Package 'GrammR'

August 11, 2014

Type Package

Title Graphical Representation and Modeling of Metagenomic Reads

Version 1.0.0

Date 2014-08-06

Author Deepak N. Ayyala, Shili Lin, Department of Statistics, The Ohio State University, Columbus, Ohio, USA.

Maintainer Deepak Nag Ayyala <ayyala.1@osu.edu>

Description Represents metagenomic samples on the Euclidean space to examine similarity amongst samples by studying clusters in the model. Given the matrix of metagenomic counts for samples, this package (1) quantifies dissimilarity between samples using Kendall's tau-distance, (2) constructs multidimensional models of different dimension, and (3) plots the models for visualization and comparison.

Depends R (>= 3.0.0)

Imports gWidgets, RGtk2, gWidgetsRGtk2, MASS, cluster, rgl, GUniFrac,ape

LazyLoad YES

License LGPL-3

NeedsCompilation yes

Repository CRAN

Date/Publication 2014-08-11 07:49:10

R topics documented:

GrammR-package *Graphical Representation and Modeling of Metagenomic Reads*

Description

An important exploratory step when analyzing metagenomic count data is visualization of the data. Researchers often restrict graphical representations of metagenomic data sets to two dimensional models for ease of presentation. However higher dimensional visualizations are known to better represent the data, providing valuable information which are otherwise not observed in two dimensional models. These graphical representations are determined by two factors: the measure of dissimilarity between samples and multidimensional scaling model used to estimate the coordinates.

UniFrac is one such measure of dissimilarity which is very popular in the metagenomic research community. The UniFrac distance between two individuals is calculated by placing the samples on a phylogenetic tree and counting the number of unshared branches between the two samples. Several extensions to the UniFrac distance have been proposed to better address the issue. However, calculation of UniFrac distances requires the phylogenetic tree information in addition to the counts. Alternatively, Kendall's τ-distance is one such measure of dissimilarity which is applicable to counts as well as relative frequencies, without the need to specify a phylogenetic tree.

A commonly used multidimensional scaling model for estimating the coordinates of the samples on a Euclidean space is Principal Coordinate Analysis (PCoA). This method is very similar to Principal Component Analysis (PCA), used for dimension reduction in multivariate statistical analysis. Goodness-of-fit of the PCoA model is measured by the percentage of variation explained using the dimensions selected. Several studies in the past reported the percent variation explained to be less than 30%. As an alternative to PCoA, metric multidimensional scaling (MDS) models can be constructed. MDS models estimate the coordinates by minimizing a stress function, providing researchers the freedom to choose the metric to be used.

GrammR provides a user-friendly interface to construct graphical representations, giving the user an option to choose the measure of dissimilarity and multidimensional scaling model to be constructed. In addition to constructing graphical models, the package also estimates the optimal number of clusters into which the data should be divided. Given prior clustering of the data constructed using attributes of the samples, the package can compare the constructed clusters to those provided apriori by calculating a misclassification error.

Details

The package provides two options to users for construction of graphical models

1. `GrammRGUI` provides a Graphical User Interface (GUI) for analyzing the count data sets. The user-friendly interface is recommended for beginners.
2. `GrammRServ` can be used as a function for analyzing data sets through the R interface without a GUI. This is recommended for large data sets which require larger run times.

Author(s)

Deepak Nag Ayyala, Shili Lin.

References

Ayyala, D. N., Lin, S. (2014) Graphical Representation and Modeling of Metagenomic Reads, Manuscript.

Lozupone, Catherine and Knight, Rob (2005) UniFrac: a New Phylogenetic Method for Comparing Microbial Communities, Applied and Environmental Microbiology, 8228-8235.

Kendall, M. G. (1938) A new measure of rank correlation, Biometrika, 30, 81-93.

`Count2Distance` *Calculating the dissimilarity matrix for metagenomic count data*

Description

This function quantifies the dissimilarity between samples or taxa. Given a $N \times k$ matrix of metagenomic counts where N equals the number of samples and k is the number of taxa/OTUs, this function returns a $N \times N$ matrix whose $(i, j)^{th}$ element gives a measure of dissimilarity between the i^{th} and j^{th} samples. If Kendall's τ-distance is specified as the measure of dissimilarity, this function also has the capability to compute dissimilarity between taxa, resulting in a $k \times k$ matrix.

Usage

```
Count2Distance(Data, Distance, Penalty = NULL, PhyTree = NULL,
 UnifOpts = NULL, Adjust = TRUE)
```

Arguments

`Data`	An $N \times k$ matrix comprising the metagenomic counts, whose rows correspond to the distinct samples.
`Distance`	Measure of dissimilarity to be used for calculating distance matrix. Possible values are c("Kendall's tau-distance", "UniFrac").
`Penalty`	Penalty to be used for ties when calculating the Kendall's tau-distance. It takes values between 0 and 1.
`PhyTree`	Rooted phylogenetic tree of R class "phylo". To be provided only when the `Distance == "UniFrac"`.

UnifOpts	Options to calculate the generalized UniFrac distance. This is a list containing two items c(Weight, Type), where Weight takes values between 0 and 1, and Type takes values in c("Unweighted", "Variance Adjusted", "Generalized").
Adjust	A logical variable. When TRUE, an infinitesimal constant (equal to .Machine$double.eps) to off-diagonal elements which are equal to zero. This is to facilitate construction of metric multidimensional models.

Value

A $N \times N$ symmetric matrix with all zeroes along the diagonal, where N is the number of samples in the data. If the transpose of the counts is provided, the function returns a $k \times k$ symmetric matrix.

Author(s)

Deepak Nag Ayyala <ayyala.1@osu.edu>

See Also

KenDist

Examples

```
data(metagencounts)
Distance <- Count2Distance(Data = metagencounts$Counts, Distance = "Kendall's tau-distance",
 Penalty = 0.5);
```

GrammRGUI *GrammR GUI for graphical modeling and visualization*

Description

A graphical user interface for GrammR, to construct and study graphical representations of metagenomic reads.

Usage

```
GrammRGUI(Direc)
```

Arguments

Direc	(Optional) The directory containing the data files and location where the constructed graphical models should be saved. If no directory is specified, the current working directory is used.

Value

The RGUI creates a window with multiple tabs. The number of tabs is determined by the modelling parameters specified by the user.

Author(s)

Deepak Nag Ayyala <ayyala.1@osu.edu>

See Also

GrammRServ

Examples

```
## Not run: GrammRGUI()
```

GrammRServ *Graphical Representation without a GUI*

Description

A non-GUI method to construct graphical representations of metagenomic count data. This function is recommended for large data sets and can be run as a background job when a user-interface is not available.

Usage

```
GrammRServ(Data = NULL, Cluster = NULL, DataType = "Counts",
DistType = "Kendall's tau-distance", PhyTree = NULL,
GunifType = NULL, GunifWeight = 0, Dim = c(2, 3, 4),
LpNorm = c(1), Penalty = 0.5, MinClust = 2)
```

Arguments

Data	Data matrix consisting of one of the following two values: • (1) metagenomic counts with the rows of the matrix representing attributes to be clustered(can be samples or taxa). • (2) measure of dissimilarity between samples or taxa.
Cluster	(Optional) The vector whose length is equal to the number of rows of Data. Values in the vector provide the cluster membership of samples determined using their attributes.
DataType	A character variable corresponding to the type of values in Data. It takes values in c("Counts", "Distance")
DistType	Measure of dissimilarity between samples to be used to calculate the distance matrix. It takes values in c("Kendall's tau-distance", "UniFrac") and is used when the DataType is equal to Counts. The default value is "Kendall's tau-distance".
PhyTree	A phylogenetic tree of class phylo to be used for calculating the UniFrac distance. This is to be provided only when DistType is set equal to "UniFrac".

GunifType	The type of UniFrac distance to be specified when calculating the UniFrac distance using GUniFrac package. It takes values in c("Unweighted", "Variance Adjusted", "Generalized").
GunifWeight	The weight parameter used in calculation of Generalized UniFrac distance. The parameter takes values between 0 and 1. For more details, see Chen et.al.(2012).
Dim	Dimension of the multidimensional scaling model to be constructed. Default value is c(2,3,4).
LpNorm	A vector valued variable which determines the norm to be used in multidimensional scaling model calculation. The default value (equal to 1) corresponds to l_1-MDS model. Principal coordinate analysis (PCoA) is performed when the value is set to two.
Penalty	A numeric value between 0 and 1 which is used as penalty for ties in calculation of Kendall's tau-distance. Default value is 0.5.
MinClust	Minimum number of clusters to be used in PAM method for estimating the optimal number of clusters. Default value is 2.

Value

Separate directories are created in the current working directory for each model constructed using all possible combinations of dimension and l_p norm specified.

1. Directories for the two dimensional models contain the average silhouette plot, true estimated model, model showing estimated clusters and (optional)model showing true clusters.
2. Directories for models of dimension greater than two contain the average silhoutte plot and subdirectories for the true model, estimated clusters model and (optional)model showing true clusters.

For all models, a text file containing the estimated cluster membership is saved in the subdirectory corresponding to the model for future validation.

Author(s)

Deepak Nag Ayyala <ayyala.1@osu.edu>

References

Chen, J., et.al. (2012) Associating microbiome composition with environmental covariates using generalized UniFrac distance, Bioinformatics, 28(16).

See Also

GrammRGUI

Examples

```
data(metagencounts)
GrammRServ(Data = metagencounts$Counts, Cluster = metagencounts$CommMemshp,
DataType = "Counts", DistType = "Kendall's tau-distance",
Dim = c(2, 3, 4), LpNorm = c(1,2), Penalty = 0.5, MinClust = 2)
```

`GraphMetagen` *Graphical model construction for metagenomic data*

Description

Given the matrix consisting of metagenomic counts or measure of dissimilarity between samples, multidimensional scaling models are constructed to visualize the samples in the Euclidean space. Clustering methods such as PAM are used to classify the samples into various clusters to study similarity amongst samples.

Usage

```
GraphMetagen(MDSdata)
```

Arguments

`MDSdata` A list containing the following items

- Contents - A matrix consisting of the metagenomic counts or dissimilarity matrix.
- Clust - (Optional) Vector comprising the cluster memberships of samples determined using other data attributes.
- DataType - Determines whether the values in `Contents` are counts or distances. Takes values in c("Counts", "Distance").
- DistType - Determines the measure of dissimilarity to be used when `Contents` contains counts. Takes values in c("Kendall's tau-distance", "UniFrac").
- PhyTree - A phylogenetic tree of class `phylo`. To be provided when `DistType` is set equal to "UniFrac".
- GUnifType - Type of generalized UniFrac distance to be calculated. Takes values in c("Unweighted", "Variance Adjusted", "Generalized")
- GUnifWeight - The weight parameter used in calculation of Generalized UniFrac distance. The parameter takes values between 0 and 1. For more details, see Chen et.al.(2012).
- Dimensions - Integer valued variable which determines the dimensions for which multidimensional scaling models should be constructed.
- Norms - The norm to be used for construction of metric multidimensional scaling models. Takes positive integer values.
- Penalty - A positive number between zero and one which determines the penalty for ties when calculating Kendall's τ-distance.
- MinClust - Minimum number of clusters to be used for estimating the optimal number of clusters. Takes values greater than 2(default value).

Value

`Name` Name of the model constructed.

Coords	A $N \times p$ matrix containing the coordinates of the samples obtained by MDS methods, where N is the number of samples and p is the dimension of the model.
ClusMem	A vector which gives the cluster membership of the samples determined using PAM.
TrueMem	The vector of true cluster membership provided to the function through Clust. If true cluster membership is not provided, it returns a value NULL.
OptimClust	A integer value giving the optimal number of clusters determined by OptimClusts.
SilPlot	A vector of length $2\sqrt{N}-1$, where N is the number of samples. It contains the average silhouette width when the number of clusters is between 2 and $2\sqrt{N}$.

Author(s)

Deepak Nag Ayyala <ayyala.1@osu.edu>

References

Chen, J., et.al. (2012) Associating microbiome composition with environmental covariates using generalized UniFrac distance, Bioinformatics, 28(16).

See Also

GrammRServ

Examples

```
data(metagencounts)
X <- list(Contents = metagencounts$Counts, Clust = metagencounts$CommMemshp,
DataType = "Counts", DistType = "Kendall's tau-distance",
Dimensions = c(2,3,4), Norms = c(1,2), Penalty = 0.5, MinClust = 2);
GraphMetagen(X);
```

KenDist *Wrapper for the C program which calculates the Kendall's τ-distance.*

Description

This function calculates the Kendall's τ-distance from the metagenomic count matrix.

Usage

```
KenDist(Data, Penalty)
```

Arguments

Data	A $N \times k$ matrix comprising the metagenomic count data. If the rows correspond to the samples and the columns correspond to taxa/OTUs, elements of the resulting distance matrix measure dissimilarity between samples.
Penalty	A number between 0 and 1 which determines the penalty for ties.

Value

A $N \times N$ symmetric dissimilarity matrix, where N is the number of samples. If

Author(s)

Deepak Nag Ayyala <ayyala.1@osu.edu>

References

Fagin, Ronald et.al. (2004) Comparing partial rankings, SIAM Journal on Discrete Mathematics, 20, 47-58.

See Also

Count2Distance

Examples

```
data(metagencounts)
Distance <- KenDist(Data = metagencounts$Counts, Penalty = 0.5)

## The result obtained in the above example is the same as
## Not run: Distance <- Count2Distance(Data = metagencounts$Counts,
Distance = "Kendall's tau-distance", Penalty = 0.5);
## End(Not run)
```

MakeGUIPlots *Construct plots for multidimensional scaling models.*

Description

Given the coordinates and estimated cluster membership of samples for a multidimensional scaling model, graphical visualizations of the estimated models are constructed. The graphical models are stored in a subdirectory within the current working directory for reuse. This function is used in GrammRGUI for constructing graphical representations and displaying them as gWidgets notebooks.

Usage

```
MakePlots2D(GraphQuant)
MakePlots3D(GraphQuant)
MakePlots4D(GraphQuant)
```

Arguments

GraphQuant	A list containing the estimated coordinates, cluster membership and average silhouette widths. The list is obtained as an outcome of GraphMetagen

Value

PlotTabs — An object of type gnotebook which contains tabs for different graphical models constructed. The constructed notebook consists of separate tabs for the average silhouette plot, estimated model, model showing estimated clusters and (optional)model showing true clusters. Plots are also saved in a directory created within the current working directory for future use.

The model tabs for two dimensional models display the graphical model, whereas for higher dimensional models, the models tabs contain buttons to display the models in the default web browser.

Author(s)

Deepak Nag Ayyala <ayyala.1@osu.edu>

See Also

Make2DPlots, Make3DPlots, Make4DPlots

MakeServPlots *Construct plots for multidimensional scaling models.*

Description

Given the coordinates and estimated cluster membership of samples for a multidimensional scaling model, graphical visualizations of the estimated models are constructed. The graphical models are stored in a subdirectory within the current working directory for reuse. This function is used in GrammRServ for constructing graphical representations.

Usage

```
Make2DPlots(GraphQuant)
Make3DPlots(GraphQuant)
Make4DPlots(GraphQuant)
```

Arguments

GraphQuant — A list containing the estimated coordinates, cluster membership and average silhouette widths. The list is obtained as an outcome of GraphMetagen

Value

This function creates a directory within the current working directory to save the graphical representations. The name of the directory created will reflect the model constructed.

For two dimensional models, plots constructed and saved in the directory are the silhouette plot, estimated model, model showing the estimated optimal clusters and (optional) the estimated model showing true clusters.

Three and four dimensional models create further subdirectories within the created model directory to save HTML files which display the models in the default web browser.

Author(s)

Deepak Nag Ayyala <ayyala.1@osu.edu>

See Also

MakePlots2D, MakePlots3D, MakePlots4D

MatrixkNorm *Calculate l_p-norm distance between samples*

Description

Given a $N \times k$ matrix X, this function calculates the l_p norm between rows.

Usage

```
MatrixkNorm(X, p)
```

Arguments

X	A $N \times k$ matrix
p	A positive integer value which determines the norm to be used. When $k = \infty$, the maximum norm is calculated.

Value

A $N \times N$ symmetric matrix, where N is the number of rows of the argument matrix X. All the diagonal elements are zeroes and the $(i, j)^{th}$ element represents the l_p-norm distance between the i^{th} and j^{th} rows, given by $\left(\sum_{s=1}^{k} |X_{is} - X_{js}|^p\right)^{1/p}$

Author(s)

Deepak Nag Ayyala <ayyala.1@osu.edu>

Examples

```
data(metagencounts)
Distance <- MatrixkNorm(metagencounts$Counts, p = 2);
```

MCError *Misclassification Error*

Description

Given the true cluster classification of the samples based on some pre-determined criterion and an estimated cluster membership determined using a clustering algorithm, this function calculates the misclassification error of the algorithm.

Usage

```
MCError(True, Est)
```

Arguments

True	A $N \times 1$ vector consisting integer values ranging between 1 and M, where N is the number of samples and M is the number of clusters in the true cluster membership .
Est	A vector whose length is the same as True, whose values range between 1 and K, where K is the estimated number of clusters.

Value

A numeric between 0 and 1. If the vectors are of unequal lengths, the function returns NA.

Author(s)

Deepak Nag Ayyala <ayyala.1@osu.edu>

Examples

```
True <- rep(seq(1,6), rep(5,6))
Est <- rep( seq(1,6), 5);
MCError(True, Est);

## Following is an example of complete mismatch, where the misclassification error is equal to 1.
True <- rep(1,10);
Est <- seq(1,10);
MCError(True, Est)
```

metagencounts *Randomly generated Metagenomic Counts*

Description

This data comprises random metagenomic counts generated using simulation model II described in the reference below. The counts are recorded for a total of 80 samples over 200 taxa. The 80 samples are divided into eight communities of equal size.

Usage

```
data(metagencounts)
```

Format

data.frame

References

Ayyala, D. N., Lin, S., (2014) Graphical Representation and Modeling of Metagenomic Reads, *Manuscript*

OptimClusts *Optimal Cluster Calculator*

Description

Given the average silhouette width obtained using partitioning around medoids(PAM) method, this function determines the optimal number of clusters to be used by calculating the maximum average silhouette width. The absolute maximum silhouette width is not a representative of the optimal number of clusters. OptimClusts calculates the optimal number as the smallest value such that the silhouette width at that value is a local maxima, and is within a neighbourhood of the global maxima.

Usage

```
OptimClusts(P, Eps)
```

Arguments

P	Vector of average silhouette widths calculated for a specified number of clusters.
Eps	A numerical value between 0 and 1 which determines the neighbourhood of the global maximum within which to search for a local maxima. It is advised to use values smaller than 10 %.

Details

The function `OptimClusts` uses the mPAM (modified PAM) algorithm described in the first reference below. For a data set with N samples (or taxa/OTUs when clustering taxa/OTUs), the value of K to be used to avoid overestimation of clusters is $\left[2\sqrt{N}\right]$, where $[x]$ is the largest integer smaller than x.

Value

An integer value between 1 and K, where K is the length of the silhouette vector P. If the minimum and maximum number of clusters specified are m and M respectively, the value represents the index of the optimal number of clusters to be used in the vector (m, M). See Details for information on the maximum number of clusters.

Author(s)

Shili Lin<shili@stat.osu.edu>

References

Ayyala, D. N., Lin, S., (2014) Graphical Representation and Modeling of Metagenomic Reads, *Manuscript*.

Peter J. Rousseeuw (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics*, **20**.

Examples

```
x <- c(0.5, 0.1, 0.6, 0.7, 0.8, 0.75, 0.77, 0.79, 0.81, 0.9)
## Not run: plot(2:10, x)
OptimClusts(x, 0.1) ## The optimal number selected is 6.
OptimClusts(x, 0.05) ## The optimal number selected is 10.
```

Summarize *Display summary of the graphical model constructed.*

Description

This function is supplementary to `MakeGUIPlots`. It constructs summary of the graphical model for displaying in the GUI. Summary includes number of samples, estimated number of clusters, cluster composition of true(if provided) and estimated clusters and the misclassification error.

Usage

```
Summarize(GraphQuant)
```

Arguments

`GraphQuant`	A list generated by `GraphMetagen`.

Value

gWidgets quantity of type `gtext` used to embed in the graphical interface.

Author(s)

Deepak Nag Ayyala <ayyala.1@osu.edu>

See Also

`GraphMetagen`, `MakePlots2D`, `MakePlots3D`, `MakePlots4D`

Index