Tutorial vignette for package 'specificity'

John L Darcy

30 August 2019

Introduction

In this example analysis, our goal is to analyze the extent to which microbes have specificity to different aspects of their environment. We will use the 'endophyte' data set (included with the specificity R package), which contains data from a survey of foliar endophytic fungi living within the leaves of native Hawaiian plants. In this context, a fungus with strong specificity would be one that occupies a narrow range of sampled environment. For example, a fungus that preferentially associates with a narrow clade of host plants (e.g. asteraceae), or prefers a narrow range of elevation (e.g. between 700 and 1200 m.a.s.l). The previous two examples were for unimodal specificity, but specificity as calculated with our software can be multimodal as well (e.g. phylogenetic specificity to both asteraceae and malvaceae). It is worth noting that a strength of this approach is that specificity is agnostic to the ideal habitat of a species: a species with strong specificity to high elevations may be just as specific to elevation as a species specific to low elevations. Indeed, a strength of this approach is that the ideal habitat does NOT need to be modeled in order to calculate specificity. Instead, the standardized effect size (SES) of specificity is a statistic that quantifies the degree to which a species occupies a more narrow range of some environmental variable than would be expected by random chance.

Software Requirements

- `R` (run on version 3.6.0, but likely works fine on earlier versions)
- Other dependencies will be automatically installed with the package.

Installation

To install specificity, open up R and run:

```
  library(devtools)
  install_github("darcyj/specificity")
```

Load 'endophyte' example data set

The endophyte dataset consists of foliar endophytic fungi sampled from leaves of native Hawaiian plants across the Hawaiian archipelago. These data can be loaded into R with:

```
# load specificity R package
suppressMessages(library(specificity))
# load endophyte data set from specificity
data(endophyte)
# check what objects are within endophyte
names(endophyte)
```

```
## [1] "metadata"  "zotutable" "supertree"
```

```
# add those objects to R namespace so we aren't typing "endophyte$" every time
attach(endophyte)
```

As you can see, there are 3 objects inside 'endophyte'.

- metadata: table of data where each row corresponds to a sample, and each column is a different metadata category (e.g. PlantGenus, Elevation, Lon=longitude, Lat=latitude)
- zotutable: table of zOTU (species) observation data, where each row is a sample, and each column is a different fungal zOTU.
- supertree: a phylogenetic tree of host plants.

Pre-processing of 'endophyte' data

Like most statistics, specificity is useless with a low sample size. In the case of specificity, this 'sample size' is the occupancy of a species, or how many samples that species was observed in. Here, we will create a new table containing only species that observed in at least 10 samples. Before we do this, we must also transform the data to proportional abundances, so that each species' observation data are a proportion of the total sample depth, instead of a proportion of the data with rare species excluded. Note that a CLR transformation could be used in place of this proportional transformation, or an ALR transformation could be used in the case where a known invariant species was added experimentally.

```
# transform to proportional abundance
zt_p <- prop_abund(zotutable, speciesRows=FALSE)
# apply occupancy threshold to remove low-occupancy species
zt_p_ovr10 <- occ_threshold(zt_p, threshold=10)
# how many species are we left with?
ncol(zt_p_ovr10)
```

```
## [1] 381
```

Specificity analysis

Calculating specificity for various data types is fairly easy with this package. In this example, we will analyze specificity to elevation, rainfall, host plant phylogeny, and geographic distance. In the function `phy_or_env_spec()`, the `n_sim` argument determines the number of simulations (i.e. permutations) to do. The default value is 1000, but so that this vignette runs more quickly we will use 300 instead. Each analysis below should take 1 or 2 minutes to complete. When you run them, you will see status updates that are ommitted in this vignette for the sake of brevity.

```
# specificity for elevation:
elev_spec <- phy_or_env_spec(zt_p_ovr10, env=metadata$Elevation, n_sim=300, n_cores=3)
```

```
## Checking inputs.

## Converting env vector to dist.

## Generating daughter seeds.

## Creating 300 permuted matrices.

## Calculating specificities for permuted matrices.

## Calculating empirical specificities.

## Calculating P-values.

## Calculating SES.

## Done.
```

```
# specificity for rainfall
rain_spec <- phy_or_env_spec(zt_p_ovr10, env=metadata$Rainfall, n_sim=300, n_cores=3)
```

```
## Checking inputs.
```

```
## Converting env vector to dist.
## Generating daughter seeds.
## Creating 300 permuted matrices.
## Calculating specificities for permuted matrices.
## Calculating empirical specificities.
## Calculating P-values.
## Calculating SES.
## Done.
```

```
# specificity for evapotranspiration
evap_spec <- phy_or_env_spec(zt_p_ovr10, env=metadata$Evapotranspiration, n_sim=300,
    n_cores=3)
```

```
## Checking inputs.
## Converting env vector to dist.
## Generating daughter seeds.
## Creating 300 permuted matrices.
## Calculating specificities for permuted matrices.
## Calculating empirical specificities.
## Calculating P-values.
## Calculating SES.
## Done.
```

```
# specificity for host phylogeny
host_spec <- phy_or_env_spec(zt_p_ovr10, hosts=metadata$PlantGenus,
    hosts_phylo=supertree, n_sim=300, n_cores=3)
```

```
## Checking inputs.
## Converting tree to dist.
## Generating daughter seeds.
## Creating 300 permuted matrices.
## Calculating specificities for permuted matrices.
## Calculating empirical specificities.
## Calculating P-values.
## Calculating SES.
## Done.
```

```
# specificity for geographic distance
geo_distmat <- distcalc(lat=metadata$Lat, lng=metadata$Lon, sampIDs=metadata$SampleID)
geo_spec <- phy_or_env_spec(zt_p_ovr10, env=geo_distmat, n_sim=300, n_cores=3)
```

```
## Checking inputs.
## Generating daughter seeds.
## Creating 300 permuted matrices.
```

```
## Calculating specificities for permuted matrices.
## Calculating empirical specificities.
## Calculating P-values.
## Calculating SES.
## Done.
```

Visualization

Our built-in visualization function takes a list of results from `phy_or_env_spec()` as input, and produces a violin plot composed of bars colored by statistical significance. The names of items within `specs_list` will be the labels used in the plot.

```
plot_specificities(specs_list=list(
    "Elevation"=elev_spec,
    "Rainfall"=rain_spec,
    "Evapotranspiration"=evap_spec,
    "Host Plant"=host_spec,
    "Geographic Distance"=geo_spec
), n_bins=60)
```