Package 'melonnpan'

February 14, 2023

Title Model-based Genomically Informed High-dimensional Predictor of Microbial Community Metabolite Profiles

Version 0.99.0

Date 2019-05-21

Description MelonnPan is a computational method for predicting metabolite compositions from microbiome sequencing data. It uses elastic net regularization to infer which sequencing features are predictive and combines these features to estimate the composite metabolome.

Depends R (>= 3.6.0), plyr, dplyr, glmnet, foreach, getopt, doParallel, vegan, data.table, ggplot2, AssocTests, optparse, tibble

Suggests knitr, rmarkdown

VignetteBuilder knitr

License MIT

LazyData true

Encoding UTF-8

RoxygenNote 7.1.1

biocViews Metagenomics, Software, Metabolomics

R topics documented:

ArcSin *Arc Sine Square Root Transformation for Proportional Data*

Description

This function applies arcsine square root transformation to proportional data.

Usage

```
ArcSin(x)
```

Arguments

x	A numerical vector of proportions (must be between 0 and 1).

See Also

SqSin

Examples

```
ArcSin(runif(100,0,1))
```

CV.ENET *Fit a regularized linear model*

Description

Fit a regularized linear model

Usage

```
CV.ENET(
  metab = metab,
  metag = metag,
  alpha = alpha,
  lambda.choice = lambda.choice,
  nfolds = nfolds,
  foldid = foldid,
  verbose = verbose,
  plot = plot,
  outputDirectory = outputDirectory
)
```

Arguments

metab	Training data of metabolite relative abundances. Should have the exact same rows (subjects/samples) as metag.
metag	Training data of microbial sequence features' relative abundances. Should have the exact same rows (subjects/samples) as metab.
alpha	Grid of alpha values between 0 and 1. Default is 'seq(0.05, 0.95, 0.05)'.
lambda.choice	Choice of optimal lambda ('lambda.min' or 'lambda.1se'). Default is 'lambda.1se'.
nfolds	Number of folds for internal cross-validation. Default is 10.
foldid	A vector of values between 1 and nfold identifying what fold each observation is in.
verbose	Should progress bar be printed. Default is TRUE.
plot	Should CV error as a function of lambda be plotted. Default is FALSE.
outputDirectory	Name of the desired output directory.

---

generateLambdaPlot *Plot CV Error As A Function of Lambda*

Description

This function plots the CV error as a function of lambda from Elastic Net Training.

Usage

```
generateLambdaPlot(CV, x, alpha, outputDirectory)
```

Arguments

CV	A glmnet object.
x	Index of glmnet fit.
alpha	Optimal alpha selected.
outputDirectory	Name of the desired output directory. Default is the current directory.

---

melonnpan.predict *Model-based Genomically Informed High-dimensional Predictor of Microbial Community Metabolite Profiles*

Description

Predict metabolites from new microbiome samples.

Usage

```
melonnpan.predict(
  metag,
  output,
  no.transform.metab = FALSE,
  no.transform.metag = FALSE,
  weight.matrix = NULL,
  train.metag = NULL,
  criticalpoint = 0.9793,
  corr.method = "pearson"
)
```

Arguments

metag	Microbial sequence features' relative abundances (matrix) for which prediction is desired. The sequence features' abundances are expected to be normalized.
output	The output folder to write results.
no.transform.metab	Should back transformation be turned off for predicted metabolites? Default is FALSE. If FALSE, it is expected that the proportional data ranging from 0 to 1 was used for training.
no.transform.metag	Should rank-based inverse normal (RIN) transformation be turned off for 'metag'? Default is FALSE.
weight.matrix	The weight matrix to be used for prediction (optional). If not provided, by default, a pre-trained weight matrix based on UniRef90 gene families from the original MelonnPan paper (Mallick et al, 2019) will be used.
train.metag	Quality-controlled training metagenomes against which similarity is desired (optional). The sequence features' abundances are expected to be normalized. If not provided, a pre-processed UniRef90 gene family training table from the original MelonnPan paper (Mallick et al. 2019) will be used.
criticalpoint	A numeric value corresponding to the significance level to find the top PCs. If the significance level is 0.05, 0.01, 0.005, or 0.001, the criticalpoint should be set to be 0.9793, 2.0234, 2.4224, or 3.2724, accordingly. The default is 0.9793 (i.e. 0.05 significance level).
corr.method	Method to correlate new metagenomes and training PCs. Default is 'pearson'.

---

melonnpan.summarize *Summarize MelonnPan results*

Description

This function returns prediction accuracy from a pair of measured and predicted metabolites tables.

Usage

```
melonnpan.summarize(
  metab = metab,
  pred = pred,
  method = method,
  correction = correction,
  cutoff = cutoff
)
```

Arguments

metab	Measured metabolite relative abundances (normalized). Should have same features and samples as pred.
pred	Predicted metabolite relative abundances (normalized). Should have same features and samples as metab.
method	Method to correlate measured and predicted metabolites. Default is 'spearman'.
correction	Multiplicity adjustment method, same as 'p.adjust'. Default is 'fdr'.
cutoff	Q-value threshold for significant prediction. Default is 0.05.

---

melonnpan.test.data *Normalized gene family abundances in the NLIBD dataset (Franzosa et al., 2019).*

Description

Normalized gene family abundance used for MelonnPan validation (Mallick et al., 2019).

Usage

```
data(melonnpan.test.data)
```

Format

A data frame of normalized gene familty abundances, as described in Mallick et al. (2019), with subjects in rows and UniRef90 IDs in columns.

References

Franzosa EA et al. (2019). Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nature Microbiology 4(2):293-305.

Mallick H, Franzosa EA, McIver LJ, Banerjee S, Sirota-Madi A, Kostic AD, Clish CB, Vlamakis H, Xavier R, Huttenhower C (2019). Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. Nature Communications 10(1):3136-3146.

`melonnpan.train` *Model-based Genomically Informed High-dimensional Predictor of Microbial Community Metabolite Profiles*

Description

Predict metabolites from microbial sequence features' abundances. Both measurements are expected to be normalized before using MelonnPan.

Usage

```
melonnpan.train(
  metab,
  metag,
  output,
  alpha = seq(0.05, 0.95, 0.05),
  lambda.choice = "lambda.1se",
  nfolds = 10,
  correction = "fdr",
  method = "spearman",
  cores = 4,
  seed = 1234,
  cutoff = 0.05,
  verbose = TRUE,
  no.transform.metab = FALSE,
  no.transform.metag = FALSE,
  discard.poor.predictions = FALSE,
  plot = FALSE,
  outputString = c("MelonnPan_Training_Summary", "MelonnPan_Trained_Weights",
    "MelonnPan_Trained_Metabolites")
)
```

Arguments

`metab`	Training data of metabolite relative abundances (normalized). Must have the exact same rows (subjects/samples) as metag.
`metag`	Training data of metagenomics sequence features' relative abundances (normalized). Must have the exact same rows (subjects/samples) as metab.
`output`	The output folder to write results.
`alpha`	Grid of alpha values between 0 and 1. Default is 'seq(0.05, 0.95, 0.05)'.
`lambda.choice`	Choice of optimal lambda ('lambda.min' or 'lambda.1se'). Default is 'lambda.1se'.
`nfolds`	Number of folds for internal cross-validation. Default is 10.
`correction`	Multiplicity adjustment method, same as 'p.adjust'. Default is 'fdr'.
`method`	Method to correlate measured and predicted metabolites ('spearman' or 'pearson'). Default is 'spearman'.
`cores`	Number of cores to use for parallel processing. Default is 4.
`seed`	Specify the arbitrary seed value for reproducibility. Default is 1234.
`cutoff`	Q-value threshold for significant prediction. Default is 0.05.

`verbose` Should detalied message be printed. Default is TRUE.

`no.transform.metab`
Should arcsine square root transformation (AST) be turned off for 'metab'? Default is FALSE. If FALSE, 'metab' must be proportional data ranging from 0 to 1.

`no.transform.metag`
Should rank-based inverse normal (RIN) transformation be turned off for 'metag'? Default is FALSE.

`discard.poor.predictions`
Should predictions with model size = 1 be discarded? Default is FALSE.

`plot` Should CV error as a function of lambda be plotted. Default is FALSE.

`outputString` Names of the three output files. Default is 'MelonnPan_Training_Summary', 'MelonnPan_Trained_Weights', and 'MelonnPan_Trained_Metabolites'.

`melonnpan.trained.model`
Weights from a pre-trained MelonnPan model in the PRISM dataset (Franzosa et al., 2019).

Description

Weight matrix (or model coefficients) based on MelonnPan training in the PRISM dataset (Franzosa et al., 2019).

Usage

```
data(melonnpan.trained.model)
```

Format

A data frame of model coefficients from a pre-trained MelonnPan model, as described in Mallick et al. (2019), with 814 UniRef90 IDs (excluding the intercept) and 80 well-predicted metabolites (unique, labeled).

References

Franzosa EA et al. (2019). Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nature Microbiology 4(2):293-305.

Mallick H, Franzosa EA, McIver LJ, Banerjee S, Sirota-Madi A, Kostic AD, Clish CB, Vlamakis H, Xavier R, Huttenhower C (2019). Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. Nature Communications 10(1):3136-3146.

`melonnpan.training.data` *Normalized gene family abundances in the PRISM dataset (Franzosa et al., 2019).*

Description

Normalized gene family abundance used for MelonnPan training (Mallick et al., 2019).

Usage

```
data(melonnpan.training.data)
```

Format

A data frame of normalized gene familty abundances, as described in Mallick et al. (2019), with subjects in rows and UniRef90 IDs in columns.

References

Franzosa EA et al. (2019). Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nature Microbiology 4(2):293-305.

Mallick H, Franzosa EA, McIver LJ, Banerjee S, Sirota-Madi A, Kostic AD, Clish CB, Vlamakis H, Xavier R, Huttenhower C (2019). Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. Nature Communications 10(1):3136-3146.

`melonnpan.visualize` *Visualization of Significant Results.*

Description

This function produces scatter plots for significantly predicted compounds.

This function produces scatter plots for significantly predicted compounds.

Usage

```
melonnpan.visualize(
  metab,
  pred,
  cohenCorrelationCutoff,
  Output_file,
  ncol = 2,
  nrow = 2
)

melonnpan.visualize(
  metab,
  pred,
  cohenCorrelationCutoff,
```

```
  Output_file,
  ncol = 2,
  nrow = 2
)
```

Arguments

metab	Measured normalized metabolite relative abundances. Should have same features and samples as pred.
pred	Predicted normalized metabolite relative abundances. Should have same features and samples as metab.
cohenCorrelationCutoff	Correlation threshold for significant prediction. Default is 0.3.
Output_file	Path to the file to write the output.
ncol	Number of columns in batch scatter plot. Default is 2.
nrow	Number of rows in batch scatter plot. Default is 2.

---

readTable *Read data using a wrapper of data.table*

Description

This function reads the input data using the data.table package but returns a matrix with row names (unlike fread).

Usage

```
readTable(Input)
```

Arguments

Input	Full path of the input data.

---

rntransform *Rank-based inverse normal tranformation*

Description

This function rank normal transforms a vector of data. The procedure is built off of that provided in the GenABEL pacakge.

Usage

```
rntransform(y, split_ties = FALSE)
```

Arguments

y	a numeric vector which will be rank normal transformed
split_ties	a binary string of FALSE (default) or TRUE indicating if tied values, of the same rank, should be randomly split giving them unique ranks.

Value

returns a numeric vector, with the same length as y, of rank normal transformed values

Examples

```
## simulate a negative binomial distribution of values
nb_data = rnbinom(500, mu = 40, size = 100)
## rank normal transform those values
rnt_data = rntransform( nb_data , split_ties = TRUE )
```

SqSin *Back Transformation of Predicted Values from A Statistical Model with ArcSine Square Root Transformed Compositional Response Variable*

Description

This function applies back transformation to the predicted values of a statistical model with arcsine square root transformed compositional response variable.

Usage

```
SqSin(x, adjust = TRUE)
```

Arguments

x	A numerical vector of arcsine square root transformed compositions or proportions.
adjust	Should values less than 0 be mapped to 0 and values greater than pi/2 be mapped to 1? Defaults to TRUE.

See Also

ArcSin

Examples

```
# This is a simple demonstration
y <- ArcSin(runif(100,0,1))
x <- rnorm(100,0,1)
z <- predict(lm(y ~ x))
SqSin(z)
```

Index