Package 'mosaic.find'

Type Package

Title Finding Rhythmic and Non-Rhythmic Trends in Multi-Omics Data (MOSAIC)

Version 0.1.1

Maintainer Jennifer Hurley <hurlej2@rpi.edu>

Description Provides a function (mosaic_find()) designed to find rhythmic and non-rhythmic trends in multi-omics time course data using model selection and joint modeling, a method called MOSAIC (Multi-Omics Selection with Amplitude Independent Criteria). For more information, see H. De los Santos et al. (2020) <doi:10.1101/2020.04.27.064147>.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

URL https://github.com/delosh653/MOSAIC

Depends R (>= 2.10)

Imports minpack.lm (>= 1.2.1)

RoxygenNote 6.1.1

Suggests knitr, rmarkdown, ggplot2

VignetteBuilder knitr

NeedsCompilation no

Author Hannah De los Santos [aut, cph],
Kristin Bennett [aut],
Jennifer Hurley [aut, cre],
R Development Core Team [aut]

Repository CRAN

Date/Publication 2020-06-10 19:50:10 UTC

R topics documented:

---

expressions_pro   *Synthetic protein expression data for 600 genes.*

Description

A dataset containing the names and protein expression values for 600 synthetically generated samples. This example data has time points from 2 to 48 hours with 2 hour resolution and 3 replicates. Random missing data is also included. Expression labels are the same as the corresponding protein dataset, expressions_rna. Synthetic data was created by randomly selecting parameters for the ECHO, ECHO Joint, ECHO Linear, ECHO Linear Joint, linear, and exponential models, then adding random noise to each expression, as described in H. De los Santos, et al. (2020). There is a 2:1 ratio of non-oscillatory to oscillatory models. See linked paper in vignette for more information.

Usage

```
expressions_pro
```

Format

A data frame with 600 rows and 73 variables (column 1: sample labels, columns to 2 to 73: numerical values for protein expression in the format TPX.Y (time point X, replicate Y)).

Details

Note the data format: its first column first column has gene labels/names, and all other columns have expression data. This expression data is ordered by time point then by replicate, and has evenly spaced time points. Any missing data has cells left blank (NA). Labels have same exact labels as corresponding RNA data.

---

expressions_rna   *Synthetic RNA expression data for 600 genes.*

Description

A dataset containing the names and RNA expression values for 600 synthetically generated samples. This example data has time points from 2 to 48 hours with 2 hour resolution and 3 replicates. Random missing data is also included. Expression labels are the same as the corresponding protein dataset, expressions_pro. Synthetic data was created by randomly selecting parameters for the ECHO, ECHO Joint, ECHO Linear, ECHO Linear Joint, linear, and exponential models, then adding random noise to each expression, as described in H. De los Santos, et al. (2020). There is a 2:1 ratio of non-oscillatory to oscillatory models. See linked paper in vignette for more information.

Usage

```
expressions_rna
```

Format

A data frame with 600 rows and 73 variables (column 1: sample labels, columns to 2 to 73: numerical values for RNA expression in the format TPX.Y (time point X, replicate Y)).

Details

Note the data format: its first column first column has gene labels/names, and all other columns have expression data. This expression data is ordered by time point then by replicate, and has evenly spaced time points. Any missing data has cells left blank (NA). Labels have same exact labels as corresponding protein data.

---

mosaic.find   *mosaic.find*

Description

Provides a function (mosaic_find()) designed to find rhythmic and non-rhythmic trends in multi-omics time course data using model selection and joint modeling, a method called MOSAIC (Multi-Omics Selection with Amplitude Independent Criteria).

Details

To read more about this project and cite us, see MOSAIC: A Joint Modeling Methodology for Combined Circadian and Non-Circadian Analysis of Multi-Omics Data by H. De los Santos et al. (2020) doi: 10.1101/2020.04.27.064147

---

mosaic_find   *MOSAIC (Multi-Omics Selection with Amplitude Independent Criteria)*

Description

Function to calculate the results for RNA and protein data using the MOSAIC (Multi-Omics Selection with Amplitude Independent Criteria) method, which uses joint modeling and model selection to find both oscillatory and non-oscillatory trends in time series data.

Usage

```
mosaic_find(rna, pro, begin, end, resol, num_reps, paired, low = 20,
  high = 28, run_all_per = F, rem_unexpr = F, rem_unexpr_amt = 70,
  rem_unexpr_amt_below = 0, is_normal = F, is_smooth = F,
  harm_cut = 0.03, over_cut = 0.15)
```

Arguments

rna	data frame of RNA expressions with the following specifications: first column has gene labels/names, and all other columns have expression data. This expression data must be ordered by time point then by replicate, and must have evenly spaced time points. Any missing data must have cells left blank (NA). Labels must have same exact labels as corresponding protein data. RNA expressions with no corresponding protein data will be removed and not calculated under the MOSAIC method.
pro	data frame of protein expressions with the following specifications: first column has gene labels/names, and all other columns have expression data. This expression data must be ordered by time point then by replicate, and must have evenly spaced time points. Any missing data must have cells left blank (NA). Labels must have same exact labels as corresponding RNA data. Protein expressions with no corresponding RNA data will be removed and not calculated under the MOSAIC method.
begin	first time point for dataset (in hours)
end	last time point for dataset (in hours)
resol	resolution of time points (in hours)
num_reps	number of replicates
paired	if replicate data, whether the replicates are related (paired) or not (unpaired)
low	lower limit when looking for rhythms, in hours. Will not be used if finding rhythms of any length within timecouse (run_all_per is TRUE).
high	upper limit when looking for rhythms, in hours. Will not be used if finding rhythms of any length within timecouse (run_all_per is TRUE).
run_all_per	boolean which indicates whether or not rhythms of any length within timecourse should be searched for.

rem_unexpr	boolean indicating whether genes with less than rem_unexpr_amt percent expression should not be considered
rem_unexpr_amt	percentage of expression for which genes should not be considered if rem_unexpr is TRUE
rem_unexpr_amt_below	
	cutoff for expression
is_normal	boolean that indicates whether data should be normalized or not
is_smooth	boolean that indicates whether data should be smoothed or not
harm_cut	postive number indicating the cutoff for a gene to be considered harmonic
over_cut	postive number indicating the cutoff for a gene to be considered repressed/overexpressed

Value

results, a data frame which contains:

Gene_Name	gene name
Best_Model_(RNA or Protein)	
	The selected model type for the RNA or protein data, from a choice of oscillatory (ECHO, ECHO Joint ECHO Linear, ECHO Linear Joint) and non-oscillatory (Linear, Exponential) models.
P_Value_(RNA or Protein)	
	Significance of MOSAIC fit for RNA or protein data, unadjusted.
BH_Adj_P_Value_(RNA or Protein)	
	Significance of MOSAIC fit for RNA or protein data, adjusted using the Benjamini-Hochberg criterion. Corrects for multiple hypothesis testing.
P_Value_Joint	Significance of MOSAIC joint fit if best model is ECHO Joint or ECHO Linear Joint, unadjusted.
BH_Adj_P_Value_Joint	
	Significance of MOSAIC joint fit if best model is ECHO Joint or ECHO Linear Joint, adjusted using the Benjamini-Hochberg criterion. Corrects for multiple hypothesis testing.
P_Value_Linear_Slope_(RNA or Protein)	
	Significance of linear slope of MOSAIC fit for RNA or protein if the best model is Linear, unadjusted.
BH_Adj_P_Value_Linear_Slope_(RNA or Protein)	
	Significance of linear slope of MOSAIC fit for RNA or protein if the best model is Linear, adjusted using the Benjamini-Hochberg criterion. Corrects for multiple hypothesis testing.
AC_Coefficient_(RNA or Protein)	
	Amplitude Change Coefficient for RNA or protein. Parameter which states the amount of amplitude change over time in the system. Used in ECHO, ECHO Joint, ECHO Linear, and ECHO Linear Joint models.
Oscillation_Type_(RNA or Protein)	
	States the expression's category based on forcing coefficient (forced, damped, harmonic) for RNA or protein. Used in ECHO, ECHO Joint, ECHO Linear, and ECHO Linear Joint models.

`Initial_Amplitude_(RNA or Protein)`

> Parameter describing initial amplitude of expression for RNA or protein data. Used in Exponential, ECHO, ECHO Joint, ECHO Linear, and ECHO Linear Joint models.

`Radian_Frequency_(RNA or Protein)`

> Parameter describing frequency of oscillations, in radians, for RNA or protein data. Used in ECHO, ECHO Joint, ECHO Linear, and ECHO Linear Joint models.

`Period_(RNA or Protein)`

> States the time for one complete oscillation, in hours, for RNA or protein. Used in ECHO, ECHO Joint, ECHO Linear, and ECHO Linear Joint models.

`Phase_Shift_(RNA or Protein)`

> Parameter describing the amount the oscillator is shifted, in radians, for RNA or protein. Used in ECHO, ECHO Joint, ECHO Linear, and ECHO Linear Joint models.

`Hours_Shifted_(RNA or Protein)`

> Desribes the amount the oscillator is shifted in hours, calculated from phase shift and fitted period, for RNA or protein. This is the time of the first peak of the oscillation, relative to 0 as determined by the time course entered by the user. Used in ECHO, ECHO Joint, ECHO Linear, and ECHO Linear Joint models.

`Growth_Rate_(RNA or Protein)`

> Parameter describing the exponential change in amplitude for RNA or protein. Used in Exponential models.

`Slope_(RNA or Protein)`

> Parameter describing the linear slope for RNA or protein. Used in Linear, ECHO Linear, and ECHO Linear Joint models.

`Equilibrium_Value_(RNA or Protein)`

> Parameter describing the center, i.e. the y-intercept at time point 0, as determined by the user supplied time course, for RNA or protein. Used in Linear, Exponential, ECHO, ECHO Joint, ECHO Linear, and ECHO Linear Joint models.

`Processed_(RNA or Protein)_(original data column name)`

> Your original data for RNA or protein, after any selected preprocessing, for specified time point (TP), using the same column names as original data.

`Fitted_(RNA or Protein)_TPX.R`

> MOSAIC's fitted data for RNA or protein for time point (TP) X, and replicate R.

Examples

```
# for more elaboration, please see the vignette
# "expressions_rna" is the example mosaic.find data frame for RNA
# "expressions_pro" is the example mosaic.find data frame for protein
# no preprocessing, looking for rhythms between 20 and 28 hours

mosaic_find(rna  = expressions_rna, pro = expressions_pro,
begin = 2, end = 48, resol = 2, num_reps = 3, paired = F,
low = 20, high = 28)
```

Index