canaper: Categorical analysis of neo- and paleo-endemism in R

Joel H. Nitta[1]

[1] University of Tokyo, Department of Biological Sciences

joelnitta@gmail.com

https://joelnitta.com

https://github.com/joelnitta/canaper

Background

- CANAPE (**C**ategorical **A**nalysis of **N**eo- and **P**aleo-endemism) is a recently developed method that provides insight into the evolutionary processes underlying endemism [1]
- CANAPE integrates a phylogenetic tree with a community (species × sites) matrix to infer if endemic areas are so because of recent speciation (neoendemism), or because they harbor old lineages that have mostly gone extinct in other areas (paleoendemism), or if they are a mixture of both
- CANAPE is currently only implemented in Biodiverse [2], a program written in perl that is used via a GUI or custom scripts.
- **`canaper` is a new R package to conduct CANAPE entirely in R**

Features

- Sparse matrix encoding of community matrices to increase computing efficiency via `phyloregion` [3]
- Simple implementation of parallel computing to increase speed via `future`

Installation and loading

`canaper` is currently available on GitHub

```
devtools::install_github("joelnitta/canaper")
```

```
# Load packages
library(canaper)
library(ape) # For handling phylogenies
library(future) # For parallel computing
library(tidyverse) # For data wrangling and visualization
```

Example: Australian *Acacia*

This demonstrates the package using the dataset of the paper where CANAPE was first published [1]: 506 species of *Acacia* in Australia distributed over 3037 sites:

```
# Phylogenetic tree
acacia$phy
#>
#> Phylogenetic tree with 510 tips and 509 internal nodes.
#>
#> Tip labels:
#>    Pararchidendron_pruinosum, Paraserianthes_lophantha, adinophylla,
       semicircinalis, aphanoclada, inaequilatera, ...
#>
#> Rooted; includes branch lengths.

# Community data matrix (in part).
# Rownames correspond to geographical coordinates
acacia$comm[1:4, 1:4]
#>                     abbreviata acanthaster acanthoclada acinacea
#> '-1025000:-1825000'          0           0            0        0
#> '-1025000:-1875000'          0           0            0        0
#> '-1025000:-1925000'          0           0            0        0
#> '-1025000:-1975000'          0           0            0        0
```

Randomization test

The first step of CANAPE is to compare the observed values of phylogenetic endemism (PE) and alternative PE (PE measured on a modified tree where all branch lengths are set equal) with those from a set of random communities. The `cpr_rand_test()` conducts the randomization, using parallel computing to increase speed. The `picante` package [4] is used to generate the random communities.

```
# Set a parallel back-end, with 4 CPUs running simultaneously
plan(multisession, workers = 4)

# Run randomization test
acacia_rand_res <- cpr_rand_test(
  acacia$comm, acacia$phy,
  n_reps = 100, n_iterations = 100000)
#> [1] "Dropping tips from the tree because they are not present in
       the community data:"
#> [1] "Pararchidendron_pruinosum" "Paraserianthes_lophantha"
#> [3] "saligna"                   "clunies-rossiae"

# Check some of the results
acacia_rand_res %>%
  slice(1:3) %>%
  select(pe_obs, pe_rand_mean, pe_obs_p_upper, pe_obs_p_lower) %>%
  as_tibble()
#> # A tibble: 3 x 4
#>      pe_obs pe_rand_mean pe_obs_p_upper pe_obs_p_lower
#>       <dbl>        <dbl>          <dbl>          <dbl>
#> 1 0.0000248     0.000130           0.09           0.91
#> 2 0.000145      0.000297           0.21           0.79
#> 3 0.000172      0.000232           0.58           0.42
```

Output summary (in part):

- `pe_obs` = observed PE
- `pe_rand_mean` = mean PE of the randomizations
- `pe_obs_p_upper` = percent of randomizations where observed PE was greater than random values
- `pe_obs_p_lower` = percent of randomizations where observed PE was lower than random values

Classify significance

The next step of CANAPE is to classify the endemism type of each site. The `cpr_classify_endem()` function does this automatically given output of `cpr_rand_test()`.

```
# Classify endemism type
acacia_canape <- cpr_classify_endem(acacia_rand_res)
# Count the results
count(acacia_canape, endem_type)
#>        endem_type    n
#> 1           mixed  101
#> 2             neo   11
#> 3 not significant 2760
#> 4           paleo   43
#> 5           super  122
```

Endemism codes:

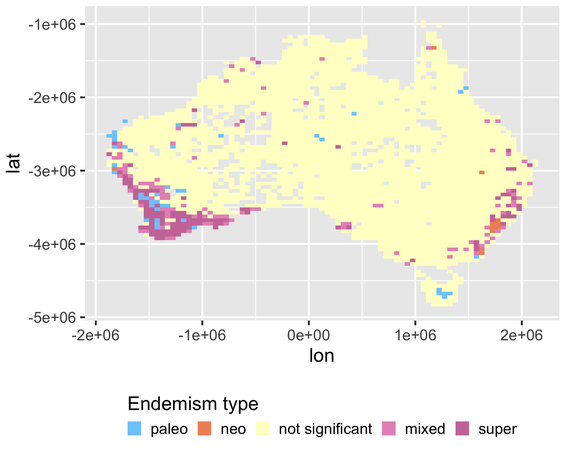
- `neo` = Neo-endemic
- `paleo` = Paleo-endemic
- `mixed` = Mix of neo and paleo
- `super` = Mixed, and highly significant ($p < 0.01$)

Visualize results

We can visualize the results of CANAPE using `ggplot2`. The `cpr_endem_cols` palette that is accessible regardless of color vision deficiency is provided.

```
# Fist do some data wrangling to make
# the results easier to plot (add lat/long columns)
acacia_canape <- as_tibble(acacia_canape, rownames = "site") %>%
  separate(site, c("lon", "lat"), sep = ":") %>%
  mutate(dplyr::across(c(lon, lat), parse_number))

# Plot the results
ggplot(acacia_canape, aes(x = lon, y = lat, fill = endem_type)) +
  geom_tile() +
  scale_fill_manual(values = cpr_endem_cols, name = "Endemism type") +
  guides(fill = guide_legend(title.position = "top")) +
  theme_gray(base_size = 24) +
  theme(legend.position = "bottom")
```



Next steps

- Implement `rand_structured` randomization algorithm of Biodiverse (should speed up randomizations ~10×)
- Submit to CRAN

Acknowledgements

Thanks to Shawn Laffan for providing help with the code.

References

[1] C.E. González-Orozco, M.C. Ebach, S. Laffan, A.H. Thornhill, N.J. Knerr, A.N. Schmidt-Lebuhn, C.C. Cargill, M. Clements, N.S. Nagalingum, B.D. Mishler, J.T. Miller, Quantifying phytogeographical regions of Australia using geospatial turnover in species composition, PLoS ONE. 9 (2014) e92558. https://doi.org/10.1371/journal.pone.0092558.

[2] S.W. Laffan, E. Lubarsky, D.F. Rosauer, Biodiverse, a tool for the spatial analysis of biological and related diversity, Ecography. 33 (2010) 643–647. https://doi.org/10.1111/j.1600-0587.2010.06237.x.

[3] B.H. Daru, P. Karunarathne, K. Schliep, phyloregion: R package for biogeographical regionalization and macroecology, Methods Ecol. Evol. 11 (2020) 1483–1491. https://doi.org/gkzsjp.

[4] S.W. Kembel, P.D. Cowan, M.R. Helmus, W.K. Cornwell, H. Morlon, D.D. Ackerly, S.P. Blomberg, C.O. Webb, Picante: R tools for integrating phylogenies and ecology, Bioinformatics. 26 (2010) 1463–1464. papers://ebd7d3df-d8f1-4be8-bcf7-4fb07ababf30/Paper/p1990.

Source code: https://github.com/joelnitta/botany_poster_2021