EEB C177 Final Project: Investigating the Fossil Occurrences of *Cetacea*

Justin Shieh

Introduction

Cetaceans, which include whales, dolphins and porpoises, have long been considered marine ecosystem engineers due to their assistance with nutrient cycling, carbon sequestration, and top-down regulation of food chain (Roman et al. 2014). As some of the biggest species to have ever existed and mammals that can go to the deepest depths in the oceans, Cetaceans have shown an incredible amount of diversity and broad geographic range. In particular, with 400 extinct and 84 extant species, whales have a fossil record that represents their successful invasion of the oceans (Slater et al. 2010). Other species like killer whales (*Orcinus orca*) are discovered in oceans across the world (Baird 2000). Given Cetacea's immense species richness and widespread inhabtiation, I thought it would be worthwhile to investigate the fossil occurrences of Cetacea using a stratigraphic record depicting the minimum and maximum ages of fossils for each species; a diversity-through-time graph; a world map with all the locations at which the fossils were discovered; and PyRate graphs showing speciation, extinction, net diversification and longevity through time. An additional point I wanted to investigate was how Cetacean body size has been correlated with speciation and extinction rates. Though I was able to obtain, parse through and match the body size data with the fossil occurrences data, I encountered extensive difficulties while trying to run the matched data through PyRate. Consequently, while I am including all the methods I used to retrieve the matched data set, there will *not* be PyRate graphs plotted with the additional body size variable. Despite this setback, the four graphical representations shown in this paper indicate important trends on net diversification and patterns in geographic range that could enhance our understanding of this critical order.

Methods and Results (with Graphical Representation)

I obtained my data on Cetacea from the online Paleobiology Database (PBDB) by downloading a .csv file containing, for each line of data, the genus name, species name, rank (as applicable), minimum age, maximum age, longitude of location of discovery and latitude of location of discovery. I then used several shell commands to filter through the data. First, I grepped for just the species. I then sorted it based on the species column and uniqued it to get a list of unique species with all their associated data.

First Function: make_species_and_period_dictionary

This function is used to create a dictionary with species as keys and the species' associated time periods for speciation as values. I created this function in the early stages of my analysis as a starting point for potentially analyzing the total amount of speciation events for each time period. I later on discovered that PyRate was a better tool for this type of analysis, so I had less of a need for this function, but it was nonetheless a useful tool for easily accessing the time period for each species.

```
def make_species_and_period_dictionary(filename):
    #Make empty dictionary
    dictionary = {}
    #Reading user-specified file and saving it in variable "all_records"
    fobj = open(filename,"r", encoding = "ISO-8859-15")
    all_records = fobj.readlines()[1:]
    #Going through "all_records" to extract species and time period
    for line in all_records:
        record_elements = line.split(",") #Splitting the line by commas
        species = record_elements[5].strip('"') #Getting species and deleting quotation
        period = record_elements[8].strip('"') #Getting time period and deleting quotat
        dictionary[species] = period#Creating key (species) and value (period) for dict
    return dictionary
```

Sample output: {'Acrophyseter deinodon': 'Messinian', 'Acrophyseter robustus': 'Serravallian', 'Aegyptocetus tarfa': 'Bartonian', 'Aetiocetus cotylalveus': 'Rupelian', 'Aetiocetus polydentatus': 'Chattian', 'Aetiocetus tomitai': 'Chattian', 'Aetiocetus weltoni': 'Chattian', ...}

Second Function: make_list_of_oldest_and_youngest_for_each_cetacea_species

This function creates an additional file that contains the species, their minimum ages, and their maximum ages. The purpose of this function is to create a file that is friendly for the language R to evaluate to produce the stratigraphic record.

```
def make_list_of_oldest_and_youngest_for_each_cetacea_species(filename, new_filename):
    #Creating output file to write to
    output_file = open(new_filename, "w")
    #Reading user-specified file and saving it in variable "all_records"
    data = open(filename, "r")
    all_records = data.readlines()[1:]
    #Going through "all_records" to extract species and time period
    for line in all_records:
        record_elements = line.split(",") #Splitting the line by commas
        species = record_elements[5].strip('"') #Getting species and deleting quotation
        minage = record_elements[10].strip('"') #Getting minimum age and deleting quota
```

```
        maxage = record_elements[11].strip('"') #Getting maximum age and deleting quota
        output_file.write(species + ", " + str(minage) + ", " + str(maxage) + "\n") #Wri
    output_file.close()
    return output_file
```

Sample output:

File containing: Argyrocetus joaquinensis, 23.03, 20.44 Pinocetus polonicus, 15.97, 13.82 Basiloterus hussaini, 41.3, 38 Sachalinocetus cholmicus, 23.03, 11.608 Praekogia cedrosensis, 7.246, 5.333 Aulophyseter rionegrensis, 7.246, 5.333 ...

Third Function: find_overlapping_species

This function generates an additional file containing the overlapping species between the species found in two user-specified files. Since I obtained the Cetacean body size data from an online scientific article that utilized a different data set from PBDB, I had to find the overlapping species between those I had body size data and those I had fossil data on. This function streamlined that process.

```
def find_overlapping_species(file_1, file_2):
    import csv

    #Reading the first user-specified file and converting to list
    data1 = open(file_1,"r")
    reader1 = csv.reader(data1)
    list1 = list(reader1)

    #Reading the second user-specified file and converting to list
    data2 = open(file_2,"r")
    reader2 = csv.reader(data2)
    list2 = list(reader2)

    #Creating empty list for overlapping species
    overlapping_species = []

    #Creating output file to write to
    overlapping_species_file = open("overlapping_species.csv","w")

    #Going through user specified files and appending to overlapping_species if the it
    for item in list1:
        if item in list2:
            overlapping_species.append(item)

    #Writing to the output file (and converting overlapping_species from list to csv i
    with overlapping_species_file as output:
```

```
        writer = csv.writer(output, lineterminator='\n')
        writer.writerows(overlapping_species)

    data1.close()
    data2.close()
    overlapping_species.close()
```

Sample output:

File containing: Kogia breviceps Kogia sima Physeter macrocephalus Platanista gangetica Delphinapterus leucas Monodon monoceros Phocoenoides dalli . . .

Below is a shell script I created to do the same thing. In the process of figuring out how to find overlapping species, I experimented with both Python and the Shell to see which would be better. Ultimately, I was able to generate a function using either.

```
#!/bin/bash
for line in `cut -d "," -f 1 $1`
do
    grep -w $line --no-messages in $2
done
```

Graphical Representations:

Stratigraphic Record of Fossil Occurrences

The R code used to obtain the graph:

```
library(ggplot2)
library(forcats)
setwd("/home/eeb177-student/Desktop/eeb-177/eeb177-final-project")
cetacea <- read.csv("/home/eeb177-student/Desktop/eeb-177/eeb177-final-project/data_for_
names(cetacea) <- c("genus","species", "minage", "maxage")
head(cetacea)
```

```
##            genus                   species  minage  maxage
## 1 Haborophocoena Haborophocoena toyoshimai  4.4665  4.4665
## 2   Eoplatanista      Eoplatanista italica 21.7350 21.7350
## 3   Ziphirostrum     Ziphirostrum recurvus  7.0980  7.0980
## 4   Huaridelphis    Huaridelphis raimondii 19.5000 19.5000
## 5    Ninoziphius  Ninoziphius platyrostris  4.4665 13.7890
## 6   Lamprolithax    Lamprolithax annectens 14.8950 14.8950
```

```
cetacea_occ <- ggplot(cetacea, aes( x = fct_reorder(species, minage, .desc = T), maxage
cetacea_occ + geom_linerange(aes(ymin = minage, ymax = maxage + 0.5)) + theme(legend.pos
```

```
## Warning: Removed 2 rows containing missing values (geom_linerange).
```

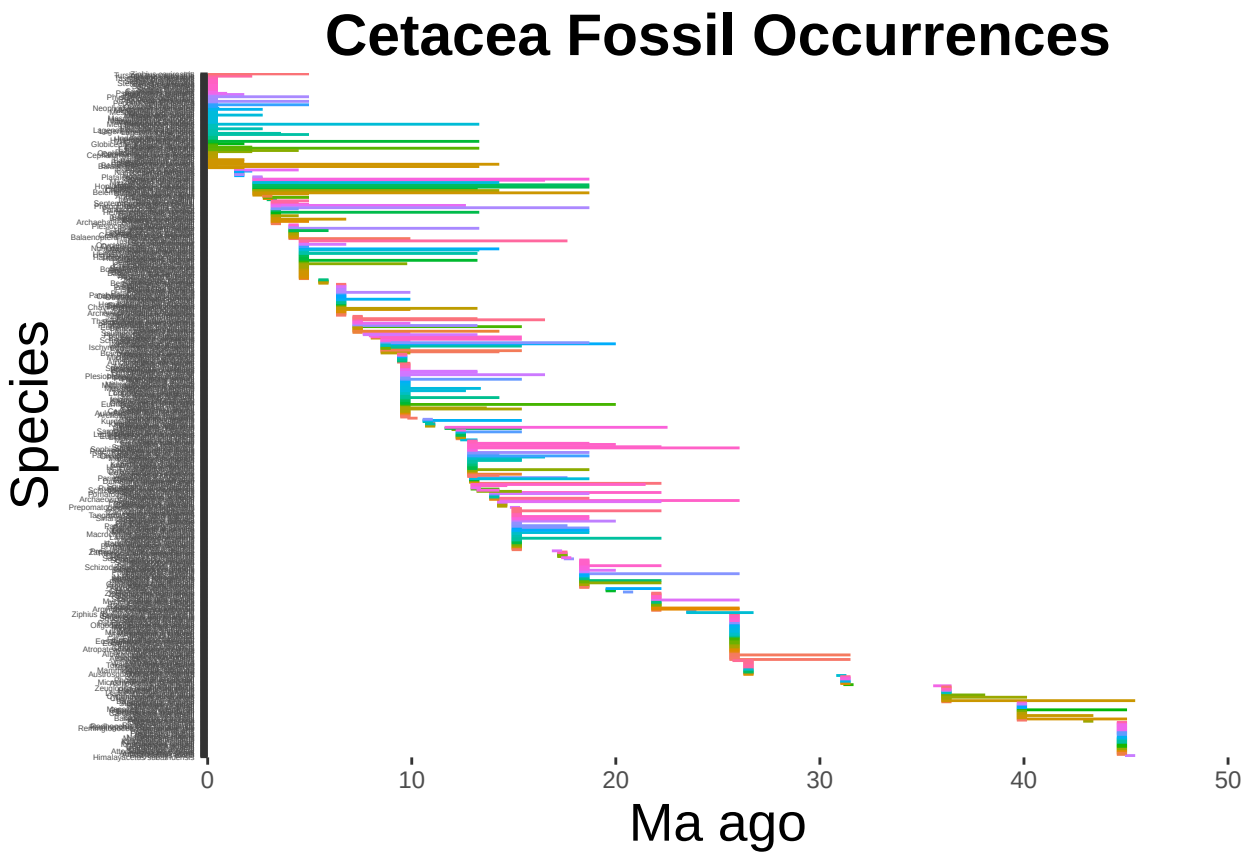

Fig 1. The stratigraphic record of fossil occurrences of Cetacean species

Diversity Through Time

The R code used to obtain the graph:

```
library(tidyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
cetacea <- read.csv("/home/eeb177-student/Desktop/eeb-177/eeb177-final-project/data_for_
names(cetacea) <- c("genus","species", "minage", "maxage")
head(cetacea)
```

```
##            genus                    species  minage  maxage
## 1 Haborophocoena Haborophocoena toyoshimai  4.4665  4.4665
## 2   Eoplatanista      Eoplatanista italica 21.7350 21.7350
## 3   Ziphirostrum     Ziphirostrum recurvus  7.0980  7.0980
## 4   Huaridelphis    Huaridelphis raimondii 19.5000 19.5000
## 5    Ninoziphius  Ninoziphius platyrostris  4.4665 13.7890
## 6   Lamprolithax    Lamprolithax annectens 14.8950 14.8950
```

```
diversity <- cetacea %>% gather(key = type, value = age, minage, maxage) %>% mutate(cour
ggplot(diversity, aes(x = age, y = diversity)) + geom_step()
```

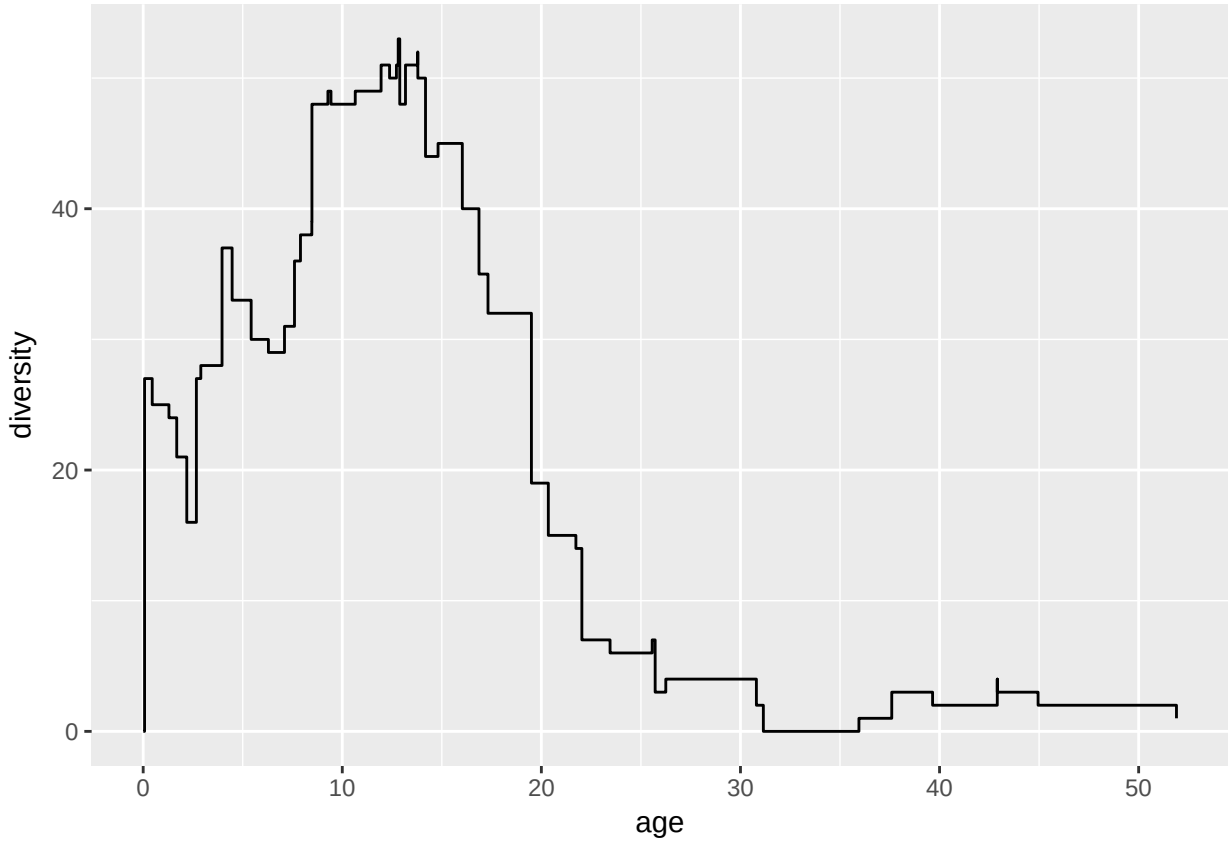

Fig 2. Cetacian diversity (in number of species) through time (in million years ago).

Geographic Distribution of Fossil Occurrences

The R code used to obtain the graph:

```
library(ggmap)
```

```
## Google Maps API Terms of Service: http://developers.google.com/maps/terms.
```

```
## Please cite ggmap if you use it: see citation("ggmap") for details.
```

```
library(maps)
library(mapdata)
```

```
world <- map_data("world")
cetacea_distribution <- read.csv("/home/eeb177-student/Desktop/eeb-177/eeb177-final-proj
names(cetacea_distribution) <- c("species", "longitude", "latitude")
cetacea_distribution$longitude<-as.numeric(cetacea_distribution$longitude)
```

```
## Warning: NAs introduced by coercion
```

```
ggplot() +  geom_polygon(data = world, aes(x=long, y = lat, group = group)) + geom_poin
```

```
## Warning: Removed 20 rows containing missing values (geom_point).
```

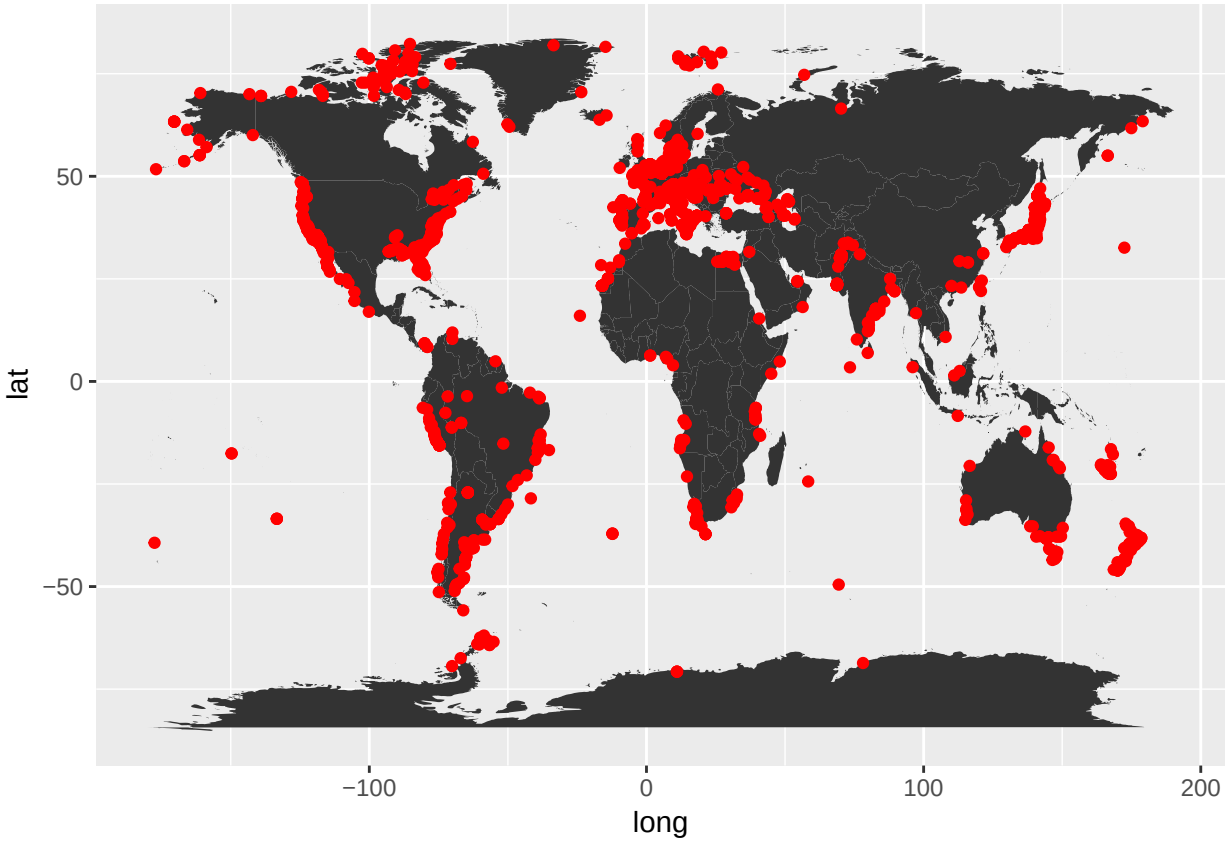

Fig 3. The distribution of fossil occurrences around the world.

PyRate Analyses

PyRate was used to obtain the analyses, with the help of the lab tutorial. The following graphs show the rates of speciation, extinction, net diversification and longevity of Cetacia through time.

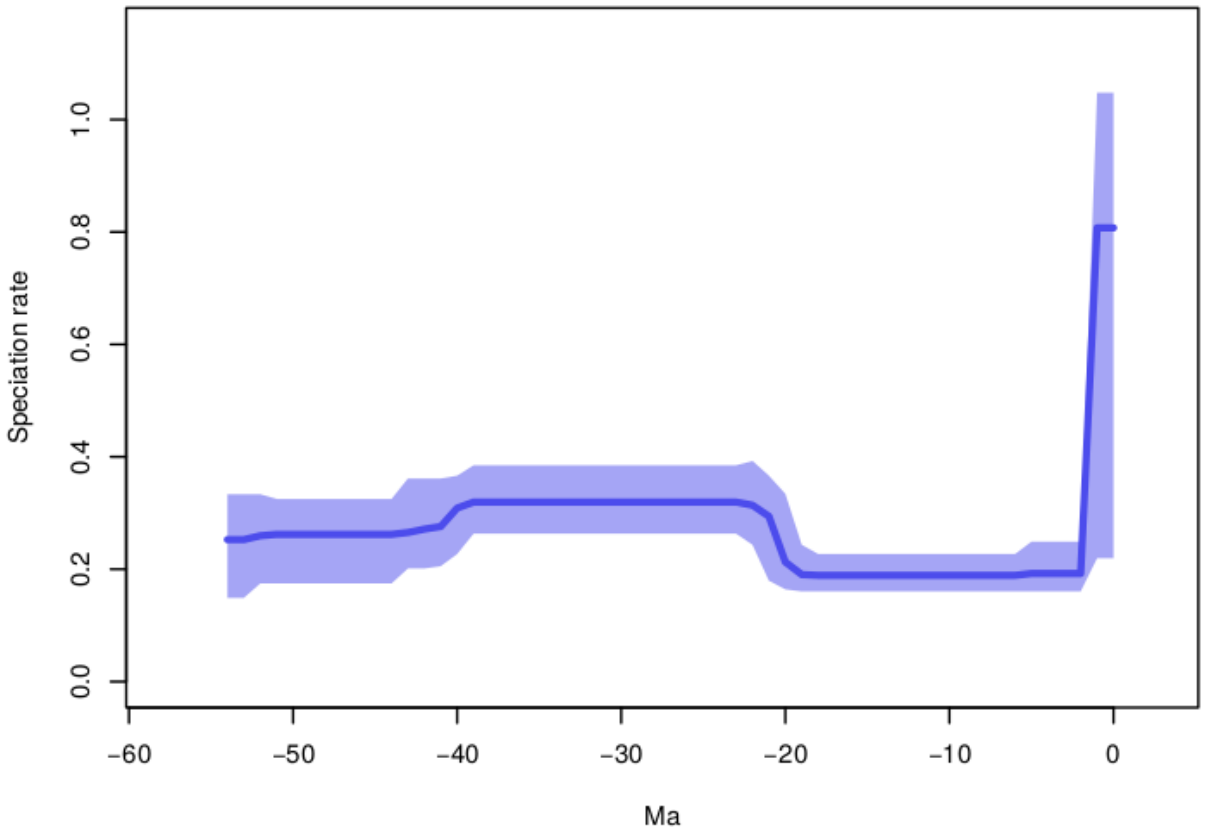

Figure 1: Cetacean Speciation Rate

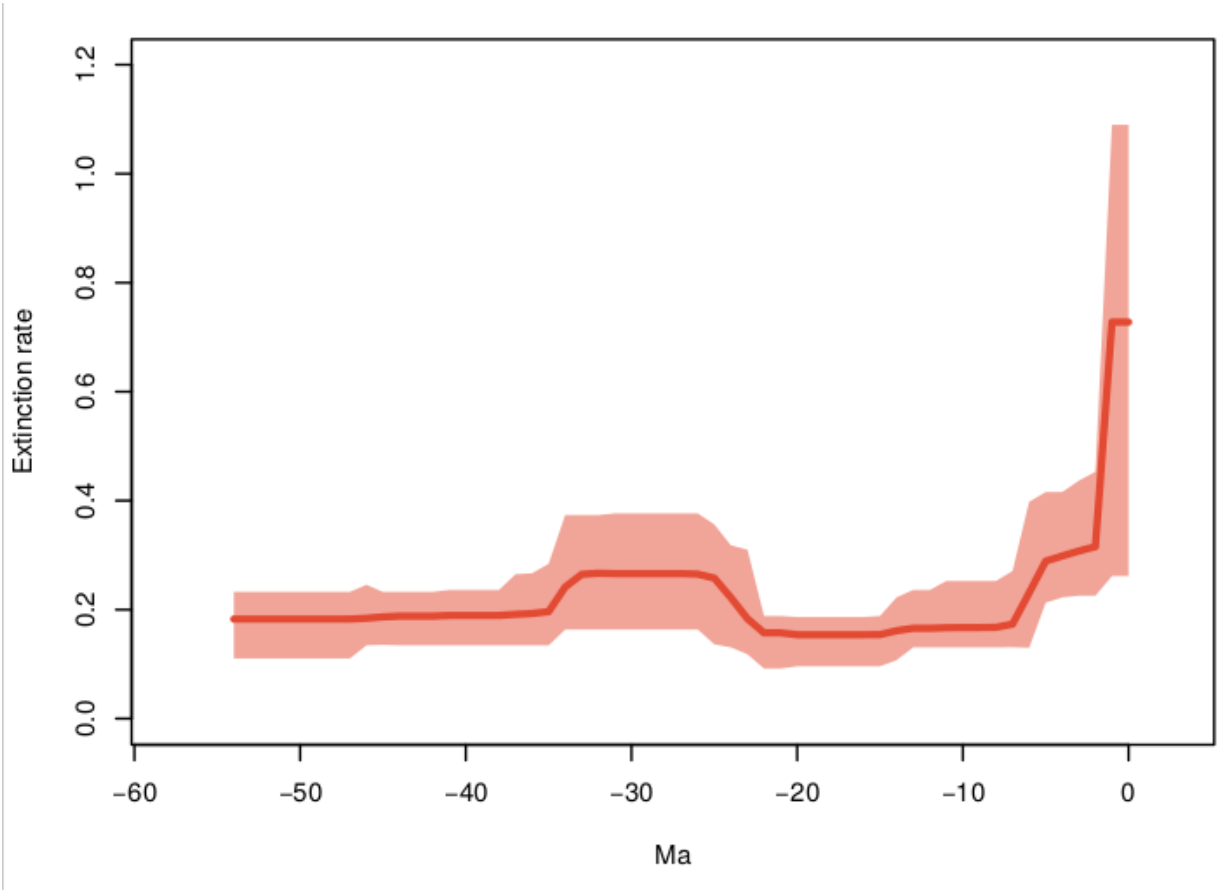

Figure 2: Cetacean Extinction Rate

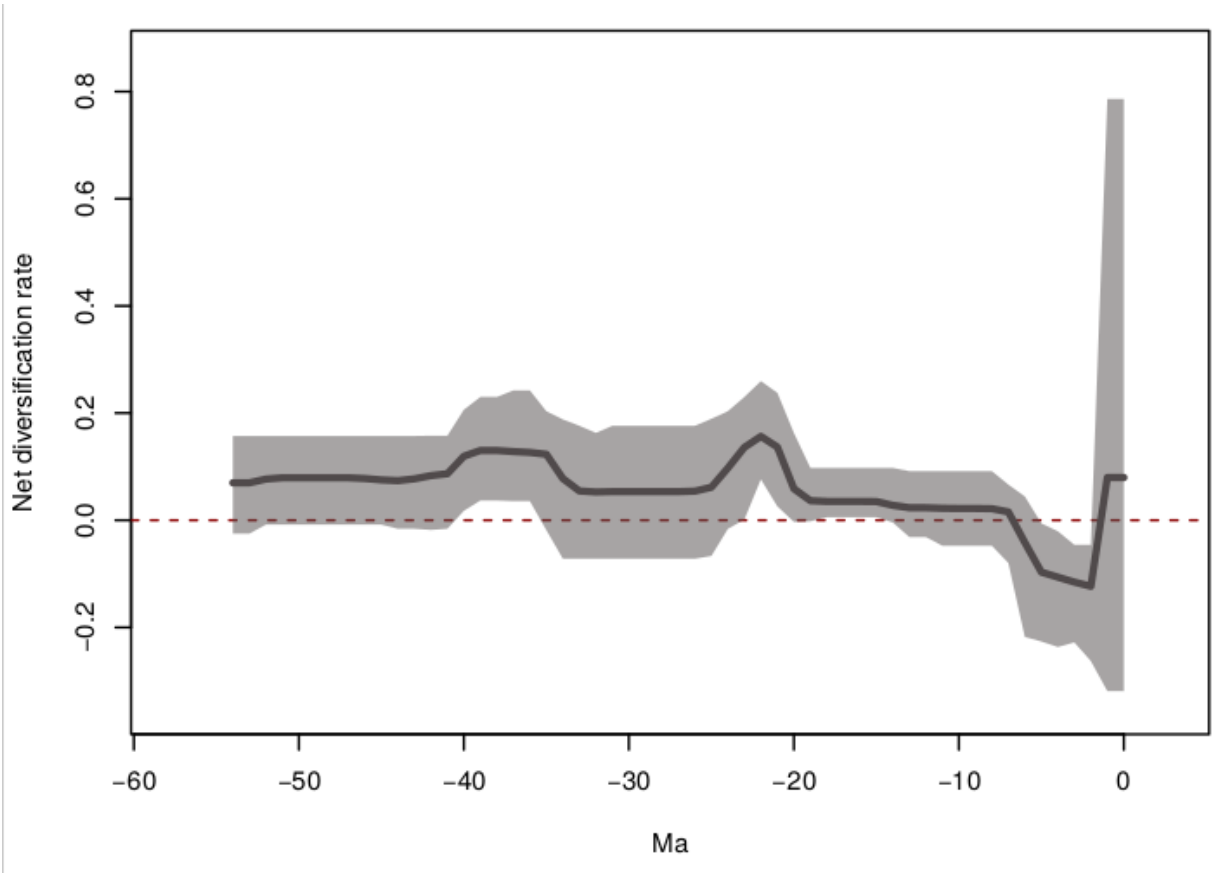

Figure 3: Cetacean Net Diversification Rate

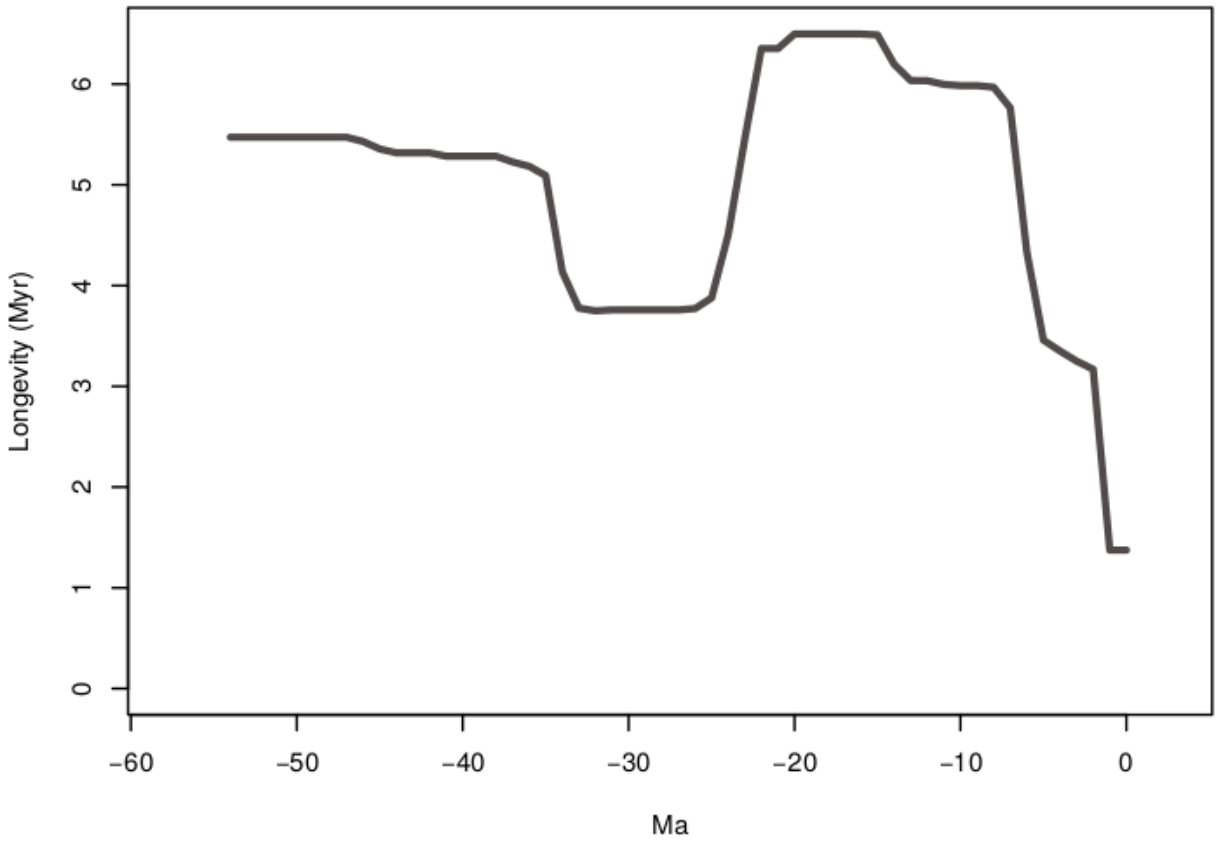

Figure 4: Cetacean Longevity Rate

Discussion and Conclusion

The stratigraphic record of fossil occurrences of Cetacean species (Fig.1) is an informative look into when each species came into existence and went extinct (if applicable). The way the graph is arranged allows for easy visualization of which species are extant as well as the ranking of the species by their minimum ages.

The plot of diversity through time (Fig.2) indicates that Cetacian diveristy reached its peak around 14 million years ago during the Miocene era. Recent discoveries in the San Joaquin Hills of Orange County, California, show an immense collection of marine mammals from the Miocene (Rivin 2010). These discoveries include toothed mysticetes - which were thought to have gone extinct by the end of the Oligocene - from three different taxas, indicating that they had a greater temporal and geographic spread than previously known and providing evidence for the high Cetacean diversity in this era (Rivin 2010). Another potential reason for the observed spike in Cetacean diversity during the Miocene could be due to the dense accumulation and preservation of Cetacean fossils as a result of repeated mass strandings (Pyenson et al. 2014). Harmful algal blooms, common in upwelling zones, are thought to be the cause of these strandings (Pyenson et al. 2014).

The geographic distribution of fossil occurrences (Fig.3) shows higher concentrations along the coastal regions of the continents throughout the world, with few occurrences scattered across the oceans. There could be two reasons for this pattern. Constrained by resource availability, cetaceans could have been more concentrated along the continental coasts and shelves due to their higher productivities (Kaschner et al. 2011). Another reason is simply the logistical challenges of discovering fossils in ocean basins.

Of all the PyRate analyses, the most informative one is the graph showing the net diversification rate of Cetacea through time (PyRate Analyses: Fig.3). Although the speciation graph (PyRate Analyses: Fig.1) shows high speciation rates in the past one million years or so, the net diversification rate takes into account the high extinction rates and more or less plateaus close to zero within the aforementioned time frame. The net diversification graph provides evidence for adequate survivorship of extant Cetacean species based on broad geological time scales and fossil occurrence data. However, whaling in the 20th century has been attributed to decimation of certain cetacean species and overall whale population decline, so future analysis of fossil occurrences on narrower geological time scales could suggest otherwise. Additionally, the fossil occurrence data itself is not always an accurate reflection of current cetacean diversity given the biases in fossil discoveries, errors in dating, etc.

In the process of evaluating Cetacean fossil occurrence data, I have improved my understanding of the historic species richness and geographic distribution of Cetacea and my skillfullness in using analytical tools like shell commands (head/tail, grep, sort, uniq, awk, sed), bash scripts, Python (loops and conditionals), R (ggplot) and PyRate. Although I encountered numerous issues along the way - particularly in trying to match data and running PyRate with the body size variable - I ultimately learned how to effectively troubleshoot both on my own using online resources like Stack Overflow and with the amazing help of my classmates, the TA and the professor.

Github link: https://github.com/jtnshieh

Bibliography

Baird, Robin W. "The killer whale." Cetacean Societies: Field Studies of Dolphins and Whales (2000): 127-153.

Kaschner, Kristin, et al. "Current and future patterns of global marine mammal biodiversity." PLoS One 6.5 (2011): e19653.

Pyenson, Nicholas D., et al. "Repeated mass strandings of Miocene marine mammals from Atacama Region of Chile point to sudden death at sea." Proceedings of the Royal Society of London B: Biological Sciences 281.1781 (2014): 20133316.

Rivin, Meredith Ann. "Early Miocene cetacean diversity in the Vaqueros Formation, Laguna Canyon, Orange County, California." CALIFORNIA STATE UNIVERSITY, FULLERTON (2010).

Roman, Joe, et al. "Whales as marine ecosystem engineers." Frontiers in Ecology and the Environment 12.7 (2014): 377-385.

Slater, Graham J., et al. "Diversity versus disparity and the radiation of modern cetaceans." Proceedings of the Royal Society of London B: Biological Sciences 277.1697 (2010): 3097-3104.