

Package ‘mSigAct’

December 3, 2022

Title Mutational Signature Activity Analysis ('mSigAct')

Version 2.3.4

Author Steve Rozen, Alvin Wei Tian Ng, Arnoud Boot, Nanhai Jiang

Maintainer Steve Rozen <steverozen@gmail.com>

Description Analyze the “activities” of mutational signatures in one or more mutational spectra. 'mSigAct' stands for mutational Signature Activity. mSigAct uses a maximum likelihood approach to estimate (conservatively) whether there is evidence that a particular set of mutational signatures is present in a spectrum. It can also determine a *minimal* subset of signatures needed to plausibly reconstruct an observed spectrum. This sparse assign signatures functionality is *deliberately biased* toward using as few signatures as possible. There is also functionality to do a maximum a posteriori estimate of signature activity, which makes use of information on the proportion of tumors in a given type that have a particular signature combined with the likelihood that a particular combination of signatures generated an observed spectrum.

License GPL-3

URL <https://github.com/steverozen/mSigAct>

BugReports <https://github.com/steverozen/mSigAct/issues>

Encoding UTF-8

Language en-US

Depends R (>= 4.0),

RoxygenNote 7.2.2

biocViews

Imports cosmicsig,
dplyr,
gtools,
ICAMS (>= 3.0.6),
nloptr,
mSigTools,
PCAWG7 (>= 0.1.0.9003),
philentropy,
quadprog,
stats,
sets,
tibble,
utils

Remotes github::steverozen/ICAMS@v3.0.6-branch,
github::steverozen/PCAWG7@v0.1.3-branch

Suggests BSgenome.Hsapiens.1000genomes.hs37d5,
devtools,
htmlwidgets,
ipc,
knitr,
profvis,
rmarkdown,
testthat (>= 2.1.0),
usethis

R topics documented:

cossim	2
DefaultManyOpts	3
deprecated_SparseAssignActivity	3
ExposureProportions	6
MAPAssignActivity	7
PresenceAssignActivity	9
ReconstructSpectrum	12
SignaturePresenceTest	13

Index	15
--------------	-----------

cossim	<i>Cosine similarity with useful argument types</i>
--------	---

Description

Cosine similarity with useful argument types

Usage

cossim(v1, v2)

Arguments

- v1 A vector or single-column matrix
- v2 A vector or single-column matrix

Examples

```
spectrum <- PCAWG7::spectra$PCAWG$SBS96[, 1, drop = FALSE]
SBS96.sigs <- cosmicSig::COSMIC_v3.2$signature$GRCh37$SBS96
exposure <- PCAWG7::exposure$PCAWG$SBS96[, 1, drop = FALSE]
reconstructed.spectrum <- ReconstructSpectrum(sigs = SBS96.sigs,
                                              exp = exposure,
                                              use.sig.names = TRUE)
cosine <- cossim(spectrum, reconstructed.spectrum)
```

DefaultManyOpts	<i>Set default options for many functions, especially nloptr</i>
-----------------	--

Description

Set default options for many functions, especially [nloptr](#)

Usage

```
DefaultManyOpts(likelihood.dist = "multinom")
```

Arguments

`likelihood.dist`

The probability distribution used to calculate the likelihood, can be either "multinom" (multinomial distribution) or "neg.binom" (negative binomial distribution).

Value

A list with the following elements

global.opts A sub-list with several options for [nloptr](#), q.v., for the global optimization phase.

local.opts A sub-list with several options for [nloptr](#), q.v., for the local optimization phase.

nbinom.size Only appearing if `likelihood.dist = "neg.binom"`. The dispersion parameter for the negative binomial distribution; smaller is more dispersed. See [NegBinomial](#).

trace If > 0 print progress messages.

global_eval_f The objective function for the global optimization phase.

local_eval_f The objective function for the local optimization phase.

local_eval_g_ineq The inequality constraint function for the local optimization phase.

likelihood.dist The probability distribution used to calculate the likelihood.

Examples

```
my.opts <- DefaultManyOpts()
my.opts$trace <- 10
```

`deprecated_SparseAssignActivity`

Find known signatures that can most sparsely reconstruct each spectrum in a catalog.

Description

Find known signatures that can most sparsely reconstruct each spectrum in a catalog.

Usage

```

deprecated_SparseAssignActivity(
  spectra,
  sigs,
  output.dir,
  max.level = 5,
  p.thresh = 0.05,
  m.opts = DefaultManyOpts(),
  num.parallel.samples = 5,
  mc.cores.per.sample = min(20, 2^max.level),
  progress.monitor = NULL,
  seed = NULL,
  max.subsets = 1000,
  drop.low.mut.samples = TRUE,
  use.sig.presence.test = FALSE,
  sig.pres.test.nbinom.size = NULL,
  sig.pres.test.p.thresh = 0.05,
  sig.pres.test.q.thresh = NULL
)

```

Arguments

<code>spectra</code>	The spectra (multiple spectra) to be reconstructed.
<code>sigs</code>	A numerical matrix, possibly an ICAMS catalog.
<code>output.dir</code>	Directory path to save the output file.
<code>max.level</code>	The maximum number of signatures to try removing.
<code>p.thresh</code>	If the p value for a better reconstruction with a set of signatures (as opposed to without that set of signatures) is > than this argument, then we can use exposures without this set.
<code>m.opts</code>	See DefaultManyOpts .
<code>num.parallel.samples</code>	The (maximum) number of samples to run in parallel. On Microsoft Windows machines it is silently changed to 1. Each sample in turn can require multiple cores, as governed by <code>mc.cores.per.sample</code> .
<code>mc.cores.per.sample</code>	The maximum number of cores to use for each sample. On Microsoft Windows machines it is silently changed to 1.
<code>progress.monitor</code>	Function called at the start of each new level (number of signatures to try excluding). Must take named arguments <code>value</code> and <code>detail</code> , and no others. Designed for a AsyncProgress progress bar function.
<code>seed</code>	Random seed; set this to get reproducible results. (The numerical optimization is in two phases; the first, global phase might rarely find different optima depending on the random seed.)
<code>max.subsets</code>	This argument provides a way to heuristically limit the amount of time spent by this function. Larger values of this argument will tend to allow longer running times. The algorithm successively tries to remove all subsets of 1 signature, 2 signatures, 3 signatures, etc., down to <code>max.level</code> . (Not every subset is tested at each level; if a subset was already found to be necessary the algorithm does

not test supersets of that subset.) If at any level the algorithm needs to test more than `max.subsets` this function will not proceed.

`drop.low.mut.samples`

Whether to exclude low mutation samples from the analysis. If `TRUE`, samples with SBS total mutations less than 100, DBS or ID total mutations less than 25 will be dropped.

`use.sig.presence.test`

Whether to use signature presence test first to filter out those signatures that are not needed in the reconstruction of the spectrum.

`sig.pres.test.nbinom.size`

Only needed when `use.sig.presence.test = TRUE`. The dispersion parameter for the negative binomial distribution used when conducting signature presence test; smaller is more dispersed. See [NegBinomial](#). If `NULL`, then use multinomial likelihood when conducting signature presence test.

`sig.pres.test.p.thresh`

Only needed when `use.sig.presence.test = TRUE`. If the p-value of signature presence test for one signature is \geq `sig.pres.test.p.thresh`, then this signature will not be used for assignment later.

`sig.pres.test.q.thresh`

Test parameter.

Value

A list with the elements:

- `proposed.assignment`: The most sparse set of signatures that can plausibly explain spectra.
- `proposed.reconstruction`: The reconstruction based on sparse assignment.
- `reconstruction.distances`: Various distances and similarities between spectra and `proposed.reconstruction`.
- `all.tested`: All tested possible ways to reconstruct each sample in spectra.
- `alt.solutions`: A tibble showing all the alternative solutions that are statistically as good as the `proposed.assignment` that can plausibly reconstruct spectra.
- `time.for.assignment`: Value from `system.time` for running `SparseAssignActivity` for each sample in spectra.
- `error.messages`: Only appearing if there are errors running `SparseAssignActivity`.

The elements `proposed.assignment`, `proposed.reconstruction`, `reconstruction.distances`, `all.tested`, `time.for.assignment` will be `NULL` if the algorithm could not find the optimal reconstruction or there are errors coming out for **all** samples.

Examples

```
## Not run:
# This is a long running example unless parallel computing is supported on your machine
indices <- grep("Lung-AdenoCA", colnames(PCAWG7::spectra$PCAWG$SBS96))
spectra <- PCAWG7::spectra$PCAWG$SBS96[, indices[1:2], drop = FALSE]
SBS96.sigs <- cosmic::COSMIC_v3.2$signature$GRCh37$SBS96
sigs.prop <- ExposureProportions(mutation.type = "SBS96",
                                cancer.type = "Lung-AdenoCA")
sigs.to.use <- SBS96.sigs[, names(sigs.prop), drop = FALSE]
sparse.out <- SparseAssignActivity(spectra = spectra,
```

```

sigs = sigs.to.use,
output.dir = file.path(tempdir(), "Lung-AdenoCA"),
max.level = ncol(sigs.to.use) - 1,
p.thresh = 0.05 / ncol(sigs.to.use),
num.parallel.samples = 2,
mc.cores.per.sample = 30,
seed = 2561)

## End(Not run)

```

ExposureProportions

Return the proportions of tumors of a given cancer type that have a particular signature

Description

Return the proportions of tumors of a given cancer type that have a particular signature

Usage

```

ExposureProportions(
  mutation.type,
  cancer.type,
  all.sigs = NULL,
  drop.sigs.no.info = TRUE,
  must.include = character(),
  must.include.prop = 0.1
)

```

Arguments

<code>mutation.type</code>	A character string, one of "SBS96", "SBS192", "ID", "DBS78".
<code>cancer.type</code>	A character string. For some common cancer types, see CancerTypes for more details.
<code>all.sigs</code>	An optional matrix of known signatures, with column names being signature ids. Only used to drop signatures not present in <code>all.sigs</code> .
<code>drop.sigs.no.info</code>	If TRUE, drop signatures not present in the column names of <code>all.sigs</code> .
<code>must.include</code>	A character vector of signature IDs that must be included, even if they have not previously been observed in that cancer type. The associated proportion is specified by <code>must.include.prop</code> .
<code>must.include.prop</code>	The value used for the expected proportion of signatures in <code>must.include</code> but not previously observed in the given <code>cancer.type</code> .

Value

A numerical vector of the proportion of tumors of type `cancer.type` with each signature for those signatures observed in `cancer.type`. The names are the signature ids.

Examples

```
cancer.types <- PCAWG7::CancerTypes()
cancer.types
sigs.prop <- ExposureProportions(mutation.type = "SBS96",
                                cancer.type = "Lung-AdenoCA")
```

MAPAssignActivity *Find Maximum A Posteriori (MAP) assignment of signature exposures that explain multiple spectra*

Description

This function also can do sparse assignment by specifying `use.sparse.assign = TRUE`.

Usage

```
MAPAssignActivity(
  spectra,
  sigs,
  sigs.presence.prop,
  output.dir,
  max.level = 5,
  p.thresh = 0.05,
  m.opts = DefaultManyOpts(),
  num.parallel.samples = 5,
  mc.cores.per.sample = min(20, 2^max.level),
  progress.monitor = NULL,
  seed = NULL,
  max.subsets = 1000,
  use.sparse.assign = FALSE,
  drop.low.mut.samples = TRUE,
  use.sig.presence.test = FALSE,
  sig.pres.test.nbinom.size = NULL,
  sig.pres.test.p.thresh = 0.05,
  sig.pres.test.q.thresh = NULL,
  save.files = TRUE
)
```

Arguments

<code>spectra</code>	The spectra (multiple spectra) to be reconstructed.
<code>sigs</code>	A numerical matrix, possibly an ICAMS catalog.
<code>sigs.presence.prop</code>	The proportions of samples that contain each signature. A numerical vector (values between 0 and 1), with names being a subset of <code>colnames(sigs)</code> . See ExposureProportions for more details.
<code>output.dir</code>	Directory path to save the output file.
<code>max.level</code>	The maximum number of signatures to try removing.

<code>p.thresh</code>	If the p value for a better reconstruction with a set of signatures (as opposed to without that set of signatures) is > than this argument, then we can use exposures without this set.
<code>m.opts</code>	See DefaultManyOpts .
<code>num.parallel.samples</code>	The (maximum) number of samples to run in parallel. On Microsoft Windows machines it is silently changed to 1. Each sample in turn can require multiple cores, as governed by <code>mc.cores.per.sample</code> .
<code>mc.cores.per.sample</code>	The maximum number of cores to use for each sample. On Microsoft Windows machines it is silently changed to 1.
<code>progress.monitor</code>	Function called at the start of each new level (number of signatures to try excluding). Must take named arguments <code>value</code> and <code>detail</code> , and no others. Designed for a AsyncProgress progress bar function.
<code>seed</code>	Random seed; set this to get reproducible results. (The numerical optimization is in two phases; the first, global phase might rarely find different optima depending on the random seed.)
<code>max.subsets</code>	This argument provides a way to heuristically limit the amount of time spent by this function. Larger values of this argument will tend to allow longer running times. The algorithm successively tries to remove all subsets of 1 signature, 2 signatures, 3 signatures, etc., down to <code>max.level</code> . (Not every subset is tested at each level; if a subset was already found to be necessary the algorithm does not test supersets of that subset.) If at any level the algorithm needs to test more than <code>max.subsets</code> this function will not proceed.
<code>use.sparse.assign</code>	Whether to use sparse assignment. If <code>TRUE</code> , arguments designed for Maximum A Posteriori assignment such as <code>sigs.presence.prop</code> will be ignored.
<code>drop.low.mut.samples</code>	Whether to exclude low mutation samples from the analysis. If <code>TRUE</code> , samples with SBS total mutations less than 100, DBS or ID total mutations less than 25 will be dropped.
<code>use.sig.presence.test</code>	Whether to use signature presence test first to filter out those signatures that are not needed in the reconstruction of the spectrum.
<code>sig.pres.test.nbinom.size</code>	Only needed when <code>use.sig.presence.test = TRUE</code> . The dispersion parameter for the negative binomial distribution used when conducting signature presence test; smaller is more dispersed. See NegBinomial . If <code>NULL</code> , then use multinomial likelihood when conducting signature presence test.
<code>sig.pres.test.p.thresh</code>	Only needed when <code>use.sig.presence.test = TRUE</code> . If the p-value of signature presence test for one signature is \geq <code>sig.pres.test.p.thresh</code> , then this signature will not be used for assignment later.
<code>sig.pres.test.q.thresh</code>	Test parameter.
<code>save.files</code>	If <code>TRUE</code> save several files for each input sample in a directory named after the sample.

Value

A list with the elements:

- `proposed.assignment`: Proposed signature assignment for spectra with the highest MAP found. If `use.sparse.assign = TRUE`, this will be the most sparse set of signatures that can plausibly explain spectra.
- `proposed.reconstruction`: Proposed reconstruction of spectra based on MAP. If `use.sparse.assign = TRUE`, this will be the reconstruction based on sparse assignment.
- `reconstruction.distances`: Various distances and similarities between spectra and `proposed.reconstruction`.
- `all.tested`: All tested possible ways to reconstruct each sample in spectra.
- `alt.solutions`: A tibble showing all the alternative solutions that are statistically as good as the `proposed.assignment` that can plausibly reconstruct spectra.
- `time.for.assignment`: Value from `system.time` for running `MAPAssignActivity1` for each sample in spectra.
- `error.messages`: Only appearing if there are errors running `MAPAssignActivity`.

The elements `proposed.assignment`, `proposed.reconstruction`, `reconstruction.distances`, `all.tested`, `time.for.assignment` will be `NULL` if the algorithm could not find the optimal reconstruction or there are errors coming out for **all** samples.

Examples

```
## Not run:
# This is a long running example unless parallel computing is supported on your machine
indices <- grep("Lung-AdenoCA", colnames(PCAWG7::spectra$PCAWG$SBS96))
spectra <- PCAWG7::spectra$PCAWG$SBS96[, indices[1:2], drop = FALSE]
SBS96.sigs <- cosmicSig::COSMIC_v3.2$signature$GRCh37$SBS96
sigs.prop <- ExposureProportions(mutation.type = "SBS96",
                                cancer.type = "Lung-AdenoCA")
sigs.to.use <- SBS96.sigs[, names(sigs.prop), drop = FALSE]
MAP.out <- MAPAssignActivity(spectra = spectra,
                            sigs = sigs.to.use,
                            sigs.presence.prop = sigs.prop,
                            output.dir = file.path(tempdir(), "Lung-AdenoCA"),
                            max.level = ncol(sigs.to.use) - 1,
                            p.thresh = 0.05 / ncol(sigs.to.use),
                            num.parallel.samples = 2,
                            mc.cores.per.sample = 30,
                            seed = 2561)

## End(Not run)
```

PresenceAssignActivity

Find minimal set of signatures that can explain multiple spectra by first using signature presence test

Description

Find minimal set of signatures that can explain multiple spectra by first using signature presence test

Usage

```

PresenceAssignActivity(
    spectra,
    sigs,
    output.dir,
    max.level = 5,
    p.thresh = 0.05,
    m.opts = DefaultManyOpts(),
    num.parallel.samples = 5,
    mc.cores.per.sample = min(20, 2^max.level),
    progress.monitor = NULL,
    seed = NULL,
    drop.low.mut.samples = TRUE,
    sig.pres.test.nbinom.size = NULL,
    sig.pres.test.p.thresh = 0.05,
    sig.pres.test.q.thresh = NULL,
    save.files = TRUE
)

```

Arguments

<code>spectra</code>	The spectra (multiple spectra) to be reconstructed.
<code>sigs</code>	A numerical matrix, possibly an ICAMS catalog.
<code>output.dir</code>	Directory path to save the output file.
<code>max.level</code>	The maximum number of signatures to try removing.
<code>p.thresh</code>	If the p value for a better reconstruction with a set of signatures (as opposed to without that set of signatures) is > than this argument, then we can use exposures without this set.
<code>m.opts</code>	See DefaultManyOpts .
<code>num.parallel.samples</code>	The (maximum) number of samples to run in parallel. On Microsoft Windows machines it is silently changed to 1. Each sample in turn can require multiple cores, as governed by <code>mc.cores.per.sample</code> .
<code>mc.cores.per.sample</code>	The maximum number of cores to use for each sample. On Microsoft Windows machines it is silently changed to 1.
<code>progress.monitor</code>	Function called at the start of each new level (number of signatures to try excluding). Must take named arguments <code>value</code> and <code>detail</code> , and no others. Designed for a AsyncProgress progress bar function.
<code>seed</code>	Random seed; set this to get reproducible results. (The numerical optimization is in two phases; the first, global phase might rarely find different optima depending on the random seed.)
<code>drop.low.mut.samples</code>	Whether to exclude low mutation samples from the analysis. If <code>TRUE</code> , samples with SBS total mutations less than 100, DBS or ID total mutations less than 25 will be dropped.
<code>sig.pres.test.nbinom.size</code>	The dispersion parameter for the negative binomial distribution used when conducting signature presence test first to filter out those signatures that are not

SignaturePresenceTest

Test whether a given signature is plausibly present in a catalog.

Description

Test whether a given signature is plausibly present in a catalog.

Usage

```
SignaturePresenceTest (
  spectra,
  sigs,
  target.sig.index,
  m.opts = DefaultManyOpts(),
  seed = NULL,
  mc.cores = 2
)
```

Arguments

spectra	The catalog (matrix) to analyze. This could be an ICAMS catalog or a numerical matrix.
sigs	A catalog of signatures from which to choose. This could be an ICAMS catalog or a numerical matrix.
target.sig.index	The index of the signature the presence of which we want to test. It can also be the signature id (e.g. "SBS22").
m.opts	See DefaultManyOpts .
seed	Random seed; set this to get reproducible results. (The numerical optimization is in two phases; the first, global phase might rarely find different optima depending on the random seed.)
mc.cores	Number of cores to use. Always silently changed to 1 on Microsoft Windows.

Value

A list of test results for each sample in `spectra`. Each sublist contains the following elements:

- `loglh.with`: The maximum log likelihood of the reconstructed spectrum using all the signatures.
- `loglh.without`: The maximum log likelihood of the reconstructed spectrum without the target signature.
- `statistic`: Likelihood ratio test statistic.
- `chisq.p`: P-value of the likelihood ratio test. The null hypothesis is we can plausibly reconstruct the spectrum without the target signature.
- `exp.with`: The exposure using all the signatures which generates the maximum log likelihood `loglh.with`.
- `exp.without`: The exposure not using the target signature which generates the maximum log likelihood `loglh.without`.

Examples

[illegible]

Index

AsyncProgress, [4](#), [8](#), [10](#)

CancerTypes, [6](#)

coessim, [2](#)

DefaultManyOpts, [3](#), [4](#), [8](#), [10](#), [13](#)

deprecated_SparseAssignActivity,
[3](#)

ExposureProportions, [6](#), [7](#)

ICAMS, [4](#), [7](#), [10](#), [13](#)

MAPAssignActivity, [7](#)

NegBinomial, [3](#), [5](#), [8](#), [11](#)

nloptr, [3](#)

PresenceAssignActivity, [9](#)

ReconstructSpectrum, [12](#)

SignaturePresenceTest, [13](#)