MiRKAT-S Package

Anna Plantinga

November 22, 2016

Overview

The R package MiRKAT-S contains functions to test an association between the microbiota (taxonomic profiles of a microbial community) and suvival outcomes via an kernel matrix. The kernel matrix is based upon distance or dissimilarity metrics designed to summarize ecological and/or microbial communities, possibly incorporating phylogenetic information.

Required packages

CompQuadForm, survival, and GUniFrac packages are required for MiRKAT-S. All three required packages are available on CRAN (https://cran.r-project.org/web/packages/). In this vignette, we use the throat data from GUniFrac for demonstration. In our demonstration, kernels are constructed from the family of UniFrac distances and the Bray-Curtis dissimilarity; however, additional options for relevant kernels are mentioned.

Using MiRKAT-S to test associations between microbiota and survival

Load dataset and prepare data

We use the throat microbiome data from package GUniFrac to demonstrate the use of MiRKAT-S. These data were generated to study the effect of smoking on the microbiota of the upper respiratory tract in 60 individuals, 28 smokers and 32 nonsmokers. As such, the original dataset includes samples from the microbiome of the nasopharynx and oropharynx on each side of the body. The dataset used here consists of the oropharynx samples from the left side of the body.

Because the dataset has a binary phenotype (smoking) rather than a measure of censored time to event outcomes, we consider smoking status and gender as covariates and generate null outcome data from the Exponential distribution. Specifically, we generate survival times as $S \sim \text{Exponential}(1 + I(\text{smoke}) + I(\text{male}))$, and censoring times as $C \sim \text{Exponential}(0.75)$. Then the outcome measure consists of $T = \min(S, C)$ and $\Delta = I(S \leq C)$. This simulation procedure results in approximately 33% censoring.

```r
# Load the data
library(MiRKATS, quietly=TRUE)
```

```
## This is vegan 2.4-1
```

```r
library(GUniFrac, quietly=TRUE)
data(throat.tree)
data(throat.otu.tab)
data(throat.meta)

# Prepare covariates
set.seed(1)
Male = rep(0, nrow(throat.meta))
Male[throat.meta$Sex == "Male"] <- 1
Smoker = rep(0, nrow(throat.meta))
```

```r
Smoker[throat.meta$SmokingStatus == "Smoker"] <- 1

# Simulate outcomes
# Here, outcome is associated with covariates but unassociated with microbiota
# 33% censoring
SurvTime <- rexp(60, (1 + Smoker + Male))
CensTime <- rexp(60, 0.75)
Delta <- as.numeric( SurvTime <= CensTime )
ObsTime <- pmin(SurvTime, CensTime)
```

Create distance matrices

Many options exist for distance or dissimilarity metrics. Below, we calculate the distances used in the manuscript: the unweighted and weighted UniFrac distances, the generalized UniFrac distance with $\alpha = 0.5$, and the Bray-Curtis dissimilarity. Choice of distance metric (and therefore kernel) is discussed following the example.

```r
unifracs <- GUniFrac(throat.otu.tab, throat.tree, alpha=c(0, 0.5, 1))$unifracs
D.weighted <- unifracs[,,"d_1"]
D.unweighted <- unifracs[,,"d_UW"]
D.generalized <- unifracs[,,"d_0.5"]
D.braycurtis <- as.matrix(vegdist(throat.otu.tab, method="bray"))
```

Convert distance matrices to kernel matrices

The D2K function in MiRKAT-S converts distance matrices to kernel matrices via

$$K = -\frac{1}{2}\left(I - \frac{11'}{n}\right) D^2 \left(I - \frac{11'}{n}\right).$$

Here, I is the identity matrix and 1 is an n-vector of ones. To ensure that K is positive semi-definite, we replace negative eigenvalues with zero. That is, we perform an eigenvalue decomposition $K = U\Lambda U$, where $\Lambda = \text{diag}(\lambda_1, ..., \lambda_n)$, and then reconstruct the kernel matrix using the nonnegative eigenvalues $\Lambda^* = \text{diag}(\max(\lambda_1, 0), ..., \max(\lambda_n, 0))$ so that $K = U\Lambda^* U$.

Alternatively, a distance matrix can be used as input to MiRKATS (argument kd) with argument distance=TRUE; the distance will be automatically converted to a kernel matrix exactly as described above.

```r
K.weighted <- D2K(D.weighted)
K.unweighted <- D2K(D.unweighted)
K.generalized <- D2K(D.generalized)
K.braycurtis <- D2K(D.braycurtis)
```

Test using a single kernel, MiRKAT-S p-values

Here, we run MiRKAT-S to test the association between the microbiota and simulated survival times, adjusting for gender and smoking status.

```r
# use kernel matrix with distance=FALSE
MiRKATS(kd = K.generalized, distance = FALSE, obstime = ObsTime, delta = Delta, covar = cbind(Male, Smok
```

```
## [1] 0.08217796
```

```
# equivalently, use distance matrix with distance=TRUE
MiRKATS(kd = D.generalized, distance = TRUE, obstime = ObsTime, delta = Delta, covar = cbind(Male, Smoke
```

```
## [1] 0.08217796
```

The argument "distance" indicates whether "kd" is a distance matrix (TRUE) or kernel matrix (FALSE). The output is the p-value for the test using Davies' exact method, which computes the p-value based on a mixture of chi-square distributions. We use a small-sample correction to account for the modest sample sizes and sparse OTU count matrices that often result from studies of the microbiome.

Test using a single kernel, permutation p-values

Because there are only 60 individuals in this dataset, permutation is a reasonable choice for p-value calculation, as MiRKAT-S may be slightly anti-conservative with very small samples. MiRKAT-S will generate a warning when permutation is not used for sample sizes $n \leq 50$. "nperm" indicates the number of permutations to perform to generate the p-value (default = 1000).

```
MiRKATS(kd = K.generalized, distance = FALSE, obstime = ObsTime, delta = Delta, covar = cbind(Male, Smok
```

```
## [1] 0.12
```

Selection of kernel(s) for testing

How to choose an appropriate distance matrix and kernel for testing is a difficult question. However, it is important, since the distance metric used to generate the kernel for MiRKAT-S strongly affects the power of the test. In particular, MiRKAT-S has highest power when the form of association between the microbiota and the outcome assumed by the kernel matches the true form of association.

In the case of the UniFrac families and the Bray-Curtis dissimilarity, the factors at play are (1) the abundance of the associated taxa and (2) whether closely related taxa (phylogenetically) tend to be related or not related to the outcome as a group. For example, the following are some of the distance metrics that have been used for studies of the microbiome:

Distance	Phylogeny?	Abundance?	Other notes	Reference
Unweighted UniFrac	Yes	No		1
Weighted UniFrac	Yes	Yes		2
Generalized UniFrac	Yes	(Yes)	Parameter alpha defines extent to which abundance is taken into account	3
Jaccard	No	No	1 - (taxa in both)/(taxa in either); typically presence/absence, but can be extended to an abundance-weighted version	4,5
Bray-Curtis	No	Yes	Similar to Jaccard, but uses counts	6

In the table above, "Yes" indicates the distance or dissimilarity metric has the feature; "(Yes)" indicates that the feature is present either in some variations of the metric or is present to some extent; and "No" indicates that the feature is not present.

Depending on which of these characteristics are expected to be present in a particular study (based on prior knowledge or intuition), an appropriate distance or dissimilarity can be selected. If the study is exploratory and strong protection of type 1 error is not needed, several distance metrics can be explored. Depending on which one(s) are highly significant, some information can be gained about the nature of any association between the microbiota and the outcome.

References

Distances and dissimilarities

1. Lozupone C & Knight R (2005). UniFrac: a new phylogenetic method for comparing microbial communities. Applied and Environmental Microbiology, 71(12), 8228-8235.
2. 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), 1576-1585.
3. 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), 2106-2113.
4. Levandowsky M & Winter D (1971). Distance between sets. Nature, 234(5323), 34-35.
5. Chao A, Chazdon RL, Colwell RK, & Shen TJ (2005). A new statistical approach for assessing similarity of species composition with incidence and abundance data. Ecology Letters, 8(2), 148-159.
6. Bray JR & Curtis JT (1957). An ordination of the upland forest communities of southern Wisconsin. Ecological Monographs, 27(4), 325-349.

MiRKAT-S method

1. Plantinga A, Zhan X, Zhao N, Chen J, Jenq RR, Wu MC (2016). MiRKAT-S: A community-level test of association between taxonomic profiles and survival times. Under revision.
2. Zhao N, Chen J, Carroll IM, Ringel-Kulka T, Epstein MP, Zhou H, Zhou JJ, Ringel Y, Li H, Wu MC (2015). Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. The American Journal of Human Genetics, 96(5), 797-807.
3. Davies RB (1980). The distribution of a linear combination of chi-2 random variables. Journal of the Royal Statistical Society Series C (Applied Statistics), 29(3), 323-333.
4. Chen J, Chen W, Zhao N, Wu MC, Schaid DJ (2016). Small sample kernel association tests for human genetic and microbiome association studies. Genetic Epidemiology, 40(1), 5-19.