Environmental filtering and niche (mis)matching of riverine invertebrate communities

NRSA DisEQ-DisEQ Pipeline

Background

The disequilibrium analytical framework measures the strength of environmental filtering and habitat matching on local communities relative to the regional pool. Local communities and environmental conditions are compared to the regional species pool and environmental conditions under which species occur to determine (1) the degree of filtering (δ) and (2) habitat matching (λ). Both environmental filtering and habitat matching are calculated using a null model to compare how taxa match the environment relative to predictions from sampling the regional pool. Prior to calculating δ and λ, two statistics are quantified: (1) Δ, the median distance between community-inferred environmental conditions (mean environmental niche of all taxa in the community) and random samples of niche values from taxa in the community; (2) Λ, the vector between inferred and observed community environmental conditions. Null distributions of Δ and Λ are generated by sampling communities of equal richness from the regional pool, comprised of taxa that could occur at the local community. Standardized niche deviations (δ and λ) are determined by comparing observed Δ and Λ to expected values in the regional pool (using z-transformations). Filtering (δ) and habitat matching (λ) values are interpreted as:

Table 1: Basic interpretations of δ and λ for the DisEQ analysis.

Metric	Negative Values < 0	Positive Values > 0
δ	Filtering	Permissiveness
λ	Matching	habitat matching

Data Management

Load Data

```
## Read in data
# Invertebrate community data
invertebrate.data <- read_csv(
    "data/NRSA-invertebrates-rel_abund.25.csv",
    show_col_types = FALSE
    )

# Physical habitat data
habitat.data <- read_csv(
    "data/NRSA-physical_habitat.csv",
    show_col_types = FALSE
    )

# Landscape data
landscape.data <- read_csv(
    "data/NRSA-landscape.csv",
    show_col_types = FALSE
    )

# Water chemistry data
water.chemistry.data <- read_csv(
    "data/NRSA-water_chemistry.csv",
    show_col_types = FALSE
    )

## Set number of cores for parallel processing
n.cores <- detectCores() - 1

## Load the workspace containing just the DisEQ analysis results
load("data_analysis/1-DisEQ_pipeline/1-DisEQ_pipeline-DisEQ_data.RData")
```

```
## Merge all data into a primary dataframe
full.data <- landscape.data %>%
    full_join(habitat.data, by = "UID") %>%
    full_join(water.chemistry.data, by = "UID") %>%
    full_join(invertebrate.data, by = "UID") %>%
    na.omit()
```

Network Centrality

```
## Network Centrality

## Calculated network centrality for each site. Network centrality was calculated
## as the mean geodesic distance between a site to all other sites in the ecoregion.
## Network centrality is a measure of how connected a site is to other sites,
## or how far organisms would have to disperse between or among sites. Site
## coordinates were subset for each ecoregion, and the subset dataframes were
## combined into a list. The list of dataframes was then supplied to the network
## centrality function, and all measures of network centrality were appended into
## a single dataframe.

## Site longitude and latitude data for each ecoregion
CPL.coord <- full.data[full.data$ecoregion == "CPL", c(23, 22)]
NAP.coord <- full.data[full.data$ecoregion == "NAP", c(23, 22)]
NPL.coord <- full.data[full.data$ecoregion == "NPL", c(23, 22)]
SAP.coord <- full.data[full.data$ecoregion == "SAP", c(23, 22)]
SPL.coord <- full.data[full.data$ecoregion == "SPL", c(23, 22)]
TPL.coord <- full.data[full.data$ecoregion == "TPL", c(23, 22)]
UMW.coord <- full.data[full.data$ecoregion == "UMW", c(23, 22)]
WMT.coord <- full.data[full.data$ecoregion == "WMT", c(23, 22)]
XER.coord <- full.data[full.data$ecoregion == "XER", c(23, 22)]

## Convert mean annual flow from CFS (cubic feet/second) to m^3/s
full.data$mean.annual.flow <- full.data$mean.annual.flow * 0.0283168

## List of long/lat dataframes
full.coord.list <- list(
    CPL.coord, NAP.coord, NPL.coord,
    SAP.coord, SPL.coord, TPL.coord,
    UMW.coord, WMT.coord, XER.coord
    )

## Define function (network centrality, or mean pairwise distance between all sites)
network.centrality <- function(j) {

    ## Calculate the mean geodesic distance (in meters) for each site
    site.centrality.bin.1 <- rowMeans(distm(j, fun = distGeo))

    ## Convert geodesic from m to km
    site.centrality <- site.centrality.bin.1/1000

    ## Set final data
    site.mean.geodesic <- tibble::tibble(site.centrality)
}

## Run the analysis for each of the 9 ecoregions
final.network.centrality <- lapply(full.coord.list, FUN = network.centrality) %>%
    bind_rows()

## Bind distance to centroid for each site to the final dataframe
final.data <- cbind(full.data, final.network.centrality)
```

```r
## Subset ecoregions for later analyses
CPL.data <- final.data %>%
    filter(ecoregion == "CPL")
NAP.data <- final.data %>%
    filter(ecoregion == "NAP")
NPL.data <- final.data %>%
    filter(ecoregion == "NPL")
SAP.data <- final.data %>%
    filter(ecoregion == "SAP")
SPL.data <- final.data %>%
    filter(ecoregion == "SPL")
TPL.data <- final.data %>%
    filter(ecoregion == "TPL")
UMW.data <- final.data %>%
    filter(ecoregion == "UMW")
WMT.data <- final.data %>%
    filter(ecoregion == "WMT")
XER.data <- final.data %>%
    filter(ecoregion == "XER")
```

Disequilibirum (DisEQ) Analyses

Filtering (δ) and habitat matching (λ) statistics for each local community were quantified as the mean of 100 random samples from the niche of each present taxon. For this study, the local community was the assemblage at a site and the regional pool was the taxa and environmental conditions in the respective ecoregion. The analysis was conducted for each of the nine regions in the NRSA. Environmental variables were centered and scaled on the mean to meet assumptions of the DisEQ analysis.

The environmental variables selected for the DisEQ analysis were:

- T_{max} annual
- T_{min} annual
- pH
- conductivity

For a full description of the DisEQ framework, please see:

> Blonder, B., et al. 2015. Linking environmental filtering and disequilibrium to biogeography with a community climate framework. *Ecology* 96: 972-985.

```
## Function to perform the DisEQ analysis
DisEQ.analysis <- function(j) {
    require(comclim)
    require(reshape2)
    require(tidyverse)

    ## Set seed
    set.seed(666)

    ## Set empty summary statistics dataframe
    DisEQ.summary.statistics <- data.frame(
        "filtering.scaled"  = numeric(),
        "filtering.P-value" = numeric(),
        "mismatch.scaled"   = numeric(),
        "mismatch.P-value"  = numeric(),
        "filtering"         = numeric(),
        "mismatch"          = numeric(),
        "Tmax.direction"    = numeric(),
        "Tmin.direction"    = numeric(),
        "pH.direction"      = numeric(),
        "cond.direction"    = numeric()
        )

    ## Data Management============================================================

    ## Format local environment data-------------------------------------
    local.environment.data.bin <- melt(
        j,
        id.vars = c("tmax.annual", "tmin.annual", "pH.lab", "cond"),
        measure.vars = colnames(j[, 67:142])
        )

    ## Add column names to local environment data
    names(local.environment.data.bin) <- c(
        "tmax.annual", "tmin.annual", "pH.lab", "cond", "taxon", "abundance"
        )
```

```r
## Remove rows with abundance = 0 so only occurrences are included in the data
local.environment.data <- filter(local.environment.data.bin, abundance > 0)

## Specify niches------------------------------------------------------------
# Values are centred and scaled for analysis
environmental.niches <- bind_cols(
    as.data.frame(scale(local.environment.data[, 1:4], center = TRUE, scale = TRUE))
    )

## Add taxon identifier to environmental niches
environmental.niches$taxon <- factor(local.environment.data$taxon)

## Specify local community---------------------------------------------------
local.community.bin <- j[, 67:142]

## Remove empty sites and taxa absent from the ecoregion
local.community <- local.community.bin[rowSums(local.community.bin) > 0,
                                                                    colSums(local.community

## Define the regional pool--------------------------------------------------
regional.pool <- as.character(local.environment.data$taxon)

## DisEQ Analysis============================================================
for (x in 1:length(j[, 1])) {

    ## Define observed environment for the local community
    observed.climate <- as.numeric(
        scale(j[, c(12, 14, 61, 52)], center = TRUE, scale = TRUE)[x, ]
        )

    ## Add variable names to the observed environment
    names(observed.climate) <- names(j[, c(12, 14, 61, 52)])

    ## Override any errors; start try()
    try ( {

    ## Input data   for communityclimte() function
    region.DisEQ <- inputcommunitydata(
        localcommunity  = colnames(local.community[x, ]),
        regionalpool    = regional.pool,
        climateniches   = environmental.niches,
        observedclimate = observed.climate
    )

    ## Run the DisEQ analysis | communityclimate() function
    region.DisEQ.results <- communityclimate(
        region.DisEQ,
        climateaxes = c("tmax.annual", "tmin.annual", "pH.lab", "cond"),
        numreplicates = 100
    )

    ## End try()
    } )
```

```
        ## Extract and append DisEQ analysis results
        DisEQ.summary.statistics[x, 1:2]  <- region.DisEQ.results@deviations$deviation_volumeMagnitude
        DisEQ.summary.statistics[x, 3:4]  <- region.DisEQ.results@deviations$deviation_mismatchgMagnitu
        DisEQ.summary.statistics[x, 5]    <- region.DisEQ.results@obsStats$volumeMagnitude
        DisEQ.summary.statistics[x, 6]    <- region.DisEQ.results@obsStats$mismatchMagnitude
        DisEQ.summary.statistics[x, 7:10] <- region.DisEQ.results@obsStats$mismatchDirections
    }


    ## Set final data-------------------------------------------------------
    DisEQ.summary.statistics$UID <- j[, 1]
    region.site.info <- j[, c(1, 25, 60, 62, 53, 41, 46, 43:44, 18, 16:17, 11, 2, 23, 22, 5, 7, 9, 143)]
    DisEQ.data <- merge(
        DisEQ.summary.statistics, region.site.info,
        by = "UID",
        all = FALSE
        )
}
```

```
## List of full dataframes
full.data.list <- list(
    CPL.data, NAP.data, NPL.data,
    SAP.data, SPL.data, TPL.data,
    UMW.data, WMT.data, XER.data
    )

## Start cluster
cluster <- makeCluster(n.cores)

## Run the DisEQ analysis
DisEQ.analysis.results <- parLapply(cluster, full.data.list, fun = DisEQ.analysis) %>%
    bind_rows()

## Stop cluster
stopCluster(cluster)
```

R Session Information

Table 2: Packages required for data management and analyses.

Package	Loaded Version	Date
broom	0.7.12	2022-01-28
comclim	0.9.5	2018-05-30
dplyr	1.0.8	2022-02-08
forcats	0.5.1	2021-01-27
geosphere	1.5-14	2021-10-13
ggplot2	3.3.5	2021-06-25
kableExtra	1.3.4	2021-02-20
knitr	1.38	2022-03-25
lattice	0.20-45	2021-09-22
permute	0.9-7	2022-01-27
purrr	0.3.4	2020-04-17
readr	2.1.2	2022-01-30
reshape2	1.4.4	2020-04-09
snow	0.4-4	2021-10-27
stringr	1.4.0	2019-02-10
tibble	3.1.6	2021-11-07
tidyr	1.2.0	2022-02-01
tidyverse	1.3.1	2021-04-15
vegan	2.5-7	2020-11-28