Package ‘philr’

March 29, 2017

Type Package
Title Phylogenetic partitioning based ILR transform for metagenomics data
Version 1.0.0
Date August 26, 2016
Author Justin Silverman
Maintainer Justin Silverman <jsilve24@gmail.com>
Description PhILR is short for Phylogenetic Isometric Log-Ratio Transform. This package provides functions for the analysis of compositional data (e.g., data representing proportions of different variables/parts). Specifically this package allows analysis of compositional data where the parts can be related through a phylogenetic tree (as is common in microbiota survey data) and makes available the Isometric Log Ratio transform built from the phylogenetic tree and utilizing a weighted reference measure.
License GPL-3
LazyData TRUE
RoxygenNote 5.0.1
Imports ape, phangorn, tidyr, ggplot2, ggtree
Depends
Suggests testthat, knitr, rmarkdown, BiocStyle, phyloseq, glmnet, dplyr
VignetteBuilder knitr
biocViews Sequencing, Microbiome, Metagenomics, Software
URL https://github.com/jsilve24/philr
BugReports https://github.com/jsilve24/philr/issues
NeedsCompilation no

R topics documented:

annotate_balance .................................................. 2
buildIlrBasep .................................................. 3
calculate.blw .................................................. 4
clrp .......................................................... 5
annotate_balance

Description

annotate a balance oriented with respect to the PhILR transform. That is, you can specify labels for
the numerator (up) and denominator (down).

Usage

annotate_balance(tr, coord, p = NULL, labels = c("+", ",-"),
offset = 0,
offset.text = 0.03, bar = TRUE, barsize = 0.01, barfill = "darkgrey",
geom = "text", ...)

Arguments

  tr             phylo object
  coord          named internal node/balance to annotate
  p              ggtree plot (tree layer), if NULL then a new plot will be created.
  labels         label for the numerator and denominator of the balance respectively
  offset         offset for bar (if bar=TRUE) from tips
  offset.text    offset of text from bar (if bar=TRUE) or from tips (if bar=FALSE)
  bar            logical, should bar for each clade be plotted
  barsize        width of bar (if bar=TRUE)
  barfill        fill of bar
  geom           geom used to draw label (e.g., 'text' or 'label')
  ...            additional parameters passed to geom_rect and specified geom

Value

ggplot object

Author(s)

Justin Silverman
References


Examples

```r
tr <- named_rtree(10)
annotate_balance(tr, 'n4', size=7)
annotate_balance(tr, 'n4', size=7, barsize=0.04, barfill='darkgreen', offset.text=0.05, color='red')
annotate_balance(tr, 'n4', bar=FALSE, size=7)
annotate_balance(tr, 'n4', bar=TRUE, size=7, labels=c('Num', 'Denom'), offset.text=.3)
annotate_balance(tr, 'n4', bar=TRUE, geom='label', size=8, offset.text=0.1)
```

### buildilrBasep

**Weighted ILR Contrast Matrix**

**Description**

Weighted ILR Contrast Matrix

**Usage**

```r
buildilrBasep(W, p)
```

**Arguments**

- `W`: sequential binary partition matrix (e.g., signary matrix; output of `phylo2sbp`)
- `p`: weights (should not be closed)

**Value**

matrix

**Author(s)**

Justin Silverman (adapted from `gsi.buildilrBase`)

**References**


**Examples**

```r
p <- seq(.1,1,by=.2)
tr <- named_rtree(5)
sbp <- phylo2sbp(tr)
buildilrBasep(sbp, p)
```
calculate.blw  Calculate Branch Length Weightings for ILR Coordinates

Description
Calculates the weightings for ILR coordinates based on branch lengths of a phylogenetic tree via a few different methods (see details).

Usage
`calculate.blw(tree, method = "sum.children")`

Arguments
- `tree`: a phylo class tree object that is binary (see `multi2di`)
- `method`: options include: (default) "sum.children" and "mean.descendants" see details for more information.

Details
ILR balances built from a binary partition of a phylogenetic tree can be imbued with branch length information. This function is helpful in calculating those weightings.

There are a number of methods for calculating these weightings, the default "sum.children" calculates the weighting for a given balance as the sum of its two direct children's branch length. An alternative that has been as yet less studied is "mean.descendants" to calculate the weighting for a given balance as the sum of its two direct children's branch lengths PLUS for each child the average distance from it to its descendant tips.

Note: That some trees contain tips with branch lengths of zero length. This can result in that tip being unreasonably downweighted, as such this function automatically adds a small pseudocount to those tips with zero length (equal to the smallest non-zero) branch length on the tree.

Value
vector of weightings for ILR coordinates produced via specified method.

Author(s)
Justin Silverman

See Also
`philr`

Examples
```r
tr <- named_rtree(50)
calculate.blw(tr, method="sum.children")[1:10]
calculate.blw(tr, method="mean.descendants")[1:10]
```
**Description**

Weighted CLR Transform

**Usage**

\[ \text{clr}(y, p) \]

**Arguments**

- \( y \): shifted data matrix (e.g., output of \texttt{shiftp})
- \( p \): weights (should not be closed)

**Details**

Note that this function will close the dataset \( y \) to 1.

**Value**

matrix

**Author(s)**

Justin Silverman

**References**


**Examples**

```r
p <- seq(.1,1,by=.2)
c <- t(rmultinom(10,100,c(.1,.6,.2,.3,.2)) + 0.65) # add a small pseudocount
x <- miniclo(c)
y <- shiftp(x, p)
clr(y, p)
```
**convert_to_long**

*Converts wide format ILR transformed data to long format*

**Description**

Converts wide format ILR transformed data (see *philr*) to long format useful in various plotting functions where long format data is required.

**Usage**

```
convert_to_long(df, labels)
```

**Arguments**

- `df`: PhILR transformed data in wide format (samples by balances) (see *philr*)
- `labels`: vector (of length `nrow(df)`) with labels to group samples by

**Value**

`df` in long format with columns

- `sample`
- `labels`
- `coord`
- `value`

**Examples**

```r
tr <- named_rtree(5)
df <- t(rmultinom(10,100,c(.1,.6,.2,.3,.2))) + 0.65  # add a small pseudocount
colnames(df) <- tr$tip.label
df.philr <- philr(df, tr, part.weights="enorm.x.gm.counts",
ilr.weights="blw.sqrt", return.all=FALSE)
convert_to_long(df.philr, rep(c('a','b'), 5))
```

---

**g.colMeans**

*Geometric Means of Columns*

**Description**

Calculates geometric mean of columns. Does not calculate WEIGHTED geometric means (vs. *g.rowMeans*)

**Usage**

```
g.colMeans(x)
```
**g.rowMeans**

### Arguments

- `x` matrix or vector

### Value

vector (geometric mean of columns)

### See Also

- `g.rowMeans`

### Examples

```r
philr:::g.colMeans(rbind(c(2,4,4), c(2,4,4)))
```

---

**g.rowMeans**  
*Weighted Geometric Means of Rows*

### Description

Calculates weighted geometric mean (see references). Note if `p=rep(1, nrow(y))` (default) then this is just the geometric mean of rows.

### Usage

```r
g.rowMeans(y, p = rep(1, nrow(y)))
```

### Arguments

- `y` shifted data matrix (e.g., output of `shiftp`)
- `p` weights (should not be closed)

### Value

vector (weighted geometric mean of rows)

### References


### See Also

- `g.colMeans`

### Examples

```r
p <- seq(.1,1,by=.2)  
c <- t(rmultinom(10,100,c(.1,.6,.2,.3,.2))) + 0.65  # add a small pseudocount  
x <- miniclo(c)  
y <- shiftp(x, p)  
philr:::g.rowMeans(y, p)
```
Weighted ILR Transform

**Description**
Calculated using weighted CLR transform (clrp)

**Usage**
\[ ilrp(y, p, V) \]

**Arguments**
- \( y \): shifted data matrix (e.g., output of shiftp)
- \( p \): weights (should not be closed)
- \( V \): weighted contrast matrix (e.g., output of buildilrBasep)

**Value**
matrix

**Author(s)**
Justin Silverman

**References**

**Examples**
```r
# Weights
p <- seq(.1,1,by=.2)

# Shifted Composition
x <- miniclo(c)
y <- shiftp(x, p)

# Contrast Matrix
tr <- named_rtree(5)
sbp <- phylo2sbp(tr)
V <- buildilrBasep(sbp, p)

ilrp(y, p, V)
```
**mean_dist_to_tips**

Mean distance from internal nodes to descendant tips

**Description**

Calculates the mean distance from each internal node to its descendant tips

**Usage**

```r
mean_dist_to_tips(tree)
```

**Arguments**

- `tree` a phylo class tree object that is binary (see `multi2di`)

**Details**

This is a function used by `calculate.blw` when `method='mean.descendants'`, there this function is called twice, once for each direct child of a given internal node and the results are summed for each node.

**Value**

vector (named if internal nodes are named)

**Examples**

```r
tr <- named_rtree(5)
mean_dist_to_tips(tr)
```

**miniclo**

small function to close (aka normalize by proportions, aka total sum scaling) a dataset to a constant k (usually taken to be 1). After closure the row sums of the dataset should sum to k.

**Usage**

```r
miniclo(c, k = 1)
```

**Arguments**

- `c` dataset to be closed
- `k` closure constant

**Value**

matrix (if c is a vector or matrix) or data.frame (if c is a data.frame)
Examples

```r
c <- matrix(c(1,2,3,1,2,3,1,2,3), nrow = 3, byrow=TRUE)
miniclo(c)
miniclo(c, k=2)
```

---

### name.balance

**Name a balance (coordinate) based on taxonomy**

#### Description

For a given ILR balance (coordinate) assigns a name to the balance based on a provided taxonomy table. This is useful for interpretation of the balances.

#### Usage

```r
name.balance(tr, tax, coord, method = "voting", thresh = 0.95, return.votes = NULL)
```

#### Arguments

- `tr`:
  - an object of class 'phylo'

- `tax`:
  - a matrix/data.frame of taxonomy, rownames should correspond to `tr$tip.labels` columns should be taxonomic levels (named) with increasing taxonomic resolution from left to right (e.g., Phylum to the left of Genus).

- `coord`:
  - the name of a balance/internal node on the tree (given as a string)

- `method`:
  - currently only 'voting' implemented. See Details.

- `thresh`:
  - threshold for assignment of taxonomy to a given part of a balance (must be greater than 0.5 if `method='voting'`; see details).

- `return.votes`:
  - whether voting results by taxonomic level should be shown for `coord`. Note: this is helpful when `name.balance` does not return a clear winner, as may be the case when a given `coord` represents more than one taxonomic lineage. votes are returned as a list indexed by `colnames(tax)` Options include:
    - `NULL` (default) only returns the combined consensus name of the balance
    - 'up' adds tallied votes for the 'up' node to the output list
    - 'down' adds tallied votes for the 'down' node to the output list
    - 'self' adds tallied votes for `coord` to the output list

#### Details

A bit of terminology:

- `coord` this is the same as the names of the balances which should be the same as the names of the internal nodes of `tr`
- 'up' this is the child node of `coord` that is represented in the numerator of the `coord` balance.
- 'down' this is the child node of `coord` that is represented in the denominator of the `coord` balance

The method 'voting' assigns the name of the each part of a balance (e.g., numerator and denominator / each child of `coord`) as follows:
1. First Subset tax to contain only descendent tips of the given child of coord
2. Second At the finest taxonomic (farthest right of tax) see if any one taxonomic label is present at or above thresh. If yes output that taxonomic label (at that taxonomic level) as the label for that child of coord. If no then move to coarser taxonomic level (leftward) and repeat.

Value
If `return.votes=NULL` returns a string of the form (ex. 'Genus_Bacteroides/Phylum_Firmicutes'). Otherwise returns a list with the above string as 'name', see Arguments for `show.votes` for other optional returned items.

Author(s)
Justin Silverman

See Also
`philr`

Examples
```r
tr <- named_rtree(40)
tax <- data.frame(Kingdom=rep('A', 40),
                 Phylum=rep(c('B','C'), each=20),
                 Genus=c(sample(c('D','F'), 20, replace=TRUE),
                         sample(c('G','H'), 20, replace=TRUE)))
rownames(tax) <- tr$tip.label
name.balance(tr, tax, 'n1')
name.balance(tr, tax, 'n34')
name.balance(tr,tax, 'n34', return.votes = c('up', 'down'))
```
**name_nodenumber_conversion**

Convert between node/tip labels and integer node numbers

**Description**

Useful if you want to convert between node labels (c), tip labels (t) and the internal integer number that identifies that node (nn). Particularly for use with plotting libraries.

**Usage**

```r
nn.to.name(tr, x)
name.to.nn(tr, x)
```

**Arguments**

- `tr` object of type `phylo`
- `x` vector of numerics or characters

**Value**

vector

**Examples**

```r
tr <- named_rtree(5)
name.to.nn(tr, 'n1')
name.to.nn(tr,c('n1','n2','t1'))
nn.to.name(tr, 1:9)
```

**philr**

Data transformation and driver of PhILR.

**Description**

This is the main function for building the phylogenetic ILR basis, calculating the weightings (of the parts and the ILR coordinates) and then transforming the data.

**Usage**

```r
philr(df, tree, sbp = NULL, part.weights = "uniform",
      ilr.weights = "uniform", return.all = FALSE)
```
Arguments

- **df**: matrix of data to be transformed (samples are rows, compositional parts are columns) - zero must be dealt with either with pseudocount, multiplicative replacement, or another method.
- **tree**: a `phylo` class tree object that is binary (see `multi2di`)
- **sbp**: (Optional) give a precomputed sbp matrix `link(phylo2sbp)` if you are going to build multiple ILR bases (e.g., with different weightings).
- **part.weights**: weightings for parts, can be a named vector with names corresponding to `colnames(df)` otherwise can be a string, options include:
  - `'uniform'` (default) uses the uniform reference measure
  - `'gm.counts'` geometric mean of parts of df
  - `'anorm'` aitchison norm of parts of df (after closure)
  - `'anorm.x.gm.counts'` `'anorm' times `'gm.counts'`
  - `'enorm'` euclidean norm of parts of df (after closure)
  - `'enorm.x.gm.counts'` `'enorm' times `'gm.counts'`, often gives good results
- **ilr.weights**: weightings for the ILR coordinates can be a named vector with names corresponding to names of internal nodes of `tree` otherwise can be a string, options include:
  - `'uniform'` (default) no weighting of the ILR basis
  - `'blw'` sum of children’s branch lengths
  - `'blw.sqrt'` square root of `'blw'` option
  - `'mean.descendants'` sum of children’s branch lengths PLUS the sum of each child’s mean distance to its descendant tips
- **return.all**: return all computed parts (e.g., computed sign matrix(sbp), part weightings (codep), ilr weightings (codeilr.weights), contrast matrix (V)) as a list (default=FALSE) in addition to returning the transformed data (`df.ilrp`). If `return.all==FALSE` then only returns the transformed data (not in list format) If FALSE then just returns list containing `df.ilrp`.

Details

This is a utility function that pulls together a number of other functions in philr. The steps that are executed are as follows:

1. Create sbp (sign matrix) if not given
2. Create parts weightings if not given
3. Shift the dataset with respect to the new reference measure (e.g., part weightings)
4. Create the basis contrast matrix from the sign matrix and the reference measure
5. Transform the data based on the contrast matrix and the reference measure
6. Calculate the specified ILR weightings and multiply each balance by the corresponding weightings

Note for both the reference measure (part weightings) and the ILR weightings, specifying `'uniform'` will give the same results as not weighting at all.

Parallelization is done through `parallel` package using type "FORK". Note parallelization is rarely needed, even for trees of upwards of 40,000 leaves.
phylo2sbp

Create Sequential Binary Partition from Phylogenetic Tree

Description

This function converts a binary phylogenetic tree to sequential binary partition to be used to then build an ILR basis for compositional metagenomic data.

Usage

phylo2sbp(tr)

Arguments

tr

a phylo tree object with n leaves

Details

The choice of orientation for a balance (i.e., which of the two descendant clades of an internal node is in the numerator or denominator of the log-ratio) is given by the default of the function phangorn::Children and that choice is used consistently throughout the philr package.

Value

a n by n-1 matrix of the sequential binary partition sign matrix

Author(s)

Justin Silverman

References

Schliep K.P. 2011. phangorn: phylogenetic analysis in R. Bioinformatics, 27(4) 592-593
**shiftp**

**See Also**

philr

**Examples**

```r
tr <- named_rtree(5)
phylo2sbp(tr)
```

---

**shiftp**  
*Shift data to origin given by p*

**Description**

Shift must be applied before transformation

**Usage**

```
shiftp(x, p)
```

**Arguments**

- **x**: closed compositional data matrix (or vector)
- **p**: weights (should not be closed)

**Value**

shifted data matrix y (no closure is applied) rows are samples, columns are parts

**Author(s)**

Justin Silverman & J. J. Egozcue

**References**


**Examples**

```r
p <- seq(.1,1,by=.2)
c <- t(rmultinom(10,100,c(.1,.6,.2,.3,.2))) + 0.65  # add a small pseudocount
x <- miniclo(c)
shiftp(x, p)
```
Index

annotate_balance, 2
build.phylo.ilr (philr), 12
buildilrBasep, 3, 8

calc.blw (calculate.blw), 4
calculate.blw, 4, 9, 14
clrp, 5, 8
cvtColor_to_long, 6

g.colMeans, 6
g.rowMeans, 6, 7
gsi.buildilrBase, 3

ilrp, 8

mean_dist_to_tips, 9
miniclo, 9
multi2di, 4, 9, 13

name.balance, 10
name.to.nn
    (name_nodenumber_conversion),
    12
name_nodenumber_conversion, 12
named_rtree, 11
nn.to.name
    (name_nodenumber_conversion),
    12

philr, 4, 6, 11, 12, 15
phylo2sbp, 3, 14, 14

shiftp, 5, 7, 8, 15