Package: Sibyl (via r-universe)

May 23, 2026

Title Sibyl

Version 1.0

Description The Sibyl package provides a convenient way of testing a range of thresholds for rarefation, evaluating the effect of different values on ordination results, integrating with commonly used packages in microbial ecology. You provide phyloseq data with count data and sample information.

License GPL (>= 3)

Language en-US

Encoding UTF-8

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.3

SystemRequirements libcurl, OpenSSL, XML2, LAPACK, Fortran, zlib, OpenGL, GLU, X11, JPEG, PNG

Depends R (>= 4.2.0)

Imports phyloseq (>= 1.38.0), clusterSim, dplyr, ggplot2, vegan, doParallel, foreach, tidyr, purrr, broom, magrittr

LazyData true

Suggests rmarkdown, ggpubr, withr, testthat (>= 3.0.0)

Config/testthat/edition 3

Config/testthat/parallel true

Config/Needs/website rmarkdown

URL https://headonpillow.github.io/Sibyl/

Config/pak/sysreqs libglpk-dev libglu1-mesa-dev libicu-dev libjpeg-dev libxml2-dev libgl1-mesa-dev libssl-dev libx11-dev zlib1g-dev

Repository https://headonpillow.r-universe.dev

Date/Publication 2026-03-24 14:20:12 UTC

RemoteUrl https://github.com/Headonpillow/Sibyl

RemoteRef HEAD

RemoteSha 61c4e15f963c8b25bc3ed897c1322a9b068e11ad

Contents

`accumulation_test` *Accumulation curve analysis*

Description

This function generates accumulation (rarefaction) curves for each sample in a given `phyloseq` object.

Usage

```
accumulation_test(input, step = 5)
```

Arguments

`input`	A `phyloseq` object.
`step`	A numeric value. The step size for drawing the accumulation curve. It influences the rarefaction curve calculation and the granularity of points plotted. Default = 5.

Details

It fits a general accumulation model using the Abundance Coverage Estimator (ACE) as an asymptote, and identifies the sequencing depth at which 75% of the ACE value is reached. It also produces a density plot showing the distribution of these 75% completion thresholds.

Sites that fail model fitting or quality-control criteria are excluded from downstream analyses, and their identities are returned for inspection.

Value

A list containing:

- `accumulation_plot`: A `ggplot` object with all sites faceted.
- `threshold_density`: A `ggplot` density/histogram plot of the 75% ACE thresholds.
- `individual_plots`: A list of `ggplot` objects, one per site.
- `nls_failed_sites`: A character vector of site names for which the nonlinear model failed to converge.

- `quality_failed_sites`: A character vector of site names for which the model converged but failed quality-control criteria (e.g. implausible threshold or poor fit).
- `fitted_table`: A data frame containing per-site model results and diagnostics, including fitted parameters, threshold estimates, and quality-control flags.

Examples

```
library(Sibyl)
# Creating a smaller subset of the data
adults_sub <- phyloseq::subset_samples(adults, location=="VK3")
# Running accumulation tests on a phyloseq object, higher step size reduces
# execution time.
accumulation_test(adults_sub, step=50)
```

`adults` *adults*

Description

This is a `phyloseq` containing example data from a 16S study performed on adult mosquitoes from Burkina Faso.

Usage

```
adults
```

Format

A `phyloseq` object with 34 samples and 5221 taxa.

A sample_data object is included, with the following columns:

sample_id A unique identifier for each sample.

location The location where the sample was collected.

Source

https://pmc.ncbi.nlm.nih.gov/articles/PMC4785398/

avg_pairwise_dist_plot *Plot average pairwise distance across thresholds*

Description

This function it's a plotting function for `avg_distances` generated from test_threshold.

Usage

```
avg_pairwise_dist_plot(avg_distances)
```

Arguments

`avg_distances`	A dataframe of average distances obtained from `test_threshold`.

Value

A `ggplot` object showing average pairwise distances across thresholds.

Examples

```
library(Sibyl)
# Creating a smaller subset of the data
adults_sub <- phyloseq::subset_samples(adults, location=="VK3")
result <- test_threshold(adults_sub,
                         repeats = 10,
                         t_min = 100,
                         t_max = 500,
                         t_step = 50,
                         group = "location",
                         verbose = FALSE)
avg_pairwise_dist_plot(result$avg_distances$repeat_number_10)
```

larvae *larvae*

Description

This is a `phyloseq` containing example data from a 16S study performed on adult mosquitoes larvae from Ethiopia.

Usage

```
larvae
```

Format

A phyloseq object with 54 samples and 722 taxa.

A sample_data object is included, with the following columns:

sample_id A unique identifier for each sample.

breeding_site_type Environmental classification of sites where mosquitoes were collected

Source

https://academic.oup.com/femsec/article/101/1/fiae161/7928135

---

repeated_rarefaction *Perform repeated rarefaction*

Description

This function performs repeated rarefaction on a phyloseq object, computes ordination, and generates a PCoA-based visualization. The same procedure is used from threshold_testing function when testing a range of thresholds.

Usage

```
repeated_rarefaction(
  input,
  repeats = 50,
  threshold = 250,
  colorb = "sample_id",
  group = "sample_id",
  cloud = TRUE,
  ellipse = FALSE,
  cores = 2,
  ...
)
```

Arguments

input	A phyloseq object.
repeats	An integer. The number of times to repeat rarefaction. A value of 1 means no repeats. If too few repeats are selected it would be not possible to draw an ellipse around the group.
threshold	An integer. The threshold value to use for rarefaction.
colorb	A string. Column name in sample_data(). Used to color sample points.
group	A string. Column name in sample_data(). Used to group the samples. The parameter is also used to draw an ellipse around the points.

cloud	A boolean. If TRUE, all the data points generated from repetitions are shown. Otherwise, only the median points of each sample repetition cloud are plotted.
ellipse	A boolean. If TRUE, confidence ellipses around sample groups are drawn.
cores	An integer. Number of cores to use for parallel processing.
...	Additional arguments are reserved to internal use.

Value

A list containing (While also showing the plot directly):

- repeats: Number of repeats.
- df_consensus_coordinates: A data frame with coordinates of the median points of the sample clouds.
- df_all: A data frame of coordinates ordered by ordination number, along with metatata.
- plot: a ggplot object.

Examples

```
library(Sibyl)
# Running this with cloud = TRUE and ellipse = TRUE will generate a plot
# where the samples belonging to the same group will be colored similarly
# and an ellipse will be drawn around the group.
repeated_rarefaction(adults,
                     repeats = 10,
                     threshold = 250,
                     group = "location",
                     colorb = "location",
                     cloud = TRUE,
                     ellipse = TRUE)

# We can run the function to highlight the spread of the single sample clouds
# too, setting the groupb parameter to the sample_id.
repeated_rarefaction(adults,
                     repeats = 10,
                     threshold = 250,
                     group = "sample_id",
                     colorb = "location",
                     cloud = TRUE,
                     ellipse = TRUE)
```

---

test_threshold *Test different rarefaction thresholds*

Description

This function uses the same steps of repeated_rarefaction in a repeated fashion and summarizes the results across multiple thresholds.

Usage

```
test_threshold(
  input,
  repeats = 50,
  t_min = 50,
  t_max = 250,
  t_step = 5,
  group = "sample_id",
  cores = 2,
  verbose = TRUE,
  ...
)
```

Arguments

input	A phyloseq object.
repeats	An integer, or a vector of integers. The number of times to repeat rarefaction. A value of 1 means no repeats. If a vector, different rarefaction thresholds will be tested sequentially.
t_min	An integer. The minimum value for the threshold testing range.
t_max	An integer. The maximum value for the threshold testing range.
t_step	An integer. The step size for the threshold testing range. A value between 0 and 1 will cause the same threshold to be tested multiple times.
group	A string. Column name in sample_data(). Used to group the samples. The parameter is also used to draw an ellipse around the points. It the context of this function, is also the value that will be used for the calculation of the Calinski-Harabasz index.
cores	An integer. Number of cores to use for parallel processing.
verbose	A logical. If TRUE, prints messages during the execution.
...	Additional arguments are reserved to internal use.

Details

Clustering performance is evaluated using the Calinski-Harabasz index. The function also calculates the average pairwise distance for each sample cloud across different rarefaction thresholds, to give a measure of sample concordance (how stable the sample is at different thresholds).

Higher index values are better, and the plateauing of the index values when testing higher thresholds indicate that the clouds of repetitions are compact enough to not being affected by increasing the threshold. The Calinski-Harabasz value takes into account both within cluster and between cluster distances, and in the scope of this function is calculated on the group parameter. A sudden drop in CH value means that samples might have been removed because they did not have enough reads to reach the threshold. In this case a warning is printed listing the samples which have been removed. In avg_distances, if samples do not have enough reads to reach t_max value, their APD are set to 0 after that threshold and a warning message is printed.

Value

A list containing:

- `index_plot`: A `ggplot` object showing a scatterplot of Calinski-Harabasz index vs. rarefaction threshold.
- `ordination_plots`: A list of lists containing the ordination plots for tested threshold values. If multiple repeat values were used, each one will be present in a different list.
- `avg_distances`: A data frame (or list of data frames if using multiple repeat values) with average pairwise distances of the cloud generated from each sample repetition attempts (for each threshold).

Examples

```
library(Sibyl)
result <- test_threshold(adults,
                         repeats = 10,
                         t_min = 100,
                         t_max = 500,
                         t_step = 50,
                         group = "location",
                         verbose = FALSE)
result$index_plot
```

Index