Package 'MiRKAT'

April 25, 2018

Version 1.0.1

Date 2018-04-25

Title Microbiome Regression-Based Kernel Association Test

Author Haotian Zheng [aut],
Xiang Zhan [aut],
Anna Plantinga [aut],
Michael Wu [aut],
Ni Zhao [aut, cre],
Jun Chen [aut]

Maintainer Ni Zhao <nzhao10@jhu.edu>

Depends R (>= 2.13.0), survival, PearsonDS, GUniFrac, MASS

Description Test for overall association between microbiome composition data with a continuous or dichotomous outcome via phylogenetic kernels. The phenotype can be univariate continuous or binary phenotypes (Zhao et al. (2015) <doi:10.1016/j.ajhg.2015.04.003>), survival outcomes (Plantinga et al. (2017) <doi:10.1186/s40168-017-0239-9>), multivariate (Zhan et al. (2017) <doi:10.1002/gepi.22030>) and structured phenotypes (Zhan et al. (2017) <doi:10.1111/biom.12684>). For all these effect, the microbiome community effect was modeled nonparametrically through a kernel function, which can incorporate the phylogenetic tree information.

License GPL (>= 2)

NeedsCompilation yes

Repository CRAN

Date/Publication 2018-04-25 19:44:09 UTC

R topics documented:

D2K *Construct kernel matrix from distance metric*

Description

Construct kernel matrix from distance matrix (matrix of pairwise distances) for microbiome data.

Usage

```
D2K(D)
```

Arguments

D	An n by n matrix giving pairwise distances or dissimilarities, where n is the sample size.

Details

The kernel matrix is constructed as $K = -(I - 11'/n)D^2(I - 11'/n)/2$, where D is the pairwise distance matrix, I is an identity matrix and 1 is a vector of 1. D^2 represents element wise square. To ensure that K to be positive semi-definite, a positive semi-definiteness correction is conducted that the negative eigen values of K are replaced by zeros.

Value

An n by n kernel or similarity matrix corresponding to the distance matrix.

Author(s)

Ni Zhao

References

Zhao, Ni, et al. "Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test." The American Journal of Human Genetics 96.5 (2015): 797-807.

Examples

```
 ############################################################
require(GUniFrac)
# Load in data and create a distance metric
data(throat.tree)
data(throat.otu.tab)
unifracs = GUniFrac(throat.otu.tab, throat.tree, alpha = c(1))$unifracs
D1 = unifracs[,,"d_1"]
# Function
K = D2K(D1)
```

KRV *Kernel RV Coefficient Test*

Description

kernel RV coefficient test to evaluate the overall association between microbiome composition and high-dimensional or structured phenotype.

Usage

```
KRV(kernel.otu, y = NULL, X = NULL, kernel.y)
```

Arguments

kernel.otu	A numerical n by n kernel matrix. It can be constructed from microbiome data, such as by transforming from a distance metric.
y	A numerical n by p matrix of p continuous phenotype variables (Default = NULL). If it is NULL, a phenotype kernel matrix must be entered for "kernel.y". No need to provide if kernel.y is a matrix.
X	A numerical n by q matrix, containing q additional covariates that you want to adjust for (Default = NULL). If it is NULL, a intercept only model was fit. Covariates can't be adjusted for if kernel.y is a matrix.
kernel.y	Either a numerical n by n kernel matrix of phenotype or a method to compute the kernel of phenotype. Gaussian kernel (kernel.y="Gaussian") can capture general relationship between microbiome and phenotypes; and linear kernel (kernel.y="linear") can be preferred if the underlying relationship is close to linear.

Details

kernel.otu should be a numerical n by n kernel matrix, where n is sample size.

When kernel.y is a method ("Gaussian" or "linear") to compute the kernel of phenotype, y should be a numerical phenotype matrix, and X (if not NULL) should be a numerical matrix of covariates. Both y and X should have n rows.

When kernel.y is a kernel matrix of phenotype, there is no need to provide X and y, and they will be ignored if provided. In this case, kernel.y and kernel.otu should both be numerical matrices with the same number of rows and columns.

Missing data is not permitted. Please remove all individuals with missing kernel.otu, y (if not NULL), X (if not NULL), and kernel.y (if a matrix is entered) prior to analysis.

Value

P-value calculated from approximated Pearson type III density

Author(s)

Haotian Zheng, Xiang Zhan, Ni Zhao

References

Zhan, X., Plantinga, A., Zhao, N., and Wu, M.C. A Fast Small-Sample Kernel Independence Test for Microbiome Community-Level Association Analysis. Biometrics. 2017 Mar 10. doi: 10.1111/biom.12684.

Examples

```
library(MASS)
library(GUniFrac)
data(throat.tree)
data(throat.otu.tab)
data(throat.meta)
attach(throat.meta)

set.seed(123)
n = nrow(throat.otu.tab)
Male = (Sex == "Male")**2
Smoker =(SmokingStatus == "Smoker") **2
anti =  (AntibioticUsePast3Months_TimeFromAntibioticUsage != "None")^2
cova = cbind(Male,  anti)

otu.tab.rff <- Rarefy(throat.otu.tab)$otu.tab.rff
unifracs <- GUniFrac(otu.tab.rff, throat.tree, alpha=c(0, 0.5, 1))$unifracs

D.weighted = unifracs[,,"d_1"]
D.unweighted = unifracs[,,"d_UW"]
D.BC= as.matrix(vegdist(otu.tab.rff , method="bray"))

K.weighted = D2K(D.weighted)
K.unweighted = D2K(D.unweighted)
K.BC = D2K(D.BC)

rho = 0.2
Va = matrix(rep(rho, (2*n)^2), 2*n, 2*n)+diag(1-rho, 2*n)
G = mvrnorm(n, rep(0, 2*n), Va)

############################################################

KRV(kernel.otu = K.weighted, y = G, X = cova, kernel.y = "Gaussian")
KRV(kernel.otu = K.weighted, kernel.y = G %*% t(G))
```

MiRKAT *Microbiome Regression-based Kernel Association Test*

Description

Test for association between microbiome composition and a continuous or dichotomous outcome by incorporating phylogenetic or nonphylogenetic distance between different microbiomes.

Usage

```
MiRKAT(y, X = NULL, Ks, out_type = "C", nperm = 999, method = "davies")
```

Arguments

y	A numeric vector of the a continuous or dichotomous outcome variable.
X	A numerical matrix or data frame, containing additional covariates that you want to adjust for (Default = NULL). If it is NULL, a intercept only model was fit.
Ks	a list of n by n kernel matrices (or a single n by n kernel matrix), where n is the sample size. It can be constructed from microbiome data through distance metric or other approaches, such as linear kernels or Gaussian kernels.
out_type	an indicator of the outcome type. "C" for the continuous outcome and "D" for the dichotomous outcome.
nperm	the number of permutations if method = "permutation" or when multiple kernels are considered. if method = "davies" or "moment", nperm is ignored.
method	a method to compute the kernel specific p-value (Default= "davies"). "davies" represents an exact method that computes the p-value by inverting the characteristic function of the mixture chisq. We adopt an exact variance component tests because most of the studies concerning microbiome compositions have modest sample size. "moment" represents an approximation method that matches the first two moments. "permutation" represents a permutation approach for p-value calculation.

Details

y and X (if not NULL) should all be numerical matrices or vectors with the same number of rows.

Ks should be a list of n by n matrices or a single matrix. If you have distance metric from metagenomic data, each kernel can be constructed through function D2K. Each kernel can also be constructed through other mathematical approaches.

Missing data is not permitted. Please remove all individuals with missing y, X, Ks prior to analysis

Parameter "method" only concerns with how kernel specific p-values are generated. When Ks is a list of multiple kernels, omnibus p-value is computed through permutation from each individual p-values, which are calculated through method of choice.

Value

indivP	p-value from each candidate kernel
omnibus_p	omnibus p value by considering multiple candidate kernels.

Author(s)

Ni Zhao

References

Zhao, N., Chen, J.,Carroll, I. M., Ringel-Kulka, T., Epstein, M.P., Zhou, H., Zhou, J. J., Ringel, Y., Li, H. and Wu, M.C. (2015)). Microbiome Regression-based Kernel Association Test (MiRKAT). American Journal of Human Genetics, 96(5):797-807

Chen, J., Chen, W., Zhao, N., Wu, M~C.and Schaid, D~J. (2016) Small Sample Kernel Association Tests for Human Genetic and Microbiome Association Studies. 40: 5-19. doi: 10.1002/gepi.21934

Davies R.B. (1980) Algorithm AS 155: The Distribution of a Linear Combination of chi-2 Random Variables, *Journal of the Royal Statistical Society. Series C* , 29, 323-333.

Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. Biom. Bull. 2, 110-114.

Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA; NHLBI GO Exome Sequencing Project-ESP Lung Project Team, Christiani DC, Wurfel MM, Lin X. (2012) Optimal unified approach for rare variant association testing with application to small sample case-control whole-exome sequencing studies. *American Journal of Human Genetics*, 91, 224-237.

Zhou, J. J. and Zhou, H.(2015) Powerful Exact Variance Component Tests for the Small Sample Next Generation Sequencing Studies (eVCTest), in submission.

Examples

```
############################################################
# Generate data
set.seed(1)
n = 100
family= "binomial"
nperm = 999
X = rnorm(n)
Z = matrix((runif(n*5) > 0.5)^2, n, 5)
K = Z %*% t(Z)
K2 = (Z %*% t(Z)+ 1)^2
Ks = list(K, K2)
y = rnorm(n)

MiRKAT(y, X = X, Ks = Ks, out_type="C", method = "davies")

################################
y = (runif(n) < 0.5)^2
MiRKAT(y, X = X, Ks = Ks, out_type="D")
```

MiRKATS *Microbiome Regression-based Kernel Association Test for Survival*

Description

Community level test for association between microbiome composition and survival outcomes (right-censored time-to-event data) using kernel matrices to compare similarity between microbiome profiles with similarity in survival times.

Usage

```
MiRKATS(kd, distance = FALSE, obstime, delta, covar = NULL, beta = NULL,
perm = FALSE, nperm = 1000)
```

Arguments

kd	A numeric n by n kernel matrix or matrix of pairwise distances/dissimilarities (where n is the sample size).
distance	Logical, indicating whether kd is a distance matrix (default = FALSE).
obstime	A numeric vector of follow-up (survival/censoring) times.
delta	Event indicator: a vector of 0/1, where 1 indicates that the event was observed for a subject (so "obstime" is survival time), and 0 indicates that the subject was censored.
covar	A vector or matrix of numeric covariates, if applicable (default = NULL).
beta	A vector of coefficients associated with covariates. If beta is NULL and covariates are present, coxph is used to calculate coefficients (default = NULL).
perm	Logical, indicating whether permutation should be used instead of analytic p-value calculation (default=FALSE). Not recommended for sample sizes of 100 or more.
nperm	Integer, number of permutations used to calculate p-value if perm==TRUE (default=1000).

Details

obstime, delta, and covar should all have n rows, and the kernel or distance matrix should be a single n by n matrix. If a distance matrix is entered (so distance=TRUE), a kernel matrix will be constructed from the distance matrix.

Missing data is not permitted. Please remove individuals with missing data on y, X or in the kernel or distance matrix prior to using the function.

The Efron approximation is used for tied survival times.

Value

P-value obtained using small sample correction

Author(s)

Anna Plantinga

References

Plantinga, A., Zhan, X., Zhao, N., Chen, J., Jenq, R., and Wu, M.C. MiRKAT-S: a distance-based test of association between microbiome composition and survival times. Microbiome, 2017:5-17. doi: 10.1186/s40168-017-0239-9

Zhao, N., Chen, J.,Carroll, I. M., Ringel-Kulka, T., Epstein, M.P., Zhou, H., Zhou, J. J., Ringel, Y., Li, H. and Wu, M.C. (2015)). Microbiome Regression-based Kernel Association Test (MiRKAT). American Journal of Human Genetics, 96(5):797-807

Chen, J., Chen, W., Zhao, N., Wu, M~C.and Schaid, D~J. (2016) Small Sample Kernel Association Tests for Human Genetic and Microbiome Association Studies. 40:5-19. doi: 10.1002/gepi.21934

Efron, B. (1977) "The efficiency of Cox's likelihood function for censored data." Journal of the American statistical Association 72(359):557-565.

Davies R.B. (1980) Algorithm AS 155: The Distribution of a Linear Combination of chi-2 Random Variables, Journal of the Royal Statistical Society Series C, 29:323-333.

Examples

```
###################################
# Generate data
require(GUniFrac)
set.seed(1)

# Throat microbiome data
data(throat.tree)
data(throat.otu.tab)
unifracs = GUniFrac(throat.otu.tab, throat.tree, alpha = c(1))$unifracs
D1 = unifracs[,,"d_1"]  # 60 subjects

# Covariates and outcomes
X <- matrix(rnorm(120), nrow=60)
S <- rexp(60, 3)
C <- rexp(60, 1)
D <- (S<=C)        # event indicator
U <- pmin(S, C)    # observed follow-up time

MiRKATS(kd = D1, distance = TRUE, obstime = U, delta = D, covar = X, beta = NULL)
```

`MMiRKAT` *Multivariate Microbiome Regression-based Kernel Association Test*

Description

Test for association between overall microbiome composition and multiple continuous outcomes.

Usage

```
MMiRKAT(Y, X=NULL, K)
```

Arguments

Y	A numerical n by p matrix of p continuous outcome variables.
X	A numerical n by q matrix or data frame, containing q additional covariates that you want to adjust for (Default = NULL). If it is NULL, a intercept only model was fit.
K	A numerical n by n kernel matrix, where n is the sample size. It can be constructed from microbiome data base on Bray-Curtis or UniFrac distance of microbiome data.

Details

Y and X (if not NULL) should all be numerical matrices with the same number of rows n.

K should be a single n by n matrix. If you have a distance matrix from metagenomic data, a kernel can be constructed through function D2K. The kernel can also be constructed through other mathematical approaches.

Missing data is not permitted. Please remove all individuals with missing Y, X, K prior to analysis

The method of generating kernel specific p-values is "davies", which represents an exact method that computes the p-value by inverting the characteristic function of the mixture chisq.

Value

P-value obtained using small-sample correction

Author(s)

Haotian Zheng, Xiang Zhan, Ni Zhao

References

Zhan, X., Tong, X., Zhao, N., Maity,A., Wu, M.C., and Chen,J. A small-sample multivariate kernel machine test for microbiome association studies. Genetic Epidemiology, 41(3), 210-220. DOI: 10.1002/gepi.22030

Examples

```
library(GUniFrac)
data(throat.tree)
data(throat.otu.tab)
data(throat.meta)
attach(throat.meta)

set.seed(123)
n = nrow(throat.otu.tab)
Male = (Sex == "Male")**2
Smoker =(SmokingStatus == "Smoker") **2
anti =  (AntibioticUsePast3Months_TimeFromAntibioticUsage != "None")^2
cova = cbind(Male,  anti)

otu.tab.rff <- Rarefy(throat.otu.tab)$otu.tab.rff
```

```
unifracs <- GUniFrac(otu.tab.rff, throat.tree, alpha=c(0, 0.5, 1))$unifracs

D.weighted = unifracs[,,"d_1"]
D.unweighted = unifracs[,,"d_UW"]
D.BC= as.matrix(vegdist(otu.tab.rff , method="bray"))

K.weighted = D2K(D.weighted)
K.unweighted = D2K(D.unweighted)
K.BC = D2K(D.BC)

Y = matrix(rnorm(n * 3, 0, 1),n ,3)

############################################################

MMiRKAT(Y = Y, K = K.weighted, X = cbind(Male, anti))
```

`throat.meta` *Meta data of the throat microbiome samples*

Description

This data set includes samples from the microbiome of the nasopharynx and oropharynx on each side of the body. It were generated to study the effect of smoking on the microbiota of the upper respiratory tract in 60 individuals, 28 smokers and 32 nonsmokers.

Usage

```
data("throat.meta")
```

Format

A data frame with 60 observations on 16 variables.

Source

Charlson ES, Chen J, Custers-Allen R, Bittinger K, Li H, et al. (2010) Disordered Microbial Communities in the Upper Respiratory Tract of Cigarette Smokers. PLoS ONE 5(12): e15216.

References

R package "GUniFrac"

Examples

```
data(throat.meta)
```

throat.otu.tab — *OTU count table from 16S sequencing of the throat microbiome samples*

Description

This data set contains 60 subjects with 28 smokers and 32 nonsmokers. Microbiome data were collected from right and left nasopharynx and oropharynx region to form an OTU table with 856 OTUs.

Usage

```
data("throat.otu.tab")
```

Format

A data frame with 60 observations on 856 variables.

Source

Charlson ES, Chen J, Custers-Allen R, Bittinger K, Li H, et al. (2010) Disordered Microbial Communities in the Upper Respiratory Tract of Cigarette Smokers. PLoS ONE 5(12): e15216.

References

R package "GUniFrac"

Examples

```
data(throat.otu.tab)
```

throat.tree — *UPGMA tree of the OTUs from 16S sequencing of the throat microbiome samples*

Description

The OTU tree is constructed using UPGMA on the K80 distance matrix of the OTUs. It is a rooted tree of class "phylo".

Usage

```
data("throat.tree")
```

Format

List of 4 data frames.

Source

Charlson ES, Chen J, Custers-Allen R, Bittinger K, Li H, et al. (2010) Disordered Microbial Communities in the Upper Respiratory Tract of Cigarette Smokers. PLoS ONE 5(12): e15216.

References

R package "GUniFrac"

Examples

```
data(throat.tree)
```

Index