Package ‘parsemsf’

December 9, 2017

Title Parse ThermoFisher MSF Files and Estimate Protein Abundances

Version 0.1.1

Description Provides functions for parsing ThermoFisher MSF files produced by Proteome Discoverer 1.4.x (see <https://thermofisher.com> for more information). This package makes it easy to view individual peptide information, including peak areas, and to map peptides to locations within the parent protein sequence. This package also estimates protein abundances from peak areas and across multiple technical replicates. The author of this package is not affiliated with ThermoFisher Scientific in any way.

URL https://github.com/benjaminjack/parsemsf/

Depends R (>= 3.2.4)

License GPL-2

Encoding UTF-8

LazyData true

RoxygenNote 6.0.1

Imports dplyr (>= 0.5.0), dbplyr, DBI, lazyeval, RSQLite (>= 1.0.0), stats, stringr (>= 1.1.0), tidyr (>= 0.6.0)

Suggests testthat (>= 1.0.2), knitr, rmarkdown, ggplot2

VignetteBuilder knitr

BugReports https://github.com/benjaminjack/parsemsf/issues

NeedsCompilation no

Author Benjamin Jack [aut, cre]

Maintainer Benjamin Jack <benjamin.r.jack@gmail.com>

Repository CRAN

Date/Publication 2017-12-09 22:00:10 UTC

R topics documented:

`make_area_table` *Make a table of peptide areas*

Description

Areas under each peptide peak that can be used downstream for quantitation. See `quantitate` for protein quantitation.

Usage

```
make_area_table(msf_file, min_conf = "High”,
  prot_regex = "^>([a-zA-Z0-9._]+)\\b”, collapse = TRUE)
```

Arguments

`msf_file` A file path to a ThermoFisher MSF file.

`min_conf` "High", "Medium", or "Low". The minimum peptide confidence level to retrieve from MSF file.

`prot_regex` Regular expression where the first group matches a protein name or ID from the protein description. Regex must contain ONE group. The protein description is typically generated from a fasta reference file that was used for the database search.

`collapse` If TRUE, peptides that match to multiple protein sequences are collapsed into a single row with multiple protein descriptions and protein IDs in the `Proteins` and `ProteinID` columns separated by semi-colons (";").

Value

A data frame containing peptide areas for peptides at or above the minimum confidence level.

`peptide_id`	a unique peptide ID
`spectrum_id`	a unique spectrum ID
`protein_desc`	protein description from reference database used to assign peptides to protein groups, parsed according to `prot_regex`
`sequence`	amino acid sequence (does not show post-translational modifications)
`area`	area under peptide peak
`mass`	peptide mass
`m_z`	mass-to-charge ratio
`charge`	peptide charge
`intensity`	peak intensity; useful if no area is available
`first_scan`	first scan in which peptide appears

Examples

```
make_area_table(parsemsf_example("test_db.msf"))
```

---

make_pep_table *Make a data frame of peptides*

Description

Extracts amino acid sequences (without post-translational modifications), assigned protein groups, and quality scores.

Usage

```
make_pep_table(msf_file, min_conf = "High",
  prot_regex = "^>([a-zA-Z0-9._]+)\\b", collapse = TRUE)
```

Arguments

`msf_file` A file path to a ThermoFisher MSF file.

`min_conf` "High", "Medium", or "Low". The minimum peptide confidence level to retrieve from MSF file.

`prot_regex` Regular expression where the first group matches a protein name or ID from the protein description. Regex must contain ONE group. The protein description is typically generated from a fasta reference file that was used for the database search.

`collapse` If TRUE, peptides that match to multiple protein sequences are collapsed into a single row with multiple protein descriptions and protein IDs in the `Proteins` and `ProteinID` columns separated by semi-colons (";").

Value

A data frame of all peptides above the confidence cut-off from a ThermoFisher MSF file.

`peptide_id` a unique peptide ID

`spectrum_id` a unique spectrum ID

`protein_id` unique protein group ID to which this peptide maps

`protein_desc` protein description from reference database used to assign peptides to protein groups, parsed according to `prot_regex`

`sequence` amino acid sequence (does not show post-translational modifications)

`pep_score` PEP score

`q_value` Q-value score

Examples

```
# Read from a path

make_pep_table(parsemsf_example("test_db.msf"))

# Retrieve full protein description

make_pep_table(parsemsf_example("test_db.msf"), prot_regex = "")

# ...which is also equivalent to...

make_pep_table(parsemsf_example("test_db.msf"), prot_regex = "^(.+)$")
```

map_peptides *Map peptides to their locations within a protein*

Description

Takes a ThermoFisher MSF file and finds the location of each peptide within its corresponding protein sequence. In cases where a single peptide maps to multiple locations within a protein sequence, only the first location is reported. If a peptide maps ambiguously to multiple proteins, all locations are reported with data from each peptide-protein combination on a separate row.

Usage

```
map_peptides(msf_file, min_conf = "High", prot_regex = "")
```

Arguments

`msf_file` A file path to a ThermoFisher MSF file.

`min_conf` "High", "Medium", or "Low". The minimum peptide confidence level to retrieve from MSF file.

`prot_regex` Regular expression where the first group matches a protein name or ID from the protein description. Regex must contain ONE group. The protein description is typically generated from a fasta reference file that was used for the database search.

Value

A dataframe containing start and stop positions (relative to the parent protein sequence) for each peptide in the database.

`peptide_id` a unique peptide ID

`spectrum_id` a unique spectrum ID

`protein_id` unique protein group ID to which this peptide maps `protein_desc` protein description from reference database used to assign peptides to protein groups, parsed according to `prot_regex`

`peptide_sequence` amino acid sequence (does not show post-translational modifications)

`pep_score` PEP score

`q_value` Q-value score

`protein_sequence` parent protein sequence

`start` start position of peptide within protein sequence

`end` end position of peptide within protein sequence

Examples

```
map_peptides(parsemsf_example("test_db.msf"))
```

`parsemsf` *parsemsf: Parse ThermoFisher MSF files and estimate protein abundances.*

Description

parsemsf: Parse ThermoFisher MSF files and estimate protein abundances.

`quantitate` *Combine technical replicates and quantitate proteins*

Description

Takes a list of thermo MSF files, parses and combines them into a single data frame, and computes areas for each protein group based on the top 3 method of quantitation.

Usage

```
quantitate(reps, normalize = T, match_peps = T, relabel = c())
```

Arguments

`reps` Vector. List of thermo MSF file names

`normalize` Boolean. Should we normalize peptide areas for technical replicate to the total areas in a given replicate?

`match_peps` Boolean. Should we quantitate only on matching peptides across technical replicates?

`relabel` Named vector for relabeling protein groups. Names correspond to a pattern or string to match (i.e. the name or ID of a protein group), and values correspond to the new value (i.e. new protein group name).

Value

A data frame containing area information for all proteins.

`protein_desc` protein description

`area_mean` average peptide area

`area_sd` peptide area standard deviation

`peps_per_rep` Number of peptides per technical replicate used to calculate `area_mean` and `area_sd`. This is typically 3 peptides, but may be less.

Examples

```
quantitate(c(parsemsf_example("test_db.msf"),
             parsemsf_example("test_db2.msf")),
           relabel = c("NP_12345.1" = "NP_1000.1"))
```

Index