

Package ‘Pbase’

April 15, 2020

Type Package

Title Manipulating and exploring protein and proteomics data

Version 0.26.0

Depends R (>= 2.10), methods, BiocGenerics, Rcpp, Gviz

Imports cleaver (>= 1.3.6), Biobase, Biostrings (>= 2.47.5), IRanges (>= 2.13.11), S4Vectors (>= 0.17.24), mzID, mzR (>= 1.99.1), MSnbase (>= 1.15.5), Pviz, biomaRt, GenomicRanges (>= 1.31.7), rtracklayer (>= 1.39.6), ensemblDb (>= 1.99.13), BiocParallel, AnnotationFilter

Suggests testthat (>= 0.8), ggplot2, BSgenome.Hsapiens.NCBI.GRCh38, TxDb.Hsapiens.UCSC.hg38.knownGene, AnnotationHub, knitr, rmarkdown, BiocStyle, EnsDb.Hsapiens.v86 (>= 2.0.0)

Description A set of classes and functions to investigate and understand protein sequence data in the context of a proteomics experiment.

License GPL-3

VignetteBuilder knitr

PackageStatus Deprecated

URL <https://github.com/ComputationalProteomicsUnit/Pbase>

BugReports <https://github.com/ComputationalProteomicsUnit/Pbase/issues>

biocViews ImmunoOncology, Infrastructure, Proteomics, MassSpectrometry, Visualization, DataImport, DataRepresentation

RoxygenNote 6.1.1

git_url <https://git.bioconductor.org/packages/Pbase>

git_branch RELEASE_3_10

git_last_commit fffa65a

git_last_commit_date 2019-10-29

Date/Publication 2020-04-14

Author Laurent Gatto [aut, cre],
Johannes Rainer [aut],
Sebastian Gibb [aut]

Maintainer Laurent Gatto <laurent.gatto@uclouvain.be>

R topics documented:

calculateHeavyLabels	2
etrid2grl	3
isReverse	4
p	5
plotAsAnnotationTrack	5
Pparams-class	6
proteinCoding-methods	7
Proteins-class	7

Index	12
--------------	-----------

calculateHeavyLabels *Calculate heavy labeled peptides*

Description

A function to calculate heavy labeled peptides for proteins stored in a [Proteins](#) object.

Usage

```
calculateHeavyLabels(proteins, peptides, maxN = 20L, nN = 4L,
  nC = 3L, endsWith = c("K", "R", "G"), ...)
```

Arguments

proteins	A Proteins object.
peptides	A named character vector containing the peptides of interest. The names must match the UniProt accession numbers of the proteins in object.
maxN	An integer, maximal length of the heavy labeled peptide.
nN	An integer, minimal number of amino acids at the N terminus.
nC	An integer, minimal number of amino acids at the C terminus.
endsWith	A character vector containing the allowed amino acids at the end of the resulting sequence (every peptide that doesn't end with one of these amino acids has to be one amino acid shorter as maxN).
...	Additional parameters passed to <code>.addOverhangs</code> .

Details

The digestion efficiency with enzymes like trypsin is below 100%. That's why spiked-in peptides for labeled quantitation have to follow the same digestion rules as the peptides of interest. Therefore it is necessary to extend the peptides of interest by a few amino acids on the N- and C-terminus. These extensions should not be a cleavage point of the used enzyme. This method provides an easy interface to find the sequences for heavy labeled peptides that could be used as spike-ins for the peptides of interest. Please see the references for a more detailed discussion.

TODO: There should be a function to find the best labels for a given protein automatically.

Value

A data.frame with 6 columns:

- Protein The Protein accession number.
- Peptide The peptide of interest.
- N_overhang The added sequence of the N-terminus.
- C_overhang The added sequence of the C-terminus.
- spikeTideResult A short description of the used creation rule.
- spikeTide The heavy labeled peptide that represents the peptide of interest best.

Author(s)

Sebastian Gibb <mail@sebastiangibb.de> and Pavel Shliha

References

The complete description of the issue: <https://github.com/sgibb/cleaver/issues/5>

Kito, Keiji, et al. A synthetic protein approach toward accurate mass spectrometric quantification of component stoichiometry of multiprotein complexes. Journal of proteome research 6.2 (2007): 792-800. <http://dx.doi.org/10.1021/pr060447s>

Examples

```
## example protein database
data(p, package = "Pbase")

## digest proteins into peptides
cleavedProteins <- cleave(p)

## find spike-ins for the peptides of interest
calculateHeavyLabels(cleavedProteins,
  peptides = c(A4UGR9 = "MEGFHIK",
              A4UGR9 = "QGNMYTLISK",
              A6H8Y1 = "GSTASNPQR"))
```

etrid2grl

From a transcript identifier to GRanges object

Description

This function takes on or more Ensembl transcript identifiers, queries Biomart and constructs a GRangesList object as would Gviz::BiomartGeneRegionTrack for a genomic region (in fact, currently most of the code has been taken from Gviz::fetchBMData and Gviz::chrName is used to validate chromosome names).

Usage

```
etrid2grl(etrid, ens, use.names = FALSE)
```

Arguments

etrid	A vector of Ensembl transcript identifiers.
ens	A instance of class Mart from biomaRt. If missing, useMart("ensembl", "hsapiens_gene_ensembl") is used.
use.names	If set to TRUE and etrid has names, then the latter are used to name the output.

Value

A GRangesList object of length length(etrid).

Author(s)

Laurent Gatto

Examples

```
id <- c("ENST00000612959", "ENST00000317091")
grl1 <- etrid2grl(id[1])
grl1
grl <- etrid2grl(id)
stopifnot(all.equal(id, names(grl)))
```

isReverse

Are all the ranges on the same strand

Description

Checks if all ranges of a GRanges object are reverse.

Usage

```
isReverse(gr)
```

```
isForward(gr)
```

Arguments

gr	A GRanges object.
----	-------------------

Value

A logical if *all* the ranges in the gr object are on the "-" (or "+" for codeisForward) strand.

Author(s)

Laurent Gatto

p

*Data accompanying the Pbase package***Description**

A small example `Proteins` test instance. This object is likely to change on a regular basis. It will be described more thoroughly when it becomes stable. The MSMS spectra that were searched against the database are available in the `pms` `MSnExp` object.

Usage

```
data(p)
data(pms)
```

See Also

The `Pbase-data` vignette.

Examples

```
data(p)
p
data(pms)
pms
```

`plotAsAnnotationTrack` *Plot gene region and annotation tracks*

Description

These functions convert ranges of peptides or exons to `AnnotationTrack` or `GeneRegionTrack` objects from the `Gviz` package and produces the corresponding plot. The `genome` argument controls whether additional ideogram and axis tracks are to be plotted. `plotAsAnnotationTrack` plots peptides that span multiple exons in red and connects them with a grey line.

Usage

```
plotAsAnnotationTrack(x, ..., genome = "hg38", plot = TRUE)

plotAsGeneRegionTrack(..., genome = "hg38", plot = TRUE)
```

Arguments

<code>x</code>	A <code>GRanges</code> object containing peptides genomics coordinates. These ranges are converted to a <code>AnnotationTrack</code> .
<code>...</code>	One or more <code>GRanges</code> instances, typically resulting from calling <code>etrid2gr1</code> , or, a single <code>GRangesList</code> . These ranges are converted to <code>GeneRegionTrack</code> instances.
<code>genome</code>	A character of length 1, giving the name of the genome. Default is "hg38". If <code>NULL</code> , no chromosome and axis tracks are displayed.
<code>plot</code>	A logical defining if the figure should be plotted. Default is <code>TRUE</code> .

Value

Used for its plotting side effects. Invisible returns a list of tracks.

Author(s)

Laurent Gatto

Pparams-class	Class "Pparams"
---------------	-----------------

Description

Pbase parametrisation infrastructure.

Objects from the Class

New Pbase parameters can be generated with the `Pparams()` constructor. `Pparams` instances control various aspects of Pbase functions, as described in the *Slots* section below. If no parameters are passed to the respective functions, default values from `Pparams()` are used.

Slots

DbFormat: The format of the protein sequence fasta database used to generate the `Proteins` object. Currently only "UniProt" is supported. "RefSeq" will be added as well as a mechanism to support arbitrary and custom fasta header.

IdFormat: The format of the identification data files used to add `pfeatures` to `Protein` instances. Currently, `mzIdentML` is supported.

IdReader: Package to be used to load the identification data. Currently one of `mzR` (via the `openIDfile` and `psms` functions) or `mzID` (via the `mzID` and `flatten` functions). Differences between these two architectures include the metadata available in the `Proteins`' `pfeatures`, speed and stability (`mzR` is much faster but less mature and currently susceptible to crashes).

verbose: A logical defining if the various functions display messages (default) or remain silent.

Methods

`show signature(object = "Pparams"): ...`

Author(s)

Laurent Gatto <lg390@cam.ac.uk>

Examples

```
Pparams()
Pparams(IdReader = "mzID")

try(Pparams(IdReader = "mzid"))
```

proteinCoding-methods *Only keep protein coding ranges*

Description

Removed all the ranges that are not protein coding. Typically used on the output of [etrid2gr1](#).

Methods

`signature(object="GRanges", mcol="character", coding="character")` Removes all the ranges that are not annotated as protein coding ranges, i.e. ranges whose `mcols()`\$`mcol` is different from coding. The default values are `mcols()``feature` and `"protein_coding"`. The method return the GRanges trimmed from all non-matching ranges.

`signature(object="GRangesList", mcol="character", coding="character")` As above but for GRanges in a GrangesList.

Proteins-class

The Proteins Class for Proteomics Data And Meta-Data

Description

The Proteins class encapsulates data and meta-data for proteomics experiments. The class stores the protein sequences as well as specific subsets of interest, typically peptides, as ranges. The Proteins instances, the sequence and peptide slots are described by their respective metadata attributes.

Objects from the Class

Objects can be created using its constructor `Proteins`. The constructor either takes a fasta file name as first argument, an [EnsDb](#) object or a named `uniprotIds` argument with valid UniProt accession numbers (not yet implemented).

The Proteins constructor with the `EnsDb` loads protein data directly from the `EnsDb` object. The additional arguments `filter`, `loadProteinDomains`, `columns` and `fetchLRG` allow to additionally specify if only proteins matching a certain criteria should be fetched, whether all protein domains should be added as pranges, optional additional annotation columns that should be retrieved and whether proteins from Locus Reference Genes (LRG) should also be retrieved from the database.

Details

An instance of class `Proteins` is characterised by one or multiple protein sequences that can be accessed as `AAStringSet` with the `aa` accessor. Sequence-specific annotation, such as accession numbers, protein and gene names, ... is available with the `acols` method. General metadata such as the data of creation of the instance are stored as a `list` returned by the `metadata` accessor, which would typically contain a `created` character that documents when the object was created, a reference genome descriptor, a `UniProtRelease` with the release data of the UniProt database and possibly others.

Each sequence of a `Proteins` instance can also be characterised by a set of specific ranges describing peptides of interest. These *peptide features* can be extracted as an `AAStringSetList`, where

each protein sequence contains 0 or more peptide features. These peptides features are encode as ranges along the original proteins sequences (a list of IRanges) that can be extracted with the `pranges` function. These peptide features have their own metadata describing for example peptide identification scores, number of missed cleavages, ... available with the `pcols` methods.

See also the Pbase-data vignette.

The Proteins constructor with argument `file` being an `EnsDb` object allows to retrieve protein sequences along with all their related protein domains from an `EnsDb` annotation database. The optional `filter` argument can be used to fetch only proteins matching the defined filtering criteria from the database. The `filter` argument takes an object extending the `AnnotationFilter` class, an `AnnotationFilterList` combining such objects or a filter expression in form of a formula. See the `AnnotationFilter` and `proteins` documentation for more details.

Additional annotation columns from the database that should be retrieved from the database and included into `acols` can be specified with the `columns` argument. The `listColumns` can be used to list all available annotation columns from the database. Ensembl protein IDs will be used as the names of the returned Proteins object. See the vignette from the `ensemldb` package for an overview of supported filters or below for some examples.

Development notes

Since version 0.2.0, `addIdentificationData` supports multiple identification file names to be added to a Proteins instance (argument renamed `filenames`) using either `mzID` or `mzR`. Added new `Pparams` parametrisation infrastructure.

See `news(package = "Pbase")` for a description of all changes.

Other possible metadata fields: `Uniprot.sw`, `biomaRt` instances.

Slots

`metadata`: Object of class "list" containing global metadata, accessed with `metadata`.

`aa`: Object of class "AAStringSet" storing the protein sequences, accessed with `aa`.

`.__classVersion__`: Object of class "Versions" documenting the class verions. Intended for developer use and debugging.

Extends

Class "`Versioned`", directly.

Methods

aa `signature(x = "Proteins")`: Returns an `AAStringSet` instance representing the sequences of the proteins.

pfeatures `signature(x = "Proteins")`: ...

pranges `signature(x = "Proteins")`: ...

metadata `signature(x = "Proteins")`: Returns a list of global metadata of the instance `x`, including data of instance creation or, if created from a set of UniProt identifiers (see constructors above), the UniProt version and UniProt.WS version number.

acols `signature(x = "Proteins")`: Returns a `DataFrame` of protein metadata.

pcols `signature(x = "Proteins")`: Returns a list of feature metadata.

avarLabels `signature(x = "Proteins")`: Returns the names of the sequences metadata.

pvarLabels `signature(x = "Proteins")`: Returns the names of the peptide feature metadata.

- seqnames** signature(x = "Proteins"): Returns the protein sequence names defined as UniProt accession numbers.
- names** signature(x = "Proteins"): Returns the protein sequence names defined as UniProt accession numbers. It is just a synonym for seqnames.
- length** signature(x = "Proteins"): Returns the number of proteins.
- [signature(x = "Proteins", i = "ANY", j = "missing"): Creates a subset of the Proteins instance.
- [[signature(x = "Proteins", i = "ANY", j = "missing"): Returns an [AAString](#) instance representing the sequence of the selected protein.
- pfilter** signature(x = "Proteins", mass = "numeric", len = "numeric", ...): ...
- cleave** signature(x = "Proteins", enzym = "character", missedCleavages = "numeric"): Cleaves all proteins using the enzym rule while allowing missedCleavages missing cleavages. Please see [cleave](#) for details.
- addIdentificationData** signature(object = "Proteins", id = "character", rmEmptyRanges = "logical", par = "Pparams"): Adds identification data from an IdentMzMI file (id) to the Proteins object. If rmEmptyRanges is TRUE proteins without any identification data are removed. See [Pparams](#) for further settings.
- addPeptideFragments** signature(object = "Proteins", filenames = "character", rmEmptyRanges = "logical", par = "Pparams"): Adds identification data from a fasta file (filenames) to the Proteins object. Please note that both fasta files (the origin of the Proteins object and the ones given in filenames) must share the same Uniprot accession numbers. If rmEmptyRanges is TRUE proteins without any identification data are removed. See [Pparams](#) for further settings.
- plot** signature(x = "Proteins", y = "missing"): Plots all proteins and associated peptides using the Gviz/Pviz infrastructure.
- show** signature(object = "Proteins"): Displays object summary as text.

Functions

- rmEmptyRanges** signature(x = "Proteins"): removes proteins with empty peptide ranges.
- proteotypic** signature(x = "Proteins"): returns a modified Proteins object. pcols(x) gains a "Proteotypic" logical column, indicating of the peptide is proteotypic or now.
- proteinCoverage** signature(pattern = "Proteins"): calculates the coverage of proteins. pcols(x) gains a "Coverage" numeric column.
- isCleave** signature(x = "Proteins", missedCleavages = "numeric"): Tests whether a Protein object was cleaved already.

Author(s)

Laurent Gatto <lg390@cam.ac.uk>, Sebastian Gibb <mail@sebastiangibb.de> and Johannes Rainer <johannes.rainer@eurac.edu>

References

Definition of the UniProt fasta comment format: <http://www.uniprot.org/help/fasta-headers>

See Also

[calculateHeavyLabels](#)

Examples

```

## Create a Protein object reading all proteins from a fasta file.
fastaFiles <- list.files(system.file("extdata", package = "Pbase"),
                        pattern = "fasta", full.names = TRUE)
p <- Proteins(fastaFiles)
p
metadata(p)

## Adding custom metadata
metadata(p, "Comment") <- "I love R"
metadata(p)

## Plotting
plot(p[1:5], from = 1, to = 30)

## Cleaving
pp <- cleave(p[1:100])
pp <- proteotypic(pp)
pp
pcols(pp[1:2])

plot(pp[1:2], from = 20, to = 30)

## Protein coverage
pp <- proteinCoverage(pp)
avarLabels(pp)
acols(pp)$Coverage
pp

## Add identification data
idfile <- system.file("extdata/Thermo_Hela_PRTC_selected.mzid",
                    package = "Pbase")
p <- addIdentificationData(p, idfile)
pranges(p)
pfeatures(p)

plot(p[1])
plot(p[1], # the first protein has 36 peptides
     fill = c(rep("orange", 13), rep("steelblue", 13)))

## Retrieve a Proteins object from an EnsDb object: first load the annotation
## database for all human genes defined in Ensembl version 86.
library(ensemldb)
library(EnsDb.Hsapiens.v86)
edb <- EnsDb.Hsapiens.v86

## Define a filter to retrieve all genes from chromosome Y
sqnf <- SeqNameFilter("Y")
## Retrieve the proteins without protein domains but specify to retrieve in
## addition the transcript biotype for the encoding transcripts and the gene
## names
prts <- Proteins(edb, filter = sqnf, loadProteinDomains = FALSE,
                columns = c("tx_biotype", "gene_name"))
aa(prts)

```

```
acols(prts)

## The listColumns method lists all available columns from the database.
listColumns(edb)

## Load all proteins from the gene ZBTB16 including all protein domains from
## the database. Here we pass the filter criteria as a formula to the method
prts <- Proteins(edb, filter = ~ gene_name == "ZBTB16")

## List available pranges
pcols(prts)

## Access the protein domains
pcols(prts)$ProteinDomains
```

Index

- *Topic **classes**
 - Pparams-class, 6
 - Proteins-class, 7
- *Topic **datasets**
 - p, 5
- *Topic **methods**
 - proteinCoding-methods, 7
- [,Proteins,ANY,ANY,ANY-method (Proteins-class), 7
- [,Proteins,ANY,ANY-method (Proteins-class), 7
- [[,Proteins,ANY,ANY-method (Proteins-class), 7

- aa (Proteins-class), 7
- aa,Proteins-method (Proteins-class), 7
- AAString, 9
- AAStringSet, 8
- acols (Proteins-class), 7
- addIdentificationData,Proteins,character-method (Proteins-class), 7
- addPeptideFragments (Proteins-class), 7
- addPeptideFragments,Proteins,character-method (Proteins-class), 7
- AnnotationFilter, 8
- AnnotationFilterList, 8
- avarLabels (Proteins-class), 7
- avarLabels,Proteins-method (Proteins-class), 7

- calculateHeavyLabels, 2, 9
- class:Proteins (Proteins-class), 7
- cleave, 9
- cleave,Proteins-method (Proteins-class), 7

- DataFrame, 8

- EnsDb, 7, 8
- etrid2gr1, 3, 5, 7

- isCleaved (Proteins-class), 7
- isForward (isReverse), 4
- isReverse, 4

- length,Proteins-method (Proteins-class), 7
- listColumns, 8

- metadata,Proteins-method (Proteins-class), 7
- metadata<-,Proteins-method (Proteins-class), 7

- names,Proteins-method (Proteins-class), 7

- p, 5
- pcols (Proteins-class), 7
- pfeatures (Proteins-class), 7
- pfeatures,Proteins-method (Proteins-class), 7
- pfilter (Proteins-class), 7
- pfilter,Proteins-method (Proteins-class), 7
- plot,Proteins,missing-method (Proteins-class), 7
- plotAsAnnotationTrack, 5
- plotAsGeneRegionTrack (plotAsAnnotationTrack), 5
- pms (p), 5
- Pparams, 9
- Pparams (Pparams-class), 6
- Pparams-class, 6
- pranges (Proteins-class), 7
- pranges,Proteins-method (Proteins-class), 7
- pranges<- (Proteins-class), 7
- pranges<- ,Proteins,CompressedIRangesList-method (Proteins-class), 7
- proteinCoding (proteinCoding-methods), 7
- proteinCoding,GRanges-method (proteinCoding-methods), 7
- proteinCoding,GRangesList-method (proteinCoding-methods), 7
- proteinCoding-methods, 7
- proteinCoverage (Proteins-class), 7
- Proteins, 2
- Proteins (Proteins-class), 7

proteins, [8](#)
Proteins, character, missing-method
 (Proteins-class), [7](#)
Proteins, EnsDb, missing-method
 (Proteins-class), [7](#)
Proteins, missing, character-method
 (Proteins-class), [7](#)
Proteins, missing, missing-method
 (Proteins-class), [7](#)
Proteins-class, [7](#)
proteotypic (Proteins-class), [7](#)
pvarLabels (Proteins-class), [7](#)
pvarLabels, Proteins-method
 (Proteins-class), [7](#)

rmEmptyRanges (Proteins-class), [7](#)

seqnames, Proteins-method
 (Proteins-class), [7](#)
show, Pparams-method (Pparams-class), [6](#)
show, Proteins-method (Proteins-class), [7](#)

Versioned, [8](#)