq_fromGene (x,sequence="protein",retries=4),USE.NAMES=FALSE)
```

---

`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**

<code>accession</code>	A character string of the transcript or protein accession.
<code>db</code>	A character string of the "nucore" or "protein" database.
<code>retries</code>	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\\_protein2mRNA](#) to obtain the transcript accessions that encode a set of protein accessions.

[refseq\\_mRNA2protein](#) to obtain the protein accessions encoded by a set of transcript accessions.

**Examples**

```
# Get the gene symbol from a set of transcript accessions
transcript = c("XM_004487701", "XM_004488493")
sapply(transcript, function(x) refseq_GeneID (x, db = "nucore", 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)
```

---

refseq_geneSymbol	<i>Get the gene symbol</i>
-------------------	----------------------------

---

### Description

refseq\_geneSymbol() Returns the gene symbol from a single Gene id. accession.

### Usage

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

### Arguments

id	A character string of the XP or XM 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 specified accession as id.

### Author(s)

Jose V. Die

### See Also

[refseq\\_XMfromXP](#) to obtain the XM ids that encode a set of XP ids.

[refseq\\_XPfromXM](#) to obtain the XP ids encoded by a set of XM ids.

### Examples

```
# Get the gene symbol from a set of XM accessions
xm = c("XM_004487701", "XM_004488493")
sapply(xm, function(x) refseq_geneSymbol (x, db = "nuccore", retries = 3), USE.NAMES = FALSE)

# Get the gene symbol from a set of XP accessions
xp = c("XP_004487758")
sapply(xp, function(x) refseq_geneSymbol (x, db = "protein", retries = 3), USE.NAMES = FALSE)
```

---

refseq\_mRNA2protein     *Get the protein accession from the transcript accession*

---

### Description

refseq\_mRNA2protein() 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_mRNA2protein(transcript)
```

### Arguments

transcript     A character string of the protein accession.

### Value

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

### Author(s)

Jose V. Die

### See Also

[refseq\\_protein2mRNA](#) to obtain the transcript ids that encode a set of proteins ids.

### Examples

```
# Get the protein id from a single transcript accession
transcript <- "XM_004487701"
refseq_mRNA2protein(transcript)

# Get the protein ids from a set of transcript accessions
transcript = c("XM_004487701", "XM_004488493")
sapply(transcript, function(x) refseq_mRNA2protein(x), USE.NAMES = FALSE)
```

---

refseq_mRNAfeat	<i>Get mRNA features</i>
-----------------	--------------------------

---

## 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\\_mRNA2protein](#) 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\_protein2mRNA *Get the transcript accession from the protein accession*

---

**Description**

refseq\_protein2mRNA() 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_protein2mRNA(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\\_mRNA2protein](#) to obtain the protein ids encoded by a set of transcript ids.

**Examples**

```
# Get the transcript id from a single protein accession
protein <- "XP_020244413"
refseq_protein2mRNA(protein)

# Get the transcript ids from a set of protein accessions
protein = c("XP_004487758", "XP_004488550")
sapply(protein, function(x) refseq_protein2mRNA(x), USE.NAMES = FALSE)
```

# Index

`extract_from_xm`, 2

`refseq_AA_length`, 4

`refseq_AA_mol_wt`, 5

`refseq_AAseq`, 3

`refseq_CDScoords`, 6, 7

`refseq_CDSseq`, 6, 7

`refseq_description`, 8

`refseq_fromGene`, 9, 13

`refseq_GeneID`, 10

`refseq_geneSymbol`, 11

`refseq_mRNA2protein`, 4, 8–10, 12, 13, 14

`refseq_mRNAfeat`, 13

`refseq_protein2mRNA`, 8–10, 12, 14

`refseq_XMfromXP`, 11

`refseq_XPfromXM`, 11

`refseqR`, 3

`refseqR-package (refseqR)`, 3