Package 'ecodive'

August 22, 2025

Type Package

Title Parallel and Memory-Efficient Ecological Diversity Metrics

Version 1.0.0

Date 2025-08-18

Description Computes alpha and beta diversity metrics using concurrent 'C' threads.
Metrics include 'UniFrac', Faith's phylogenetic diversity, Bray-Curtis
dissimilarity, Shannon diversity index, and many others.
Also parses newick trees into 'phylo' objects and rarefies feature tables.

URL https://cmmr.github.io/ecodive/, https://github.com/cmmr/ecodive

BugReports https://github.com/cmmr/ecodive/issues

License MIT + file LICENSE

Encoding UTF-8

LazyData true

Depends R (>= 3.6.0)

RoxygenNote 7.3.2

Config/Needs/website rmarkdown

Config/testthat/edition 3

Imports parallel, utils

Suggests knitr, rmarkdown, testthat (>= 3.0.0)

VignetteBuilder knitr

NeedsCompilation yes

Author Daniel P. Smith [aut, cre] (ORCID:
<https://orcid.org/0000-0002-2479-2044>),
Alkek Center for Metagenomics and Microbiome Research [cph, fnd]

Maintainer Daniel P. Smith <dansmith01@gmail.com>

Repository CRAN

Date/Publication 2025-08-22 17:40:15 UTC

Contents

`bray_curtis` *Bray-Curtis*

Description

Bray-Curtis beta diversity metric.

Usage

```
bray_curtis(counts, weighted = TRUE, pairs = NULL, cpus = n_cpus())
```

Arguments

`counts`	An OTU abundance matrix where each column is a sample, and each row is an OTU. Any object coercible with `as.matrix()` can be given here, as well as `phyloseq`, `rbiom`, `SummarizedExperiment`, and `TreeSummarizedExperiment` objects.
`weighted`	If `TRUE`, the algorithm takes relative abundances into account. If `FALSE`, only presence/absence is considered.

`pairs`	Which combinations of samples should distances be calculated for? The default value (`NULL`) calculates all-vs-all. Provide a numeric or logical vector specifying positions in the distance matrix to calculate. See examples.
`cpus`	How many parallel processing threads should be used. The default, `n_cpus()`, will use all logical CPU cores.

Value

A `dist` object.

Calculation

In the formulas below, `x` and `y` are two columns (samples) from `counts`. `n` is the number of rows (OTUs) in `counts`.

$$D = \frac{\sum_{i=1}^{n} |x_i - y_i|}{\sum_{i=1}^{n} (x_i + y_i)}$$

```
x <- c(4, 0, 3, 2, 6)
y <- c(0, 8, 0, 0, 5)
sum(abs(x-y)) / sum(x+y)
#>  0.6428571
```

References

Sorenson T 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content. Kongelige Danske Videnskabernes Selskab, 5.

Bray JR and Curtis JT 1957. An ordination of the upland forest communities of southern Wisconsin. Ecological Monographs, 27(4). doi:10.2307/1942268

See Also

Other beta_diversity: `canberra()`, `euclidean()`, `generalized_unifrac()`, `gower()`, `jaccard()`, `kulczynski()`, `manhattan()`, `unweighted_unifrac()`, `variance_adjusted_unifrac()`, `weighted_normalized_unifrac()`, `weighted_unifrac()`

Examples

```
# Example counts matrix
ex_counts

# Bray-Curtis weighted distance matrix
bray_curtis(ex_counts)

# Bray-Curtis unweighted distance matrix
bray_curtis(ex_counts, weighted = FALSE)

# Only calculate distances for A vs all.
bray_curtis(ex_counts, pairs = 1:3)
```

`canberra` *Canberra*

Description

Canberra beta diversity metric.

Usage

```
canberra(counts, weighted = TRUE, pairs = NULL, cpus = n_cpus())
```

Arguments

`counts`	An OTU abundance matrix where each column is a sample, and each row is an OTU. Any object coercible with `as.matrix()` can be given here, as well as `phyloseq`, `rbiom`, `SummarizedExperiment`, and `TreeSummarizedExperiment` objects.
`weighted`	If `TRUE`, the algorithm takes relative abundances into account. If `FALSE`, only presence/absence is considered.
`pairs`	Which combinations of samples should distances be calculated for? The default value (`NULL`) calculates all-vs-all. Provide a numeric or logical vector specifying positions in the distance matrix to calculate. See examples.
`cpus`	How many parallel processing threads should be used. The default, `n_cpus()`, will use all logical CPU cores.

Value

A `dist` object.

Calculation

In the formulas below, `x` and `y` are two columns (samples) from `counts`. `n` is the number of rows (OTUs) in `counts`.

OTUs must be removed if they are absent from both samples.

$$D = \frac{1}{n}\sum_{i=1}^{n}\frac{|x_i - y_i|}{x_i + y_i}$$

```
x <- c(4, 0, 3, 0, 6)[-4]
y <- c(0, 8, 0, 0, 5)[-4]
sum(abs(x-y) / (x+y)) / length(x)
#>  0.7727273
```

References

Lance GN and Williams WT 1967. A general theory of classificatory sorting strategies II. Clustering systems. The computer journal, 10(3). doi:10.1093/comjnl/10.3.271

See Also

Other beta_diversity: `bray_curtis()`, `euclidean()`, `generalized_unifrac()`, `gower()`, `jaccard()`, `kulczynski()`, `manhattan()`, `unweighted_unifrac()`, `variance_adjusted_unifrac()`, `weighted_normalized_unifrac()`, `weighted_unifrac()`

Examples

```
# Example counts matrix
ex_counts

# Gower weighted distance matrix
canberra(ex_counts)

# Gower unweighted distance matrix
canberra(ex_counts, weighted = FALSE)

# Only calculate distances for A vs all.
canberra(ex_counts, pairs = 1:3)
```

---

chao1 *Chao1*

Description

Chao1 alpha diversity metric.

A non-parametric estimator of the number of unobserved species in a sample. The Chao1 index estimates total species richness based on the number of species that occur only once (singletons) and twice (doubletons) in the sample.

Usage

```
chao1(counts, cpus = n_cpus())
```

Arguments

`counts`	An OTU abundance matrix where each column is a sample, and each row is an OTU. Any object coercible with `as.matrix()` can be given here, as well as `phyloseq`, `rbiom`, `SummarizedExperiment`, and `TreeSummarizedExperiment` objects.
`cpus`	How many parallel processing threads should be used. The default, `n_cpus()`, will use all logical CPU cores.

Value

A numeric vector.

Calculation

Prerequisite: all counts are whole numbers.

In the formulas below, x is a single column (sample) from counts. n is the total number of non-zero OTUs, a is the number of singletons, and b is the number of doubletons.

$$D = n + \frac{a^2}{2b}$$

```
x <- c(1, 0, 3, 2, 6)
sum(x>0) + (sum(x==1) ^ 2) / (2 * sum(x==2))
#>  4.5
```

Note that when x does not have any singletons or doubletons ($a = 0, b = 0$), the result will be NaN. When x has singletons but no doubletons ($a > 0, b = 0$), the result will be Inf.

References

Chao A 1984. Non-parametric estimation of the number of classes in a population. Scandinavian Journal of Statistics, 11:265-270.

See Also

Other alpha_diversity: faith(), inv_simpson(), shannon(), simpson()

Examples

```
# Example counts matrix
ex_counts

# Chao1 diversity values
chao1(ex_counts)

# Low diversity
chao1(c(100, 1, 1, 1, 1)) # Inf

# High diversity
chao1(c(20, 20, 20, 20, 20)) # NaN

# Low richness
chao1(1:3) # 3.5

# High richness
chao1(1:100) # 100.5
```

euclidean *Euclidean*

Description

Euclidean beta diversity metric.

Usage

```
euclidean(counts, weighted = TRUE, pairs = NULL, cpus = n_cpus())
```

Arguments

`counts`	An OTU abundance matrix where each column is a sample, and each row is an OTU. Any object coercible with `as.matrix()` can be given here, as well as `phyloseq`, `rbiom`, `SummarizedExperiment`, and `TreeSummarizedExperiment` objects.
`weighted`	If `TRUE`, the algorithm takes relative abundances into account. If `FALSE`, only presence/absence is considered.
`pairs`	Which combinations of samples should distances be calculated for? The default value (`NULL`) calculates all-vs-all. Provide a numeric or logical vector specifying positions in the distance matrix to calculate. See examples.
`cpus`	How many parallel processing threads should be used. The default, `n_cpus()`, will use all logical CPU cores.

Value

A `dist` object.

Calculation

In the formulas below, `x` and `y` are two columns (samples) from `counts`. `n` is the number of rows (OTUs) in `counts`.

$$D = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

```
x <- c(4, 0, 3, 2, 6)
y <- c(0, 8, 0, 0, 5)
sqrt(sum((x-y)^2))
#>  9.69536
```

References

Gower JC, Legendre P 1986. Metric and Euclidean Properties of Dissimilarity Coefficients. Journal of Classification. 3. doi:10.1007/BF01896809

Legendre P, Caceres M 2013. Beta diversity as the variance of community data: dissimilarity coefficients and partitioning. Ecology Letters. 16(8). doi:10.1111/ele.12141

See Also

Other beta_diversity: `bray_curtis()`, `canberra()`, `generalized_unifrac()`, `gower()`, `jaccard()`, `kulczynski()`, `manhattan()`, `unweighted_unifrac()`, `variance_adjusted_unifrac()`, `weighted_normalized_unifrac()`, `weighted_unifrac()`

Examples

```
# Example counts matrix
ex_counts

# Euclidean weighted distance matrix
euclidean(ex_counts)

# Euclidean unweighted distance matrix
euclidean(ex_counts, weighted = FALSE)

# Only calculate distances for A vs all.
euclidean(ex_counts, pairs = 1:3)
```

`ex_counts` *Example counts matrix*

Description

Genera found on four human body sites.

Usage

```
ex_counts
```

Format

A matrix of 4 samples (columns) x 6 genera (rows).

Source

Derived from The Human Microbiome Project dataset. `https://commonfund.nih.gov/hmp`

ex_tree — *Example phylogenetic tree*

Description

Companion tree for `ex_counts`.

Usage

```
ex_tree
```

Format

A `phylo` object.

Details

`ex_tree` encodes this tree structure:

```
      +----------44---------- Haemophilus
  +-2-|
  |   +----------------68---------------- Bacteroides
  |
  |             +---18---- Streptococcus
  |      +--12--|
  |      |      +--11-- Staphylococcus
  +--11--|
         |      +-----24----- Corynebacterium
         +--12--|
                +--13-- Propionibacterium
```

faith — *Faith's PD*

Description

Faith's phylogenetic diversity metric.

A higher value indicates a greater amount of evolutionary history represented within the community, suggesting higher biodiversity in terms of evolutionary relationships.

Usage

```
faith(counts, tree = NULL, cpus = n_cpus())
```

Arguments

counts
: An OTU abundance matrix where each column is a sample, and each row is an OTU. Any object coercible with `as.matrix()` can be given here, as well as `phyloseq`, `rbiom`, `SummarizedExperiment`, and `TreeSummarizedExperiment` objects.

tree
: A `phylo`-class object representing the phylogenetic tree for the OTUs in `counts`. The OTU identifiers given by `colnames(counts)` must be present in `tree`. Can be omitted if a tree is embedded with the `counts` object or as `attr(counts, 'tree')`.

cpus
: How many parallel processing threads should be used. The default, `n_cpus()`, will use all logical CPU cores.

Value

A numeric vector.

Calculation

Given n branches with lengths L and a sample's abundances on each of those branches coded as 1 for present or 0 for absent:

$$\sum_{i=1}^{n} P_i \times L_i$$

References

Faith DP 1992. Conservation evaluation and phylogenetic diversity. Biological Conservation, 61:1-10. doi:10.1016/00063207(92)912013

See Also

Other alpha_diversity: `chao1()`, `inv_simpson()`, `shannon()`, `simpson()`

Examples

```
# Example counts matrix
ex_counts

# Faith diversity values
faith(ex_counts, tree = ex_tree)
```

generalized_unifrac *Generalized UniFrac*

Description

Generalized UniFrac beta diversity metric.

Usage

```
generalized_unifrac(
  counts,
  tree = NULL,
  alpha = 0.5,
  pairs = NULL,
  cpus = n_cpus()
)
```

Arguments

counts	An OTU abundance matrix where each column is a sample, and each row is an OTU. Any object coercible with `as.matrix()` can be given here, as well as `phyloseq`, `rbiom`, `SummarizedExperiment`, and `TreeSummarizedExperiment` objects.
tree	A `phylo`-class object representing the phylogenetic tree for the OTUs in `counts`. The OTU identifiers given by `colnames(counts)` must be present in `tree`. Can be omitted if a tree is embedded with the `counts` object or as `attr(counts, 'tree')`.
alpha	How much weight to give to relative abundances; a value between 0 and 1, inclusive. Setting `alpha=1` is equivalent to `weighted_normalized_unifrac()`.
pairs	Which combinations of samples should distances be calculated for? The default value (`NULL`) calculates all-vs-all. Provide a numeric or logical vector specifying positions in the distance matrix to calculate. See examples.
cpus	How many parallel processing threads should be used. The default, `n_cpus()`, will use all logical CPU cores.

Value

A `dist` object.

Calculation

Given n branches with lengths L, a pair of samples' abundances (A and B) on each of those branches, and abundance weighting $0 \leq \alpha \leq 1$:

$$D = \frac{\sum_{i=1}^{n} L_i(\frac{A_i}{A_T} + \frac{B_i}{B_T})^{\alpha} |\frac{\frac{A_i}{A_T} - \frac{B_i}{B_T}}{\frac{A_i}{A_T} + \frac{B_i}{B_T}}|}{\sum_{i=1}^{n} L_i(\frac{A_i}{A_T} + \frac{B_i}{B_T})^{\alpha}}$$

See `vignette('unifrac')` for details and a worked example.

References

Chen J, Bittinger K, Charlson ES, Hoffmann C, Lewis J, Wu GD, Collman RG, Bushman FD, Li H 2012. Associating microbiome composition with environmental covariates using generalized UniFrac distances. Bioinformatics, 28(16). doi:10.1093/bioinformatics/bts342

See Also

Other beta_diversity: `bray_curtis()`, `canberra()`, `euclidean()`, `gower()`, `jaccard()`, `kulczynski()`, `manhattan()`, `unweighted_unifrac()`, `variance_adjusted_unifrac()`, `weighted_normalized_unifrac()`, `weighted_unifrac()`

Examples

```
# Example counts matrix
ex_counts

# Generalized UniFrac distance matrix
generalized_unifrac(ex_counts, tree = ex_tree)

# Only calculate distances for A vs all.
generalized_unifrac(ex_counts, tree = ex_tree, pairs = 1:3)
```

`gower` *Gower*

Description

Gower beta diversity metric.

Usage

```
gower(counts, weighted = TRUE, pairs = NULL, cpus = n_cpus())
```

Arguments

`counts`	An OTU abundance matrix where each column is a sample, and each row is an OTU. Any object coercible with `as.matrix()` can be given here, as well as `phyloseq`, `rbiom`, `SummarizedExperiment`, and `TreeSummarizedExperiment` objects.
`weighted`	If `TRUE`, the algorithm takes relative abundances into account. If `FALSE`, only presence/absence is considered.
`pairs`	Which combinations of samples should distances be calculated for? The default value (`NULL`) calculates all-vs-all. Provide a numeric or logical vector specifying positions in the distance matrix to calculate. See examples.
`cpus`	How many parallel processing threads should be used. The default, `n_cpus()`, will use all logical CPU cores.

Value

A `dist` object.

Calculation

Each row (OTU) of `counts` is rescaled to the range 0-1. In cases where a row is all the same value, those values are replaced with `0`.

```
counts                      scaled recounts
     A B C  D                      A   B   C D
OTU1 0 0 0  0    ->        OTU1 0.0 0.0 0.0 0
OTU2 0 8 9 10    ->        OTU2 0.0 0.8 0.9 1
OTU3 5 5 5  5    ->        OTU3 0.0 0.0 0.0 0
OTU4 2 0 0  0    ->        OTU4 1.0 0.0 0.0 0
OTU5 4 6 4  1    ->        OTU5 0.6 1.0 0.6 0
```

In the formulas below, `x` and `y` are two columns (samples) from the scaled counts. `n` is the number of rows (OTUs) in `counts`.

$$D = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i|$$

```
x <- c(0, 0, 0, 1, 0.6)
y <- c(0, 0.8, 0, 0, 1)
sum(abs(x-y)) / length(x)
#>  0.44
```

References

Gower JC 1971. A general coefficient of similarity and some of its properties. Biometrics. 27(4). doi:10.2307/2528823

Gower JC, Legendre P 1986. Metric and Euclidean Properties of Dissimilarity Coefficients. Journal of Classification. 3. doi:10.1007/BF01896809

See Also

Other beta_diversity: bray_curtis(), canberra(), euclidean(), generalized_unifrac(), jaccard(), kulczynski(), manhattan(), unweighted_unifrac(), variance_adjusted_unifrac(), weighted_normalized_unifrac(), weighted_unifrac()

Examples

```
# Example counts matrix
ex_counts

# Gower weighted distance matrix
gower(ex_counts)

# Gower unweighted distance matrix
gower(ex_counts, weighted = FALSE)

# Only calculate distances for A vs all.
gower(ex_counts, pairs = 1:3)
```

---

inv_simpson — *Inverse Simpson*

Description

Inverse Simpson alpha diversity metric.

Usage

```
inv_simpson(counts, cpus = n_cpus())
```

Arguments

counts	An OTU abundance matrix where each column is a sample, and each row is an OTU. Any object coercible with `as.matrix()` can be given here, as well as `phyloseq`, `rbiom`, `SummarizedExperiment`, and `TreeSummarizedExperiment` objects.
cpus	How many parallel processing threads should be used. The default, `n_cpus()`, will use all logical CPU cores.

Value

A numeric vector.

Calculation

Pre-transformation: drop all OTUs with zero abundance.

In the formulas below, x is a single column (sample) from `counts`. p are the relative abundances.

$$p_i = \frac{x_i}{\sum x}$$

$$D = 1 / \sum_{i=1}^{n} p_i \times \ln(p_i)$$

```
x <- c(4, 0, 3, 2, 6)[-2]
p <- x / sum(x)
1 / sum(p * log(p))
#>  -0.7636352
```

References

Simpson EH 1949. Measurement of diversity. Nature, 163. doi:10.1038/163688a0

See Also

Other alpha_diversity: `chao1()`, `faith()`, `shannon()`, `simpson()`

Examples

```
# Example counts matrix
ex_counts

# Inverse Simpson diversity values
inv_simpson(ex_counts)

# Low diversity
inv_simpson(c(100, 1, 1, 1, 1)) # 1.08

# High diversity
inv_simpson(c(20, 20, 20, 20, 20)) # 5

# Low richness
inv_simpson(1:3) # 2.57

# High richness
inv_simpson(1:100) # 75.37
```

jaccard — *Jaccard*

Description

Jaccard beta diversity metric.

Usage

```
jaccard(counts, weighted = TRUE, pairs = NULL, cpus = n_cpus())
```

Arguments

`counts`	An OTU abundance matrix where each column is a sample, and each row is an OTU. Any object coercible with `as.matrix()` can be given here, as well as `phyloseq`, `rbiom`, `SummarizedExperiment`, and `TreeSummarizedExperiment` objects.
`weighted`	If `TRUE`, the algorithm takes relative abundances into account. If `FALSE`, only presence/absence is considered.
`pairs`	Which combinations of samples should distances be calculated for? The default value (`NULL`) calculates all-vs-all. Provide a numeric or logical vector specifying positions in the distance matrix to calculate. See examples.
`cpus`	How many parallel processing threads should be used. The default, `n_cpus()`, will use all logical CPU cores.

Value

A `dist` object.

Calculation

In the formulas below, `x` and `y` are two columns (samples) from `counts`. `n` is the number of rows (OTUs) in `counts`.

$$b = \frac{\sum_{i=1}^{n} |x_i - y_i|}{\sum_{i=1}^{n} x_i + y_i}$$

$$D = \frac{2b}{1 + b}$$

```
x <- c(4, 0, 3, 2, 6)
y <- c(0, 8, 0, 0, 5)
bray <- sum(abs(x-y)) / sum(x+y)
2 * bray / (1 + bray)
#>  0.7826087
```

References

Jaccard P 1908. Nouvellesrecherches sur la distribution florale. Bulletin de la Societe Vaudoise des Sciences Naturelles, 44(163). doi:10.5169/seals268384

See Also

Other beta_diversity: `bray_curtis()`, `canberra()`, `euclidean()`, `generalized_unifrac()`, `gower()`, `kulczynski()`, `manhattan()`, `unweighted_unifrac()`, `variance_adjusted_unifrac()`, `weighted_normalized_unifrac()`, `weighted_unifrac()`

Examples

```
# Example counts matrix
ex_counts

# Jaccard weighted distance matrix
jaccard(ex_counts)

# Jaccard unweighted distance matrix
jaccard(ex_counts, weighted = FALSE)

# Only calculate distances for A vs all.
jaccard(ex_counts, pairs = 1:3)
```

`kulczynski` *Kulczynski*

Description

Kulczynski beta diversity metric.

Usage

```
kulczynski(counts, weighted = TRUE, pairs = NULL, cpus = n_cpus())
```

Arguments

`counts`	An OTU abundance matrix where each column is a sample, and each row is an OTU. Any object coercible with `as.matrix()` can be given here, as well as `phyloseq`, `rbiom`, `SummarizedExperiment`, and `TreeSummarizedExperiment` objects.
`weighted`	If `TRUE`, the algorithm takes relative abundances into account. If `FALSE`, only presence/absence is considered.
`pairs`	Which combinations of samples should distances be calculated for? The default value (`NULL`) calculates all-vs-all. Provide a numeric or logical vector specifying positions in the distance matrix to calculate. See examples.
`cpus`	How many parallel processing threads should be used. The default, `n_cpus()`, will use all logical CPU cores.

Value

A `dist` object.

Calculation

In the formulas below, `x` and `y` are two columns (samples) from `counts`. `n` is the number of rows (OTUs) in `counts`.

$$t = \sum_{i=1}^{n} min(x_i, y_i)$$

$$D = 1 - 0.5(\frac{t}{\sum_{i=1}^{n} x_i} + \frac{t}{\sum_{i=1}^{n} y_i})$$

```
x <- c(4, 0, 3, 2, 6)
y <- c(0, 8, 0, 0, 5)
t <- sum(pmin(x,y))
1 - (t/sum(x) + t/sum(y)) / 2
#>  0.6410256
```

References

Kulcynski S 1927. Die Pflanzenassoziationen der Pieninen. Bulletin International de l'Académie Polonaise des Sciences et des Lettres, Classe des Sciences Mathématiques et Naturelles, Série B: Sciences Naturelles.

See Also

Other beta_diversity: `bray_curtis()`, `canberra()`, `euclidean()`, `generalized_unifrac()`, `gower()`, `jaccard()`, `manhattan()`, `unweighted_unifrac()`, `variance_adjusted_unifrac()`, `weighted_normalized_unifrac()`, `weighted_unifrac()`

Examples

```
# Example counts matrix
ex_counts

# Kulczynski weighted distance matrix
kulczynski(ex_counts)

# Kulczynski unweighted distance matrix
kulczynski(ex_counts, weighted = FALSE)

# Only calculate distances for A vs all.
kulczynski(ex_counts, pairs = 1:3)
```

manhattan *Manhattan*

Description

Manhattan beta diversity metric.

Usage

```
manhattan(counts, weighted = TRUE, pairs = NULL, cpus = n_cpus())
```

Arguments

counts	An OTU abundance matrix where each column is a sample, and each row is an OTU. Any object coercible with `as.matrix()` can be given here, as well as `phyloseq`, `rbiom`, `SummarizedExperiment`, and `TreeSummarizedExperiment` objects.
weighted	If `TRUE`, the algorithm takes relative abundances into account. If `FALSE`, only presence/absence is considered.
pairs	Which combinations of samples should distances be calculated for? The default value (`NULL`) calculates all-vs-all. Provide a numeric or logical vector specifying positions in the distance matrix to calculate. See examples.
cpus	How many parallel processing threads should be used. The default, `n_cpus()`, will use all logical CPU cores.

Value

A `dist` object.

Calculation

In the formulas below, `x` and `y` are two columns (samples) from `counts`. `n` is the number of rows (OTUs) in `counts`.

$$D = \sum_{i=1}^{n} |x_i - y_i|$$

```
x <- c(4, 0, 3, 2, 6)
y <- c(0, 8, 0, 0, 5)
sum(abs(x-y))
#>  18
```

References

Paul EB 2006. Manhattan distance. Dictionary of Algorithms and Data Structures. https://xlinux.nist.gov/dads/HTML/manhattanDistance.html

See Also

Other beta_diversity: bray_curtis(), canberra(), euclidean(), generalized_unifrac(), gower(), jaccard(), kulczynski(), unweighted_unifrac(), variance_adjusted_unifrac(), weighted_normalized_unifrac(), weighted_unifrac()

Examples

```
# Example counts matrix
ex_counts

# Manhattan weighted distance matrix
manhattan(ex_counts)

# Manhattan unweighted distance matrix
manhattan(ex_counts, weighted = FALSE)

# Only calculate distances for A vs all.
manhattan(ex_counts, pairs = 1:3)
```

---

n_cpus *Number of CPU Cores*

Description

A thin wrapper around `parallel::detectCores(all.tests = TRUE, logical = TRUE)` which falls back to 1 when the number of CPU cores cannot be detected, or when the system does not support `pthreads`. Consider using `parallely::availableCores()` in place of `n_cpus()` for more advanced interrogation of system resources.

Usage

```
n_cpus()
```

Value

A scalar integer, guaranteed to be at least 1.

Examples

```
n_cpus()
```

`rarefy` *Rarefy OTU counts.*

Description

Sub-sample OTU observations such that all samples have an equal number. If called on data with non-integer abundances, values will be re-scaled to integers between 1 and `depth` such that they sum to `depth`.

Usage

```
rarefy(
  counts,
  depth = 0.1,
  n_samples = NULL,
  seed = 0,
  times = NULL,
  cpus = n_cpus()
)
```

Arguments

`counts`	An OTU abundance matrix where each column is a sample, and each row is an OTU. Any object coercible with `as.matrix()` can be given here, as well as `phyloseq`, `rbiom`, `SummarizedExperiment`, and `TreeSummarizedExperiment` objects.
`depth`	How many observations to keep per sample. When `0 < depth < 1`, it is taken as the minimum percentage of the dataset's observations to keep. Ignored when `n_samples` is specified. Default: `0.1`
`n_samples`	The number of samples to keep. When `0 < n_samples < 1`, it is taken as the percentage of samples to keep. If negative, that number of samples is dropped. If `0`, all samples are kept. If `NULL`, then `depth` is used instead. Default: `NULL`
`seed`	An integer seed for randomizing which observations to keep or drop. If you need to create different random rarefactions of the same data, set the seed to a different number each time.
`times`	How many independent rarefactions to perform. If set, `rarefy()` will return a list of matrices. The seeds for each matrix will be sequential, starting from `seed`.
`cpus`	How many parallel processing threads should be used. The default, `n_cpus()`, will use all logical CPU cores.

Value

An integer matrix.

Examples

```
# Create an OTU matrix with 4 samples (A-D) and 5 OTUs.
counts <- matrix(
  data     = c(4,0,3,2,6,0,8,0,0,5,0,9,0,0,7,0,10,0,0,1),
  nrow     = 5,
  dimnames = list(paste0('OTU', 1:5), LETTERS[1:4]) )
counts
colSums(counts)

counts <- rarefy(counts, depth = 14)
counts
colSums(counts)
```

`read_tree` *Read a newick formatted phylogenetic tree.*

Description

A phylogenetic tree is required for computing UniFrac distance matrices. You can load a tree from a file or by providing the tree string directly. This tree must be in Newick format, also known as parenthetic format and New Hampshire format.

Usage

```
read_tree(newick, underscores = FALSE)
```

Arguments

`newick`	Input data as either a file path, URL, or Newick string. Compressed (gzip or bzip2) files are also supported.
`underscores`	If `TRUE`, underscores in unquoted names will remain underscores. If `FALSE`, underscores in unquoted named will be converted to spaces.

Value

A `phylo` class object representing the tree.

Examples

```
tree <- read_tree("
    (A:0.99,((B:0.87,C:0.89):0.51,(((D:0.16,(E:0.83,F:0.96)
    :0.94):0.69,(G:0.92,(H:0.62,I:0.85):0.54):0.23):0.74,J:0.1
    2):0.43):0.67);")
class(tree)
```

shannon — *Shannon*

Description

Shannon alpha diversity metric.

The index considers both the number of different OTUs (richness) and how evenly the observations are distributed among those OTUs (evenness).

Usage

```
shannon(counts, cpus = n_cpus())
```

Arguments

`counts`	An OTU abundance matrix where each column is a sample, and each row is an OTU. Any object coercible with `as.matrix()` can be given here, as well as `phyloseq`, `rbiom`, `SummarizedExperiment`, and `TreeSummarizedExperiment` objects.
`cpus`	How many parallel processing threads should be used. The default, `n_cpus()`, will use all logical CPU cores.

Value

A numeric vector.

Calculation

Pre-transformation: drop all OTUs with zero abundance.

In the formulas below, x is a single column (sample) from `counts`. p_i is the proportion of the i-th OTU in the total community.

$$p_i = \frac{x_i}{\sum x}$$

$$D = -\sum_{i=1}^{n} p_i \times \ln(p_i)$$

```
x <- c(4, 0, 3, 2, 6)[-2]
p <- x / sum(x)
-sum(p * log(p))
#>  1.309526
```

References

Shannon CE, Weaver W 1949. The Mathematical Theory of Communication. University of Illinois Press.

See Also

Other alpha_diversity: chao1(), faith(), inv_simpson(), simpson()

Examples

```
# Example counts matrix
ex_counts

# Shannon diversity values
shannon(ex_counts)

# Low diversity
shannon(c(100, 1, 1, 1, 1)) # 0.22

# High diversity
shannon(c(20, 20, 20, 20, 20)) # 1.61

# Low richness
shannon(1:3) # 1.01

# High richness
shannon(1:100) # 4.42
```

---

simpson *Simpson*

Description

Simpson alpha diversity metric.

Gauges the uniformity of species within a community. A Simpson index of `0` indicates that one or a few high abundance OTUs dominate the community, which is indicative of low diversity.

Usage

```
simpson(counts, cpus = n_cpus())
```

Arguments

counts	An OTU abundance matrix where each column is a sample, and each row is an OTU. Any object coercible with `as.matrix()` can be given here, as well as `phyloseq`, `rbiom`, `SummarizedExperiment`, and `TreeSummarizedExperiment` objects.
cpus	How many parallel processing threads should be used. The default, `n_cpus()`, will use all logical CPU cores.

Value

A numeric vector.

Calculation

Pre-transformation: drop all OTUs with zero abundance.

In the formulas below, x is a single column (sample) from `counts`. p are the relative abundances.

$$p_i = \frac{x_i}{\sum x}$$

$$D = 1 - \sum_{i=1}^{n} p_i \times \ln(p_i)$$

```
x <- c(4, 0, 3, 2, 6)[-2]
p <- x / sum(x)
1 - sum(p * log(p))
#>  2.309526
```

References

Simpson EH 1949. Measurement of diversity. Nature, 163. doi:10.1038/163688a0

See Also

Other alpha_diversity: `chao1`(), `faith`(), `inv_simpson`(), `shannon`()

Examples

```
# Example counts matrix
ex_counts

# Simpson diversity values
simpson(ex_counts)

# Low diversity
simpson(c(100, 1, 1, 1, 1)) # 0.075

# High diversity
simpson(c(20, 20, 20, 20, 20)) # 0.8

# Low richness
simpson(1:3) # 0.61

# High richness
simpson(1:100) # 0.99
```

`unweighted_unifrac` *Unweighted UniFrac*

Description

Unweighted UniFrac beta diversity metric.

Usage

```
unweighted_unifrac(counts, tree = NULL, pairs = NULL, cpus = n_cpus())
```

Arguments

`counts`	An OTU abundance matrix where each column is a sample, and each row is an OTU. Any object coercible with `as.matrix()` can be given here, as well as `phyloseq`, `rbiom`, `SummarizedExperiment`, and `TreeSummarizedExperiment` objects.
`tree`	A `phylo`-class object representing the phylogenetic tree for the OTUs in `counts`. The OTU identifiers given by `colnames(counts)` must be present in `tree`. Can be omitted if a tree is embedded with the `counts` object or as `attr(counts, 'tree')`.
`pairs`	Which combinations of samples should distances be calculated for? The default value (`NULL`) calculates all-vs-all. Provide a numeric or logical vector specifying positions in the distance matrix to calculate. See examples.
`cpus`	How many parallel processing threads should be used. The default, `n_cpus()`, will use all logical CPU cores.

Value

A `dist` object.

Calculation

Given n branches with lengths L and a pair of samples' abundances (A and B) on each of those branches:

$$D = \frac{\sum_{i=1}^{n} L_i(|A_i - B_i|)}{\sum_{i=1}^{n} L_i(max(A_i, B_i))}$$

Abundances in A and B are coded as `1` or `0` to indicate their presence or absence, respectively, on each branch.

See https://cmmr.github.io/ecodive/articles/unifrac.html for details and a worked example.

References

Lozupone C, Knight R 2005. UniFrac: A new phylogenetic method for comparing microbial communities. Applied and Environmental Microbiology, 71(12). doi:10.1128/AEM.71.12.8228-8235.2005

See Also

Other beta_diversity: `bray_curtis()`, `canberra()`, `euclidean()`, `generalized_unifrac()`, `gower()`, `jaccard()`, `kulczynski()`, `manhattan()`, `variance_adjusted_unifrac()`, `weighted_normalized_unifrac()`, `weighted_unifrac()`

Examples

```
# Example counts matrix
ex_counts

# Unweighted UniFrac distance matrix
unweighted_unifrac(ex_counts, tree = ex_tree)

# Only calculate distances for A vs all.
unweighted_unifrac(ex_counts, tree = ex_tree, pairs = 1:3)
```

---

variance_adjusted_unifrac

Variance Adjusted UniFrac

Description

Variance Adjusted UniFrac beta diversity metric.

Usage

```
variance_adjusted_unifrac(counts, tree = NULL, pairs = NULL, cpus = n_cpus())
```

Arguments

`counts`	An OTU abundance matrix where each column is a sample, and each row is an OTU. Any object coercible with `as.matrix()` can be given here, as well as `phyloseq`, `rbiom`, `SummarizedExperiment`, and `TreeSummarizedExperiment` objects.
`tree`	A `phylo`-class object representing the phylogenetic tree for the OTUs in `counts`. The OTU identifiers given by `colnames(counts)` must be present in `tree`. Can be omitted if a tree is embedded with the `counts` object or as `attr(counts, 'tree')`.
`pairs`	Which combinations of samples should distances be calculated for? The default value (`NULL`) calculates all-vs-all. Provide a numeric or logical vector specifying positions in the distance matrix to calculate. See examples.

cpus — How many parallel processing threads should be used. The default, `n_cpus()`, will use all logical CPU cores.

Value

A `dist` object.

Calculation

Given n branches with lengths L and a pair of samples' abundances (A and B) on each of those branches:

$$D = \frac{\sum_{i=1}^{n} L_i \frac{|\frac{A_i}{A_T} - \frac{B_i}{B_T}|}{\sqrt{(A_i + B_i)(A_T + B_T - A_i - B_i)}}}{\sum_{i=1}^{n} L_i \frac{\frac{A_i}{A_T} + \frac{B_i}{B_T}}{\sqrt{(A_i + B_i)(A_T + B_T - A_i - B_i)}}}$$

See `vignette('unifrac')` for details and a worked example.

References

Chang Q, Luan Y, Sun F 2011. Variance adjusted weighted UniFrac: a powerful beta diversity measure for comparing communities based on phylogeny. BMC Bioinformatics, 12. doi:10.1186/1471210512118

See Also

Other beta_diversity: `bray_curtis()`, `canberra()`, `euclidean()`, `generalized_unifrac()`, `gower()`, `jaccard()`, `kulczynski()`, `manhattan()`, `unweighted_unifrac()`, `weighted_normalized_unifrac()`, `weighted_unifrac()`

Examples

```
# Example counts matrix
ex_counts

# Variance Adjusted UniFrac distance matrix
variance_adjusted_unifrac(ex_counts, tree = ex_tree)

# Only calculate distances for A vs all.
variance_adjusted_unifrac(ex_counts, tree = ex_tree, pairs = 1:3)
```

weighted_normalized_unifrac — *Normalized UniFrac*

Description

Normalized UniFrac beta diversity metric.

Usage

```
weighted_normalized_unifrac(counts, tree = NULL, pairs = NULL, cpus = n_cpus())
```

Arguments

`counts`	An OTU abundance matrix where each column is a sample, and each row is an OTU. Any object coercible with `as.matrix()` can be given here, as well as `phyloseq`, `rbiom`, `SummarizedExperiment`, and `TreeSummarizedExperiment` objects.
`tree`	A `phylo`-class object representing the phylogenetic tree for the OTUs in `counts`. The OTU identifiers given by `colnames(counts)` must be present in `tree`. Can be omitted if a tree is embedded with the `counts` object or as `attr(counts, 'tree')`.
`pairs`	Which combinations of samples should distances be calculated for? The default value (`NULL`) calculates all-vs-all. Provide a numeric or logical vector specifying positions in the distance matrix to calculate. See examples.
`cpus`	How many parallel processing threads should be used. The default, `n_cpus()`, will use all logical CPU cores.

Value

A `dist` object.

Calculation

Given n branches with lengths L and a pair of samples' abundances (A and B) on each of those branches:

$$D = \frac{\sum_{i=1}^{n} L_i|\frac{A_i}{A_T} - \frac{B_i}{B_T}|}{\sum_{i=1}^{n} L_i(\frac{A_i}{A_T} + \frac{B_i}{B_T})}$$

See `vignette('unifrac')` for details and a worked example.

References

Lozupone CA, Hamady M, Kelley ST, Knight R 2007. Quantitative and Qualitative β Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities. Applied and Environmental Microbiology, 73(5). doi:10.1128/AEM.0199606

See Also

Other beta_diversity: `bray_curtis()`, `canberra()`, `euclidean()`, `generalized_unifrac()`, `gower()`, `jaccard()`, `kulczynski()`, `manhattan()`, `unweighted_unifrac()`, `variance_adjusted_unifrac()`, `weighted_unifrac()`

Examples

```
# Example counts matrix
ex_counts

# UniFrac weighted distance matrix
weighted_normalized_unifrac(ex_counts, tree = ex_tree)

# Only calculate distances for A vs all.
weighted_normalized_unifrac(ex_counts, tree = ex_tree, pairs = 1:3)
```

`weighted_unifrac` *Weighted UniFrac*

Description

Weighted UniFrac beta diversity metric.

Usage

```
weighted_unifrac(counts, tree = NULL, pairs = NULL, cpus = n_cpus())
```

Arguments

`counts`	An OTU abundance matrix where each column is a sample, and each row is an OTU. Any object coercible with `as.matrix()` can be given here, as well as `phyloseq`, `rbiom`, `SummarizedExperiment`, and `TreeSummarizedExperiment` objects.
`tree`	A phylo-class object representing the phylogenetic tree for the OTUs in `counts`. The OTU identifiers given by `colnames(counts)` must be present in `tree`. Can be omitted if a tree is embedded with the `counts` object or as `attr(counts, 'tree')`.
`pairs`	Which combinations of samples should distances be calculated for? The default value (`NULL`) calculates all-vs-all. Provide a numeric or logical vector specifying positions in the distance matrix to calculate. See examples.
`cpus`	How many parallel processing threads should be used. The default, `n_cpus()`, will use all logical CPU cores.

Value

A `dist` object.

Calculation

Given n branches with lengths L and a pair of samples' abundances (A and B) on each of those branches:

$$D = \sum_{i=1}^{n} L_i | \frac{A_i}{A_T} - \frac{B_i}{B_T} |$$

See `vignette('unifrac')` for details and a worked example.

References

Lozupone CA, Hamady M, Kelley ST, Knight R 2007. Quantitative and Qualitative β Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities. Applied and Environmental Microbiology, 73(5). doi:10.1128/AEM.0199606

See Also

Other beta_diversity: `bray_curtis()`, `canberra()`, `euclidean()`, `generalized_unifrac()`, `gower()`, `jaccard()`, `kulczynski()`, `manhattan()`, `unweighted_unifrac()`, `variance_adjusted_unifrac()`, `weighted_normalized_unifrac()`

Examples

```
# Example counts matrix
ex_counts

# Weighted UniFrac distance matrix
weighted_unifrac(ex_counts, tree = ex_tree)

# Only calculate distances for A vs all.
weighted_unifrac(ex_counts, tree = ex_tree, pairs = 1:3)
```

Index