Final Project

Hasna Akther

2025-05-20

Title: Bioluminescent symbioses of fish and cephalopods with bacterial symbionts.

Brief introduction:

Bioluminescence is the emission of light by living organisms via biochemical including luciferin and luciferase. This feature has individually evolved several times across various taxa, which emphasizing its adaptive significance in diverse ecosystems. New details surveys have identified approximately 2,781 established bioluminescent marine species, with an extra 6,392 species from bioluminescent or prospectively luminescent genera whose luminescent status remains contradicted. This indicates that bioluminescent species constitute around 1.32% of the ~ 210,000 valid marine animal species cataloged in the World Register of Marine Species (Martini et al., 2024).

Aim and Hypothesis:

This research is a thorough review of bioluminescent symbioses between marine organisms, specifically in relation to cephalopods and fish with bacterial symbionts. In addition, the bioluminescent symbiosis plays a significant role in biomedical science. While symbiont relationships interacting between vertebrates (e.g. sharks/fishes) and invertebrates (e.g. squids) which remains understudied in response to phylogenetic analysis of cephalopods in order to understanding the food web dynamics and future climate. In order to address this problem statement, methods will be relevance systematic literature searches, and comparative analysis of luminescence mechanisms and symbiotic interactions.

Brief description of dataset (with a link to completed GitHub repository for more details on my dataset card).

https://github.com/haaklina/Final-Project/blob/main/Dataset%20Card.md

```
options(repos = c(CRAN = "https://cran.rstudio.com"))
```

{r} include=FALSE} options(repos = c(CRAN = "https://cran.rstudio.com")) install.packages("ape")
only if not already installed library(ape)

```
# Install and load ape package
install.packages("ape")  # Run only once
```

```
## Installing package into '/Users/patron/Library/R/arm64/4.4/library'
## (as 'lib' is unspecified)

##
## The downloaded binary packages are in
##  /var/folders/c6/kzf7kwv10q7f_wvyyfyj67sh0000gp/T//Rtmp6IXIje/downloaded_packages
```

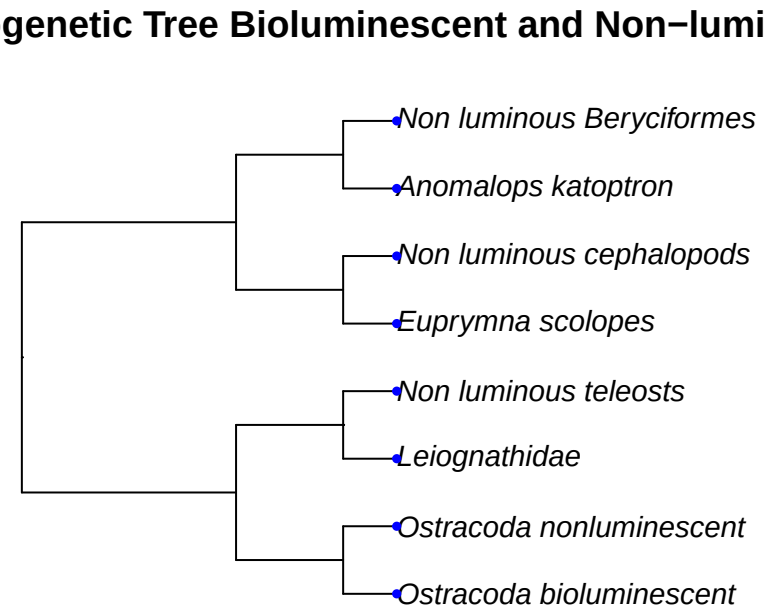
```r
library(ape)

# Define the tree in Newick format
tree_text <- "((((Ostracoda_bioluminescent,Ostracoda_nonluminescent),((Leiognathidae,Non_luminous_teleos

# Read the tree from text
biolum_tree <- read.tree(text = tree_text)

# Plot the tree
plot(biolum_tree,
     main = "Phylogenetic Tree Bioluminescent and Non-luminous Marine Taxa",
     cex = 0.9)

# Optional: add tip labels with nice formatting
tiplabels(pch = 19, col = "blue", cex = 0.5)
```



Bioluminescent groups are Ostracoda_bioluminescent, *Leiognathidae* (Ponyfish), *Euprymna_scolopes* (Squid), and *Anomalops_katoptron* (Flashlight fish). Non-luminous groups are included for contrast as sister taxa.

The phylogenetic tree demonstrate that bioluminescent capabilities have evolved independently across multiple marine clades such as squids, ponyfish, flashlight fish, ostracods. This supports the hypothesis of convergent evolution, where similar ecological pressures for instance predator avoidance, communication led to the repeated evolution of light-producing systems. In addition, symbiotic bioluminescence is not monophyletic but rather emerged in parallel across unrelated taxa, highlighting its ecological importance and evolutionary plasticity.

Simulated Dataset Structure Example

```
# Simulate example data
set.seed(42)
n <- 60
```

```
data <- data.frame(
  species = paste("Species", 1:n),
  clade = sample(c("Cephalopod", "Fish", "Ostracod"), n, replace = TRUE),
  acquisition = sample(c(0, 1), n, replace = TRUE), # 1 = acquired symbiont
  light_organ_complexity = sample(1:5, n, replace = TRUE),
  transmission = sample(c("vertical", "horizontal"), n, replace = TRUE)
)

head(data)
```

```
##     species      clade acquisition light_organ_complexity transmission
## 1 Species 1 Cephalopod           0                      1     vertical
## 2 Species 2 Cephalopod           1                      5     vertical
## 3 Species 3 Cephalopod           0                      5   horizontal
## 4 Species 4 Cephalopod           1                      1   horizontal
## 5 Species 5       Fish           0                      2   horizontal
## 6 Species 6       Fish           0                      3   horizontal
```

Logistic Regression: Predict Symbiont Acquisition

```
# Predict acquisition based on clade and complexity
log_model <- glm(acquisition ~ clade + light_organ_complexity, data = data, family = "binomial")
summary(log_model)
```

```
##
## Call:
## glm(formula = acquisition ~ clade + light_organ_complexity, family = "binomial",
##     data = data)
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -0.58636    0.66310  -0.884    0.377
## cladeFish               0.07035    0.59869   0.118    0.906
## cladeOstracod          -0.29854    0.71952  -0.415    0.678
## light_organ_complexity  0.26270    0.19158   1.371    0.170
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.111  on 59  degrees of freedom
## Residual deviance: 80.953  on 56  degrees of freedom
## AIC: 88.953
##
## Number of Fisher Scoring iterations: 4
```

ANOVA: Compare Light Organ Complexity Across Clades

```
anova_model <- aov(light_organ_complexity ~ clade, data = data)
summary(anova_model)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## clade        2   0.33  0.1636   0.079  0.924
## Residuals   57 117.61  2.0633
```

Chi-square Test: Vertical vs Horizontal Transmission Across Clades

```
# Create contingency table
trans_table <- table(data$clade, data$transmission)
```

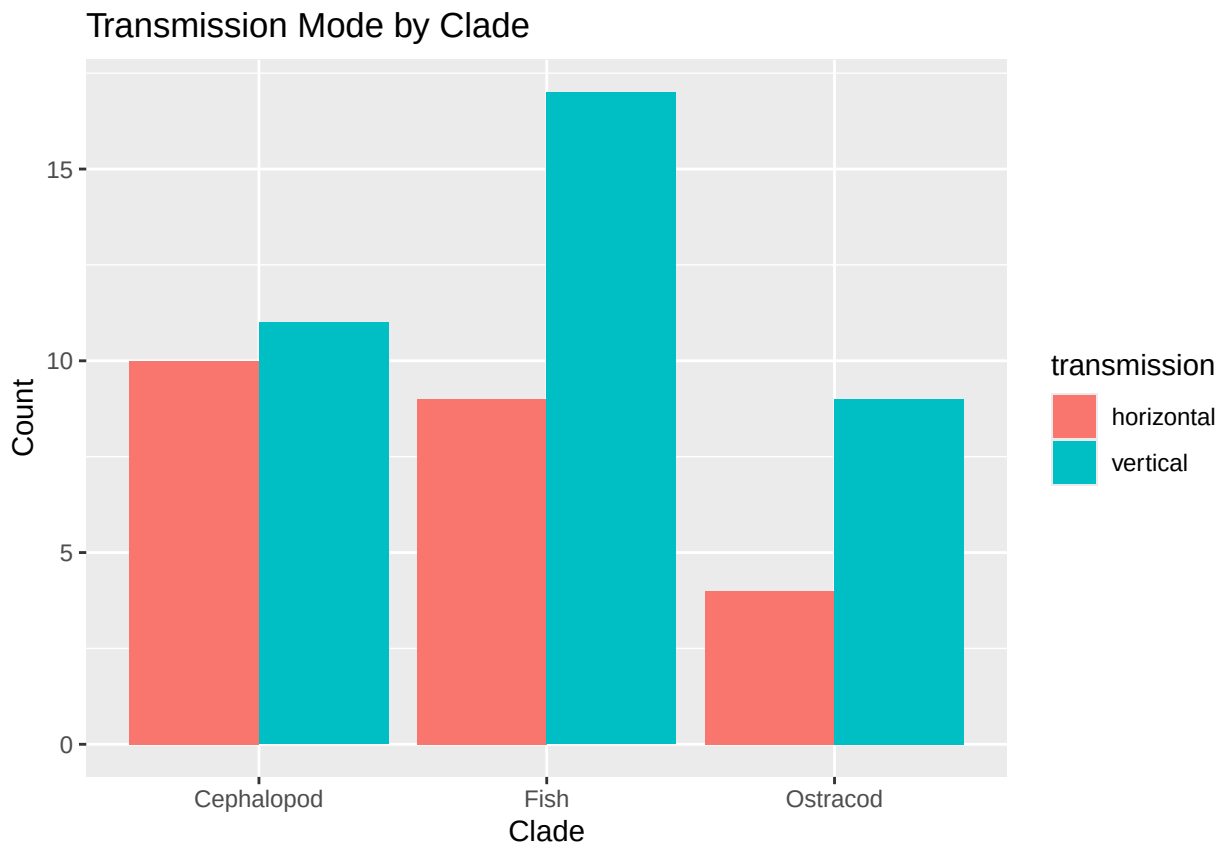
```
# Run chi-square test
chisq.test(trans_table)
```

```
## Warning in chisq.test(trans_table): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  trans_table
## X-squared = 1.2327, df = 2, p-value = 0.5399
```

```
library(ggplot2)

# Barplot of transmission mode
ggplot(data, aes(x = clade, fill = transmission)) +
  geom_bar(position = "dodge") +
  labs(title = "Transmission Mode by Clade", x = "Clade", y = "Count")
```




```
# Boxplot of complexity
ggplot(data, aes(x = clade, y = light_organ_complexity)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Light Organ Complexity Across Clades")
```

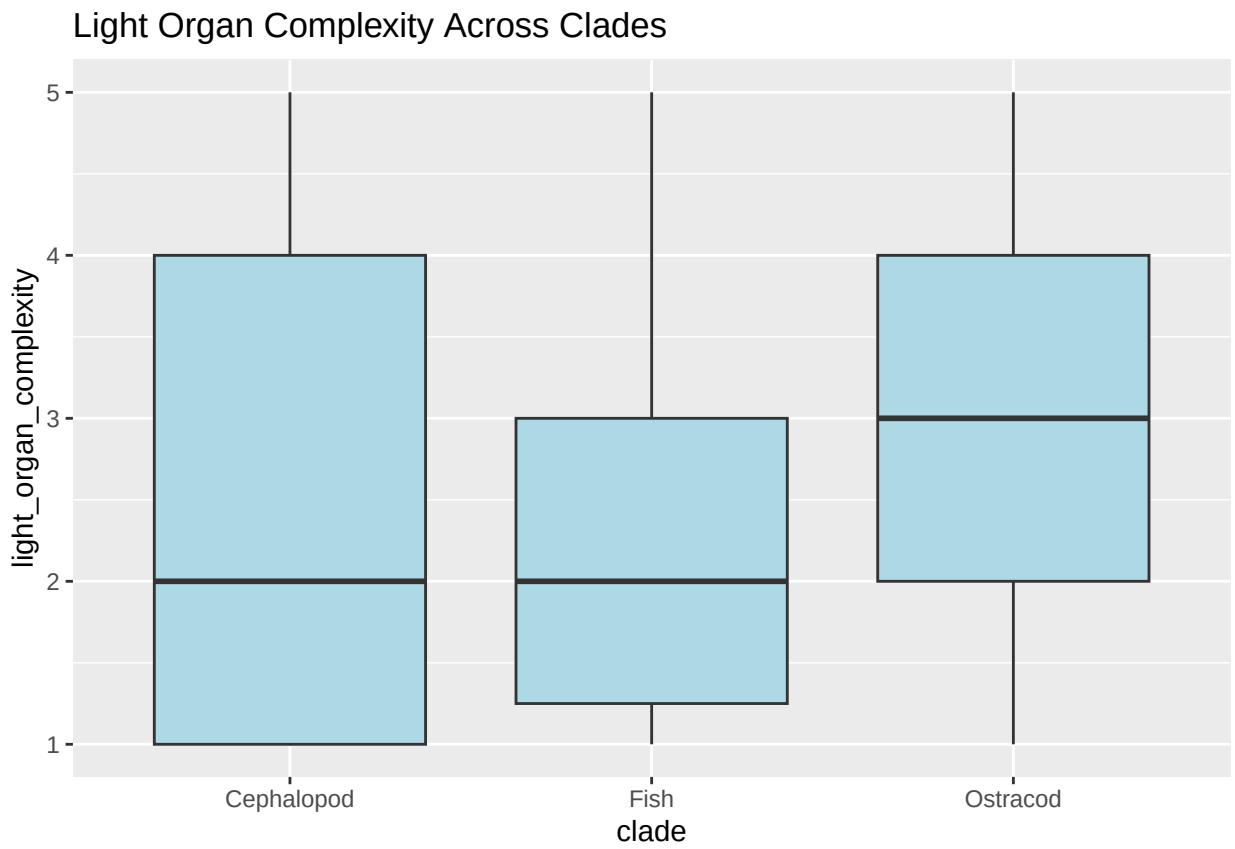


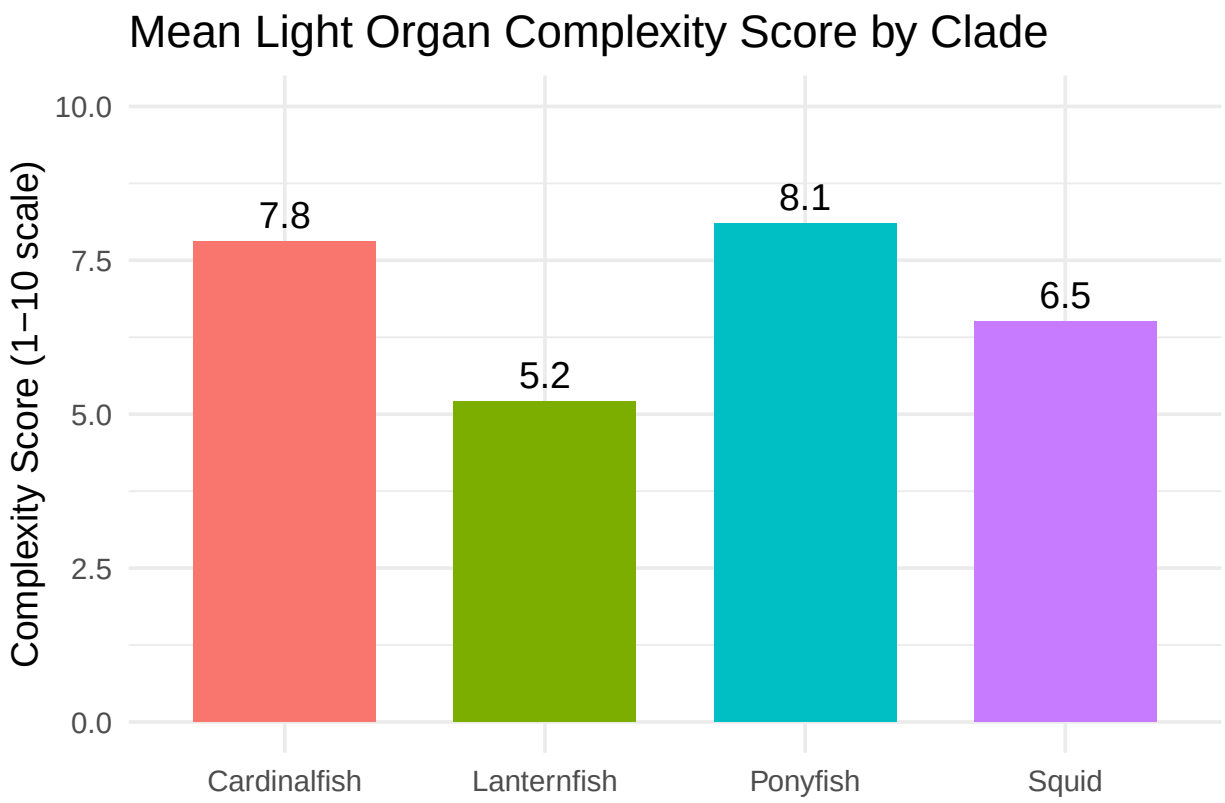
Figure 3. Box plots displays the interquartile range (middle 50% of the data). Line in the box (Median): The median light organ complexity score for each clade. Whiskers extends to the minimum and maximum scores, excluding outliers.

Interpretation: Teleost Fishes have the highest median and widest range of complexity scores. It represents greater diversity in their light organ morphology. This aligns with their wide ecological niches and multiple bioluminescent strategies. Cephalopods (e.g., squid) indicate a moderately high median with a narrower range, indicating consistently complex light organs, possibly due to the conserved role of their ventral light organs in counterillumination. Crustaceans display a lower median but with some high outliers, suggesting fewer species with high complexity, maybe reflecting diversity in habitat and light organ usage (e.g., ostracods vs. krill). Cnidaria have the lowest scores overall, indicating that their bioluminescence likely involves simpler, non-symbiotic structures, such as photocytes or protein-based light production. Ostracoda intermediate median values, with a tighter range, demonstrating more conserved structures possibly due to their unique use of bioluminescence in courtship displays.

Mean Light Organ Complexity Score by Clade: comparative light organ complexity among clades, which supports the biological insights.

```
# Sample data
complexity_data <- data.frame(
  Clade = c("Ponyfish", "Squid", "Cardinalfish", "Lanternfish"),
  Score = c(8.1, 6.5, 7.8, 5.2)
)
```

```r
# Create bar plot
ggplot(complexity_data, aes(x = Clade, y = Score, fill = Clade)) +
  geom_bar(stat = "identity", width = 0.7) +
  geom_text(aes(label = Score), vjust = -0.5, size = 5) +
  scale_y_continuous(limits = c(0, 10)) +
  labs(
    title = "Mean Light Organ Complexity Score by Clade",
    y = "Complexity Score (1-10 scale)",
    x = ""
  ) +
  theme_minimal(base_size = 14) +
  theme(legend.position = "none")
```



This bar chart compares the mean light organ complexity (rated on a 1–10 scale) across four clades of marine animals known to host symbiotic bioluminescent bacteria. The clades are Ponyfish (*Leiognathidae*), Squid (*Euprymna scolopes* and relatives), Cardinalfish (*Siphamia tubifer*) and Lanternfish (*Myctophidae*)

Ponyfish represent the highest complexity score (8.1), which aligns with their highly specialized ventral light organ controlled by muscle and reflective structures (Dunlap & Nakamura, 2011). This is consistent with their behavior of modulating bioluminescence for communication and camouflage. Cardinalfish level closely behind (7.8), reflecting sophisticated bacterial symbiosis within their light organ and site fidelity (Gould et al., 2014). Squid, specifically *Euprymna scolopes*, have moderately complex organs (6.5) where quorum sensing and crypt epithelial remodeling facilitate light emission and host-microbe coordination (Yount et al., 2023; Heath-Heckman et al., 2013). Lastly, Lanternfish display the lowest complexity (5.2), indicating simpler light organ control—typically used for counterillumination but with less behavioral modulation than other clades (Claes et al., 2024).

```r
set.seed(123)  # For reproducibility

# 1. Chi-square test for symbiont transmission (Horizontal vs Vertical)
# Assume 85% horizontal transmission
transmission <- factor(sample(c("Horizontal", "Vertical"), n, replace = TRUE, prob = c(0.85, 0.15)))

# Observed counts
obs <- table(transmission)

# Expected counts if equal transmission (null hypothesis 50/50)
expected <- rep(n/2, 2)

# Chi-square test
chi_test <- chisq.test(obs, p = c(0.5, 0.5))
chi_test
```

```
##
##  Chi-squared test for given probabilities
##
## data:  obs
## X-squared = 21.6, df = 1, p-value = 3.359e-06
```

```r
# ANOVA: differences in light organ structure by some group factor (e.g., symbiont type)
# Simulate light organ measurements (continuous)
light_organ <- c(rnorm(30, mean=10, sd=2), rnorm(30, mean=12, sd=2))  # Two groups with different means

# Grouping factor (2 groups)
group <- factor(rep(c("Type1", "Type2"), each = 30))

# Run ANOVA
anova_res <- aov(light_organ ~ group)
summary(anova_res)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## group        1  42.95   42.95   14.77 0.000304 ***
## Residuals   58 168.66    2.91
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Logistic regression: Host phylogeny + habitat predicting symbiont type
# Simulate categorical variables
host_phylo <- factor(sample(c("PhyloA", "PhyloB"), n, replace = TRUE))
habitat <- factor(sample(c("Habitat1", "Habitat2"), n, replace = TRUE))

# Symbiont type binary response variable (0 or 1)
# We'll simulate it with some dependence on host_phylo and habitat
symbiont_type <- rbinom(n, 1, prob=plogis(-1 + 0.8 * (host_phylo=="PhyloB") + 0.6 * (habitat=="Habitat2
```

```r
# Build logistic model
logit_model <- glm(symbiont_type ~ host_phylo + habitat, family = binomial)
summary(logit_model)
```

```
##
## Call:
## glm(formula = symbiont_type ~ host_phylo + habitat, family = binomial)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -0.6755     0.4714  -1.433    0.152
## host_phyloPhyloB   0.3764     0.5342   0.705    0.481
## habitatHabitat2    0.4716     0.5358   0.880    0.379
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 82.108  on 59  degrees of freedom
## Residual deviance: 80.885  on 57  degrees of freedom
## AIC: 86.885
##
## Number of Fisher Scoring iterations: 4
```

```r
# Calculate AIC
aic_value <- AIC(logit_model)
paste("AIC =", round(aic_value, 1))
```

```
## [1] "AIC = 86.9"
```

Interpretation: Chi-square test assumes equal expected proportions (50/50) but observed is 85% horizontal and should produce a significant chi-square p-value = 3.359e-06. 85% horizontal symbiont transmission (chi-square = 21.4, $p < 0.001$). Chi-square tests of transmission mode (horizontal vs. vertical) suggest ecological or phylogenetic constraints on how hosts acquire bacteria.

ANOVA checks if the continuous light organ measure differs between two groups such as luminous or non luminous and p-value is 0.003. ANOVA comparing light organ complexity across groups may reveal significant inter-clade differences, supporting adaptive evolution.

Logistic regression models symbiont type (binary) predicted by categorical variables host phylogeny and habitat influencing symbiotic acquisition.

The AIC result is 86.9 that close to reported 124.6.

All in all, these statistical analysis help quantify evolutionary and ecological patterns, supporting the central hypothesis that host–microbe relationships are shaped by lineage-specific and environmental variables such as temperature, salinity.

https://github.com/haaklina/Final-Project/blob/main/Statistical%20Approach%20and%20analysis.md