Package 'OligoDistiller'

July 28, 2024

Type Package

Title Oligonucleotide LC-MS data processing

Version 0.1

Date 2022-10-01

Author Youzhong LIU <liu-youzhong@hotmail.com>

Maintainer Youzhong LIU <liu-youzhong@hotmail.com>

Depends R (>= 3.6), shiny, tidyverse

Imports BRAIN, OrgMassSpecR, tibble, purrr, tidyr, dplyr, tools, plyr, stringr, shinyjs, V8, pracma, DT, plotly

Description Functions and webtool for metabolite/impurity profiling of oligonucleotide therapeutics

License GNU

LazyData true

RoxygenNote 7.2.3

Suggests knitr, rmarkdown, RMassBankData, gplots

VignetteBuilder knitr

R topics documented:

`annotate_scan_mix` *Targeted followed by non-targeted screening from a deconvoluted oligonucleotide spectra*

Description

The function first searches a complex deconvoluted oligonucleotide spectra against a user provided oligonucleotide impurity/metabolite database, annotating and scoring isotopic pattern matches. It then searches DNA/RNA-like isotope patterns from the rest of deconvoluted oligonucleotide spectra. It provides the monoisotopic molecular weight, average, intensity and envelope likeness of all features detected.

Usage

```
annotate_scan_mix(
  scan_processed_aggregated,
  MSMS = F,
  ntheo = 10,
  formula_flp = "C192H239O117N73P18S4F8",
  cpd_flp = "Demo A",
  transformation_list = NULL,
  mdb = NULL,
  bblock = "DNA",
  min_overlap = 0.6,
  max_msigma = 5,
  max_mmw_ppm = 10,
  baseline = 1000
)
```

Arguments

`scan_processed_aggregated`
: Data frame representing deconvoluted NMS with the true molecular weight scale. Output of the function process_scan.

`MSMS`
: Boolean. TRUE if the spectrum is MS/MS.

`ntheo`
: Integer. Estimated isotope envelop size in number of isotope peaks.

`formula_flp`
: Character. Neutral elemental formula of the main compound (e.g. full length product). Used when a transformation list from the main compound is defined.

`cpd_flp`
: Character. Name of the main compound. Used to label compounds in the output table. Use the separator & if you expect multiple main compound in your sample.

`transformation_list`
: Data frame. Transformation list defining the mass and elemental formula difference from the main compound. Should contain following columns: ID, CPD (IDs and names of transformation products), Plus_Formula, Minus_Formula (Elemental formula difference), Delta.AVG.MW, Delta.MONO.MW (Mass difference for average and mono molecular weight)

`mdb`
: Data frame. You can directly provide all expected compounds to be annotated from your mixture without defining FLP or compound names.

bblock	Character. Either "DNA" or "RNA". Should reflect the main nucleic acid composition of the strand. Used for monoisotopic peak prediction by Pointless algorithm.
min_overlap	Double between 0 and 1. The minimum matching score between experimental and theoretical isotope envelops (known compounds from database/transformation list or unknown predicted by Pointless).
max_msigma	Double between 1 and 50. The maximum-allowed deviation between the shapes of experimental and theoretical isotope pattern. Should set higher for noisy or MS/MS data.
max_mmw_ppm	Double between 1 and 50. The maximum allowed ppm error between masses in the NMS and theoretical molecular weight of oligonucleotide features. Depend on experimental mass deviation and deconvolution bias.
baseline	Numeric. Estimated baseline level (noise) of input spectrum. Depending on instrument and acquisition method. Baseline of MS/MS spectrum is 100 for most instruments.

Author(s)

Youzhong Liu, <liu-youzhong@hotmail.com>

Examples

```
## Not run:

## Example of MS1 data:

data("Strand_A")

scan.deconvoluted = process_scan(scan.A,
polarity = "Negative", baseline = 1000, mz_error = 0.01,
min_charge = 3, max_charge = 12,
min_mz = 500, max_mz = 1200, min_mw = 4000, max_mw = 10000,
mw_gap = 1.1, mw_window = 10)
SCAN_NMS = scan.deconvoluted$scan_processed_aggregated

data("TRANS") # Transformation list for potential oligonucleotide degradants

scan.deconvoluted.annotated = annotate_scan_mix(SCAN_NMS, MSMS = F, ntheo = 10,
formula_flp = "C189H238O119N66P18S4F8",cpd_flp = "Demo A", transformation_list = transformation_list, mdb = NU
bblock = "RNA", min_overlap = 0.6, max_msigma = 5, max_mmw_ppm = 10, baseline = 1000)

view(scan.deconvoluted.annotated$feature)

## Example of MS2 data:

data("Strand_B_MS2")

scan.deconvoluted = process_scan(scan2_B, MSMS = T,
polarity = "Negative", baseline = 100, mz_error = 0.02,
min_charge = 1, max_charge = 10,
min_mz = 500, max_mz = 1200, min_mw = 0, max_mw = 7000,
mw_gap = 1.1, mw_window = 6)
SCAN_NMS = scan.deconvoluted$scan_processed_aggregated
```

```
seq = "OH-Am*-Af*-Cm*-Af-Um-Uf-Gm-Af-Gm-Cf-Gm-Af-Um-Gf-Um-Cf-Cm-Am*-Cm-OH"
mDB = predict_esi_frag(seq) # Prediction of product ions based on sequence

scan.deconvoluted.annotated = annotate_scan_mix(SCAN_NMS, MSMS = T, ntheo = 6,
         formula_flp = "",  cpd_flp = "", transformation_list = NULL, mdb = mDB,
       bblock= "RNA", min_overlap = 0.4, max_msigma = 20, max_mmw_ppm = 10, baseline = 50)

view(scan.deconvoluted.annotated$feature)


## End(Not run)
```

`annotate_scan_targeted` *Annotating a deconvoluted oligonucleotide spectra*

Description

The function searches a complex deconvoluted oligonucleotide spectra against a user provided oligonucleotide impurity/metabolite database, annotating and scoring isotopic pattern matches. Alternatively, user can provide the FLP (full length product) formula and a list of bio or chemical transformations.

Usage

```
annotate_scan_targeted(
  scan_processed_aggregated = NULL,
  formula_flp = "C192H239O117N73P18S4F8",
  cpd_flp = "Demo A",
  transformation_list = NULL,
  mdb = NULL,
  ntheo = 12,
  min_overlap = 0.6,
  max_msigma = 3,
  max_mmw_ppm = 10,
  baseline = 1000
)
```

Arguments

`scan_processed_aggregated`
: Data frame representing deconvoluted NMS with the true molecular weight scale. Output of the function process_scan.

`formula_flp`
: Character. Neutral elemental formula of the main compound (e.g. full length product). Used when a transformation list from the main compound is defined.

`cpd_flp`
: Character. Name of the main compound. Used to label compounds in the output table. Use the separator & if you expect multiple main compound in your sample.

`transformation_list`
: Data frame. Transformation list defining the mass and elemental formula difference from the main compound. Should contain following columns: ID, CPD (IDs and names of transformation products), Plus_Formula, Minus_Formula (Elemental formula difference), Delta.AVG.MW, Delta.MONO.MW (Mass difference for average and mono molecular weight)

`mdb`
: Data frame. You can directly provide all expected compounds to be annotated from your mixture without defining FLP or compound names.

`ntheo`
: Integer. Estimated isotope envelop size in number of isotope peaks.

`min_overlap`
: Double between 0 and 1. The minimum matching score between experimental and theoretical isotope envelops.

`max_msigma`
: Double between 1 and 50. The maximum-allowed deviation between the shapes of experimental and theoretical isotope pattern. Should set higher for noisy data.

`max_mmw_ppm`
: Double between 1 and 50. The maximum allowed ppm error between masses in the NMS and theoretical molecular weight of oligonucleotide features. Depend on experimental mass deviation and deconvolution bias.

`baseline`
: Numeric. Estimated baseline level (noise) of input spectrum. Depending on instrument and acquisition method. Baseline of MS/MS spectrum is 100 for most instruments.

Author(s)

Youzhong Liu, <liu-youzhong@hotmail.com>

Examples

```
## Not run:

## Example of MS1 data:

data("Strand_A")

scan.deconvoluted = process_scan(scan.A,
polarity = "Negative", baseline = 1000, mz_error = 0.01,
min_charge = 3, max_charge = 12,
min_mz = 500, max_mz = 1200, min_mw = 4000, max_mw = 10000,
mw_gap = 1.1, mw_window = 10)
SCAN_NMS = scan.deconvoluted$scan_processed_aggregated

data("TRANS") # Transformation list for potential oligonucleotide degradants

scan.deconvoluted.annotated = annotate_scan_targeted(SCAN_NMS, formula_flp = "C189H238O119N66P18S4F8",
cpd_flp = "Demo A", transformation_list = transformation_list, mdb = NULL,
ntheo = 10, min_overlap = 0.6, max_msigma = 5, max_mmw_ppm = 10, baseline = 1000)

head(scan.deconvoluted.annotated$feature)

## Example of MS2 data:

data("Strand_B_MS2")

scan.deconvoluted = process_scan(scan2_B, MSMS = T,
polarity = "Negative", baseline = 100, mz_error = 0.02,
min_charge = 1, max_charge = 10,
```

```
min_mz = 500, max_mz = 1200, min_mw = 0, max_mw = 7000,
mw_gap = 1.1, mw_window = 6)
SCAN_NMS = scan.deconvoluted$scan_processed_aggregated

seq = "OH-Am*-Af*-Cm*-Af-Um-Uf-Gm-Af-Gm-Cf-Gm-Af-Um-Gf-Um-Cf-Cm-Am*-Cm-OH"
mDB = predict_esi_frag(seq) # Prediction of product ions based on sequence

scan.deconvoluted.annotated = annotate_scan_targeted(SCAN_NMS, formula_flp = "",
cpd_flp = "", transformation_list = NULL, mdb = mDB,
ntheo = 6, min_overlap = 0.4, max_msigma = 20, max_mmw_ppm = 10, baseline = 50)


## End(Not run)
```

`annotate_scan_untargeted` *Non-targeted screening from a deconvoluted oligonucleotide spectra*

Description

The function searches DNA/RNA-like isotope patterns from a deconvoluted oligonucleotide spectra. It provides the monoisotopic molecular weight, average, intensity and envelope likeness of all features detected.

Usage

```
annotate_scan_untargeted(
  scan_processed_aggregated,
  bblock,
  ntheo = 12,
  min_overlap = 0.6,
  max_msigma = 10,
  max_mmw_ppm = 10,
  baseline = 1000
)
```

Arguments

`scan_processed_aggregated`	Data frame representing deconvoluted NMS with the true molecular weight scale. Output of the function process_scan.
`bblock`	Character. Either "DNA" or "RNA". Should reflect the main nucleic acid composition of the strand. Used for monoisotopic peak prediction by Pointless algorithm.
`ntheo`	Integer. Estimated isotope envelop size in number of isotope peaks.
`min_overlap`	Double between 0 and 1. The minimum matching score between experimental and theoretical isotope envelops (known compounds from database/transformation list or unknown predicted by Pointless).
`max_msigma`	Double between 1 and 50. The maximum-allowed deviation between the shapes of experimental and theoretical isotope pattern. Should set higher for noisy or MS/MS data.

max_mmw_ppm	Double between 1 and 50. The maximum allowed ppm error between masses in the NMS and theoretical molecular weight of oligonucleotide features. Depend on experimental mass deviation and deconvolution bias.
baseline	Numeric. Estimated baseline level (noise) of input spectrum. Depending on instrument and acquisition method. Baseline of MS/MS spectrum is 100 for most instruments.

Author(s)

Youzhong Liu, <liu-youzhong@hotmail.com>

Examples

```
## Not run:

## Example of MS1 data:

data("Strand_A")

scan.deconvoluted = process_scan(scan.A,
polarity = "Negative", baseline = 1000, mz_error = 0.01,
min_charge = 3, max_charge = 12,
min_mz = 500, max_mz = 1200, min_mw = 4000, max_mw = 10000,
mw_gap = 1.1, mw_window = 10)
SCAN_NMS = scan.deconvoluted$scan_processed_aggregated

scan.deconvoluted.annotated = annotate_scan_untargeted(SCAN_NMS, ntheo = 10,
bblock = "RNA", min_overlap = 0.6, max_msigma = 5, max_mmw_ppm = 10, baseline = 1000)

head(scan.deconvoluted.annotated$feature)

## End(Not run)
```

---

deconvolution1 *Deconvoluting a mixed envelop to two pure components*

Description

The function decomposes a mixed envelop into two theoritical isotope envelopes. It quantifies the two components by decomposing.

Usage

```
deconvolution1(
  scan_df,
  theor_ID_cmpd1,
  theor_ID_cmpd2,
  n_theor_peaks,
  expected_charge_range,
  matching_mass_accuracy,
  noise_threshold,
  deduplicate_fun
)
```

Arguments

scan_df	Data frame with two columns, mz and intensity, representing all peaks in a single mass scan at one selected retention time
n_theor_peaks	Numeric. How many theoretical peaks should be returned by BRAIN
expected_charge_range	Numeric, which charge state should be analysed? e.g. 5:11
matching_mass_accuracy	Numeric, relative (in Da) mass tolerance for assigning observed peaks to theoretical peaks
noise_threshold	Numeric, keep for the analysis only those peaks with intensities greater or equal than this threshold
deduplicate_fun	Character, "max" by default. How to deduplicate multiple observed peaks assigned to one theoretical peak - "max" or "sum"
ac_cmpd1	List used by the BRAIN algorithm to compute theoretical isotope distributions. The lists have to follow this format: list(C = 1, H = 2, F = 3, N = 4, O = 5, P = 6, S = 7)), where numbers 1-7 should be replaced with actual values
ac_cmpd2	Same as previous input, another expected theoritical envelop list

Value

- by_charge: list, with elements like z5, z6, ..., each storing estimate (estimated proportion of cmpd2, 0-1 scale), se (standard error of the proportion estimate), p_value (null hypothesis: estimate=0), significance (1 = significant, 0 = insignificant), mpcse (pearson chi-square goodness-of-fit, similar to the pointless4dna paper)
- by_charge: not yet available

Author(s)

Piotr Prostko, <piotr.prostko@uhasselt.be>

---

display_coverage *Display sequence coverage*

Description

The function returns a ggplot of labelled fragments and internal fragments on an oligonucleotide sequence

Usage

```
display_coverage(
  scan.deconvoluted.annotated,
  seq,
  int.frag = F,
  return.plot = T
)
```

Author(s)

Youzhong Liu, <YLiu186@ITS.JNJ.com>

`predict_esi_frag` *Prediction of ESI fragments second function*

Description

The function creates fragment ions for a sequence given with additional base loses and internal fragments

Usage

```
predict_esi_frag(test_seq = "OH-Ad-Ad-Ad-Ad-Ad-Ad-OH")
```

Author(s)

Youzhong Liu, <YLiu186@ITS.JNJ.com>

`predict_esi_frag_basic` *Prediction ESI fragments first function*

Description

The function creates fragment ions for a sequence given

Usage

```
predict_esi_frag_basic(test_seq = "OH-Ad-Ad-Ad-Ad-Ad-Ad-OH")
```

Author(s)

Youzhong Liu, <YLiu186@ITS.JNJ.com>

process_scan *Converting multi-charged oligonucleotide mass spectra to true molecular weight scale*

Description

The function determines charges state based on peak spacing. It also creates a deconvoluted molecular weight spectra by combining multiple charge state of the same isotopic species.

Usage

```
process_scan(
  test.scan = NULL,
  polarity = c("Positive", "Negative"),
  MSMS = F,
  baseline = 100,
  min_charge = 3,
  max_charge = 12,
  min_mz = 500,
  max_mz = 1500,
  min_mw = 4000,
  max_mw = 12000,
  mz_error = 0.02,
  mw_gap = 1.1,
  mw_window = 10
)
```

Arguments

test.scan	A two-column data frame, representing mass peak (m/z) and intensity of the input unprocessed spectrum.
polarity	Character. Either "Negative" or "Positive".
MSMS	Boolean. TRUE if the spectrum to be deconvoluted is a MS/MS spectrum.
baseline	Numeric. Estimated baseline level (noise) of input spectrum. Depending on instrument and acquisition method. Baseline of MS/MS spectrum is 100 for most instruments.
min_charge	Integer. Absolute minimum charge state of oligonucleotide species of interest in the spectrum. Should be at least 1.
max_charge	Integer. Absolute maximum charge state of oligonucleotide species of interest in the spectrum. Should be below 30 for our algorithm.
mz_error	Numeric. Estimated mass measurement error (Da).
mw_gap	Numeric. Estimated gap between closely-located oligonucleotide isotope envelops. The default value of 1.1 Da is adapted for most applications.
mw_window	Numeric. Estimated isotope envelop size (Da)
min_mz/max_mz	Numeric. Mass range of input spectrum to be deconvoluted. Zone with low spectrum quality should be excluded.
min_mw/max_mw	Numeric. Molecular weight range of output spectrum from deconvolution. Should cover compounds of interest. min_mw could be set as 0 for MS/MS spectra to cover small fragments.

Value

- scan_processed:Data frame of input spectrum along with charge state and neutral molecular weight estimation for each mass peak. z = 0 on a mass peak in the output table means that the charge state cannot be confidently assigned by our algorithm either because of its low intensity or its potential charge state is outside the user-specified range
- scan_processed_aggregated:Data frame representing NMS with the true molecular weight scale. NMS is obtained by aggregating multiple charge states of the same isotope peak. The mass of each peak in the NMS is the averaged NMW, and the intensity is the sum across all user-defined charge states.

Author(s)

Youzhong Liu, <liu-youzhong@hotmail.com>

Examples

```
## Not run:

data("Strand_A")

scan.deconvoluted = process_scan(scan.A,
polarity = "Negative", baseline = 1000, mz_error = 0.01,
min_charge = 3, max_charge = 12,
min_mz = 500, max_mz = 1200, min_mw = 4000, max_mw = 10000,
mw_gap = 1.1, mw_window = 10)

head(scan.deconvoluted$scan_processed_aggregated)
head(scan.deconvoluted$scan_processed)

## End(Not run)
```

---

process_scan_maldi *Converting maldi oligonucleotide mass spectra*

Description

The function processes and envelops maldi scan

Usage

```
process_scan_maldi(
  test.scan,
  polarity = c("Positive", "Negative"),
  baseline = 1000,
  min_mz = 100,
  max_mz = 2500,
  mw_gap = 1.1,
  mw_window = 10
)
```

Author(s)

Youzhong Liu, <YLiu186@ITS.JNJ.com>

`reconstruct_scan_annotated` *Reconstruct a theoretical mass spectra based on oligonucleotide features detected*

Description

The function provides a way to check the detected features against an original spectrum.

Usage

```
reconstruct_scan_annotated(
  scan0,
  scans.deconvoluted.annotated,
  polarity = "Negative",
  mode = c("targeted", "untargeted", "mixed"),
  mz_error = 0.01,
  input_charges = 5:12,
  bblock = "C21 H26.4 O13.2 N7.3 P2 S0.4 F0.9",
  ntheo = 12
)
```

Author(s)

Youzhong Liu, <YLiu186@ITS.JNJ.com>

Index