Diversity and dataframe formating

Trevor Eakes

May 8, 2016

Contents

##Dataframes and stuff
Two dataframes will be crucial to all future analysis.
One contains phytoplankton abundances and presence in sample stations called Kuroshio_Phytoplankton. This will become
The other is of all physical concurrent measurements also at those stations called KuroAllData. This will become
For Kuroshio_Phytoplankton we will need to convert the species counts from/L to per centiliter, which is the actual scale of measurements of abundance
For Adiv.abiotic we start with the chemical and physical measuremnts at the station and will add new vectors such as diversity measurements, Chlorophyll, Catagorical groups, and total cell count.
It could be useful to do this for both the test dataset, which only includes diatom and dinoflagellates, and the all phytoplankton dataset called original. Original also has abundances of large phytoplankton taxa identified to the class level.
###Required Packages:
utils sparcl ggplot2

Helper Functions

I made some functions to make this whole process a little easier that are in the file Niffty TrevFunctions.R. We will run this script.

Here are the functions from that file that we are using:

1. Create.Diversity.I() will calculate Richness, Simpson diversity, Simpson evenness, Hill number 1 (Shannon-Wiener diversity) and Shannon evenness for every row of a data frame with species abundances/sample. It's great.

2. PotentialST calculates potential temperature and it follows the oceanographic standard formula for doing this which is long and complicated.

3. AbioticCluster is a convenient clustering protocol which will use Euclidean distance to cluster a given dataframe. Whatever is put into the formula will be clustered so pick wisely. It spits out three dendrograms with three seperate clustering methods, the complete method, the single method and the Aver method. THis helps the user pick which method to use when presenting their data. It will also cut the tree into the specified number of clades for each dendrogram and create a vector with the clade identity of each row input into the function. Pretty nifty.

*Note the distance calculation method is easily changed for those looking to use a Curtis-Bray, Jaccard or some other dissimilarity matrix.

4. SplitData is a convenience function for easily splitting a dataframe into new dataframes based upon a single factor vector within the dataframe.

You will need to change this code to reflect where the file is sourced.

```
source('NifftyTrevFunctions.R', encoding = 'UTF-8')
```

Adiv.abiotic: the phsical measurements for every sample with biological indices. You will need the file test.csv & origin.csv saved in the directory to run this script. Download the files and add it to the directory. ###You will need to change the read.csv functions to reflect your file directory

```
require(utils)
All <- read.csv('KuroAlldata.csv')
bigtest <-  read.csv('Allbigstuff.csv')
require(sparcl)
```

```
## Loading required package: sparcl
```

```
require(ggplot2)
```

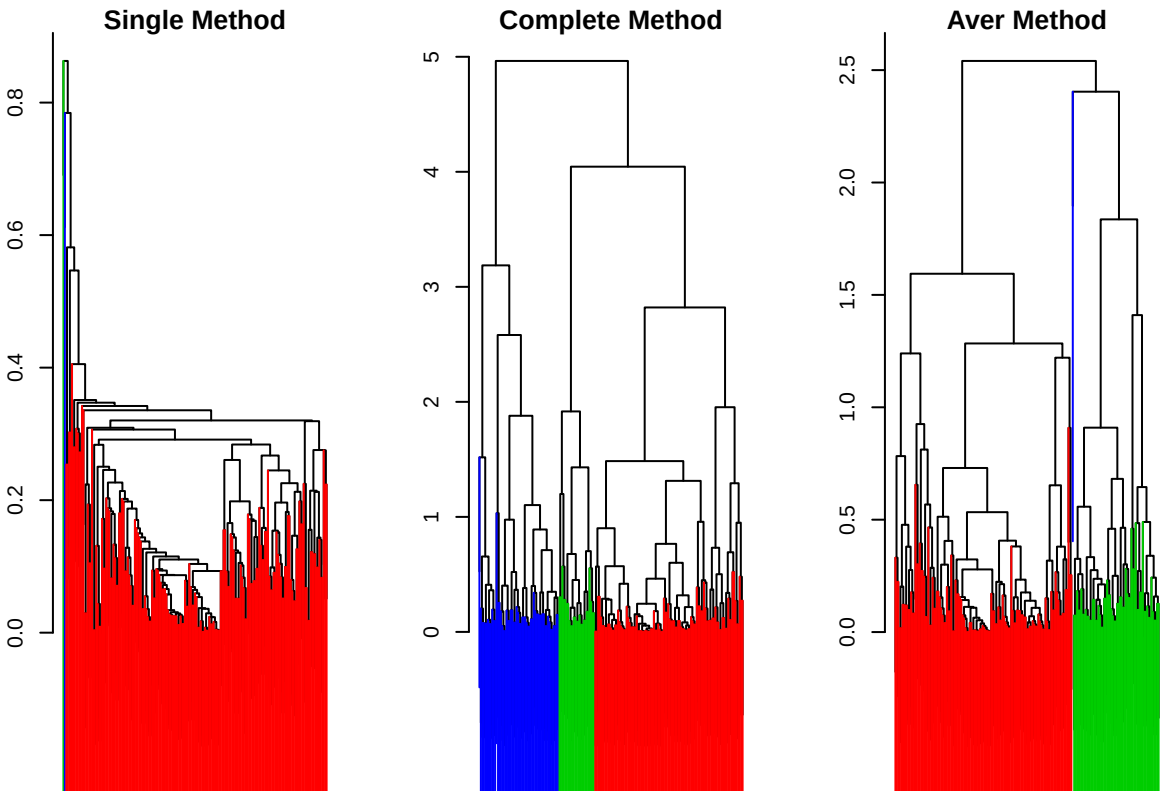
```
## Loading required package: ggplot2
```

```
All <-All[complete.cases(All$Diatoms..cells.l.),]
#selected diatoms..cells.l. because any rows reading NA
All <- cbind("rows"=c(1:190), All)
#would not match Kuroshio phytoplankton. Also changes the name to Al.

#Calculating Potential Density and Temperature
  d <- All$depth..m. #depth from the data set
  t <- All$T.C.        #Temperature from the data set
  s <- All$S #Salinity from the data set
  div.abiotic <- PotentialST(d,t,s )
  #function for calculating the potential density and temperature

  #Expanding the data set; more things to analyze
  #|This data set has the generalized large classes of micro plankton
#found in the samples. It has a total cells per liter for all cells
  bigtest<- bigtest[,4:ncol(bigtest)]
  TotA <-apply(bigtest, 1, function(x) sum(x))
  #This is all cells for all microplankton
  Adiv.abiotic <- cbind(div.abiotic, TotA)
  #creates a new master dataset with additional information
  Adiv.abiotic <- div.abiotic[-c(186:190),]
  #removes the last 5 rows because they have missing data

#Clustering data based on abiotic factors, automated
Abiotic <- as.data.frame(cbind(Adiv.abiotic$S, Adiv.abiotic$Theta ))
#Creates the desired dataframe with all the abiotic factors to be clustered
  AbioticCluster(Abiotic, 3)
```


Single Method
Complete Method
Aver Method
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
5
0.0
0.5
1.0
1.5
2.0
2.5

```
#Creates three different method dendrograms and three seperate clade lists based on
#each of those methods. Clustering has been seperated in 3 clades
#in this formula too. Formula may be found in Nifty Functions.

Adiv.abiotic <- Adiv.abiotic[-c(97:99),]
#Taking out rows 97:99 because the centiliter cell count is not in whole numbers
Adiv.abiotic<- cbind("cluster"=clades[-c(97:99),2], Adiv.abiotic)
#Joining the complete clustering factor to the main dataframe.
```

---

centitests

Adding diversity from both diatom and dinoflagellates or from all phytoplankton.

Also converting cell volumes to centi litters. A,B,C transects are .08, D and E are .1 ###You will need to change the read.csv function to reflect the directory you have saved the file in

---

```
cent <- read.csv('Kuroshio_Phytoplankton.csv') #diatom and dinoflagellates
#centitest for just diatoms and dinoflagellates
Cent <- cent[-c(97:99, 186:190),] #trim it as needed
cent <- Cent[,c(3:length(Cent))]
cent <- cbind(Adiv.abiotic$lon, cent)
#add longitude to break it up by transect
cent <- SplitData(cent, 1, "centi")
```

```
#Using my function to split dataframes based off of catagorical variables.

centitest <- rbind(cent$centi143.5[2:ncol(cent$centi143.5)]/80,
cent$centi144[2:ncol(cent$centi144)]/80, cent$centi144.5[2:ncol(cent$centi144.5)]/80,
cent$centi145[2:ncol(cent$centi145)]/100, cent$centi145.5[2:ncol(cent$centi145.5)]/100)

centitest <- cbind("rows"=c(1:nrow(centitest)), centitest)
#adding row numbers to keep track of things
centitestN <- centitest[which(rowSums(centitest[,2:ncol(centitest)])!=0),]
#getting rid of samples with no species for purposes of Neutral Test
centitest <- cbind(Cent[,1:2], centitest) #add row information
write.csv(centitest, "centitest.csv") #saving the file
centiDiverse <- Create.Diversity.I(centitest, 5, ncol(centitest))
 centiDiverseN <- Create.Diversity.I(centitestN, 1, ncol(centitestN))
 #creates new dataframe of diversity indices

centiDiverseN <- cbind("rows"=centitestN$rows, centiDiverseN)
#adding rows back into the picture
Adiv.abioticN <- merge(centiDiverseN, Adiv.abiotic, by="rows")
Adiv.abiotic <- cbind(centiDiverse, Adiv.abiotic)

#Chosing to add centiDiverse to Adiv.abiotic instead of the diversity dataframe
#for all phytoplankton because it is my main interest.
```

Origin

Creating a dataframe with diversity of the original phytoplankton abundance including large taxa

```
origin <- read.csv('origin.csv')
# diatom and dinoflagellates plus flow cyctometer identified taxa
#centitest for all organisms
origi <- origin[-c(97:99, 186:190), ]
#get rid of unneeded rows, but leave arround for latter extraction of station info
cent <- origi[,c(5:ncol(origi))]
#get rid of station info
cent <- cbind(Adiv.abiotic$lon, cent)
#add longitude
Cent <- SplitData(cent, 1, "centi")
#split by longitiude
#make centiliter conversions grouped by longitude
centiorigin <- rbind(Cent$centi143.5[2:ncol(Cent$centi143.5)]/80,
Cent$centi144[2:ncol(Cent$centi144)]/80,Cent$centi144.5[2:ncol(Cent$centi144.5)]/80,
Cent$centi145[2:ncol(Cent$centi145)]/100,
Cent$centi145.5[2:ncol(Cent$centi145.5)]/100)

centiorigin <- cbind(origi[, c(2:4)], centiorigin)
write.csv(centiorigin, "centiorigin.csv")
centiODiverse <- Create.Diversity.I(centiorigin, 4, ncol(centiorigin))
```