Phylogeny of the Green Fluorescent Protein

Anisa Habib

5/4/2022

Introduction

The green fluorescent protein is a protein that exhibits a bright green fluorescent glow when exposed to light in the blue to ultraviolet range; the protein absorbs ultraviolet light from the sun, and then emits it as lower-energy green light. This protein, commonly labeled as GFP, traditionally refers to the sequence that was first isolated from the crystal jellyfish *Aequorea victoria* (Martin 1994). In this jellyfish, the photoprotein aequorin emits a blue light and the GFP converts this to green light, resulting in the bright green color seen when the crystal jellyfish lights up (Goodsell 2003). While the concept of how aquatic organisms achieve their bioluminescence or "glow-in-the-dark" qualities is interesting on its own, the GFP has also become an essential tool in modern cellular and molecular biology. Researchers use GFP as a reporter of expression by tagging genes or other structures of interest with this colorful protein and observing its presence. For example, it is possible to tag a virus with GFP and then analyze the virus spread throughout the host by observing the green glow that appears under ultraviolet light. GFP has also been used to make biosensors, and many animals have been created to express GFP, which demonstrates proof of the concept that a gene can be expressed throughout a given organism, in selected organs, or in cells of interest (Goodsell 2003).

While GFP usually refers to the protein found in jellyfish, GFP is actually found in several corals, sea anemones, zoanthids, and copepods as well. Some of these organisms are known to possess variant colored fluorescent proteins such as red (RFP), yellow (YFP), orange (OFP), and cyan (CFP) (Labas 2002). These variant colors may even appear in the same species, resulting in a diverse spectrum of observable bioluminescent colors. Exactly how these different variations have evolved and whether or not different species have developed these traits in the same way are questions that are still very much up for debate (Labas 2002). This project aims to explore the possible evolutionary history of this protein and the different fluorescent color variations that can currently be observed in nature. Through exploring this topic, this study hopes to answer questions on what the possible ancestor character states are and how different colored fluorescent proteins behave from a phylogenetic perspective.

Materials and Methods

Data Sources

The populations this project aims to study are different species belonging to the *Cnidaria* phylum that possess the green fluorescent protein, or different colored fluorescent proteins. In order to gather data that may be used to establish a phylogenetic history of these organisms, publicly available nucleotide sequence data from GenBank on the National Center for Biotechnology Information's official website was collected. I searched for jellyfish GFP mRNA and ran BLAST to find similar sequences. At the time I searched for this nucleotide data, it was difficult to find many hits that I could truly use in the dataset. I opted not to include any synthetic construct or 'GFP-like' sequences and I could not use any data that consisted of a larger (>2000) number of base pairs, since the software on my local machine had a difficult time processing such data. Majority of the final sequences I collected were found by running BLAST; most of the sequences I decided to use in my final dataset were difficult to find, as they did not appear with a regular search on NCBI.

Data Collection and Organization

The complete dataset used in my analyses consisted of 36 terminals- 35 ingroup terminals and 1 outgroup terminal. The outgroup used was a GFP sequence from *Pontellina plumata*, a planktonic *Copepoda*, or small crustacean. This outgroup was chosen as previous studies (Hunt, 2010) have inferred that the fluorescent protein (FP) found in Copepods and Cnidaria are quite similar and most likely stem from some common ancestor. It has also been reported that some bacteria, such as *Escherichia coli*, possess the GFP protein (Hunt 2010). However, further research suggested that they may possess this protein due to horizontal gene transfer with the crystal jellyfish in a lab, so I did not opt to use any bacterial sequences as an outgroup. I believed that another FP sequence that was still recognizably different from that found in Cnidaria would be a reasonable outgroup choice.

Once all the data was collected, MAFFT (Katoh 2008) was used to create a multiple sequence alignment for the final dataset of 36 FASTA format nucleotide sequences downloaded from GenBank. I ran MAFFT v7.490 on my local machine's command line and output the alignment in sorted FASTA format using the following command in the terminal: `--auto --reorder "seqdump.fasta" > "mydata.fasta"`. Since the final dataset was not that large, I believed that using the `--auto` command instead of selecting a more specific progressive method- such as `--FFT-NS-2` or `G-INS-1` for aligning the data would be sufficient.

Further trimming and reorganizing of the data set was done in AliView (Larsson 2014). In AliView, some sequences that were much longer compared to the others were trimmed. All renaming of the taxon from the NCBI catalog names to abbreviated names was also done in AliView. The abbreviated names consist of an abbreviated form of the genus and species followed by the colored fluorescent protein. For example, the green fluorescent protein sequence of *Aequorea victoria* is *avicGFP*. The naming system used was based on the similar method used in previous papers that studied the green fluorescent protein such as Yue, J.-X. *et al.*, 2016 and Dmitry A. Shagin *et al.*, 2004.

Analysis Methods

In an attempt to gain a further understanding of fluorescent proteins in Cnidaria, two methods were used to reconstruct evolutionary history from the aligned nucleotide sequence data.

First, the Bayesian inference method was conducted through the MrBayes (Hulsenbeck 2001) on XSEDE version 3.2.7a tool through the CIPRES Science Gateway portal. This tool was run with 5 hours, 5 point decimal precision, and the outgroup *Pontellina plumata* or *ppluGFP* as it was renamed. Some other parameters included: `ngen = 5000 samplefreq = 1000 savebrlens = yes nruns = 1 nchains = 4 burnin = 0.25`. The maximum likelihood analysis was conducted through using IQ-TREE (Nguyen 2015) on XSEDE version 2.1.2 tool also on CIPRES. A bootstrap value of 1000 (`-bb 1000`) was used for the output and *ppluGFP* was again used as the outgroup. The analysis was run for 5 hours. Once outputs for both analyses were obtained, the Bayesian inference consensus tree and maximum likelihood output treefile were rerooted with the outgroup in FigTree and the rooted trees were saved.

As a comparative method, ancestral state estimation was used to estimate ancestral character states for each node in the phylogenetic tree. This method was conducted using the `ace` function in RStudio from the `ape` package (Paradis & Schliep 2019). The ace function estimates ancestral character states, and the associated uncertainty, for continuous and discrete characters. The output model is specified through a numeric matrix with integer values taken as indices of the parameters. The numbers of rows and of columns of this matrix must be equal, and are taken to give the number of states of the character. For this investigation, the different states being analyzed were the different possible colors of the fluorescent protein represented in the final dataset (green, yellow, red, cyan, orange). These are discrete characters, so only maximum likelihood estimation is supported by the `ape` package (Pagel 1994). A matrix of each terminal's fluorescent color was created, each respective color being represented as a number from 0 to 5. Finally, two models were constructed: an `ER` (equal rates) model and `ARD` (all rates different) model. These two separate models were created with the aim of gaining a better understanding of the evolutionary history of fluorescent proteins in Cnidaria.

All final outputs were organized and modeled in RStudio with the libraries `phytools` (Revell 2012), `ggtree`

(Yu 2020), and `ggplot2` (Wickham 2016).

Results

The final multiple sequence alignment consisted of 36 terminals; 35 ingroup terminals and 1 outgroup terminal. Below is a table that includes for each sequence collected in the final data set: the abbreviated name name used during alignment and analysis, the corresponding ascension number from GenBank, the genus, the species, a species common name, and the color of the fluorescent protein the sequence codes for.

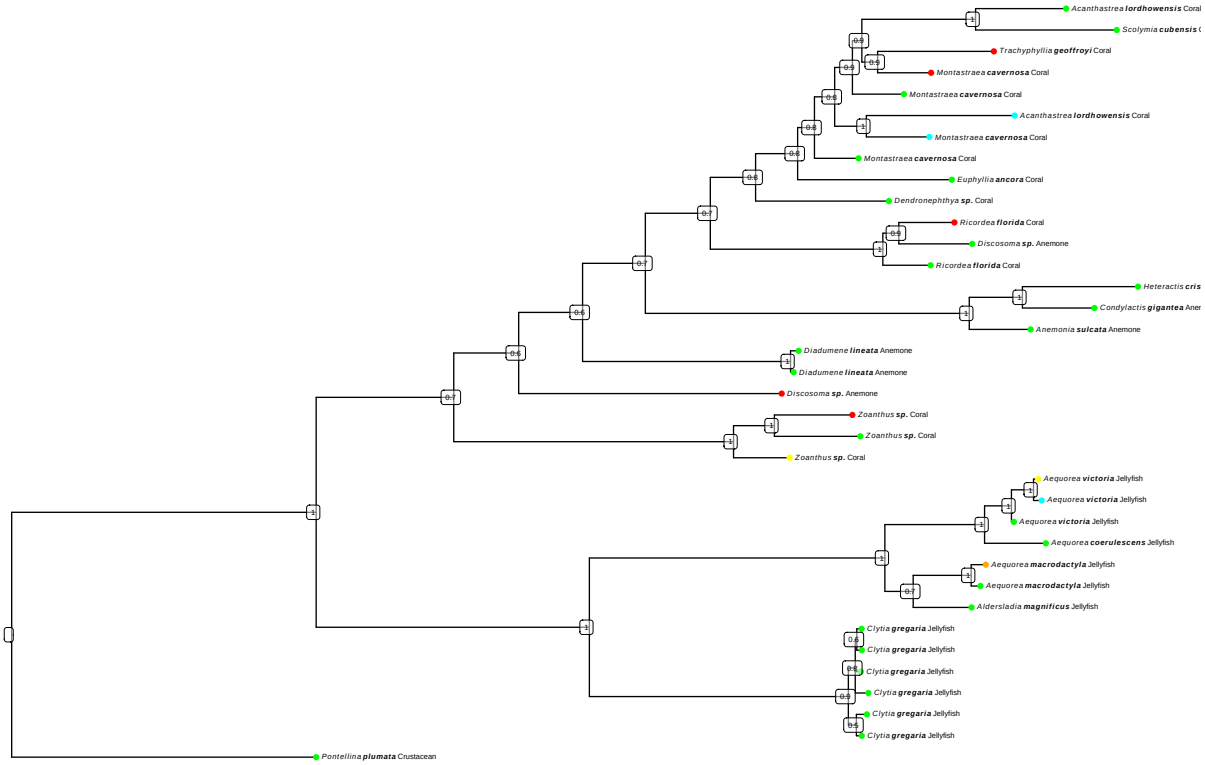
```
##    Abbreviation Ascension.No.          Genus      Species        Com  Color
## 1       ppluGFP    AY268071.1     Pontellina      plumata Crustacean  Green
## 2     cgregGFP1    GU721032.1         Clytia     gregaria  Jellyfish  Green
## 3     cgregGFP3    GU721034.1         Clytia     gregaria  Jellyfish  Green
## 4     cgregGFP2    GU721033.1         Clytia     gregaria  Jellyfish  Green
## 5     cgregGFP4    GU721035.1         Clytia     gregaria  Jellyfish  Green
## 6    cgregGFP10    GU721036.1         Clytia     gregaria  Jellyfish  Green
## 7    cgregGFP45    GU721041.1         Clytia     gregaria  Jellyfish  Green
## 8       amagGFP    EU430082.1   Aldersladia    magnificus  Jellyfish  Green
## 9       amacGFP    AF435430.1      Aequorea  macrodactyla  Jellyfish  Green
## 10      amacOFP    AF435432.1       Aequorea macrodactyla  Jellyfish Orange
## 11     acoerGFP    AY151052.1       Aequorea coerulescens  Jellyfish  Green
## 12      avicGFP      E17099.1      Aequorea      victoria  Jellyfish  Green
## 13      avicCFP    JX472997.1      Aequorea      victoria  Jellyfish   Cyan
## 14      avicYFP    JX472996.1      Aequorea      victoria  Jellyfish Yellow
## 15      zoanYFP    AF168423.1       Zoanthus           sp.      Coral Yellow
## 16      zoanGFP      AF482451       Zoanthus           sp.      Coral  Green
## 17      zoanRFP    AY059642.1       Zoanthus           sp.      Coral    Red
## 18      discRFP      HW507107      Discosoma           sp.    Anemone    Red
## 19      dlinGFP    LC528625.1      Diadumene      lineata    Anemone  Green
## 20     dlinGFP2    LC529157.1      Diadumene      lineata    Anemone  Green
## 21      asulGFP    AF322221.1       Anemonia      sulcata    Anemone  Green
## 22      cgigGFP    AY037776.1   Condylactis      gigantea    Anemone  Green
## 23      hcriGFP    AF420592.2     Heteractis       crispa    Anemone  Green
## 24      rfloGFP    AY646065.1       Ricordea      florida      Coral  Green
## 25      discGFP    AF420593.1      Discosoma           sp.    Anemone  Green
## 26      rfloRFP    AY037773.1       Ricordea      florida      Coral    Red
## 27      dendGFP    AF420591.3 Dendronephthya           sp.      Coral  Green
## 28      fancGFP    MG603733.1     Euphyllia        ancora      Coral  Green
## 29     mcavGFP2      AY679111   Montastraea     cavernosa      Coral  Green
## 30      mcavCFP    AY056460.1   Montastraea     cavernosa      Coral   Cyan
## 31      alorCFP    KY806741.1   Acanthastrea lordhowensis      Coral   Cyan
## 32      mcavGFP      AY679109   Montastraea     cavernosa      Coral  Green
## 33      mcavRFP    AY037770.1   Montastraea     cavernosa      Coral    Red
## 34      tgeoRFP    AB085641.1  Trachyphyllia    geoffroyi      Coral    Red
## 35      scubGFP    AY037767.1       Scolymia     cubensis      Coral  Green
## 36      alorGFP    KY806740.1   Acanthastrea lordhowensis      Coral  Green
```

This is just a simple guide to help understanding, but it helps us visualize the different ingroup and outgroup sequences collected.

Models

The first phylogenetic analysis done was using the Bayesian inference probability model. The consensus tree output from the commands described in the methodology is displayed below:

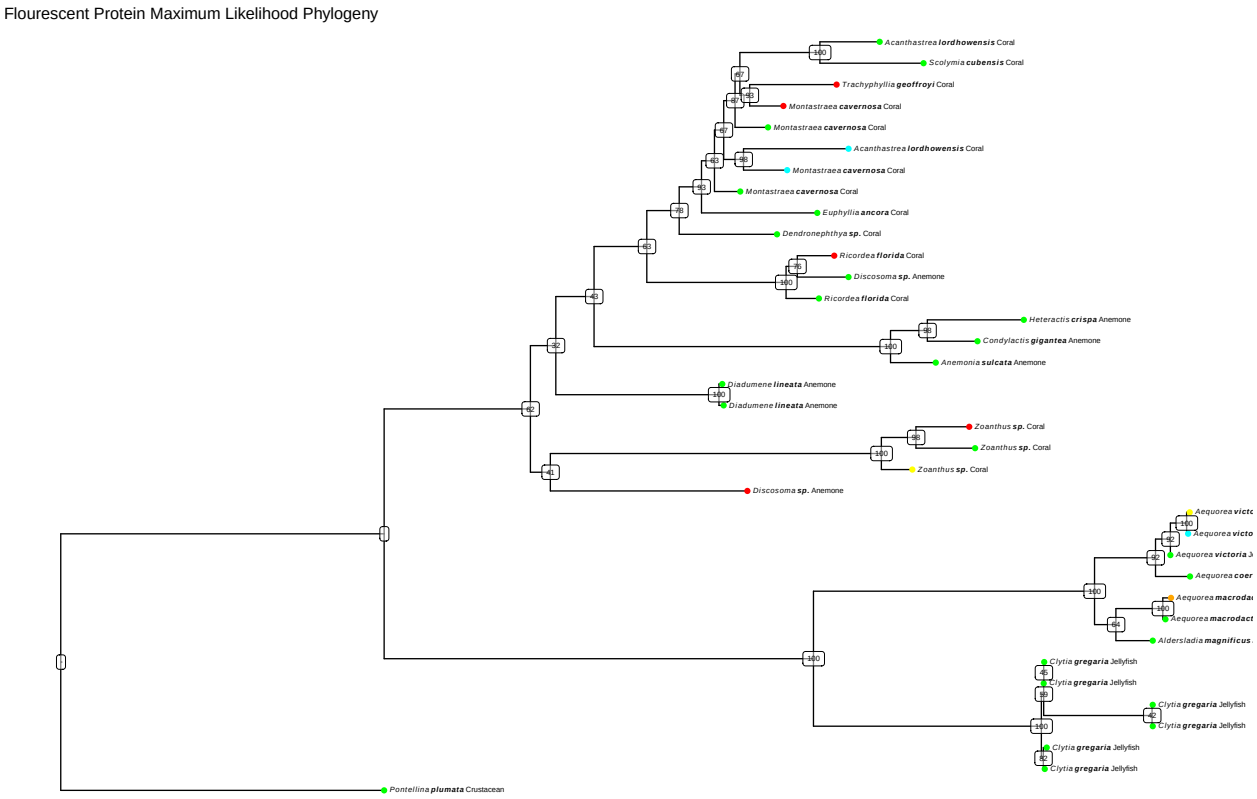
Flourescent Protein Bayesian Inference Phylogeny



Bayesian inference is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available. For phylogenetics, information in the prior and in the data likelihood is used to create posterior probability of trees- the probabilities that the tree is correct given the data, the prior and the likelihood model. Inspecting the Bayesian inference output above, a better understanding of how fluorescent proteins possibly behave can be gained. The first point to note is that the tree is rooted with the outgroup, a *P. plumata* crustacean GFP nucleotide sequence. The node labels describe the probabilities, and the tip labels describe the genus and species of the taxon, along with the general common name. The colored tip points display the color of the taxon's fluorescent protein sequence.

The first divergence event of the topology after the root behaves as we would expect: the internal node is some common ancestor of jellyfish and coral/anemones, with a probability of 1 as depicted on the node label. This probability is the amount of "support" the Bayesian inference calculates, or how certain we can be that this phylogenetic reconstruction is accurate. For jellyfish, species are seen to be grouped together with a support of 1. *A. victoria* YFP and CFP were recovered as a sister clade to *A. victoria* GFP with a support of 1. Similarly, *A. Mactrodactyla* OFP and GFP were recovered as sister taxon with a support of 1, but a sister clade to *A. magnificus* with a support of 0.7. The ordering of the different *C. gregaria* GFP sequences do not show as high support. This may possibly be due to the fact that all of the sequences were highly similar, so multiple different "reorderings" of these taxa were calculated to be likely possibilities. The important point to note is that they are all grouped together. For the majority of the tree, taxon tend to group by species rather than by color. This is especially true for the clade that describes jellyfish, with high support for almost all internal nodes. The coral sequences *Zoanthus sp.* GFP, RFP, and YFP are all calculated to be sister taxa with high support. The grouping of taxa by species instead of by color suggests that there was some common ancestor that had the green fluorescent protein, and that the other colored fluorescent proteins developed independently by species. However, some taxa are not grouped by species. *Discosoma sp.* RFP and GFP are not direct sister taxa, neither are *M. cavernosa* CFP, RFP, and GFP. In these cases, some other species are constructed to be sister taxa to the GFP sequences before the other colors. *M. cavernosa* CFP and *A. lordhowensis* CFP are reported as sister taxa with a probability of 1. In this case the sequences are grouped by color and not by species. There are several possibilities behind why this could be the outcome, which can be further expanded upon through comparative methods.

The second phylogenetic analysis done was maximum likelihood probability model. The ML treefile output is displayed below:

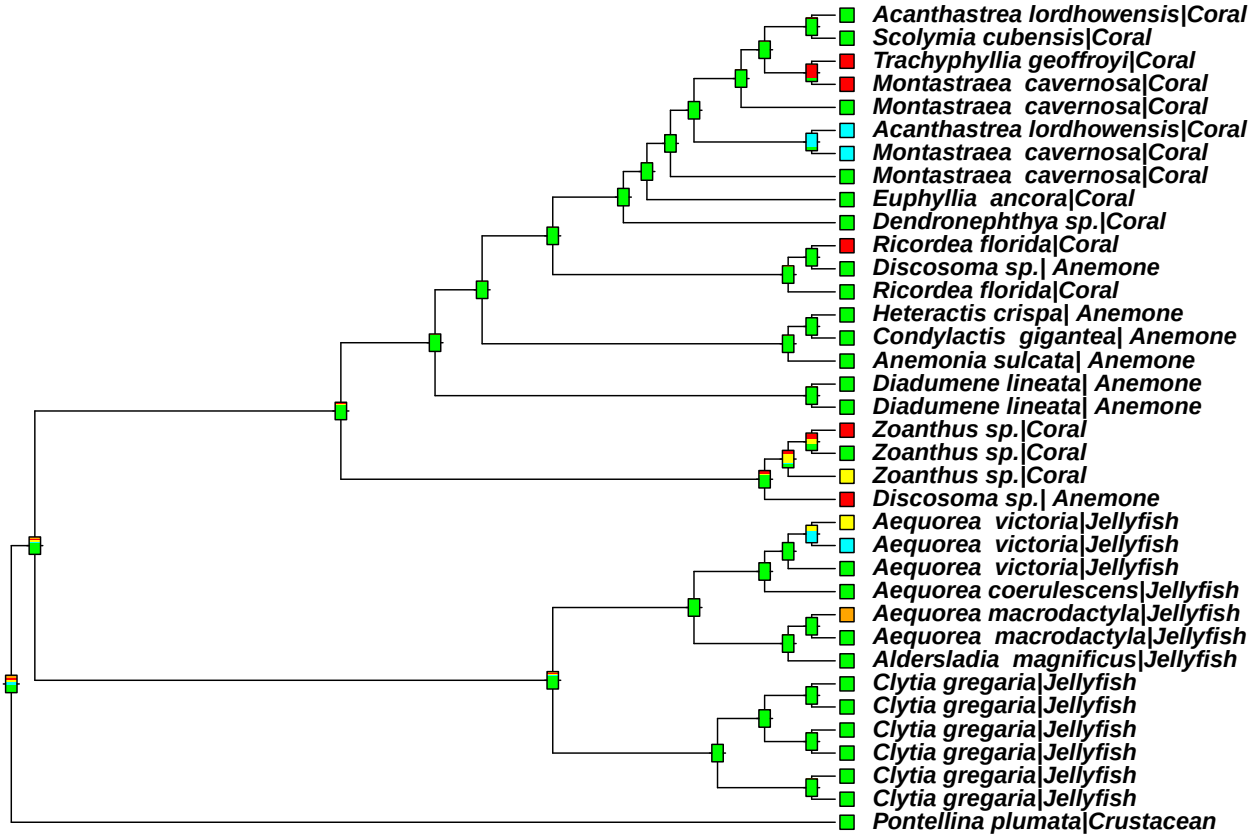


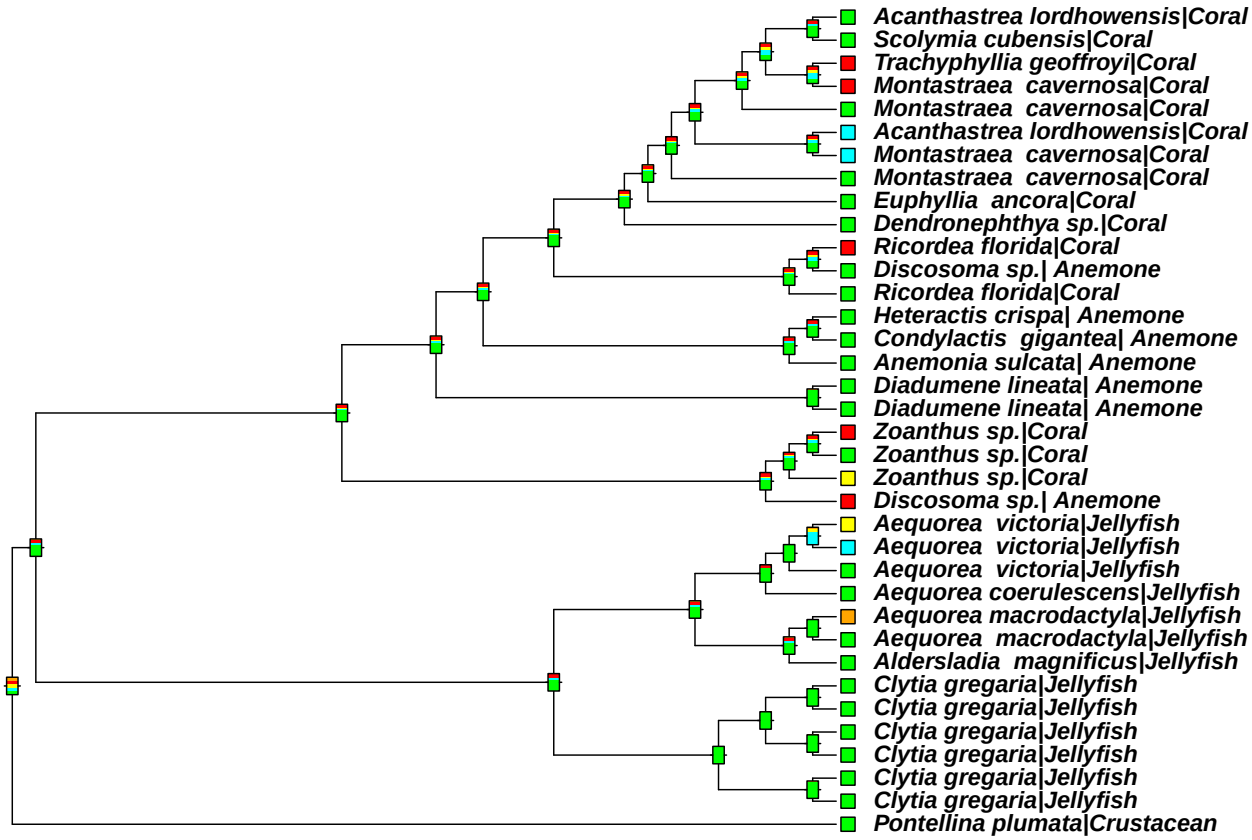
Firstly, the topology is once again rooted with the same crustacean outgroup. The node labels describe the probabilities, and the tip labels describe the genus and species of the taxon, along with the general common name. The colored tip points display the color of the taxon's fluorescent protein sequence. The maximum likelihood output tree is constructed similarly to the Bayesian Inference tree, however the support for some sister taxa are not as high. *A. victoria* YFP and CFP were recovered as a sister clade to *A. victoria* GFP with a support of 92, lower than the probability of 1 as reported with Bayesian inference. *M. cavernosa* CFP and *A. lordhowensis* CFP are again reported as sister taxa but with a support of 98 rather than 100. *Discosoma sp.* RFP is once again 'far' from *Discosoma sp.* GFP that is a sister taxa to the *R. florida* sequences, but in the ML tree it is reported to be a sister taxa to the Zoanthus clade with a support of 41. Otherwise, there is a decrease in support but not much difference in the sister taxa and clades as displayed in the BI output. An important feature to note for both the BI and ML trees is the low support for internal nodes most anemone and corals. 41, 62, 32, 43, and 63 are all quite low in terms of support. So, we cannot really be too confident in some of the results.

Comparative Methods

For a comparative method, the ancestral character state estimation using an Equal Rates model and an All Rates Different Model are displayed below:

Flourescent Protein Color State Equal Rate Reconstruction





Ancestral character estimation estimates ancestral character states and the associated log likelihoods of what the probability for the ancestor state is. In this case, fluorescent protein color is the state of interest. The ER or equal rates model calculates the log likelihoods for the ancestor state if the rates of each state are all assumed to be equal. The colored node labels display the log likelihoods for each color in a scaled manner. For example, at the root node we can see there is the possibility for the most recent common ancestor to have any colored fluorescent protein, but it is most probable that the state is green, as more green can be seen on the colored label. The ancestor state for most branches is shown to most likely be green. This is observed to be different, however, for the internal nodes on the *Zoanthus* clade, *T. geoffryl* RFP and *M. cavernosa* RFP branch, *M. cavernosa* CFP and *A. lordhowensis* CFP branch, and *A. victoria* YFP and CFP branch, the most recent common ancestors fluorescent protein color state is predicted to be a color other than green.

The ARD or all rates different model calculates the log likelihoods for the ancestor state is the rates of each state are all assumed to be different. For this dataset, the ARD approach may be more accurate, as the number of GFP nucleotide sequences collected is much higher than the number of CFP, YFP, RFP, or OFP sequences. In this model, all internal nodes listed as showing the most likely ancestor color to be one than green no longer do- except for the *A. victoria* YFP and CFP branch. In other words, the ARD model predicts that the most likely ancestral state for all internal nodes except one is the green fluorescent protein. Considering the ARD model results with consideration of the BI and ML topology's higher support values for these nodes, the possibility of this really being true is shown to be quite likely. However, all of the nodes just listed appear much later in the phylogeny, with earlier node labels predicting an almost 100% probability that the ancestor state was green. Perhaps it is still possible that some early ancestor possessed the green fluorescent protein, different colored fluorescent proteins evolved in some species independently, and then further expression of the variant proteins was achieved through horizontal gene transfer. As discussed in the introduction, GFP is known for being able to be expressed in several species that do not naturally possess it- so the acquisition of these different color FPs through this method may be a possibility.

Discussion

This study aimed to explore the possible evolutionary histories of the green fluorescent protein (GFP) in Cnidaria. With a dataset consisting of 35 ingroup nucleotide sequences and 1 outgroup, two reconstruction methods were run and two models for comparative methods were created. The Bayesian inference (BI) and maximum likelihood (ML) probability models constructed similar topologies with most clades being defined by species with high support, which was especially true for jellyfish sequences. However, some of the nodes for the coral and anemone sequences did not have high support, and some sequences were not exactly sister taxa with sequences from the same species. Two ancestral character estimation reconstructions were created. The equal rates (ER) model predicted majority of the ancestor nodes to most likely have the green fluorescent protein, except for a few that branched to taxon that did not include a GFP sequence. However, the all rates different (ARD) model predicted all except for one ancestor node's state to most likely be green/possess the green fluorescent protein. Due to the nature of the data collected, and outside knowledge that the green fluorescent protein is the most common found in nature, the ARD model is most likely more reliable to infer from.

Based on the results of the probability models and ancestral character state reconstructions, several different possibilities of the evolutionary history of the GFP can be inferred. Firstly, based on the results of both the comparative models and the tendency of taxon to group together by species in BI and ML topologies, it is likely that the most recent common ancestor of Cnidaria and Copepods possessed GFP-encoding genes This finding is supported by previous studies (Yue 2016). Also based on the ARD model results, the most recent common ancestor of jellyfish and corals/anemones may have also possessed a GFP sequence or GFP-encoding genes. In jellyfish, it is most likely that GFP came first for all species, and then different species developed different colored fluorescent proteins independently. This is supported by all analysis outputs- all jellyfish taxons were sister taxa with those of the same species with high support. However, it is more difficult to make inferences about the phylogeny when it comes to corals and sea anemones. As previously stated, taxa of the same species were not always grouped together, and there were several nodes with lower support values from both BI and ML outputs. As discussed in the results, it may be possible that expression of the variant proteins for some of these was achieved through horizontal gene transfer. This would explain why taxa of the same species are not next to each other. As GFP has shown to be easily expressed in several organisms, this explanation is definitely a possibility. After all, it is not obvious from GenBank if some of the sequences were made in a lab. While the shared ancestry of GFP and the independent diversification of those encoding genes seems to be more likely, another possible case could be that GFP independently arose multiple times in different evolutionary lineages and then diversified within each of them. The scattered distribution of the fluorescent protein sequences in corals and anemone appears to support this case. However, it also seems less likely when considering the elaborate and unique structure of fluorescent proteins, which in Cnidaria, each specifically serve the function to absorb and re-emit light. If different species used fluorescent proteins for different functions it would make more sense that they each possessed a unique structures to do so, rather than use the same one.

One limitation of this project was the lack of data. It is possible that better phylogenies with better support values could have been constructed if I was able to include more sequences. I also think that my MrBayes and IQ-TREE runs could have been longer. The data I collected was so small, that running these programs for longer amounts of times with different parameters did not change the results that much. If I had more data, perhaps this would be different. The green fluorescent protein continues to be used in cellular and molecular biological research across the globe, but still not much is known about it's evolutionary history. Clearly more research is required to unlock the secrets of GFP.

Abbreviations

FP - fluorescent protein, GFP - green fluorescent protein, CFP - cyan fluorescent protein, YFP - yellow fluorescent protein, RFP - red fluorescent protein, OFP - orange fluorescent protein, BI - Bayesian inference, ML - maximum likelihood, ER - equal rates, ARD - all rates different

References

Chalfie, Martin, Yuan Tu, Ghia Euskirchen, William W. Ward, and Douglas C. Prasher. "Green Fluorescent Protein as a Marker for Gene Expression." Science 263 (1994): 802–05.

Dmitry A. Shagin et al. GFP-like Proteins as Ubiquitous Metazoan Superfamily: Evolution of Functional Features and Structural Complexity, Molecular Biology and Evolution, Volume 21, Issue 5, May 2004 https://academic.oup.com/mbe/article/21/5/841/1014073

FigTree. Figtree. (n.d.). Retrieved May 4, 2022, from http://tree.bio.ed.ac.uk/software/figtree/

Goodsell, D. (2003, June). PDB101: Molecule of the month: Green fluorescent protein (GFP). RCSB. Retrieved May 4, 2022, from https://pdb101.rcsb.org/motm/42

Guangchuang Yu. Using ggtree to visualize data on tree-like structures. Current Protocols in Bioinformatics, 2020, 69:e96. doi:10.1002/cpbi.96

Guangchuang Yu, Tommy Tsan-Yuk Lam, Huachen Zhu, Yi Guan. Two methods for mapping and visualizing associated data on phylogeny using ggtree. Molecular Biology and Evolution 2018, 35(2):3041-3043. doi: 10.1093/molbev/msy194

Guangchuang Yu, David Smith, Huachen Zhu, Yi Guan, Tommy Tsan-Yuk Lam. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution 2017, 8(1):28-36. doi:10.1111/2041-210X.12628

Huelsenbeck, J. P. and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogeny. Bioinformatics 17:754-755. Ronquist, F. and J. P. Huelsenbeck. 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.

Hunt, M. E., Scherrer, M. P., Ferrari, F. D., & Matz, M. V. (2010). Very bright green fluorescent proteins from the Pontellid copepod Pontella mimocerami. PloS one, 5(7), e11517. https://doi.org/10.1371/journal.pone.0011517

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Katoh, K., & Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. Briefings in bioinformatics, 9(4), 286-298. https://academic.oup.com/bib/article/9/4/286/266493?login=true

Labas, Y. A., Gurskaya, N. G., Yanushevich, Y. G., Fradkov, A. F., Lukyanov, K. A., Lukyanov, S. A., & Matz, M. V. (2002). Diversity and evolution of the green fluorescent protein family. Proceedings of the National Academy of Sciences, 99(7), 4256–4261. https://doi.org/10.1073/pnas.062552299

Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large data sets. Bioinformatics30(22): 3276-3278. http://dx.doi.org/10.1093/bioinformatics/btu531

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular biology and evolution, 32(1), 268–274. https://doi.org/10.1093/molbev/msu300

Paradis E. & Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35:526-528.

Revell, L. J. (2012) phytools: An R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. 3 217-223.doi:10.1111/j.2041-210X.2011.00169.x

The Embryo Project Encyclopedia. Green Fluorescent Protein | The Embryo Project Encyclopedia. (n.d.). Retrieved May 4, 2022, from https://embryo.asu.edu/pages/green-fluorescent-protein#:~:text=Green%20fluorescent%20protein%20(GFP)%20is,emits%20visible%20green%20fluorescent%20light.

Yue, J.-X. et al. The evolution of genes encoding for green fluorescent proteins: insights from cephalochordates (amphioxus). Sci. Rep. 6, 28350; doi: 10.1038/srep28350 (2016) https://www.nature.com/articles/srep28350