2021

Why are Glycines 31, 33, and 35 Highly Conserved in all Fluorescent Proteins?

Justin Nwafor
Connecticut College, jnwafor1@hotmail.com

Why are Glycines 31, 33, and 35 Highly Conserved in all Fluorescent Proteins?

Justin Nwafor

Connecticut College, Department of Chemistry

New London, Connecticut 06320

Spring 2021

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

As I write this last section of my thesis at 6:27 AM on May 3, 2021, all I can think about is how this is the only place on this whole document where I could possibly put a joke without facing major scrutiny. I will not do any such thing, unless I have space at the bottom of this page to do so.

I'd like to first thank God and my family for being one of my biggest pillars of support and inspiration. In all fairness and reality, they are the ones that I work so hard for. To Marc Zimmer, my research/academic advisor, thank you for giving me the space to grow as a researcher and a person in your lab over the past four years. To the friends I've made over the years, thank you for keeping my time at Conn all the more exciting (even though you all made me age 40 years). To Paula Orbe and all the professors of the chemistry, mathematics, and physics departments, thank you for your support and allowing me to really see the deep rooted connections and elegance of the world around me.

To my past labmates, Franceine Welcome and Sercan Durmus, thank you for all the laughs and late night lab adventures. To my current labmates, Sophia Moroney and Admirabilis Kalolella, although we never really had the chance to work together in the lab due to the pandemic, I know you two are very bright students and I am glad I know I will leave the lab in good hands. To Christian Salguero, my partner in crime, my best friend, words cannot even explain how much I appreciate our friendship; love you like one of my brothers, you dork.

Lastly, I thank those who have even gotten to this point in reading this thesis. Odds are you're not here by choice, but hopefully the read is not too painful. Enjoy! I guess now it is time to prove that Christian and I actually did something other than goof around in Hale for three academic years and a summer.

ABSTRACT

Since the discovery of the Green Fluorescent Protein (GFP), fluorescent proteins (FPs) have been used as biomarkers to monitor biological phenomena across many scientific disciplines. All naturally-occurring FPs consist of an 11-stranded β-barrel with a non-canonical α-helix, which contains the chromophore, running through the central axis of the protein. Glycines 31, 33, and 35 are highly conserved across the fluorescent proteins found in the PDB. These three residues are of interest due to them all being located in the second strand of the β-barrel, but having no direct involvement in chromophore formation. This led to a presumption that the glycines likely allowed space in a correctly folded β-barrel for the chromophore to form.

In this study, molecular dynamics simulations of G31A, G33A, and G35A single point mutants of wild-type GFPs with immature (pre-cyclized) chromophores were used to investigate how mutations to these residue positions could affect chromophore formation. Four additional mutant simulations were performed to investigate the hydrophobic pocket that contains G35. This was done by examining the hydrogen bond network in the central α-helix, water migration through the β-barrel, aromatic rescue interactions, and main chain interactions among the N-terminus β-sheets. The simulations show that if the β-barrel folds correctly, mutating the conserved glycines does not result in hindrance or prevention of chromophore formation.

Through experimental analysis, it was found that the G3XA mutants were prone to misfold and aggregate, suggesting that these glycines play a crucial role in the folding pathway of fluorescent proteins. Computationally, this was confirmed as the introduced mutations caused reduced main chain interactions among the N-terminus β-sheets.

INTRODUCTION

GFP History

Green Fluorescent Protein (GFP) was first discovered in the 1960's by Dr. Osamu Shimomura while studying bioluminescence of the crystal jellyfish, *Aequorea victoria* (Fig.1).[1] Shimomura first found that the molecule responsible for *Aequorea* luminescence was aequorin. Aequorin is a monomeric photoprotein that consists of an apoprotein, apoaequorin, and a chromophore made of coelenterazine, a luciferin, and molecular oxygen. In the photo-organs of the jellyfish, Ca^{+2} binds to aequorin, causing an oxidation of coelenterazine to coelenteramide, which yields light (λ_{max} = 470 nm), carbon dioxide, and a blue photoprotein that consists of the oxidized coelenteramide and apoaequorin.[2]

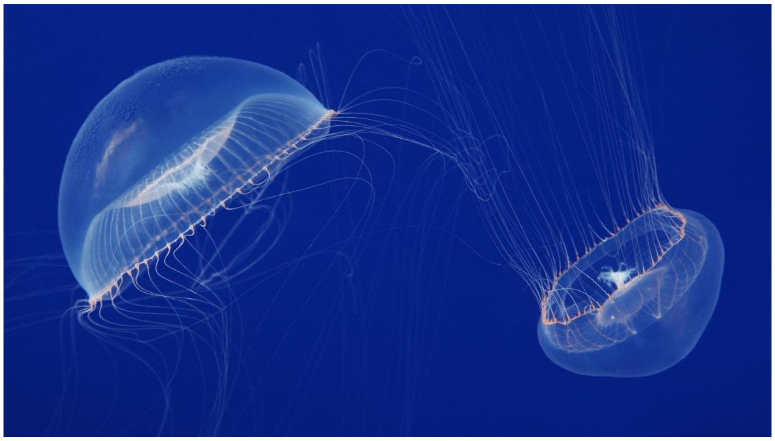


Figure 1. Crystal Jellyfish, *Aequorea victoria*, courtesy of the Monterey Bay Aquarium

When the reaction happens *in vitro*, it gives off blue light, meaning that the green light given off by the crystal jellyfish could not be completely explained by aequorin photochemistry.

In 1969, Shimomura found that there was a protein present that turned blue light to green light and appropriately called it “green fluorescent protein”. Shimomura then spent the next ten years investigating the structure of the GFP chromophore, which he published in a paper in June of 1979.[3]

Attempts to clone GFP by Doug Prasher and Bill Ward, postdocs in the Cormier Lab group, were made in order to resolve the issue of having to accumulate massive numbers of jellyfish to only receive minimal amounts of the protein for investigation. Ward sequenced *Aequorea* aequorin and GFP, and Prasher ended up cloning the aequorin. Prasher’s attempt to clone GFP resulted in a non-fluorescent apoGFP, which led Ward to think that formation of the chromophore was most likely a nonspontaneous process and that it could be not used as a tracer molecule, as Prasher had first thought.[4]

Martin Chalfie, a specialist in neurobiology and genetics at Columbia University, and Ghia Euskirchen, a rotation student that was working under Chalfie at the time, were the first to correctly isolate and express the GFP gene. This was done by using polymerase chain reaction (PCR) to amplify the coding gene for the protein, instead of using the same method that Prasher had used to cut the gene out; this left extra nucleotides that preceded the GFP gene, preventing the protein, and subsequently the chromophore, from forming correctly. Euskirchen was the first to successfully express the GFP gene in *Escherichia coli*. Later on, Chalfie was able to express GFP in touch neurons of *Caenorhabditis elegans*, a small nematode that is widely used in genetic analysis laboratories.[5] The most important finding from this work was that the GFP chromophore does form autocatalytically. Since Chalfie’s expression of GFP in *C. elegans*, fluorescent proteins have been used in many different organisms as a tracer molecule to monitor different phenomena and make visualization of different cell functions, processes, gene expression, etc. in different organisms much easier. GFP is also a much easier way to image cellular functions because the protein readily makes its own chromophore, so its DNA code can be easily tagged onto the DNA of the protein under investigation and it will fluoresce without any further manipulation.

Roger Tsien, a biochemist at University of California San Diego, engineered multiple variants of fluorescent proteins by performing structural changes in the chromophore. These structural changes resulted in different chromophores, which led to different excitation and emission maxima.[6,7] He also looked into other structural aspects of GFP like protein folding and pH impact. Shimomura, Chalfie and Tsien were awarded the 100th Nobel Chemistry Prize in 2008.

Fluorescent Proteins in Nature

______The first fluorescent proteins that were studied were the GFPs that came from *Aequorea victoria*; consequently, fluorescent proteins were only described from certain species of jellyfish. Over time, other organisms that contained fluorescent proteins were discovered. There was an emergence of fluorescent proteins that originated from coral reefs during the early 2000's. These proteins were of very high interest because from all of the GFPs and GFP variants that were studied or engineered, none of them were able to emit light at wavelengths longer than 529 nm.[8] So when the coral reef fluorescent proteins were discovered, it marked a significant expansion of emission wavelengths that could be used for biological imaging applications. One of the proteins that came from this group is drFP583, or DsRed, which has an emission maxima of 583 nm[9], has since become an experimental standard for marker proteins in cellular biology studies.[10]

Although fluorescent proteins are very widely used, the natural function/purpose in animals that bear fluorescent proteins are still debated. This is not the case for reef coral fluorescent proteins, which most likely serve the purpose of changing the light environment for the symbiotic algae. It was thought that since the signals these proteins emitted were so strong, they had to serve some role for the organism. Other hypotheses of natural function included that

(i) it could be an effective, metabolically inexpensive way to produce color patterns in lower depths of the sea, where most color wavelengths are not present besides blue light, or (ii) that the protein could serve a physiological function and that the fluorescence is an unrelated side product. Recent experiments done by Steven Haddock and Casey Dunn suggested that fluorescent proteins might have the natural function of being an attractant for their prey because of the high amount of contrast that the longer wavelengths of light would cause in the monochromatic environment of the sea.[11]

Uses of Fluorescent Proteins

Fluorescent probes (small molecules) and fluorescent proteins have been used in a multitude of scientific and medical research fields. They are being found, developed, and used as *in vivo* sensors for many different types of target molecules and ions (ex. Ca^{2+} and ethylene).[12,13] For example, the Dodani group at the University of Texas at Dallas characterized a fluorescent protein found in the jellyfish species *Phialidium* called phiYFP which can serve as a turn-on yellow fluorescent protein sensor for chloride, which they hope will be a useful tool for imaging chloride dynamics in the cell.[14] A lab at the Goethe University of Frankfurt has developed a superfolder variant of the GFP variant pHluorin, a fluorescent protein that is one of the easiest and most convenient tools to use to measure intracellular pH.[15] The original pHluorin has been used in many studies with varying organisms to measure pH, but it did not have a high fluorescence intensity or pH sensitivity to perform *in vivo* pH measurements of the endoplasmic reticulum in the yeast species *Saccharomyces cerevisiae*. This was likely because of protein misfolding due to the environment of the organelle. The superfolder variant caused the emission intensity to increase significantly at a wavelength of 508 nm across all pH values tested, which led to the conclusion that this protein could be useful to study pH changes under certain growth conditions and mutant strains.

3D Structure of Fluorescent Proteins

The crystal structure of GFP was simultaneously solved by the Tsien/Remington and Philips research groups in the 1990's.[16,17] The Phillips group solved it with wild type GFP and the Tsien/Remington group solved it with an enhanced GFP S65T variant. Although fluorescent proteins come from many different species, they all have the same basic structure consisting of an 11-stranded β-barrel, which is unique to fluorescent proteins, with an α-helix running through the axis of the barrel.[18] All fluorescent proteins are around 30 Å in diameter and 40 Å in height.[19] The C and N termini of GFP are on the same side of the β-barrel and are relatively close together. Lids composed of short helices are on each end of the barrel to protect the chromophore, located in the middle of the α-helix, from quenching by bulk solvent (Fig. 2).

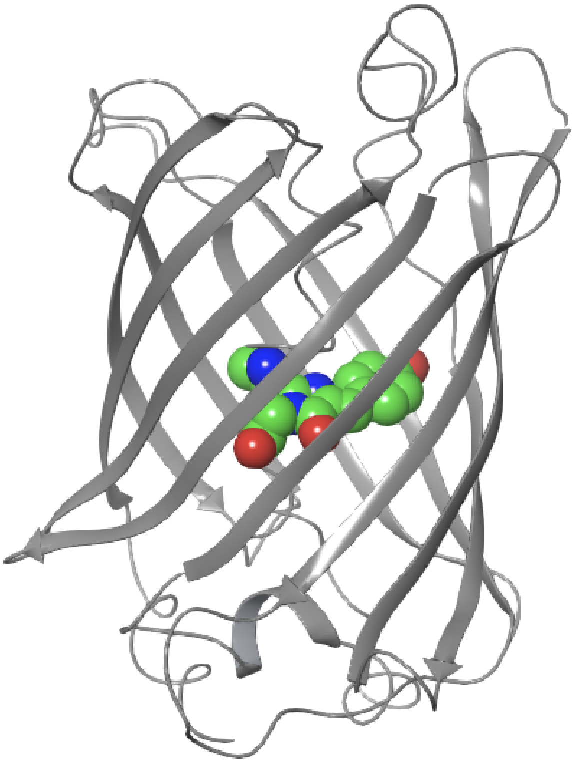


Figure 2. Crystal structure of GFP (PDB: 1EMB). The chromophore is shown in CPK representation.

Conserved Residues

There are over 200 marine organisms that contain fluorescent proteins. Some amino acids are highly conserved across all fluorescent protein structures (Fig. 3). Many of these conserved residues are located at the ends of the β- barrel, specifically in the β- turns and the lids. These residues include the 89th, 91st, 196th, 20th, 23rd, 27th, 53rd, 55th, 101st,102nd, 104th, 127th, 130th, 134th, and 136th residue positions (based on avGFP sequence). Some of these listed residues are glycines, which makes sense for effective flexibility, but there are also larger residues like phenylalanines at the 27th, 55th, and 130th residue positions. Previous proposals by the Zimmer group suggest the conservation of these lid residues were due to a potentially unknown protein-protein binding function.[18]

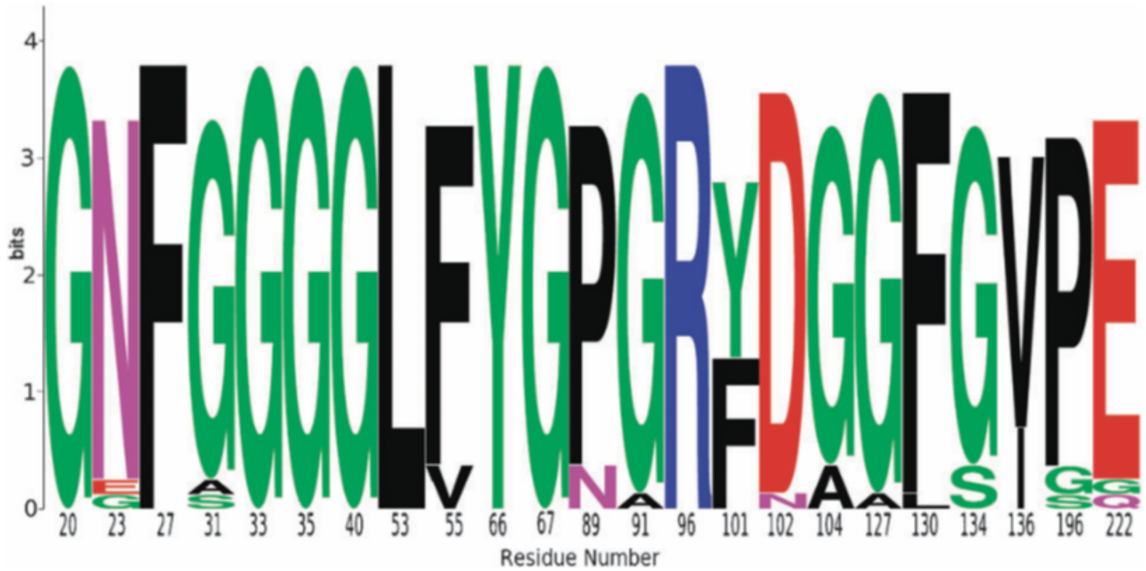


Figure 3. Sequence Logo (SeqLogo) of the most conserved residues in wild-type GFP structures.[18] The size of the letter represents the frequency of residues found in the listed residue positions (according to avGFP sequence numbering).

One study published in 2016 that examined the local fitness landscape of avGFP by investigating the effect single/multiple mutations have on fluorescence, found that there was a narrow fitness peak for the fluorescent protein. About 75% of the mutations had a negative effect on fluorescence, but while most resulted in a small decrease in fluorescence; only about one tenth of the single mutations resulted in a large decrease in fluorescence. Genotypes containing multiple mutations were more likely to have weak fluorescence or no fluorescence.[20] The mutations that had the most effect on its fluorescence were usually located near sites coding for residues in close proximity to the chromophore.

There are some residues that are conserved in the central portion of all fluorescent proteins. They are responsible for the chromophore formation. These include the 66th, 67th, 96th, and 222nd residue positions. The conservation of the 66th residue is interesting because in all wild-type fluorescent protein structures, a tyrosine is present in this position, but other aromatic amino acids can take the same position with successful chromophore formation, but the protein will emit a different color.[19] For example, a F66[18] or W66[19] mutant will produce a cyan fluorescent protein while a H66 mutation will form a blue fluorescent protein. Chromophore formation still occurs with G66, L66, and S66 mutants, but they result in non-fluorescent proteins because of the lack of the aromatic group.

The three glycines at the 31st, 33rd, and 35th residue positions are of interest because they are the only highly conserved residues located in the β-strands of GFP that are not involved in chromophore formation. The Matzke research group in Taiwan were studying GFP loss of function mutations and found that a G35S mutation did result in weak fluorescence and very weak protein accumulation.[21] G31D and G33D mutants were also expressed and both resulted in no fluorescence nor protein accumulation, suggesting that these residues may serve a role in facilitating GFP folding and stability. In regards to this study, this would suggest that mutations to the glycines of interest would be likely to result in either a non-fluorescent or low fluorescent protein with low to no protein accumulation and that the mutations do not affect chromophore formation.

Chromophore Formation

The formation of GFP and GFP-like chromophores is a spontaneous process which is a result of protein folding. The folding of the protein causes the amide nitrogen of the 67th residue, which is always a glycine, to come in close proximity to perform a nucleophilic attack on the carbonyl carbon of the 65th residue, forming a five membered imidazolone ring. This is followed by dehydration of the carbonyl oxygen of the 65th residue and oxidation of the α and β-carbon bonds of the 66th residue, resulting in a conjugated ring system (Fig. 4, left).[22,23] The absolute conservation of the 67th residue is important to chromophore formation because only glycine has the flexibility to form the kinked α-helix conformation needed for the nucleophilic attack to occur. Any other amino acid in place of the 67th position would result in impairment of chromophore synthesis.[19]

The presence of the 96th (always an arginine) and 222nd (typically a glutamic acid) residue serve a catalytic role in the formation of the GFP chromophore. R96 plays the role of the catalytic acid and E222 is the catalytic base (Fig 4, right).

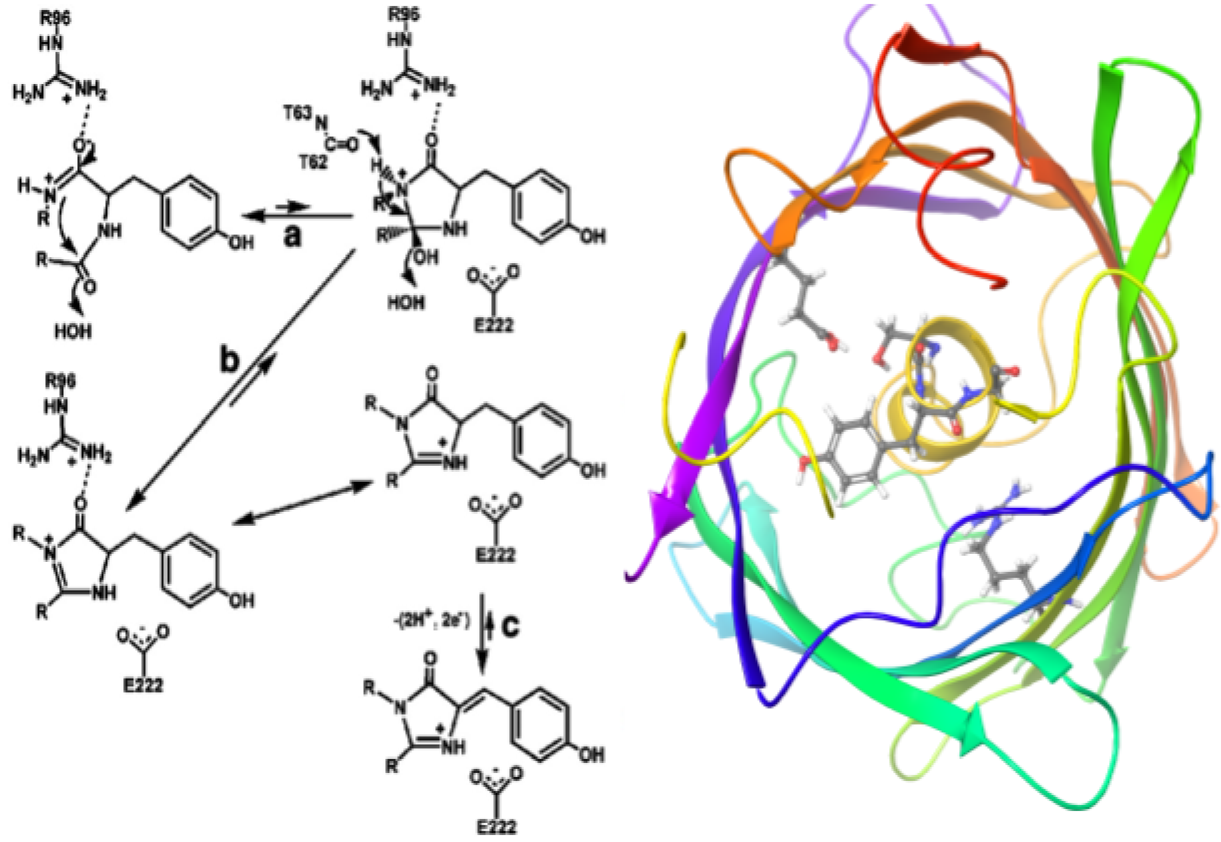


Figure 4. (Left) GFP chromophore formation mechanism. (Right) Spatial orientation of chromophore forming residues within the M96R PDB2AWJ β-barrel. E222 (top left of barrel), the precyclized chromophore tripeptide (middle), and R96 (bottom right) are shown in a ball-and-stick representation. Parts of the α-helix and lids were removed for visual clarity.

The Barondeau group found that a R96M mutation resulted in successful chromophore formation, however chromophore maturation took a significantly longer time due to the lack of the positive charge character. The positive charge of the arginine side chain points into the β-barrel at the carbonyl oxygen of the 66th residue position, pushing the 65th and 67th residues closer together for cyclization to occur at a much higher rate.[24]

Fluorescent Protein Folding

The crystal jellyfish typically live at temperatures lower than room temperature, which likely explains why GFP never evolved to fold efficiently at higher temperatures.[25] Since there are plenty of experimental organisms that have internal environments warmer than 25°C, it is desirable to have fluorescent proteins that can fold properly and efficiently at higher temperatures. This prompted the production of multiple GFP mutants, an example being a F99S/M153T/V163A triple mutant that increased the amount of fluorescent proteins that matured properly at 37°C and its ability to diffuse within the cell. However, it did not increase the brightness of the protein compared to GFP matured at optimal conditions.

The way in which fluorescent proteins fold is critical to fluorescence. Folding of GFP occurs by a set of disordered transition states, where the native state can be reached and fluorescence is returned for a short amount of time, or the protein goes into a state of slow equilibration.[26] A study examining fluorescent protein folding through experimental and multicanonical molecular dynamics simulations proposed a folding model that contains multiple pathways; including multiple kinetic and equilibrium intermediates that can act as potential energy traps, preventing the protein from adopting a native state (Fig. 5).[27] In most of these intermediates, the N-terminus β-strands, β(1-3), stay intact.

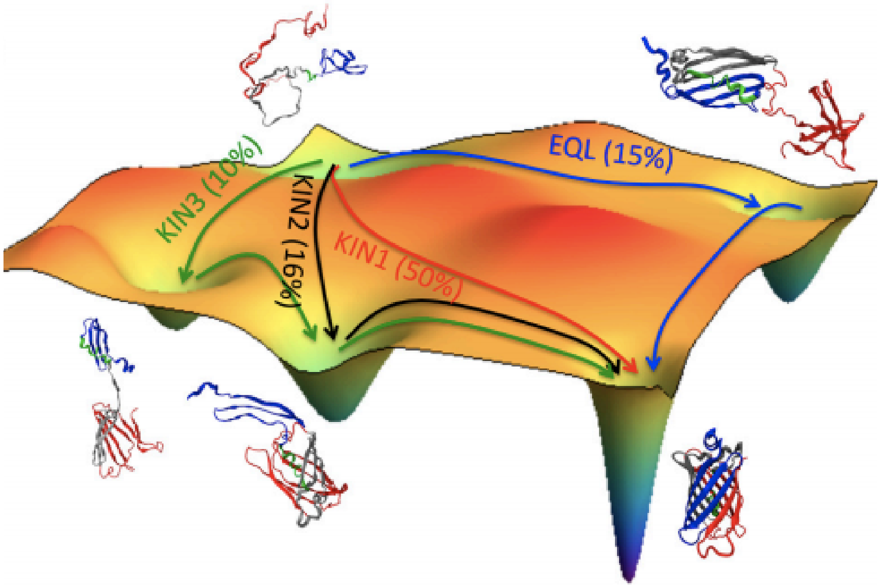


Figure 5. Folding landscape model of GFP.[27]

Fluorescent proteins can adopt many different partially folded states, so understanding how GFP unfolds is of equal importance to its folding. GFP unfolding starts with disruption of chromophore fluorescence. This is followed with the unfolding of β11 then unfolding of β(7-11). During the unfolding and refolding process, intermediates do not fluoresce, which can be attributed to internal rearrangements when the protein is at its native end-to-end length. Nitrogen NMR experiments have been used to examine the dynamics of GFP on a ps to ns timescale and shown the backbones for most of the GFP β-barrel are rigid. Conformational dynamics studies showed that the 7th, 8th, and 10th β-strands have higher degrees of flexibility than the rest of the protein, which is in agreement with molecular dynamics simulations on the protein.[28]

Glycine in β-Sheets

Glycine is the most simple amino acid due to its side chain being a single hydrogen. This allows glycine to possess a much higher degree of flexibility than any other amino acid. Statistical analyses showed the distribution of amino acids around different sections of proteins are not equal. The first experiments that examined β-sheet stability involved single point mutations in a peptide or small protein that were exposed to solvent with each of the naturally-occurring amino acids to observe the effects on the β-sheet. It was found that aromatic and β-branched residues stabilize the β-sheet the most while glycine greatly destabilizes the β-sheet.[29,30] However, even with the high destabilizing effect of glycines to β-sheet structure, they still do occur in β-sheets.

Surveys of these structures show that glycines found in β-sheets tend to be found in a cross-stranded pair with an aromatic residue. In antiparallel strands, the backbones of the two residues will have much more direct hydrogen bonding. This will cause the aromatic residue to form a positive gauche rotamer which will lead to the protection of the backbone, increasing its stability (Fig. 6).[30]

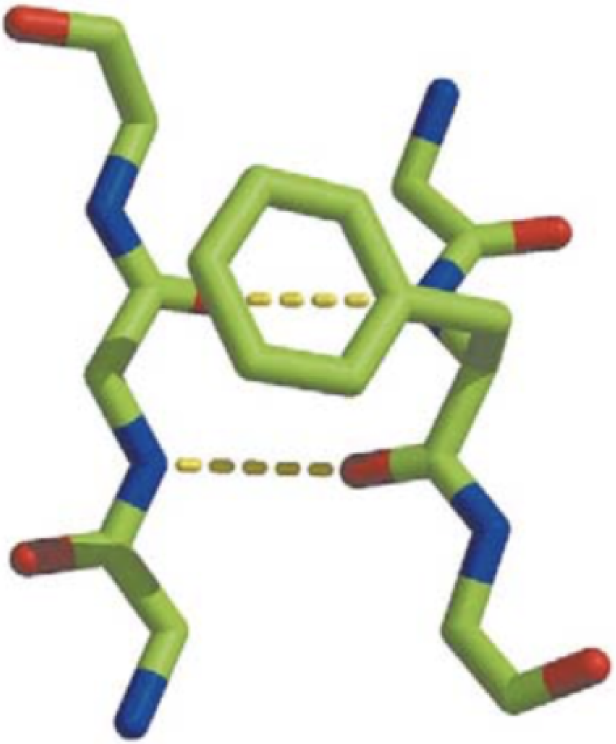


Figure 6. Antiparallel β-sheet backbone hydrogen bonding of cross-stranded residues 82 and 94 in PDB1PLC.[30] Note the aromatic of the phenylalanine bending to cover the space made by the lack of the side chain in the glycine.

Glycine in Fluorescent Proteins

Upon inspection of the amino acid sequence in wild type GFP (PDB1EMB), it can be seen that glycines mostly appear in areas where the protein bends; likely because of its flexibility, it allows the protein to make tighter turns without increasing stress on the system (Fig. 7). But, only few glycines are found within the β-sheets around the protein, and those residues are usually located near the end of the β-sheet. The only glycines that are seen in the middle of the β-sheet are the GXGXG residues of the 2nd β-sheet that are under investigation in this study.

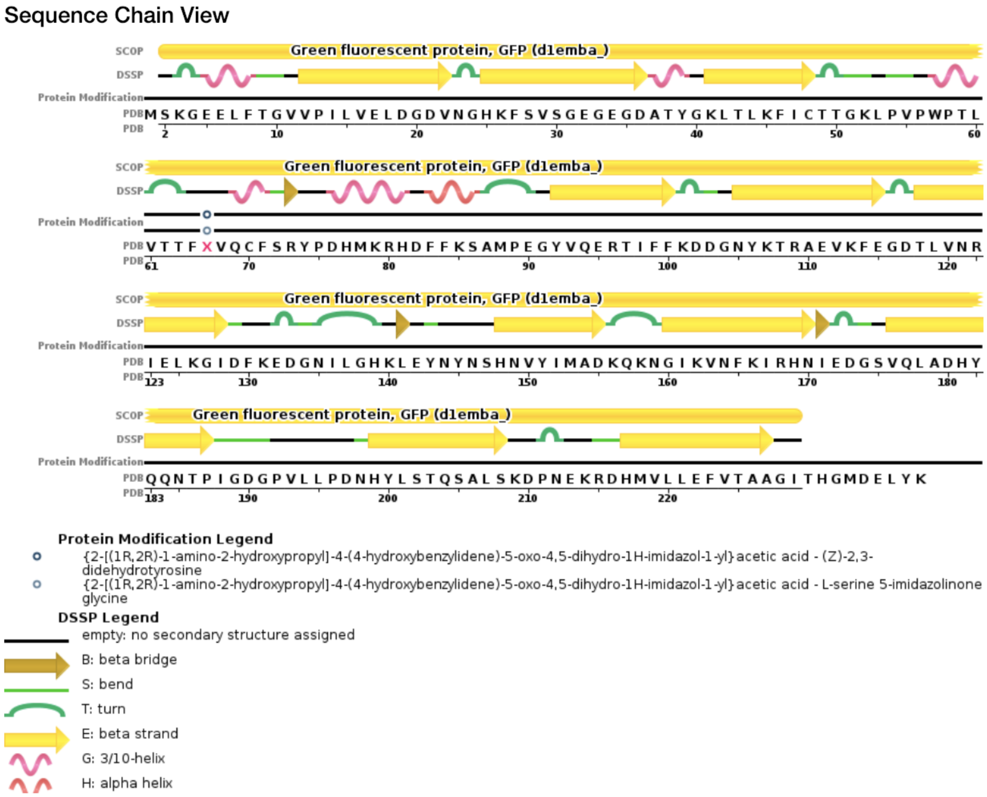

Sequence Chain View
SCOP
DSSP
Protein Modification
Green fluorescent protein, GFP (d1emba_)
PDB M S K G E E L F T G V V P I L V E L D G D V N G H K F S V S G E G E G D A T Y G K L T L K F I C T T G K L P V P W P T L
PDB 2 10 20 30 40 50 60
PDB V T T F X V Q C F S R Y P D H M K R H D F F K S A M P E G Y V Q E R T I F F K D D G N Y K T R A E V K F E G D T L V N R
PDB 61 70 80 90 100 110 120
PDB I E L K G I D F K E D G N I L G H K L E Y N Y N S H N V Y I M A D K Q K N G I K V N F K I R H N I E D G S V Q L A D H Y
PDB 123 130 140 150 160 170 180
PDB Q Q N T P I G D G P V L L P D N H Y L S T Q S A L S K D P N E K R D H M V L L E F V T A A G I T H G M D E L Y K
PDB 183 190 200 210 220
Protein Modification Legend
{2-[(1R,2R)-1-amino-2-hydroxypropyl]-4-(4-hydroxybenzylidene)-5-oxo-4,5-dihydro-1H-imidazol-1-yl}acetic acid - (Z)-2,3-didehydrotyrosine
{2-[(1R,2R)-1-amino-2-hydroxypropyl]-4-(4-hydroxybenzylidene)-5-oxo-4,5-dihydro-1H-imidazol-1-yl}acetic acid - L-serine 5-imidazolinone glycine
DSSP Legend
empty: no secondary structure assigned
B: beta bridge
S: bend
T: turn
E: beta strand
G: 3/10-helix
H: alpha helix

Figure 7. Amino acid sequence and residue location in PDB1EMB, courtesy of the RSCB

In "De Novo design of fluorescence activating β-barrel", the difficulties and methods on how to design β-barrels from scratch were discussed.[31] This involved taking backbones that had the most interstrand hydrogen bonding, connecting them with short loops and optimizing them to get the lowest energy sequences. The designed sequences were expressed in *E. coli* and it was found that almost all of the designs were insoluble or oligomeric. Upon examination of this result, it was found that significant amounts of backbone hydrogen bond interactions at the ends of the barrels were distorted or broken. To combat this issue, different methods were examined to make a uniform β-barrel backbone that did not have any loop structures and valines at every residue position as a placeholder. Upon relaxation of the structure with heavy hydrogen bond constraints, two different structural strains were observed: steric strain and residues adopting unfavourable twists due to their chirality.[31]

In order to reduce these strains, some of the residues were replaced with glycines due to their smaller size and achirality, which allows twists that would be unfavourable for other residues, while maintaining the hydrogen bond pattern of the β-sheet. However, with the relieved strains on the structure, introduction of the glycines formed irregular torsions in the sheets.[31]

This Work

In this study, computational simulations and analysis were used in order to look at the effects of mutations to G31, G33, and G35 on the GFP structure. These three glycines are of interest since they are highly conserved residues in the β-strands of GFP and are not directly involved in chromophore formation. They are also interesting because glycines in β-sheets are rare, even more so three glycines in a GXGXG sequence. This study was done by taking the crystal structure of precyclized GFP intermediate (PDB2AWJ) and making mutations to its amino acid sequence so that it would be the same as wild-type GFP (PDB1EMB). After these mutations were made, the previously described single point mutations were made and then the resulting structures were put through molecular dynamics simulations under standard conditions until it reached a stable state; this was done with all three described mutations along with a baseline reference structure that did not have any mutations of the 31st, 33rd, and 35th residue positions.

The data from the molecular dynamics simulations were then analyzed by looking at a variety of geometric parameters associated with autocatalytic chromophore formation and β-barrel formation.

EXPERIMENTAL METHODS

1. *Software*

The starting crystal structures were retrieved from the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB).[32] Molecular visualizations were done using Maestro[33] and pyMOL[34]. Molecular Mechanics (MM) and conformational search calculations were done through MacroModel.[35] Molecular Dynamics (MD) simulations and simulation analyses were done through Desmond[36] on 32-core processor computers. All MM and MD calculations were done using the OPLS3 force field.[37]

2. *Generation of Immature wtGFP and G3XA Mutant Structures*

Residues in the PDB2AWJ structure were mutated in order to have a wtGFP (PDB1EMB) residue sequence. Once the mutations to the intermediate were graphically made, a 25,000 step Large Scale Low Mode conformational search[35] was done to find the lowest energy structure. A 10,000 step energy minimization was performed on the resulting structure. To validate the mutated 2AWJ structure, the protein was overlapped with the chromophore forming residues of PDB2AWJ and the matured chromophores and the 96th residue of both PDB2AWK and PDB1EMB.

To superimpose the structures, all four of the structures were uploaded to Maestro. All three PDB structures (1EMB, 2AWJ, and 2AWK) were completely aligned using the quick align function.[33] After aligning the structures, only residues 64-67 and 96, or if the chromophore is matured, CRO66 and residue 96, were displayed. With the superposition function, all heteroatoms of Y66 and G67, the carbonyl carbon of the 65th residue, and the amino acid backbone of the 96th residue were superimposed.[33] For superimpositions of the engineered wtGFP structure and the two structures that contain mature chromophores, atoms that form the five-membered imidazole ring come from both the 65th and 67th residue positions, so those atoms had to be superpositioned appropriately.

The G3XA variants were generated by taking the engineered immature wtGFP structure described above and computationally performing the desired single point mutations, then performing a second round of conformational searches (25,000 step LMCS) and minimizations in order to obtain the lowest energy conformation to be used for molecular dynamics simulations.

3. wtGFP and G3XA Variant Simulations

Desmond[38] is integrated with the Schrödinger molecular modeling programs, allowing for the use of functions found in MacroModel to prepare the protein structures to be chemically correct prior to running a molecular dynamics simulation.[38] Model systems for all MD simulations in this study were made with the SPC solvation model and orthorhombic box shape of size 10 Å x 10 Å x 10 Å. For MD simulations, all simulations were performed with the NPT ensemble class, at a temperature of 300 K, and a pressure of 1.01325 bar. For the engineered wtGFP, G33A, and G35A structures, 100 ns MD simulations were performed. A 200 ns molecular dynamics simulation was performed on the mutated G31A mutant.

4. Hydrophobic Pocket Simulations

Four GFP variants (G35/F71L, G35V/F71, G35/F71Y, G35V/F71L) were engineered following the same procedure as the G3XA mutants. Introduction of the mutations typically resulted in atomic overlaps across residues, but this was mediated using the Desmond Minimization function once the water box was assembled. All 4 of the mutants were subjected to 200 ns MD simulations under the previously described conditions. Another 200ns wtGFP simulation was also conducted to serve as the baseline measurement for later analyses and comparison.

RESULTS & DISCUSSION

1. *Engineered wtGFP Structure Validation*

PDB2AWJ was used as a starting structure due to the fact that it is a GFP variant with an R96M mutation and an immature chromophore. Although arginine and methionine are similar in size, the methionine side chain lacks the positive charge character that R96 has, which plays an important catalytic role in chromophore cyclization. This results in a significantly reduced chromophore formation rate, forming a mature chromophore and fluorescing after three months, instead of a few hours for wtGFP.[24] The crystal structure of the R96M mutant with a mature chromophore can be seen in the PDB2AWK structure.

The slowed maturation kinetics of the PDB2AWJ structure allowed for crystal structure determination of a precyclized GFP intermediate.[24] Therefore, graphically mutating the sequence of PDB2AWJ to the wtGFP sequence results in a realistic estimation of an immature wtGFP structure that could be used as a benchmark for comparison with other immature GFP variants.

For the superimposition of the engineered wtGFP and the original 2AWJ structure, the Root-Mean Squared Deviation (RMSD) value was 0.8625 Å with the maximum difference being 1.7347 Å between the carbonyl oxygens of G67 (Fig. 8A). The superimposition of the engineered wtGFP with 2AWK had an RMSD of 0.5911 Å with the largest difference being 1.5486 Å between the carbons in the tyrosine ring that are next to the carbon with the hydroxyl group (Fig. 8B). The overlap of 2AWJ with 1EMB gave an RMSD value of 0.6729 Å with the largest difference being 1.4808 Å between the carbonyl carbon of S65 (which in 1EMB is the carbon between the two nitrogens of the five membered ring in its chromophore structure) (Fig. 8C).

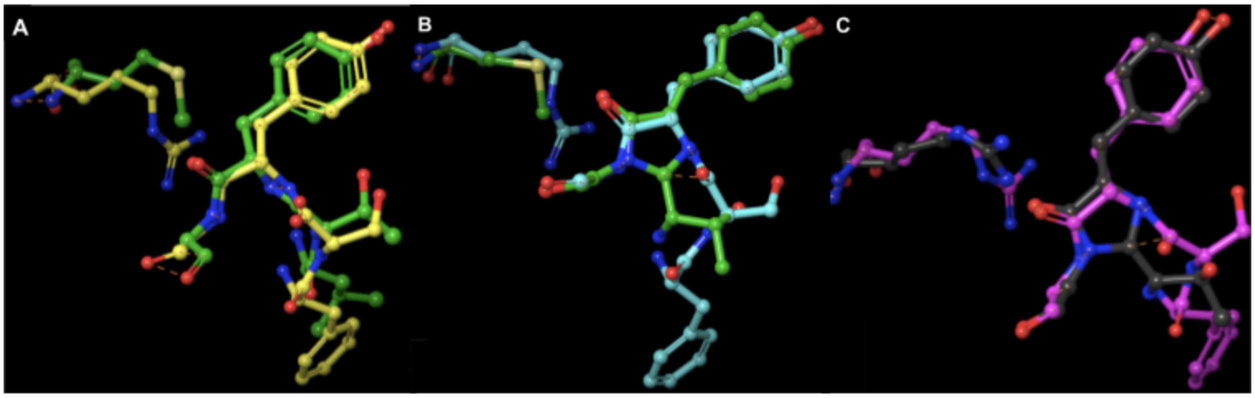


Figure 8. (A) Superimposition of a 25000 step large scale low mode (LSLM) conformational search on a mutated precyclized GFP intermediate 2AWJ structure (yellow) and the default 2AWJ structure. Note the L64F, T65S, and M96R mutations made on the mutated 2AWJ structure. (B) Superimposition of the mutated 2AWJ structure (light blue) with the default 2AWK structure (green) containing the mature chromophore. (C) Superimposition of the mutated 2AWJ structure (fuschia) and 1EMB (graphite).

Due to the fairly low deviation between the engineered precyclized wtGFP intermediate and the actual R96M intermediate, the matured R96M structure, and mature wtGFP structure, it was determined that the engineered structure would serve as a valid baseline for subsequent comparison amongst the G3XA variants. This structure was assumed to undergo chromophore formation kinetics like wtGFP (avGFP).

2. *Structural Comparison to validate the MD Simulations*

RMSD measurements were calculated by comparing the starting structure to all subsequent structures to see whether the protein had reached an equilibrium throughout the simulation. After the first 100 ns MD simulation on the G31A mutated structure, the RMSD consistently increased throughout the simulation, reaching a peak of about 2.25 Å (Fig. 9A). This suggests that over this time period, the protein was undergoing large conformational changes. Since the protein would not be in a stable conformation over this time period, the last frame of the simulation was taken and used to start another 50ns MD simulation, extending it to a total simulation time of 150 ns. For the first 25ns of this simulation, the RMSD stayed fairly consistent, fluctuating between 1.0 Å and 1.2 Å. From 25 to about 37 ns, the RMSD values consistently rises, then decreases until the 40 ns mark, where the RMSD value then began to steadily fluctuate between 1.4 Å and 1.6 Å (Fig. 9B). The last frame of the first 50 ns simulation was taken so that another 50 ns simulation would be performed (totaling 200 ns). The first 20 ns of this simulation were fairly consistent, staying at an RSMD value of around 1.0 Å. From the 20 ns to the 40ns mark, the RMSD value rises and then starts to fluctuate between 1.3 Å and 1.6 Å for the last 10ns of the simulation (Fig. 9C). From the data given in the RMSD calculations, the last 100 ns of the simulation were used to perform structural analyses.

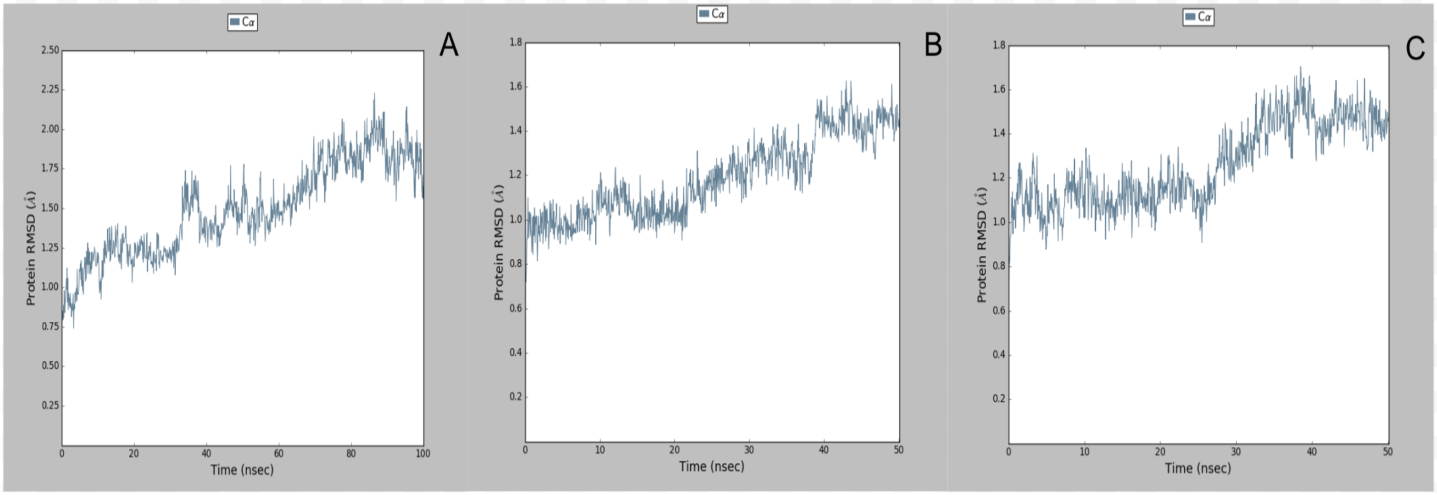


Figure 9. RMSD graphs of (A) the first 100 ns of MD simulation on the G31A mutated structure. (B) 50 ns MD simulation (starting with the last frame of the first 100 ns MD simulation). (C) Second 50 ns MD simulation (starting from the last frame of the first 50 ns MD simulation).

The G33A simulation quickly reached equilibrium by the 5 ns mark of the simulation, with the RMSD value consistently fluctuating between 1.0 and 1.4 Å for the rest of the simulation time (Fig. 10).

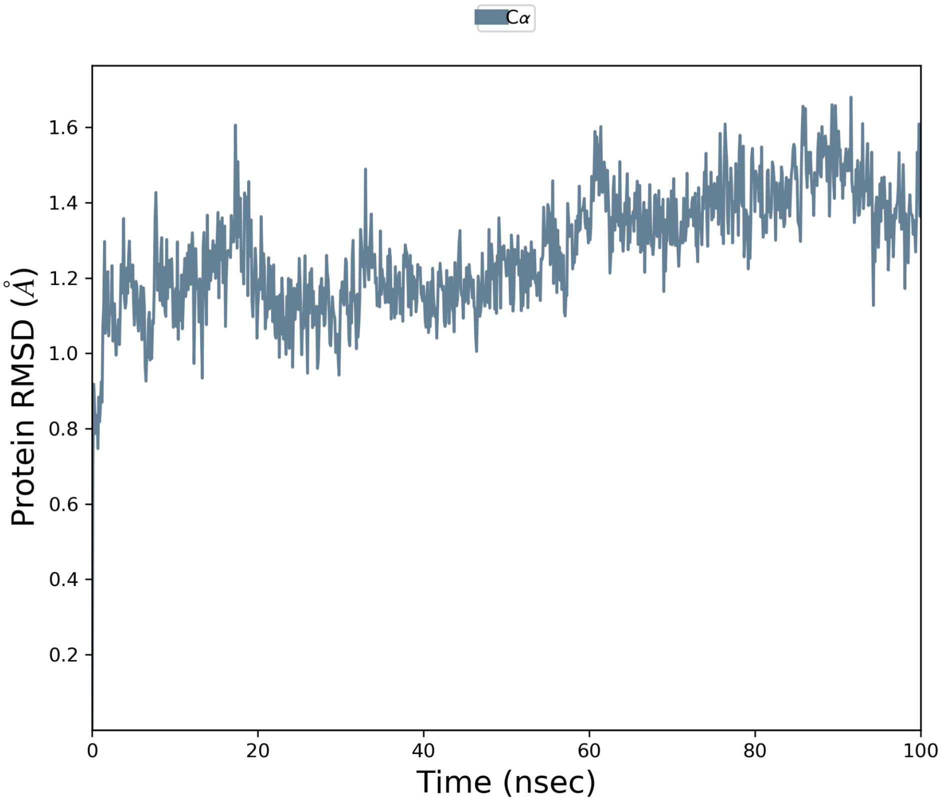


Figure 10. RMSD graph of the 100 ns G33A MD simulation.

The G35A simulation began significant structural changes over the first 15 ns of the simulation; equilibrium was reached after 20 ns, where the RMSD mostly fluctuates between 1.2 and 1.4 Å for the rest of the simulation (Fig. 11)

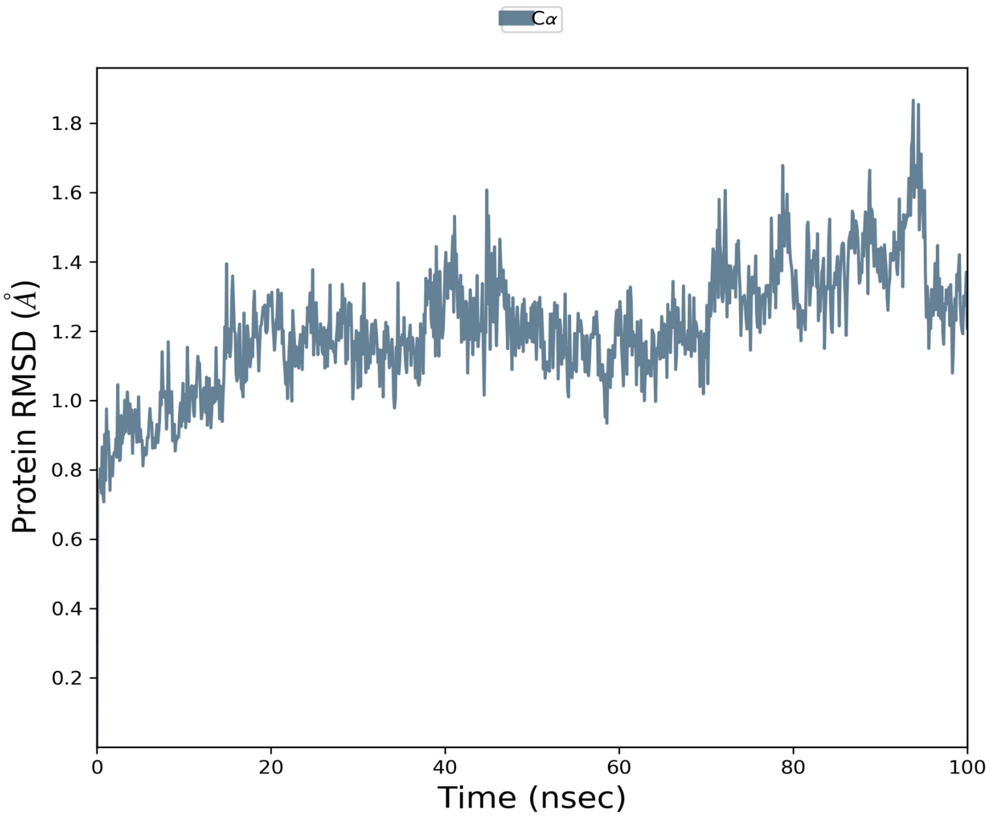
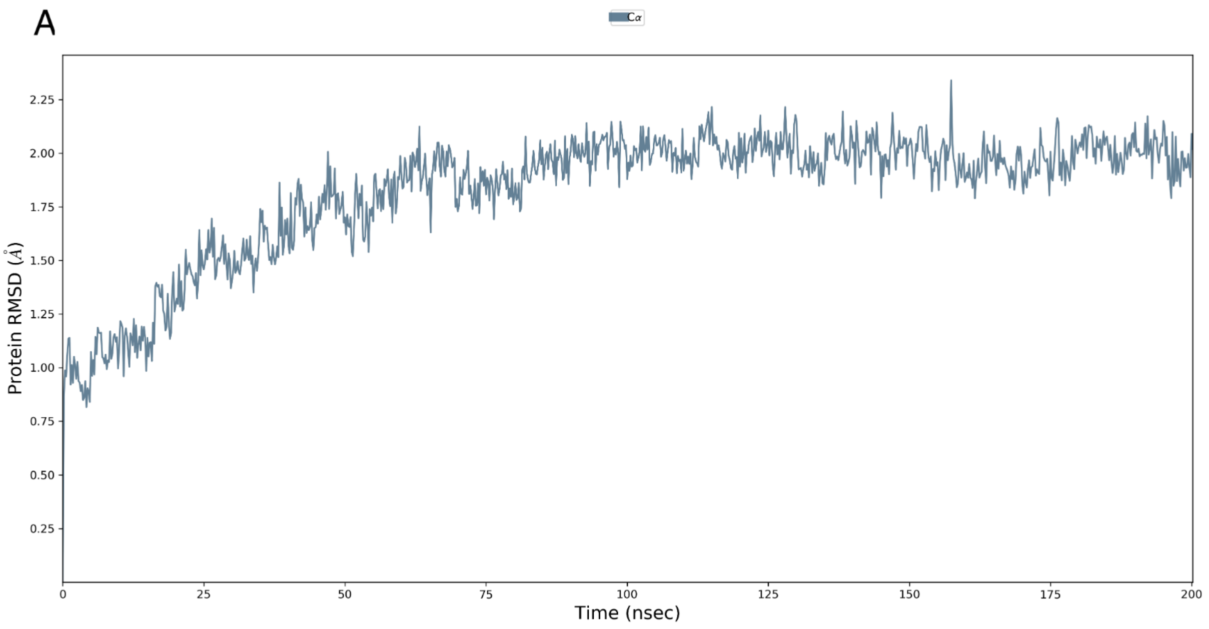
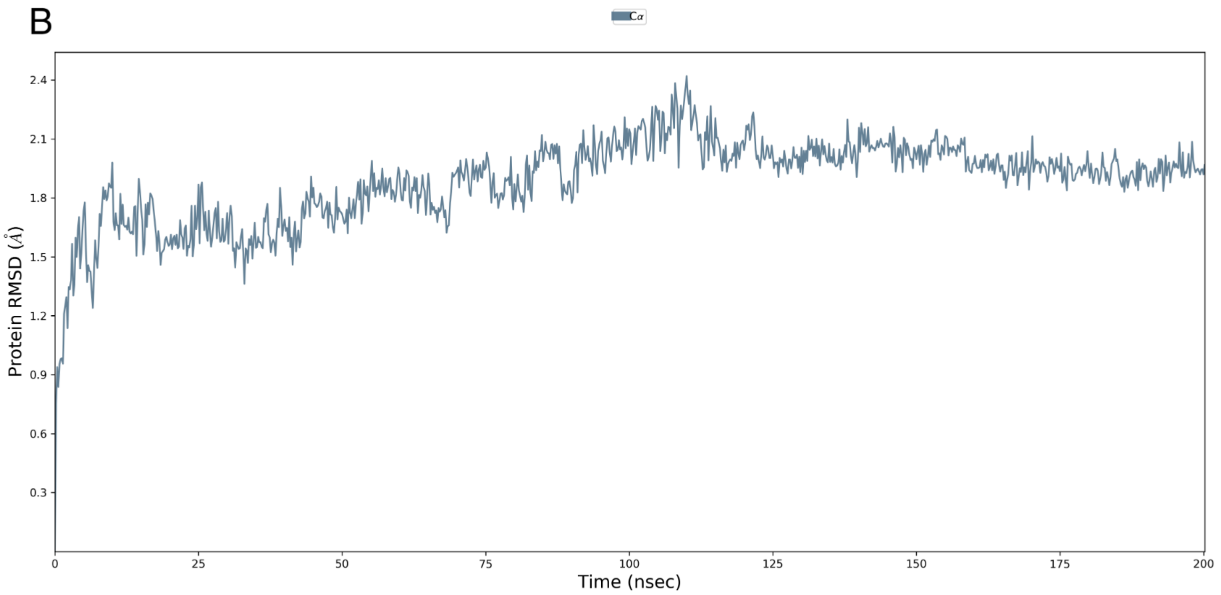


Figure 11. RMSD graph of the 100 ns G35A MD simulation

In the hydrophobic pocket mutant calculations, all four simulations reached an equilibrium within the simulation time. The G35/F71L simulation began with a significant amounts of structural change, noted by the consistently increasing RMSD until equilibrium is reached at the 38 ns mark, where the RMSD consistently fluctuates between 1.75 and 2.0 Å (Fig. 12A). The G35V/F71 simulation initially starts with a rapid increase in RMSD, but after about 12 ns, structural equilibrium is reached with fluctuation between 1.7 and 1.95 Å (Fig. 12B). In both the G35/F71Y (Fig. 12C) and G35V/F71L (Fig. 12D) simulations, the simulation began with a large structural change, but then equilibrium is quickly reached after about 5 ns, where the RMSD values fluctuate between 1.0-2.0 Å and 1.25-1.75 Å, respectively.


A
Cα
Protein RMSD (Å)
2.25
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0
25
50
75
100
125
150
175
200
Time (nsec)


B
Cα
Protein RMSD (Å)
2.4
2.1
1.8
1.5
1.2
0.9
0.6
0.3
0
25
50
75
100
125
150
175
200
Time (nsec)

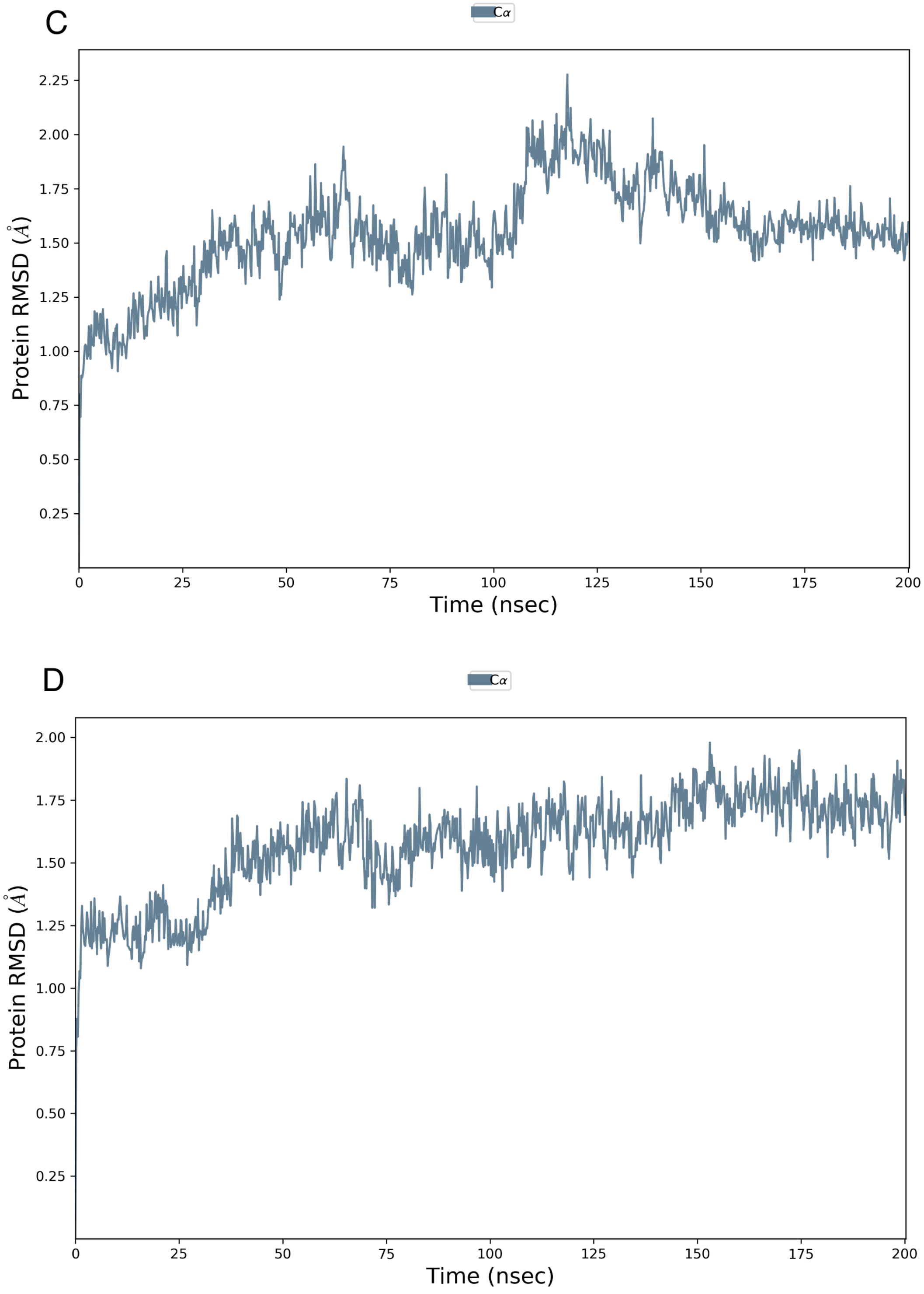


Figure 12. RMSD graphs of (A) G35/F71L 200 ns MD simulation. (B) G35V/F71 200 ns MD simulation. (C) G35/F71Y 200 ns MD simulation. (D) G35V/F71L 200 ns MD simulation.

3. *Simulation Measurements*

The simulation event analysis panel was used to measure and compare the distances of residues that have a role in chromophore formation, hydrogen bonding interactions in the central α-helix, cross-stranded aromatic residue rotations over the GXGXG motif of the second β-sheet, and water migration through the □-barrel of the engineered wtGFP intermediate and the G31A mutant.

3.1. Tight Turn Distance

The first step of chromophore formation requires a nucleophilic attack on the carbonyl carbon of the 65th residue position by the amide nitrogen of G67 (Fig. 13) . For this attack to be more likely to occur, the two atoms must be in close proximity to each other.[39]

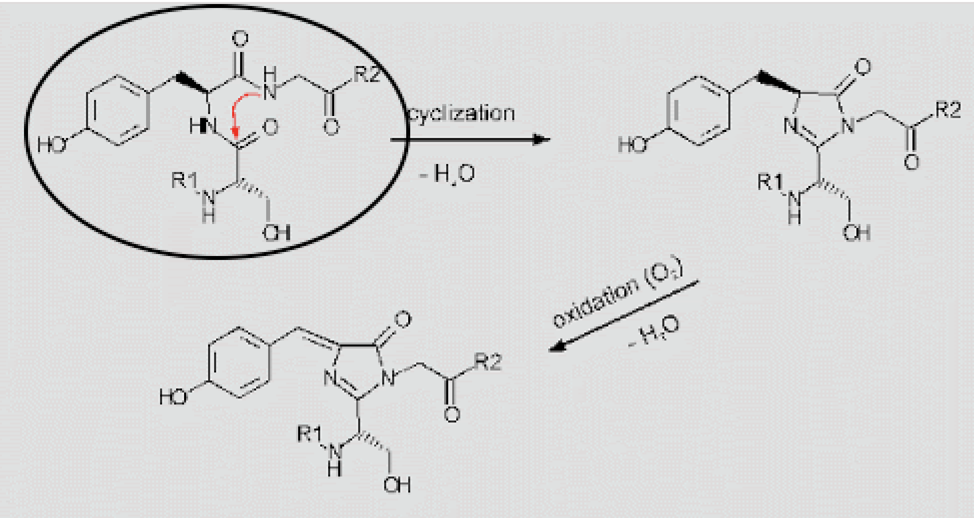


Figure 13. Chromophore formation mechanism. The precyclized chromophore structure with the tight turn attack is circled on the top left.

Measurements of the distance between these two atoms were taken over the course of the MD simulations (Fig. 14).

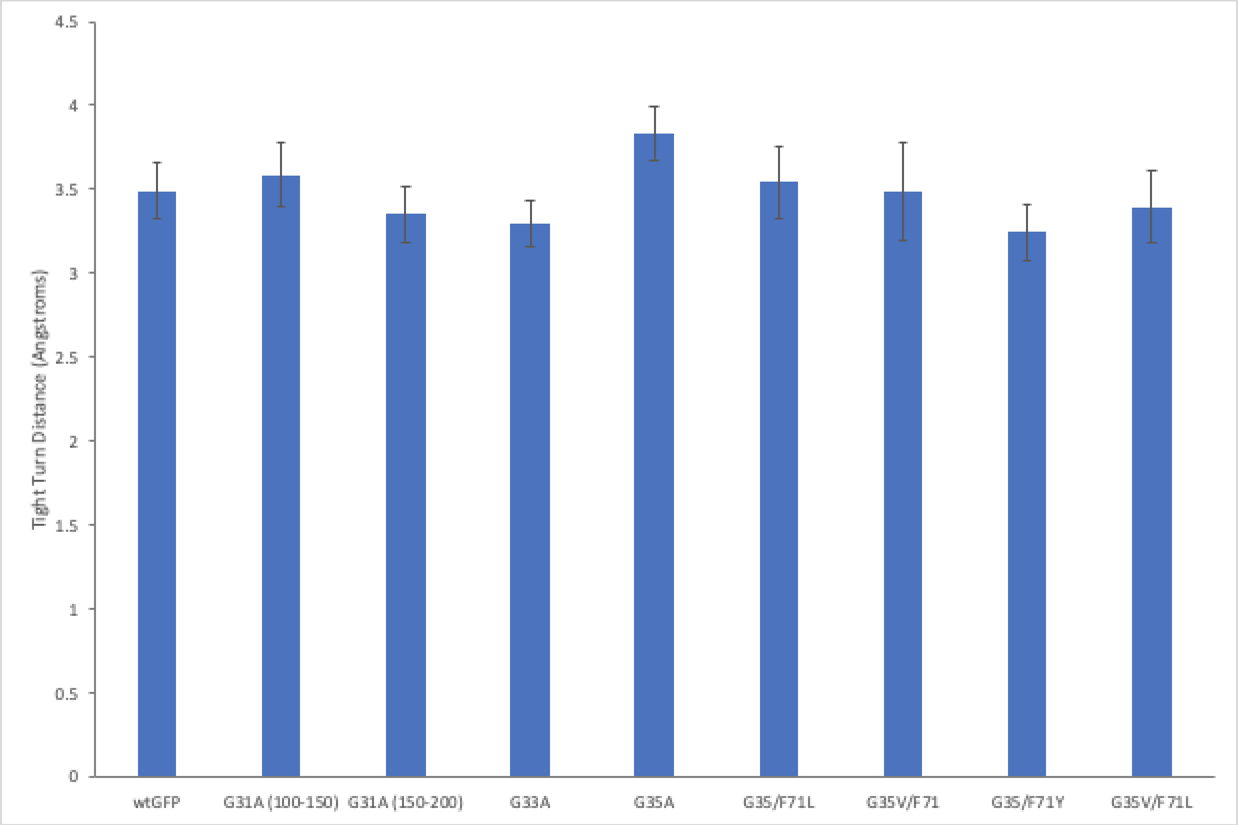


Figure 14. Graphical representation of the tight turn distance measured over the course of the simulation for the engineered immature benchmark wtGFP and all mutant MD simulations.

It was observed that the tight turn distance of each mutant was not significantly different from that of the wtGFP simulation, suggesting that the mutations at these positions do not have a significant structural effect on the tight turn.

3.2. α-Helical Interactions

Distortion of the α-helix is required to adopt the tight turn conformation when forming the chromophore. This results in all fluorescent proteins having a noncanonical α-helix because an α-helix with typical *i* and *i+4* hydrogen bonding amongst the main chain would be so stable, that it would be too energetically costly to break the intra-strand hydrogen bonding interactions involving residues 65,66, and 67 that adopt the tight turn conformation before cyclizing (Fig.

15).[22] Having a noncanonical α-helix results in a higher energy structure, lowering the activation energy required for cyclization to occur.

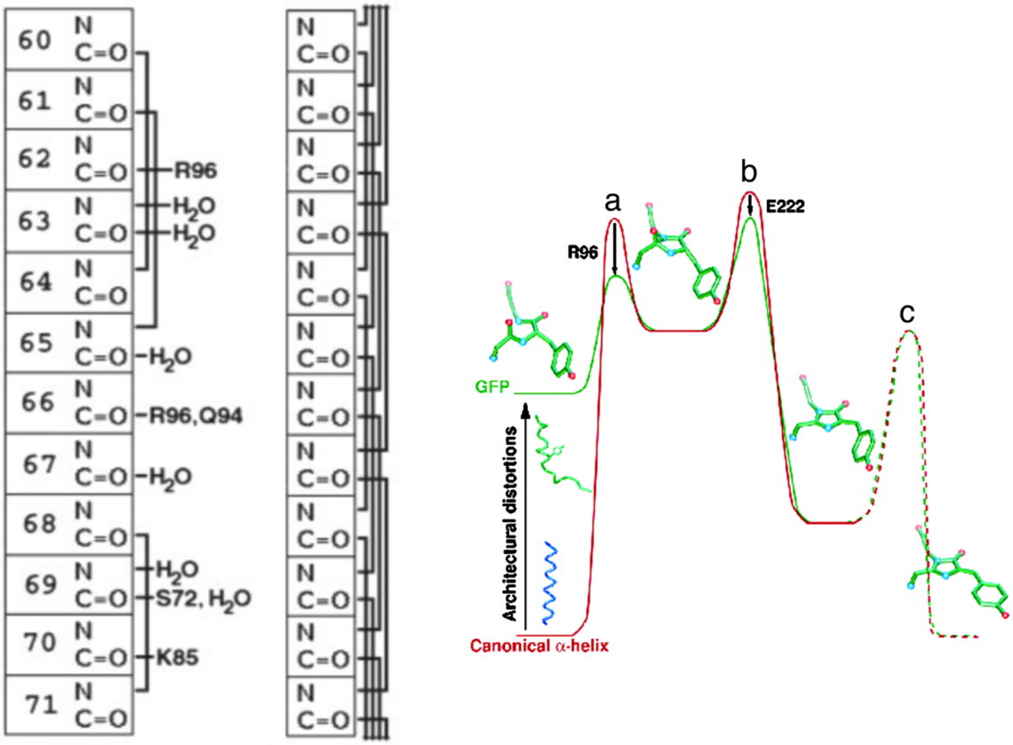


Figure 15. (Left) Main chain hydrogen bond interactions of the non-canonical α-helix in GFP compared to hydrogen bonding interactions of a canonical α-helix.[22] Note that the GFP chromophore is formed from residues 65-67. (Right) Energy profiles of chromophore formation with a canonical α-helix (red) and for the non-canonical helix of GFP (green).[22]

Measurements of main chain hydrogen bonds amongst residues in the central α-helix were measured over the course of the simulation to see if any significant changes in hydrogen bond interactions were observed upon making the glycine to alanine mutations. Table 1 below shows a summary of the hydrogen bonding analyses.

Table 1. Main chain interactions in the α-Helix for all mutant simulations

R_1-R_2	60-64	61-65	62-96	66-96	66-94	68-71	69-72	70-85
wtGFPimm								
avg distance (Å)	4.739	3.546	5.411 (N^+) 4.232 (N^0)	8.872 (N^+) 7.519 (N^0)	4.036	3.283	4.318	4.047
std deviation	0.327	0.303	0.373 (N^+) 0.362 (N^0)	0.555 (N^+) 0.531 (N^0)	1.089	0.291	0.331	0.447
R_1-R_2	60-64	61-65	62-96	66-96	66-94	68-71	69-72	70-85
G31A (100-150 ns)								
avg distance (Å)	3.638	2.911	4.062 (N^+) 5.145 (N^0)	3.103 (N^+) 4.935 (N^0)	2.895	3.082	3.347	3.563
std deviation	0.265	0.159	0.495 (N^+) 0.483 (N^0)	0.326 (N^+) 0.472 (N^0)	0.229	0.222	0.327	0.710
G31A (150-200 ns)								
avg distance (Å)	4.034	2.936	4.686 (N^+) 4.898 (N^0)	3.010 (N^+) 4.419 (N^0)	3.053	3.090	3.366	5.407
std deviation	0.374	0.162	0.668 (N^+) 0.442 (N^0)	0.287 (N^+) 0.531 (N^0)	0.628	0.224	0.367	0.339
G33A								
avg distance (Å)	4.143	3.897	4.085 (N^+) 3.054 (N^0)	7.898 (N^+) 5.668 (N^0)	6.222	3.400	3.212	3.300
std deviation	0.362	0.311	0.398 (N^+) 0.321 (N^0)	0.340 (N^+) 0.363 (N^0)	0.721	0.287	0.229	0.341
G35A								
avg distance (Å)	4.179	3.000	4.259 (N^+) 5.779 (N^0)	6.697 (N^+) 4.933(N^0)	4.713	3.358	4.218	5.238
std deviation	0.315	0.213	0.398 (N^+) 0.456 (N^0)	0.507 (N^+) 0.438 (N^0)	1.047	0.307	1.180	1.098
G35V								

avg distance (Å)	3.358	3.149	3.935(N^+)
			4.677(N^0)
				3.028(N^+)
				4.365(N^0)
					4.340	5.563	5.970	5.770

std deviation	0.360	0.294	0.467 (N^+) 0.458(N^0)	0.407(N^+) 0.574 (N^0)	1.356	0.779	2.276	2.614
G35V/F71L								

avg distance (Å)	4.150	3.285	5.072(N^1)
			3.833(N^0)
				8.133(N^+)
				7.444(N^0)
					3.620	3.293	3.578	4.317

std deviation	0.291	0.266	0.590(N^1)
			0.404(N^0)
				0.621(N^+)
				0.614(N^0)
					1.016	0.319	0.856	0.586

F71L								

avg distance (Å)	4.691	3.814	6.275(N^1)
			6.858(N^0)
				3.544(N^+)
				5.606(N^0)
					3.544	3.149	3.026	3.346

std deviation	1.019	0.744	1.298(N^1)
			1.307(N^0)
				0.741(N^+)
				0.816(N^0)
					0.741	0.364	0.201	0.766

F71Y								

avg distance (Å)	4.001	3.015	3.825(N^1)
			4.599(N^0)
				3.313(N^+)
				4.940 (N^0)
					2.864	3.224	3.666	3.956

std deviation	1.049	0.765	0.492(N^1)
			0.385(N^0)
				0.4917(N^+)
				0.633 (N^0)
					0.191	0.339	0.863	1.007

As with the tight turn distance measurements, some main chain interactions tended to fluctuate more heavily than others, but no significant structural changes between the wild type and the mutants were observed, suggesting that the mutations do not affect the non-canonical α-helical hydrogen bonding pattern, which is required for chromophore formation.

3.3. Aromatic - Glycine Cross Strand Interactions

Measurements were made to examine the distance of the aromatic group of F71 to the α-carbon of the 35th residue in each of the G31A, G33A, and G35A single-point mutant structures (Fig. 16, Table 2). This measurement was examined due to the tendency of aromatic residues being able to provide a stabilizing effect for glycine residues located in β-sheets by performing rotations of the side chains, protecting the backbone from bulk solvent.[30] The idea was that due to the flexibility of the GFP α-helix, it might be possible for the aromatic group of F71 to get in closer proximity to the backbone of the 35th residue and stabilize the backbone not through a cross stranded β-sheet interaction, but through an α-helical/β-sheet interaction.

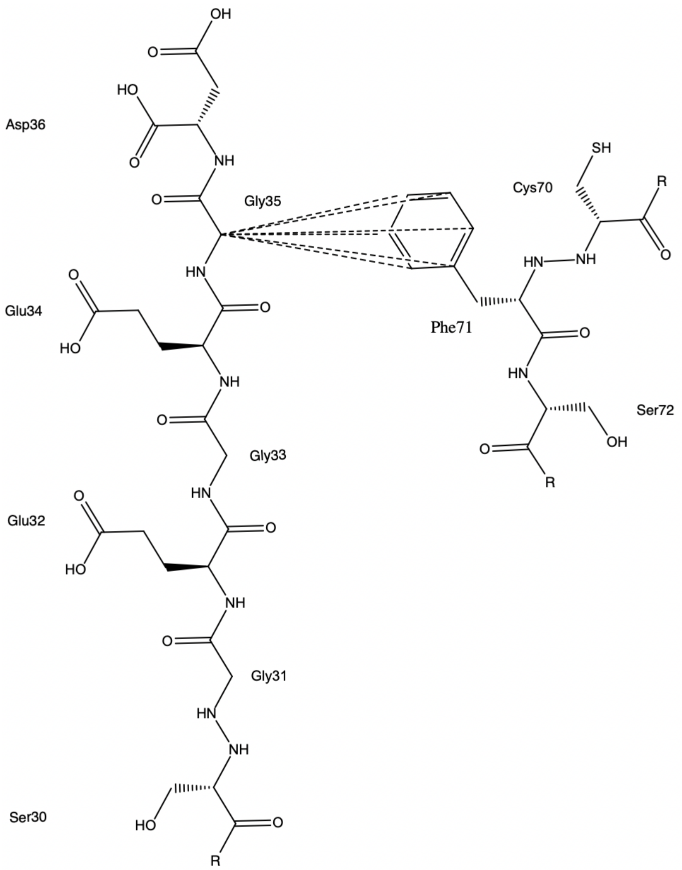


Figure 16. 2-D Representation of GXGXG motif in the second β- sheet (left) of GFP with F71 of the α-helix (right) to visualize hydrogen bond measurements over the course of the simulation. The hydrogen bond distances that were measured are represented as dashed lines.

Table 2. Distance of F71 Phenyl Ring to G35 α-Carbon for G3XA Simulations

	Avg. F71- C_αG35 Distance (Å)	**Std. Deviation**
wtGFPimm	4.4796	0.4912
G31A (100-150)	5.0152	0.7044
G31A (150-200)	5.1193	0.7864
G33A	4.4332	0.4039
G35A	5.6834	0.4112

This increased distance between the residues could cause formation of water channels directly toward the chromophore forming residues, which would cause significant decreases in maturation kinetics.

3.4. Water Migration through GFP β-Barrel

Prior research has shown that water migration through the β-barrel is an important aspect in chromophore formation of fluorescent proteins due to the presence of highly conserved water molecules in the crystal structure of GFP and fluorescent proteins. These waters contribute to the intricate hydrogen bonding network around the chromophore and may play other roles during chromophore formation. These waters interact with R96 and E222, GFP's highly conserved catalytic residues (Fig. 17). Water migration is also important for fluorescent proteins to control the internal environment of the protein, especially because of the very different environments the chromophore is in when it is precyclized opposed to when it is matured.

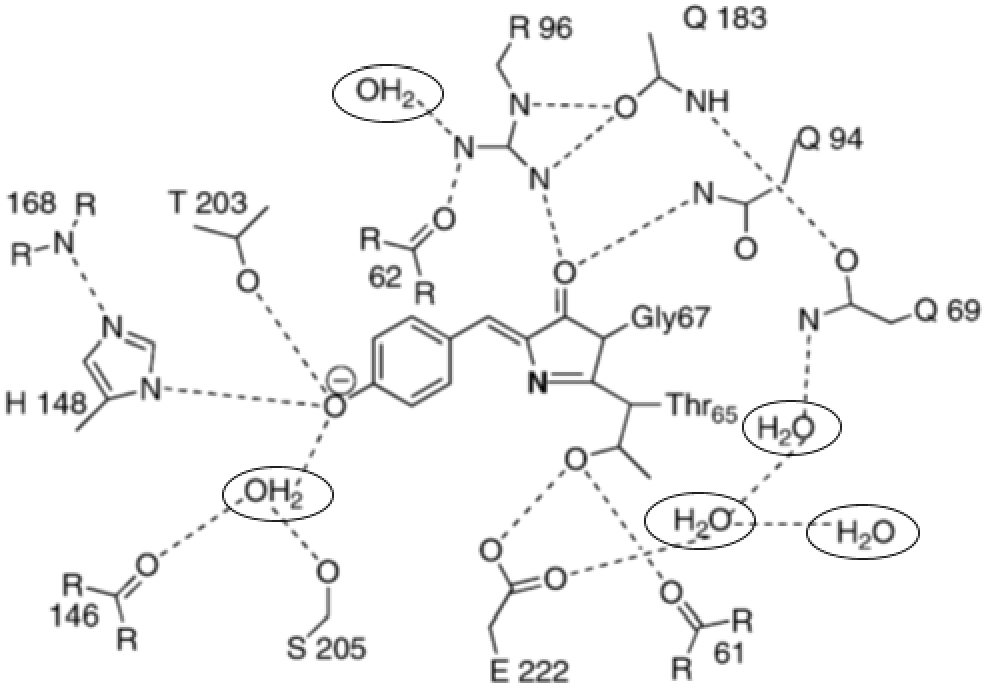


Figure 17. Schematic representation of the hydrogen bond network around the GFP chromophore. Hydrogen bonds that include water molecules are circled.[4]

A previous study by the Zimmer group proposed that the structure of TurboGFP contained a water filled pore leading from the exterior of the structure to the chromophore that increased chromophore formation speeds.[40] This channel was lined by the 136th, 137th, 156th, 197th, and 198th residues. The study investigated what effect water diffusion through the β-barrel of GFP had on chromophore formation kinetics. This was done by performing 50 ns molecular dynamics simulations of avGFP, TurboGFP, their precyclized intermediates, and a V197L TurboGFP mutant; the idea being that the larger leucine residue would increase the steric bulk around the water channel, reducing the amount of water and oxygen migrating into the β-barrel.

As expected, the mutated TurboGFP showed much slower maturation kinetics. It also showed that water diffusion through the barrel is much more common in precyclized structures compared to that of the GFP structures with a mature chromophore.

In this study, water migration was recorded through all of the simulations to see where and when water had either migrated in or out of the β-barrel of the mutant structures and compared this with the engineered wtGFP. This was done by first determining which water molecules were in the barrel in the first frame of each simulation. After determining which waters were in the barrel, the simulation was then examined ten frames after the initial frame; if the waters escaped the barrel by this frame, they were deemed to be outside of the barrel to begin with. The accepted water molecules were then put into CPK representation and examined throughout the whole simulation to see how the water moves around the β-barrel and see where in the structure they escape from.

For the reverse process, the β-barrel was examined to see which waters were present at the end of the simulation. These were then put into CPK representation to see where and when these water molecules had migrated into the β-barrel. Tables 3-11 shows all the analyses of water migration through the performed MD simulations.

Table 3. Water Migration Analysis of the 100 ns wtGFPimm MD simulation

Atom #	SPC #	Description
75152₸	23841	Begins close to the nitrogen of the backbone in I188 and G189. Leaves between frames 192-193, between the bend after the 9th β-sheet and the turn in the α-helix. Between G191 and A87.
75182	23851	Begins near N146 (looks like H-Bonding between the nitrogens of its backbone and side chain and H169 of the 8th β-sheet). Leaves from the space between the helix before the 7th β-sheet and 8th β-sheet closer to the side with N/C termini. Left between the 506th and 507th frames.
75104₸	23825	Near S72 (bend of helix).
75221₸	23864	Begins closest to S65 **(Getzoff)**. For the first roughly 30 frames, the water molecule is moving but after that, it stays close to S202 for the rest of the simulation.
75956	24109	Begins closest to H148 of the 7th β-sheet and T203 (H-Bonding preventing it from leaving, one of the other waters had to find a different route out of the barrel). By 150 frames, the water moves around the α-helix, passing the glycines of interest, and then starts hydrogen bonding with Q183 and is in close proximity for the rest of the simulation.
75569	23980	Starts closest to another water and along with Q69 **(Getzoff)** near the bottom of α-helix (near N/C termini). Doesn't stay in any one position for too long, but it stays near the region where the chromophore forms.
75905₸	23822	Begins near T63 **(Getzoff)** in the α-helix. Stays in close proximity T63 and T108 of the 5th β-sheet.
75503₸	23958	Begins close to Q69 and G67. After about 4 ns, the water molecule moves closer to V163 and Q183 for about 50 ns. Then moves between Q183 and T62 and stays there for about 10 ns then moves back to the previous position (V163 and Q183) for the rest of the simulation.
75680₸	24017	Starts closest to T63 **(Getzoff)**. Stays in very close proximity to or in the same position relative to T63 for the duration of the simulation.
75098₸	23823	Begins closest to side chain oxygen of T62 and amide nitrogen of the same residue. Stays close to the residue for the duration of the simulation. This water is likely H-bonding between the side chains of R96 and T62 for the whole simulation
75275	23882	Starts closest to side chain nitrogen of R96, the carbonyl oxygens of Q94, and the hydroxyl group of T108. Over the first 15 ns, this water moves from its position in the beginning of the simulation towards T63. Stays between 94 and 63 for about 20 ns. After this it starts to move closer to the N/C Termini end of the protein by the end of simulation.

75143	23838	Starts closest to K131 and L137 (opposite the N/C termini). The water stays in the same position the whole simulation (most likely because of its H-bonding to the backbone so there's less flexibility opposed to H-bonding to a side chain.

Table 4. Waters Present at Frame 1 of the G31A MD simulation (100-150ns)

Atom #	SPC #	Description
4842	404	This water molecule started off neither completely in or out of the barrel (near N146 and the 10th β-strand. It then moves closer towards the center of the barrel and stays close to T62 for the rest of the simulation.
23449	8273	Starts between the 9th and 10th β-strands, leaves the β-barrel at frame 13 between S208 and M218.
18201	4857	Located close to the middle of the lid, leaves the β-barrel at frame 31.
4419	263	Located near the previously described water, but it was closer to S72, leaves the β-barrel at frame 42 of the simulation.
4470	280	Located between the 1st and 2nd β-strand near G10 and A37, leaves at frame 18.

Table 5. Waters Present at Frame 1001 of the G31A MD simulation (100-150ns)

Atom #	SPC #	Description
5958‡	776	This water entered the β-barrel between F100 and N135 at frame 984.
5307	559	Located in the middle of the lid between residues L137, K131, and D103. Entered at frame 3.
32634	9668	Entered through between the 10th and 11th β-strands (near H169 and Y145) at frame 664. Located near T62 at the end of the simulation.
10944	2438	Entered through the same gap that was previously described at frame 969. Moved close to the α-helix and was near N149 and V150 by the end of the simulation.
29901‡	8757	It entered through the gap that is located between E5 and Q80 of the lid at frame 981.
17775	4715	Entered the same gap as the previously described water at frame 995, it is near A37 at the end of the simulation.
13461‡	3277	Entered the same gap as the previously described water at frame 999.

27462[₸]	7944	Entered in the between lid residues E5, L194, and D82 at frame 841.
25653	7341	Entered through the gap formed by A37 and R73 at frame 818, stayed in this position for 190 frames and then moved towards Q80.
16884	4418	Entered through the gap at G4 and S86 at frame 976, left and came back through the gap lined by E5,T38, and K79 at frame 999. Final position near Y75.
15468	3946	Entered through the top lid at frame 999.
18822[₸]	50664	Entered through the Q80 and E5 gap.
31326[₸]	9232	Entered through the lid in between D197 and D82 at frame 999.

Table 6. Waters Present at Frame 1 of the G31A MD simulation (150-200ns)

Atom #	SPC #	Description
11728	2702	Starts close to the N-terminus (near G10 and A37). G10 is on a turn while A37 is on a 3_{10} helix. Leaves from this space between frames 93 and 94 (around 4.60 ns into the simulation).
28909[₸]	8429	Starts between K131 and ASN 135 (on turn before 7th β-sheet). Definite H-bonding with the oxygen of the water and amino groups of each side chain and possible H-bonding with amide nitrogen of the N135 backbone.

Table 7. Waters Present at Frame 1001 of the G31A MD simulation (150-200ns)

Atom #	SPC #	Description
21277[₸]	5885	Entered at frame 804 of the simulation, positioned close to K131.
7003	1127	Entered between D102 and G134 at frame 986 of the simulation.
19771	5383	Entered through the gap near L53 on the lid of the β-barrel (opposite of the termini) at frame 53, between V55 and H217 by the end of the simulation.
8464	1614	Entered between N146 and S205 at frame 470. Close to the same position by the end of the simulation.
21265	5881	Entered between N146 and S205 at frame 480. Close to the same position by the end of the simulation.
18586	4998	Entered between N146 and S205 at frame 101. Close to the same position by the end of the simulation.

29761	8713	Entered between N146 and S205 at frame 20. Between Y66 and L44 by the end of the simulation.
17623	4667	Entered through a gap in the lid of the barrel (termini side) lined by D82, N198, and G228 at frame 355. Near Y66 by the end of the simulation.
22585	6321	Entered through a gap formed by V163 and I152 at frame 578, near H152 at the end of the simulation.
17176	4518	Entered between H181 and T38 (close to the 2nd β-sheet) at frame 679 , near C70 by the end of the simulation.
13981	3453	Entered between K85 and G4 at frame 928, near R73 at the end of the simulation.
15667[‡]	4015	Entered through G4 and S86 at frame 10. Minimal movement for the rest of the simulation.
24821	7073	Entered between E5 and K85 at frame 960. Near H81 by the end of the simulation.

Table 8. Waters Present at Frame 1 of the G33A MD simulation

Atom #	SPC #	Description
57679	15027	The starting position for this water was between Y143 and H169, within the first 20 frames of the simulations, it flowed out. It never re-entered the barrel.
57238	17880	The starting position for this water was between Y143 and H169, within the first 20 frames of the simulations, it flowed out. It never re-entered the barrel.
57451	17951	The starting position for this water was between Y143 and H169, within the first 20 frames of the simulations, it flowed out. It never re-entered the barrel.
57682	18028	The starting position for this water was between Y143 and H169, within the first 20 frames of the simulations, it flowed out. It never re-entered the barrel.
57532	17978	Positioned near D149 and S205 in the first frames, moving towards the α-helix as the simulation progressed. It stayed near the α-helix until frame 295, where it left through the gap between N146 and A206.
57151	17851	Positioned near the α-helix and β-strands, between L60 and H181. Roams around this area until frame 55, where it leaves through the gap formed by Y145 and N170.

57445*†	17949	Positioned near the α-helix between T62, T59 and I167. Then moves towards the V61 & L60 of the α-helix.
57625	18009	Positioned near the L201 and Y66, between the α-helix and β-strand. Moves closer to the α-helix near S65 and roams around the area between G67 and C70, normally staying closer to S65. It then slowly moves out of the protein through the gap between S147 and A206 in frame 181. It does not enter the protein after this.
57154*†	17852	Positioned between T62 and V61. The water stayed near the V61 of the α-helix.
57277	17893	Positioned near the S65 of the α-helix. It moved towards the Q69 and C70. It was later pushed off and it moved towards the strands for a bit after coming closer again to the α-helix.
57559	17987	Positioned near G67 of the α-helix. It moves closer to the α-helix, roaming the area of the chromophore forming residues. It started to head out, moving towards the lids of the barrel.
57736	18046	Positioned near the T108 and E124. Water moves towards the Y66 and roams around the chromophore forming residues.
57531*†	17911	Started near the R96, then moved a little towards the chromophore forming region towards the G67.
57160*†	17884	Positioned near the E5, L85, and C70.
57187*†	17863	Positioned near the F84, D197.

Table 9. Waters present at frame 1001 of the G33A MD simulation

Atom #	SPC #	Description
14062	3488	Ended near the opening of S205 and other waters, but the others did not completely enter the structure. Entered at frame 986 between the gap S197 and S205. Where it moved towards the α-helix.
26608	7670	Ended near α-helix close to T62. Entered through the gap formed by S147 and S205 at frame 977.
12805	3069	Ended near the α-helix by P58. Entered in the 266 frame between the gap friend by S205 and Y143, and it moved towards the L60, where it stayed near it for the rest of the simulation.
22537	6513	Ended near the lid residues by N144. Entered at frame 226 between the gap N144 and H169. It moved towards the general direction of E142, where it stayed at.

Table 10. Waters Present at Frame 1 of the G35A MD simulation

Atom #	SPC #	Description
55639	17318	• Begins nearest to Phe 8 on α-helix position of strand 1 • Leaves at frame 4 from termini capped end of barrel
55195	17170	• Begins closest to Phe 84 on the α-helix that runs through the center of barrel • Leaves at frame 403 (40.3ns) from the termini capped end of barrel
55222	17179	• Begins closest to Phe 83 on the α-helix that runs through the center of barrel • Leaves at frame 695 (69.5ns) from the termini capped end of barrel
55312[‡]	17209	• Begins closest to Phe 71 near α-helix that runs down the center of the barrel • Moves toward Gly 67 at frame 142 (14.2ns), but never leaves cavity during the simulation
55660[‡]	17325	• Begins closest to Phe 71 near α-helix that runs down the center of the barrel • Moves toward Gly 67 at frame 142 (14.2ns), but never leaves cavity during the simulation
55771[‡]	17362	• Begins closest to Gly 67 near α-helix that runs down the center of the barrel • Remains there over the course of the simulation, never leaving the barrel
55186[‡]	17167	• Begins closest Leu 60 near α-helix that runs down the center of the barrel • Moves closer to Ala 179 at frame 792 (79.2ns) and stays there inside barrel for the remainder of the simulation
55567	17294	• Begins closest to Ser 205 on Strand 9 • Leaves at frame 467 (46.7ns) through strands 9 and 10
55189	17168	• Begins closest Thr 62 near α-helix • Leaves through strands 7 and 8 at frame 8
55486	17267	• Begins closest to His 169 near strands 7 and 8 • Leaves through strands 7 and 8 at frame 8
55234[‡]	17183	• Begins closest to Asn 135 on α-helix on the end of the barrel without the termini • Remains there over the course of the simulation, never leaving the barrel
55273	17196	• Begins closest to Tyr 145 between strands 6 and 7 • Leaves through strands 6 and 7 at frame 2
55594[‡]	17303	• Begins simulation closest to Gln 69 on α-helix that runs down the center of the barrel • Remains there over the course of the simulation, never leaving the barrel
55246	17187	• Begins closest to Asn 170 on strand 6 on the end of the barrel without termini • Leaves at frame 229 (22.9ns) between strands 7 and 8

Table 11. Waters Present at Frame 1001 of the G35A MD simulation

Atom #	SPC #	Description
53521	16612	• At frame 1001, this water molecule is closest to K101 on the end of the barrel without termini • Water molecule enters GFP at frame 882 through the loops next to K101
22708	6341	• At final frame, water molecule is closest to Ile 171 between strands 6 and 7 • Enters barrel at frame 767 through the end of the barrel without termini
16060	4125	• At frame 1001, water molecule is closest to S147 between strands 5 and 6 • Enters cavity at frame 928 from end of GFP without termini through strands 5 and 6
37009	11108	• At final frame 1001, water molecule is closest to F84 on α-helix that runs through the center of barrel on the end without the termini • Enters cavity at frame 335 next to the on α-helix that runs through the center of barrel on the end with termini closest to R73
38710	11675	• At last frame, water molecule is closest to Y66 on the on α-helix that runs through the center of barrel, right in the center • Enters cavity at frame 237 through a gap between strands 6 and 7
19321	5212	• At final frame, water molecule is closest to D36 on the second strand • Enters GFP at frame 951 through the top of GFP with termini
32431	9582	• At final frame, water molecule is closest to Lys 85 on α-helix that runs through the center of barrel on the end with the termini • Enters barrel at frame 874 through the termini capped end of GFP
23893	6736	• At frame 1001, water molecule is closest to Ile 188 at the termini capped end of GFP • Enters GFP at frame 888 through termini capped end

₸ - Stays in the same position for the whole/rest of simulation.

It was observed that the β-barrel of the wtGFP simulation had less overall water migration than that of the G31A simulation. However, the wtGFP structure did have more water molecules near the chromophore region of the protein, making main chain interactions with residues that are known in the literature to hydrogen bond with waters. Although more water migration was observed in the G31A simulation, most of the waters were seen entering the protein near the lids of the protein rather than through the β-barrel near the chromophore, which would be expected due to the many polar residues exposed to the bulk solvent.

The G33A and G35A simulations also were observed to have more water migration into the β-barrel than the wtGFP simulation. The increased rate of water migration occurred in the same manner as the G31A mutant, with most of the waters migrating into the β-barrel near the lids of the protein. Unlike the G31A simulations, these structures were observed to have more of the conserved waters around the chromophore, suggesting that the G31A mutation clogs the space above the chromophore. This could prevent waters from flowing into the barrel to make some of the main chain interactions with the α-helix, along with the waters that are needed for the catalytic residues to form the chromophore.

3.5. MOLEonline water channel location

MOLEonline is a web-based interactive application used for calculating and characterizing channels within biomacromolecular structures.[41] This is done by (i) inputting a PDB or mmCIF file and (ii) implementation of the MOLE 2.0 algorithm. The algorithm undertakes a seven step process to perform the following actions:

- Computing the Delaunay Triangulation/Voronoi diagram
- Approximation of the molecular surface and identification of cavities
- Identification of possible start and end points of the channels
- Computing and filtering of channels.[42]

MOLEonline was used to qualitatively (Figs. 18,19) and quantitatively (Tables 12-16) compare the location and dimensions of possible water channels in the β-barrels of wtGFP and the three G3XA variants and to investigate if (i) the predicted channels are in the same location in which the water molecules migrate through the β-barrel in the MD simulations and (ii) if there is an association between the dimensions of the predicted channels and the number of water molecules moving in and out of the protein. For an even comparison, across each MD simulation, the 1st, 501st, and 1001st frames of each simulation had their protein structure extracted to a PDB file. These structures were then put through the MOLEonline server in order to gauge what channels were being formed or closed over the course of the simulation.

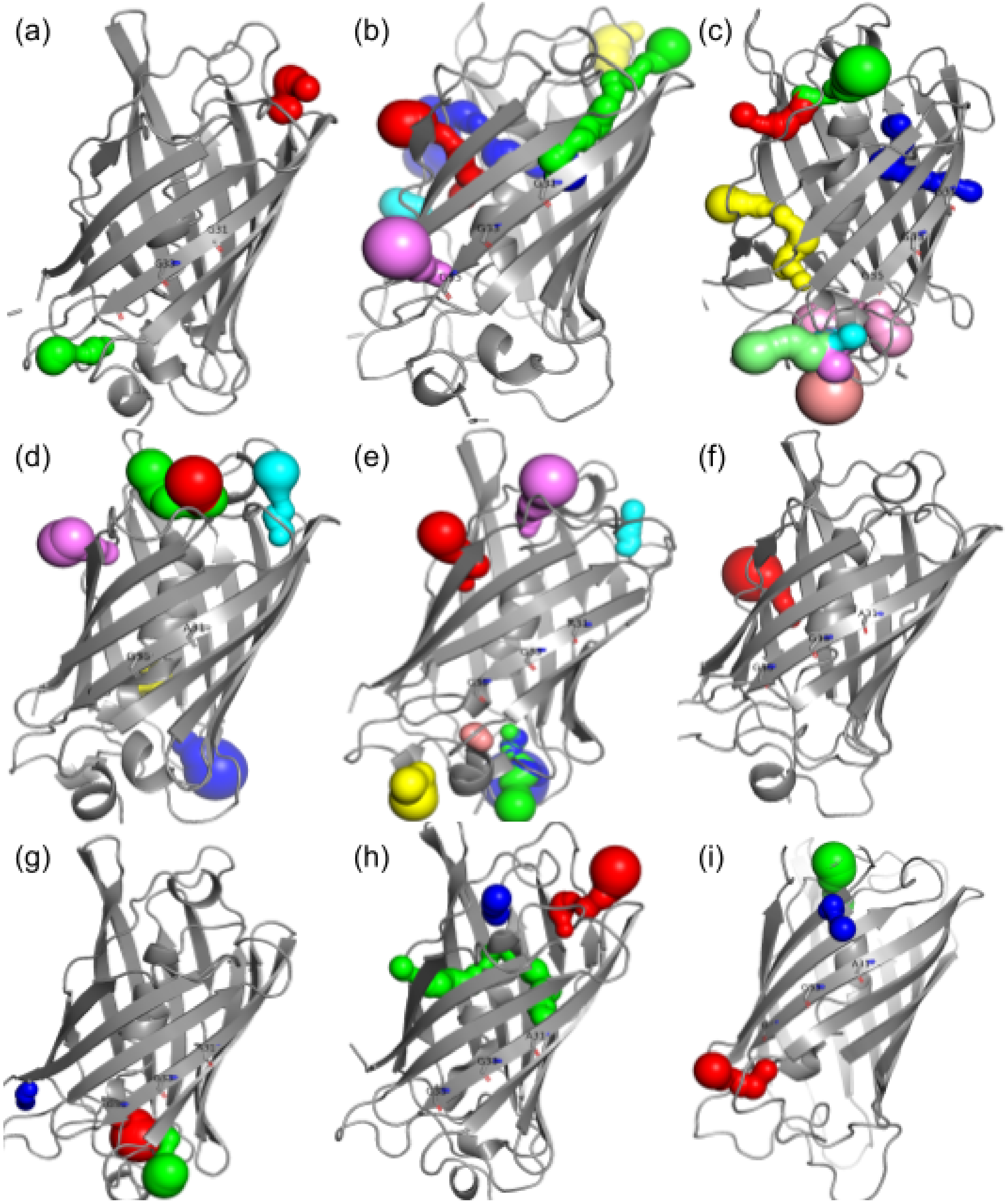


Figure 18. Visualization of predicted water channels in and through the β-barrels of wtGFPimm and the G31A variant MD simulations. (a-c) Scaled dimensions and locations for the 1st, 501st, and 1001st frames of the engineered wtGFP 100ns MD simulation. (d-f) Dimensions and location of predicted water channels for the 100-150 ns MD simulation of the G31A mutant. (g-i) Dimensions and location of predicted water channels for the last 150-200 ns MD simulation of the G31A mutant. The GXGXG residues are labeled for positional reference.

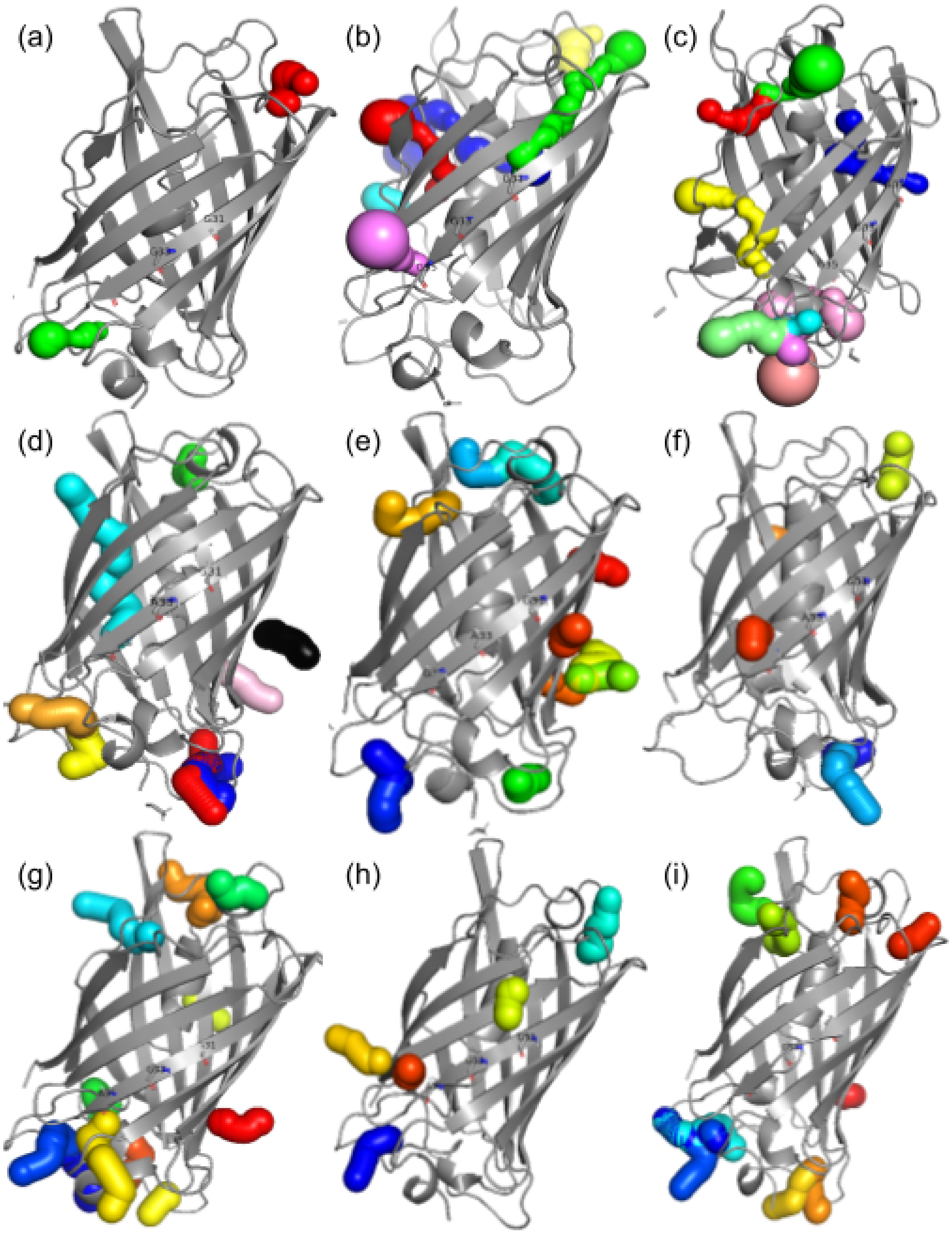


Figure 19. Visualization of predicted water channels in and through the β-barrels of wtGFPimm, G33A, and G35A variant MD simulations. (a-c) Scaled dimensions and locations for the 1st, 501st, and 1001st frames of the engineered wtGFP 100ns MD simulation. (d-f) Dimensions and location of predicted water channels for the MD simulation of the G33A mutant. (g-i) Dimensions and location of predicted water channels for the MD simulation of the G35A mutant. The GXGXG residues are labeled for positional reference.

Table 12. Predicted Channel Dimensions of the wtGFP MD simulation

Frame	Name	Length (Å)	Bottleneck (Å)	Lining Residues	Description
1	T1C3	8.2	1.5	V22, H25, P54, L137, ***V22***	Located at the end opposite of the C and N Termini of the β-barrel. Near the 1st β-sheet and the turn after the 6th β-sheet. Has an almost horseshoe-type shape.
1	T2C4	8.2	1	D36, A37, T38, Y39, ***F8**, **T9**, **G10**, **A37***	Located near the C and N termini (lined by the helices near the 1st and second β-sheets.
501	T1C2	17	1.1	V61, T62, S65, Y66, N144, Y145, T203, S205, L207, L220, ***V61***, ***T62***, ***Y145***, ***S205***, ***A206***, ***L207***	Located near the α-helix, exiting between the turn of the 7^{th} β-sheet and the 10^{th} β-sheet.
501	T2C2	29	1	L18, V22, N23, H25, F27, V29, L53, P54, V55, T57, L60, T63, F64, I123, L125, E132, L137, ***V22***, ***N23***	Located near the α-helix, exits between the turn of the 1^{st} β-strand and the turn after the 6^{th} β-sheet.
501	T3C2	39.5	0.8	L60, V61, T62, T63, F64, R96, I98, Y106, I123, L125, Y145, N146, H148, R168, H169, N170,H181, ***R168***, ***H169***, ***P58***, ***T59***, ***L60*** ***N144***, ***Y145***, ***N146***, ***K166***	Located near the α-helix, exits near the 7th β-sheet. It looks as if it is connected T2C2, but they're going in opposite directions.
501	T4C3	7.7	2	D129, K131, D133, ***D102***, ***D103***, ***D129***, ***G134***	This tunnel is almost completely outside of the β-barrel. It is located opposite of the C and Termini end of the protein near the 4th and

					6th β-strands. It’s also very compact unlike the other tunnels described previously.
501	T5C4↥	9.4	1.1	D36, L42, T43, F71, ***E34***, ***G35***, ***D36***, ***K41***, ***L42***	Lined by residues on the 2nd and 3rd β-sheet and F71 of the α-helix. G35’s backbone also lines this cavity.
501	T6C5	3.8	1.3	S147, N149, T203, ***H148***, ***L201***, ***S202***, ***T203***	This tunnel is only lined by β-sheet residues, it does not go far enough into the barrel to be attracted by the α-helix. Exits between the 10th and 7th β- strands.
1001	T1C1	11	1.1	P58, L141, Y145, N146, H169, N170, ***E142, N144, Y145, N146, R168, H169, N170***	Only lined by α-helix residue (P58). Almost kidney shaped. Leaves between the turn of the 6th β-sheet and the residues before the 5th β-sheet.
1001	T2C1	12.7	1	P58, H139, L141, Y143, H169, K209, ***L141, E142 , N170***	Seems to be connected to tunnel T1C1. It is lined by some of the same residues as T1C1, but it moves out more in the opposite direction, allowing it to be lined with residues that are near the turn of the 10th and 11th residue (i.e. K209).
1001	T3C1	26.9	1.1	V16, E17, L18, F27, V29, F46, L53, V55, L60, T63, R96 , T108, I123, L125, ***V16 E17, L18, V29, S30, E124***	This tunnel is wide enough and twists within the β-barrel, resulting in it being lined with residues that are located on 6 different β-sheets, along with the α-helix. Exits between the 1st and 2nd β-sheets, slightly above G31.
1001	T4C2	24.3	1	Y66, Q69, S72, Y74, F84, V150, Y151, I152, I161, V163, F165, N185, H199, L201, ***Y66***	Closer to the C and N termini of the barrel, very close to the α-helix so more of those residues line this tunnel. Exits through 7th and 8th β-sheets.

1001	T5C3	9.9	1.2	E5, F8, T9, A37, T38, K85, ***G4, F8, T9, A37, T38***	Smaller tunnel and it is almost completely outside of the barrel. It’s located almost right next to the C Termini of the protein.
1001	T6C3	10.7	1.2	E5, F8, A37, T38, K85, ***G4, F8, T9, A37, T38***	There’s almost complete overlap with this tunnel and T5C3. They differ at the directions that they point at the outermost point of the tunnels. This tunnel points back up between the two helices while T5C3 points to the side near the C termini.
1001	T7C3	14.4	1.3	K3, E5, F8, K85, S86, L194, ***G4, E5, K79, Q80, D82, S86***	This tunnel looks connected to both T6C3 and T5C3, but it points in a separate direction. This one actually does pass by the C termini while the other two just approach it. This tunnel is also much wider than the two that were previously described.
1001	T8C3	15.6	1.2	E5, F8, A37, T38, R73, K79, K85, ***G4, F8, A37, T38, Y74, D76***	This tunnel also looks connected to the three previously described tunnels, but this one points opposite the direction of T5C3 for a longer length which allows it to be lined with residues like Y74.
1001	T9C4	21.5	1.2	E90, K156, N159, P187, P192, V193, L195, ***S86, A87, G189, D190, G191, P192, V193***	This tunnel is located on the bend after the 9th β-sheet. The middle of it sits on the helix and then each side the tunnel goes out of the protein. The side closest to the C termini gets in fairly close proximity to T7C3.

Table 13. Predicted Channel Dimensions of 2AWJ G31A MD simulation (100-150 ns)

Frame	Name	Length (Å)	Bottleneck (Å)	Lining Residues	Description
1	T6C9	8.9	1.1	V22, H25, F27, P54, V55, L137, ***V22, H25, P54***	This tunnel is located on the end of the barrel opposite of the C and N termini. It is lined by residues of the end and turn of the first β-strand, the turn between the sixth and seventh β-strands, and residues that are close in sequence to the α-helix in the barrel.
1	T3C3	6.5	1.6	F83, A154, P187,V193, L195, **K156, K158, G160**	This tunnel is located on the lid of the β-barrel on the side of K158 and G191. It is on the turns of the seventh and eighth β-sheet. The bottleneck bends towards the residue 154.
1	T4C4	16	1	Y66, Q69, F84, I152, M153, I161, V163, Q183, N185, L201, **Y151, I152, M153, K162**	The tunnel travels through the residue in between residue I152 and V163. It is in between the 6th and 7th β-sheet and the bottleneck bends towards L201. It does not get sufficiently close to the α-carbon.
1	T5C6	6.8	1.2	N144, Q204, A206, L207, **Y145, S205, A206, L207**	The tunnel slightly enters the β-barrel through the gap between L207 and Y145. It is a gap between the turn of the 6th and 7th β-sheet and the 10th β-sheet.
1	T2C2	11.3	1.6	K52, P56, W57, P58, H139, Y143, E172, D216, ***L53, V55, P56, W57, L141, E142***	The tunnel is in one of the lids of the β-barrel on the side of E142. It seems to enter the protein slightly before leaving yet again. It overlaps with the tunnel T1C2

1	T1C2	3.4	2.2	K52, W57, H139, D216, ***L53, V55***	The closest residue is L53, and it overlaps with T2C2. It seems not to enter the protein.
501	T3C2	20.7	1	F83, A87, Y92, N159, P187, V193, ***A87, G91, K156, P187, I188, G189, D190, G191***	The tunnel goes through the V193 and I188. The bottleneck is bending towards F84.
501	T2C2	13.2	1.1	F83, A87, E90, Y92, P187, V193, ***S86, A87, G91, P187, I188, G191, P192***	It is perpendicular to the tunnel T3C2 and travels in the same manner as it.
501	T4C3	8.9	1.6	K3, E5, K79, L194	This tunnel is located in the center of the lid that contains G4. It does not enter the protein, it remains entirely outside.
501	T1C1	11	1.1	V61, N144, Y145, S205, L207, L220, ***Y145, N146, S205, L207***	This tunnel enters the protein through a gap provided by Y145 and A206. It bends towards the α-helix slightly, but does not get too close to it.
501	T5C4	10.2	1.1	P56, P58, H139, Y143, H169, K209, ***P56, T59, L141, E142, Y143***	The widest part of the tunnel is outside the protein on the lid that contains E142. It enters in between W57 and E142. P58 is the residue that is closest to the end of the tunnel.
1001	T1C3	10.9	1.1	Y66, H148, N149, V150, F165, R168, ***S147, H148, N149, K166***	The tunnel is located between residues V150 and F165.

Table 14. Predicted Channel Dimensions of 2AWJ G31A MD simulation (150-200 ns)

Frame	Name	Length (Å)	Bottleneck (Å)	Lining Residues	Description
1	T2C6	9.1	1	V12, P13, F114, D117, L119, **L7, V11, P13, D117, L119**	The tunnel is located between the D117 and V12. Thus is it between the second β-sheet and the turn of the 5th β- sheet.
1	T3C7	4	1.1	M78, H81, H199, I229, **N198, G228**	The tunnel is located between G228, N198 and H199. It is between the 10th and the 11th β-sheet
1	T1C5	8.1	1.5	F83, N159, P187, V193, P196, **K158, V193, L194**	The tunnel is located between L194, T186, which means that it is close to the 9th β-strand and the turn that connects both the 9th and the 10th β-strand.
501	T1C1	15.1	1	V22, H25, P54, V55, Y106, F130, E132, L137**, V22, N23, P54**	This tunnel is located in between the barrel lids in between N23 and L53. The bottleneck resides within the protein, but as the tunnel moves outward it is more wide.
501	T3C4	5.5	1.2	P56, P58, T59, Y143, H169, T59, **L141, E142, Y143**	The tunnel is located between W57 and E142, meaning that it does not reside in any of the β-sheets, but in the turns that are located in the lid. This tunnel is in the same lid as the T1C1 of this frame.
501	T2C1	44.1	0.9	V16, L18, V29, F46, L60, T62, T63, F64, S65, Y66, I98, F100, Y106, Y108, I123, L125, Y145, S147, H181, T203, Q204,	The entry/exit is in between the residues Q204 and N146, in between the 10th and 7th β-sheet. It loops around the α-helix on the side of V162, Y182, I98 , L125, L18, ending in S30. It seems to have an equal

				S205, E222, **S30, T59, T62, E124, N146, S147, T203, Q204, S205**	distance between the α-helix and the β-sheet.
1001	T2C3	9.6	1.1	P56, P58, T59, H139, L141, Y143, H169, **W57, T59, L141, E142**	The tunnel is located L141 and P58. It is located in the lid and the entry to the protein is located near the α-helix, the widest part of the tunnel is located near 141.
1001	T3C4	9.7	0.9	E32, K45, I47, R215, H217, **F46, M218**	The tunnel is outside the protein and it does not enter anywhere. The closest the tunnel is in the protein is at F46 and M218.
1001	T1C1	11.5	1.2	V11, V12, D36, A37, T38, **E6, L7, F8, T9, G10, D36, A37, T38**	It is located in the other lid, opposite to the first tunnel described for this frame. It is in between G10 and A37. It goes slightly in towards the bottleneck

Table 15. Predicted Channel Dimensions of 2AWJ G33A MD Simulation

Frame	Name	Length (Å)	Bottleneck (Å)	Lining Residues	Description
1	T1C1	28.3	0.9	L42, V61, Y66, Q69, S72, N144, Y145, T203, S205, A206, L207, L220, E22, V224, **S65, Y145, L207**	Comes in between strands 10 and 7 opposite the N/C termini, then goes all the way toward the chromophore tripeptide and other parts of the α-helix.
1	T2C3	17.1	1.3	A87, E90, N159, P187, G189, D190, G191, V193, **S86, G189, D190, G191**	Mainly lined by loop regions between β(9-10) and β(7-8) on the same side as the N/C termini. Somewhat horseshoe shaped.
1	T3C3	18.8	1.2	A87, P89, E90, N159, P187, G189, D190, G191, V193, **S86, G189, G191**	Overlaps significantly with T2C3. Close to the N-terminus.

1	T4C5	7.1	1.2	K3, E5, K79, D82, L194, **K79**	Small tunnel lined by residues of the N-terminus, loop between β(9-10) and the end of the α-helix.
1	T5C5	13.1	1.1	E5, A37, T38, R73, Y74, P75, K79, D82, K85, **A37**	Slight overlap with T4C5, also located close to N-terminus, closer to the end of the α-helix, and interacts with residues of the β-turn between strands 2 and 3.
1	T6C6	6.5	1.3	R109, A110, E111, R122, I123, E124, **E111, R122, E124**	Lined by residues pointing outward toward bulk solvent. Ends at space between strands 5 and 6, almost aligned with the chromophore tripeptide, but does not go into the barrel to interact with those residues.
1	T7C7	5.8	1.1	D102, D103, K131, G134, N135, I136, N177, **K101, D102, N135**	Lined by residues of loops between strands 4,5, and the lid opposite the N/C termini.
1	T8C8	6.1	1.3	V93, E95, K158, T186	Tunnel points toward the space between β-strands 4 and 9, but does not actually enter the β-barrel.
501	T1C2	7.1	1.8	E5, T9, A37, T38, K79, Q80, **E5, F8**	Somewhat bean-shaped, lined by residues close to the N-terminus, the β-turn between strands 2 and 3, and residues in the loop between the α-helix and strand 4.
501	T2C2	8.4	1.8	E5, E6, T9, K79, Q80	Some overlap with T1C2, points further out into the bulk solvent.
501	T3C3	7.4	1.8	D102, N135, I171, S175, Q177, **K101, V176**	Small tunnel lined by residues on loops between strands 4 and 5 and the lid opposite the N/C termini.
501	T4C3	13.8	1.1	D102, D103, F130, K131, K131, G134, N135, Q177, **D102,**	Some overlap with T3C3, points in the opposite direction when exiting the

				D129	β-barrel.
501	T5C6	8.2	1	E90, P187, G189, D190, G191, V193, **S86, G189, D190, G191**	Lined by residues on the loop between strands 9/10 and the loop between the α-helix and strand 4.
501	T6C7	6.4	1.4	E111, K113, V120, R122, **E111**	Points toward the space between strands 5 and 6, but does not enter the β-barrel.
501	T7C7	11.1	1.4	R109, A110, E111, K113, V120, R122, E124, **E111, R122**	Overlap with T6C7, goes in the opposite direction on the way out of the β-barrel.
501	T8C9	11.9	0.8	P58, Y143, N144, Y145, H169, L207, **E142, Y143, N144, Y145**	Lined by residues at the top of the α-helix (opposite N/C termini). Exits the β-barrel between strands 7 and 10.
501	T9C10	6.6	1.2	V93, E95, Q184, N185, T186, **E95, Q184**	V-shaped, points toward the space between strands 4 and 9, but never enters the barrel.
501	T10C11	4	1.1	L15, E17, S30, R122, **V16**	Points between strands 1 and 6, but does not enter the β-barrel. Also interacts with S30 of the strand 2 and is in line with the chromophore tripeptide.
501	T11C12	5.5	1.1	K107, K126, G127, I128, **K126**	Points toward the space between strands 5 and 6 opposite the N/C termini, but does not enter the β-barrel.
1001	T1C3	7.1	1.5	N159, P187, D190, V193, **G189, G191**	Located in the loop region of strands 9 and 10 and strand 8 (N159), on the same side as the N/C termini.
1001	T2C3	16.7	1.6	K3, A87, P89, E90, P187, G189, D190, G191, V193, **S86, M88, G191**	Some overlap with T1C3, goes in the opposite direction of T1C3 out of the barrel.
1001	T3C4	9.4	1	V22, H25, P54, E132, L137	Interacts with the β-turns of β(1-2), β(6-7), and the loop region of the α-helix (between β3 and helix)

1001	T4C10	5.6	1.5	F99, K101, L178, A179, D180, **F99**	Tunnel points toward space between strands 4 and 9 opposite N/C termini, but does not enter the barrel.
1001	T5C12	4.2	1.9	V11, E34, K41, T43	Hovers over G35. Points right into the space between strands 2 and 3, but does not enter the β-barrel.

Table 16. Predicted Channel Dimensions of 2AWJ G35A MD Simulation

Frame	Name	Length (Å)	Bottleneck (Å)	Lining Residues	Description
1	T1C1	12.8	1.3	K3, E5, K79, D82, K85, S86, L194, **G4, E5, K79**	Located right next to the N-terminus and the loop between the α-helix and β4. Most of the tunnel is parallel with the bottom of the barrel.
1	T2C1	13.6	1.3	E5, T9, A37, T38, Y74, K79, D82, K85, **F8, A37, Y74**	Some overlap with T1C1, and go out the opposite side of the barrel. Both T2C1 and T1C1 combined to have a horseshoe shape around the N-terminus.
1	T3C2	16.5	1.2	P58, Y143, Y145, N146, I167, R168, H169, N170, V176, L207, **N144, R168, N170**	This tunnel enters between β7 and β8, and penetrates the barrel directly to P58, at the top of the helix.
1	T4C3	5.5	3.5	K52, H139, K209, D216	Very large tunnel that interacts with lids regions of the side opposite to the N/C termini.
1	T5C5	22.3	0.7	Y74, F83, F84, I152, M153, A154, I161, L195, P196, D197, N198, H199, **F83,**	Long tunnel that enters between β7 and β10 on the side of the N/C termini. Penetrates into the barrel, interacting with the loop after the helix.

				A154, P196, D197, N198	
1	T6C7	5.5	N/A	T97, F99, Y182	Points towards space between β4 and β9, but does not come close to the barrel at all.
1	T7C8	3.6	2.2	A87, P89, E90, G189, G191, P192, **S86, P192**	Located between the B-termini of β9 and β10 and the loop of the helix on the same side of the N/C termini.
1	T8C9	15.8	0.6	L7, T9, G10, A35, A37, F71, D117, **L7, T9, G10, A35, A37**	This tunnel goes right into the hydrophobic pocket that G35 us typically in.
1	T9C10	16.9	0.9	K101, D102, D103, N135, I136, L141, I171, S175, Q177, **K101, V176**	Interacts with β9, the β4 and β5 turn, and lids opposite to the N/C termini.
1	T10C11	8.4	0.9	K156, N159, V193, L195, **V193**	Located between the loops of β7/β8 and β9/β10 on the same side of the N/C termini. Tunnel points straight up into the barrel, but the tunnel is very much short.
1	T11C12	6	1.2	E111, K113, R122, **V120**	Pointing at the space between β5/β6 (supposed to be strands but it is a loop), on the same side as N/C termini.
501	T1C1	8.7	1.9	E5, T38, Y74, K79, K85, **A37, Y74**	Located on the same side as N/C termini, interacts with residues on the β-turn of β2/β3. The loop following the α-helix, and loop of the N-terminus.
501	T2C3	9.6	0.8	V22, H25, K52, P54, V55, L137, **V22, L137**	Located on the turn of β1/β2, interacting with the loop prior to α-helix (opposite to N/C termini) and the loop between

					β6/β7.
501	T3C4	3.5	1.8	E32, K45, I47, E213	This tunnel is lined by residues that point outward towards bulk solvent, on the 2nd and 3rd β-sheets and loop between β10/β11 (opposite to N/C termini).
501	T4C5	7.4	1.2	Y39, R73, Q204, F223, T225, **Y39, G40, V224**	Funnel shaped tunnel that points into the space between β3/β11. The same side as the N/C termini, but does not go far unto the β-tunnel.
501	T5C7	3.8	2.1	V11, E34, A35, D36, K41, T43	This tunnel points right into the space between β2/β3 where our G35A simulation is, but it does not go into the β-barrel.
1001	T1C1	10.9	1.8	F8, A37, T38, R73, P75, K79, K85, **F8, A37, Y74, D76**	Located near the β-turn of strands 2,3, and the loop region, immediately following the α-helix, and some of the loop following the N-terminus. Runs somewhat parallel with the bottom of the protein.
1001	T2C1	13	1.8	E5, E6, F8, T9, A37, T38, R73, K79, K85, **F8, A37**	Overlaps with T1C1 (fairly perpendicular to each other), points more towards the N-terminus on the way out towards the bulk solvent.
1001	T3C1	19.5	1	E5, T38, R73, Y74, P75, K79, D82, K85, **A37, Y74, D76, K79, H81**	In the same place T1C1 and T2C1, but there’s more overlap with this tunnel and T1C1. Narrower size and longer, having more interactions with residues of the post-α-helix loops regions.
1001	T4C2	10.2	1.3	E142, N144, N146, R168, N170, **R168**	Horseshoe shaped with both ends pointing into the space between β7 and β8 (and the loop that follows

					them), opposite to the N/C termini.
1001	T5C2	13.3	1	P58, V61, Y143, Y145, I167, H169, L207, S208, M218, **Y143, Y145, Y146**	Another horseshoe type tunnel near T4C2, but this tunnel is inside the barrel. One end points out towards the same space that T4C2 points into, and the other points out to the space between β7 and β10 opposite to the N/C termini.
1001	T6C3	9.5	1.8	K3, A87, E90, G191, V193, **G4, S86, P192**	Located right next to the N-terminus and the loop residues between β9 and β10.
1001	T7C3	10.1	1.4	A87, E90, P187, G189, P190, G191, V193, **G189, G191**	Overlap with T6C3, go in the same direction, These tunnels are basically stacked on top of each other.
1001	T8C6	6.8	1.4	K52, L53, W57, H139, Y143, D216	This tunnel interacts with residues near the top of the α-helix (W57). Does not go for enough to interact with chromophore tripeptide.
1001	T9C8	5	1.5	H25, F27, T50, L53, P54, **K26, T50, K52**	Interacts with β2, β3, and the loop prior to α-helix (opposite N/C termini). Points right in the β-barrel through the vertical axis.
1001	T10C12	3.4	2.2	E95, K158, Q184, T186	Tunnel points into space between β4/β9, right under R96, but it does not go into the β-barrel.

***Bold** lining residues indicate interaction with the backbone of the named residue.

Quantitatively, it is clear that the G3XA mutants resulted in a decreased rate of water migration into the β-barrel toward the chromophore, as many more predicted channels enter the chromophore region of the protein in the wtGFP simulation. The G3XA mutations may cause the methyl group of the alanine side chain to participate in steric clashing with other residues that point inward toward the β-barrel. This could lead to downstream disruptions of the hydrogen bonding patterns amongst other secondary structures in the protein over the course of the simulation, allowing for different channels to open and close. In all the G3XA simulations, many of the predicted water channels enter near the lids of the protein, but since this is composed of mainly loop structures that are highly exposed to bulk solvent, it would be expected that many waters would be interacting with these residues.

Qualitative analysis of the predicted water channels over the course of the simulations was done by observing the location of the water channels within the β-barrel over the simulation time. It was also observed that more water channels appeared near the lids of the barrel, further suggesting that these alanine mutations may cause significant amounts of space around the chromophore to be taken up, preventing water molecules from migrating into this region. This corresponds with the quantitative water migration results as most waters were described to not be near the catalytic or chromophore-forming residues in the G3XA simulations.

3.6. β-Sheet Interactions

Expression of the three G3XA mutants in the wet lab by Professor Tanya Schneider led to the finding that these mutants were all susceptible to misfolding and aggregation. This observation led to the idea that instead of chromophore formation, G31, G33, and G35 could play a significant role in the initial folding of the β-barrel. Since the N-terminus β-sheets stay intact through most pre-folded intermediates, the hydrogen bond distances between sheets 1 and 2, as well as sheets 2 and 3 were examined for each mutant simulation to see if remnants of structural effect from these mutations could be observed in a fully folded β-barrel (Table 17).

Table 17. Hydrogen Bond Distances across β(1-3) in G3XA Simulations

	G31A (100 - 150ns)	G31A (100 - 150ns)	G31A (150 - 200ns)	G31A (150 - 200ns)	G33A	G33A	G35A	G35A	wtGFP	wtGFP
	Avg (Å)	SD	Avg (Å)	SD	Avg (Å)	SD	Avg (Å)	SD	Avg (Å)	SD
L41CO - D36NH	2.6792	0.6415	2.4109	0.5911	2.1002	0.2483	2.2321	0.3019	2.0512	0.1868
E34CO - T43NH	2.4111	0.6908	2.0165	0.2182	1.9685	0.1640	2.0538	0.1952	2.0023	0.1790
E34NH- T43CO	2.6463	0.8218	2.1583	0.2757	2.0497	0.1856	2.0587	0.2256	2.0301	0.1747
L45NH - E32CO	2.1728	0.4868	2.1469	0.3194	2.3117	0.3632	2.2523	0.3140	2.1230	0.2384
L45CO - E32NH	2.3259	0.4949	2.3996	0.4111	2.5832	0.5980	2.4307	0.4495	2.3074	0.3244
I47NH - S30CO	1.9489	0.1681	1.9343	0.1486	1.8778	0.1546	1.8828	0.1575	1.8994	0.1401
I14CO - S30NH	2.0664	0.1808	2.0256	0.1701	2.0061	0.1503	2.0079	0.1573	2.0425	0.1596
A31CO -V16NH	3.0753	0.9130	2.5835	0.6149	N/A	N/A	N/A	N/A	N/A	N/A
A31NH -V16CO	2.8252	0.6864	2.6654	0.5162	N/A	N/A	N/A	N/A	N/A	N/A
H25NH -V22CO	2.5303	0.6867	2.2230	0.2998	2.3450	0.3910	2.3459	0.3488	2.1522	0.2916
H25CO- V22NH	2.1590	0.2357	2.0487	0.1909	2.1852	0.2372	2.1895	0.2481	2.1590	0.2183
V29NH - L18CO	1.9173	0.1882	1.9281	0.1892	1.8879	0.1580	1.9444	0.1882	1.9780	0.2016
V29CO - L18NH	2.3562	0.4192	2.4975	0.4567	2.0155	0.2013	2.0215	0.2461	2.0395	0.2201
G31NH- V16CO	N/A	N/A	N/A	N/A	2.5747	0.6116	2.2521	0.3916	2.1886	0.3191
G31CO- V16NH	N/A	N/A	N/A	N/A	2.6427	0.7767	2.1283	0.3049	2.1522	0.2916
I14CO- A33NH	N/A	N/A	N/A	N/A	2.2903	0.5172	N/A	N/A	N/A	N/A
I14NH - A33CO	N/A	N/A	N/A	N/A	1.9426	0.1839	N/A	N/A	N/A	N/A
F27CO	2.2968	0.2481	2.0170	0.1880	2.3748	0.2714	2.2989	0.2558	2.1289	0.2130

-G20NH										
F27NH - G20CO	2.0615	0.2046	2.1699	0.2246	2.1260	0.2283	2.0915	0.2085	2.3746	0.2731
G35CO-V12NH	1.9378	0.1911	1.9279	0.1878	2.0896	0.3482	N/A	N/A	1.9454	0.1898
G35NH-V12CO	2.1830	0.2489	2.1712	0.2697	2.2805	0.3447	N/A	N/A	2.1107	0.2113
G35CO-V12NH	N/A	N/A	N/A	N/A	N/A	N/A	1.9204	0.1679	N/A	N/A
G35NH-V12CO	N/A	N/A	N/A	N/A	N/A	N/A	2.1514	0.1981	N/A	N/A

When compared to the wtGFP simulation, the N-terminus β-strands of the G3XA mutants tended to have a larger amount of separation, indicated by the larger average distance of the interstrand hydrogen bond distances, which would make sense to accommodate the steric strain caused by the alanine side chain (Fig. 20).

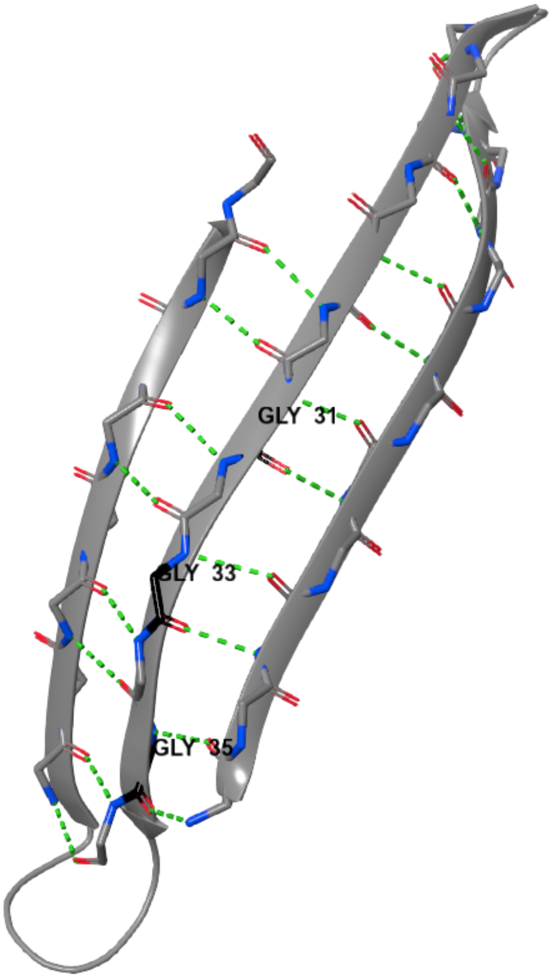


Figure 20. Visualization of the backbone hydrogen bonds (green dashed lines) of the N-terminus β-strands that were measured over all mutant simulations. G31, G33, and G35 of the wtGFP N-terminus strands are labeled for locational reference.

3.7. Hydrophobic Pocket Mutants

Upon examination of glycines 31, 33 and 35, it was observed that G35 is in the middle of a fairly large hydrophobic pocket, possibly interacting with F71 through a H-π interaction to maintain hydrophobic packing. To investigate this, four additional GFP variants were engineered and MD simulations were performed:

- G35/F71L: to investigate how the GFP structure would be affected by allowing space within the pocket.

- G35V/F71: to investigate how the GFP structure would be affected by significantly increasing the steric bulk in the pocket.
- G35/F71Y: to investigate how the GFP structure would be affected by introducing charge character in the pocket.
- G35V/F71L: to see if simultaneously increasing the size of one residue while decreasing the size of the other will cause maintenance of stability.

Table 18 summarizes the data results of examining the interactions of the N-terminus strands in these four mutants.

Table 18. Hydrogen Bond Distances across β(1-3) in Hydrophobic Pocket Simulations

	G35/F71L	G35/F71L	G35/F71Y	G35/F71Y	G35V/F71	G35V/F71	G35V/F71L	G35V/F71L	wtGFP	wtGFP
	Avg (Å)	SD	Avg (Å)	SD	Avg (Å)	SD	Avg (Å)	SD	Avg (Å)	SD
H25NH-V22CO	2.3708	0.4711	4.2686	1.8843	2.1462	0.2240	2.3196	0.3408	2.1590	0.2916
H25CO-V22NH	2.0995	0.2196	2.1546	0.2493	2.3045	0.3373	2.1432	0.2212	2.1590	0.2183
F27NH-G20CO	2.0567	0.2031	2.0743	0.2145	2.1061	0.2199	2.0957	0.2317	2.3746	0.2731
F27CO-G20NH	2.2558	0.2447	2.2935	0.2636	2.2911	0.2397	2.3205	0.2829	2.1289	0.2130
V29NH-L18CO	1.9240	0.1804	1.8914	0.1650	1.9841	0.1926	1.9773	0.1983	1.9780	0.2016
V29CO-L18NH	2.2748	0.3307	2.1845	0.3003	2.1206	0.2711	2.1588	0.3112	2.0395	0.2201
G31NH-V16CO	2.2785	0.3919	2.5359	0.9936	2.0481	0.2419	5.2081	0.4742	2.1886	0.3191
G31CO-V16NH	2.1016	0.3347	2.4322	0.7980	1.9987	0.2029	3.5411	0.5474	2.1522	0.2916
I47CO - S30NH	1.9772	0.1596	2.0311	0.1730	2.0319	0.1636	2.0398	0.1707	2.0425	0.1596
I47NH - S30CO	1.8992	0.1554	1.9443	0.1854	1.9278	0.1521	1.9215	0.1533	1.8994	0.1401

L45CO-E32NH	2.2052	0.3466	2.1875	0.3071	2.3183	0.4321	2.0648	0.2583	2.3074	0.3244
L45NH-E32CO	2.0901	0.2832	2.0857	0.2658	2.3377	0.5862	3.0888	0.5518	2.1230	0.2384
E34NH-T43CO	2.1241	0.2313	2.4578	0.7993	2.1788	0.4875	7.5276	0.9802	2.0301	0.1790
E34CO-T43NH	1.9961	0.1829	2.6536	0.8737	2.2030	0.7329	5.5800	0.5803	2.0023	0.1790
G35NH-V12CO	2.1886	0.2217	3.1598	0.8277	N/A	N/A	N/A	N/A	2.1107	0.2113
G35CH-V12NH	1.9271	0.1838	4.0419	1.8843	N/A	N/A	N/A	N/A	1.9454	0.1898
V35NH-V12CO	N/A	N/A	N/A	N/A	1.9669	0.1862	2.0892	0.2286	N/A	N/A
V35CO-V12NH	N/A	N/A	N/A	N/A	2.0694	0.2018	1.8852	0.1582	N/A	N/A

No significant structural effects were observed in the G35/F71L and G35V/F71 mutant simulations, further supporting the hyperstability of GFP. The G35/F71Y and G35V/F71L simulations were observed to result in several significant deviations (> 1.0 Å avg. difference) from the wtGFP simulation. For the G35/F71Y mutant, the effect can be explained by the increased size of the tyrosine residue, along with the unfavorable electrostatic interactions due to the introduction of the *para*-hydroxy substituent in the tyrosine side chain. In the G35V/F71L mutant, having the two alkyl side chains also caused steric strain, leading to distortions that prevented the backbone of the residues 32-36, which are typically part of β2, from making the necessary hydrogen bonds with strands 1 and 3.

When quantitatively comparing the G35/F71Y and G35V/F71L mutant simulations, it can be seen that the G35V/F71L simulation had a higher amount of structural change in terms of the number of affected backbone interactions and the magnitude of the differences to the wtGFP simulation. The distortion in the G35/F71Y mutant was likely to have been partially mitigated by the small size of the glycine side chain, allowing for some movement within the pocket without completely compromising structural integrity. In the case of the double mutant, the increased distortion is likely caused by there being many more rotational degrees of freedom in these larger alkyl side chains. This allows for these residues to adopt many more rotamers, which then results in more steric clashes that can only be accommodated by either weakening or completely breaking some of the interstrand hydrogen bonds.

CONCLUSION

Analysis of the fluorescent proteins (FPs) structures deposited in the PDB revealed that residues that are conserved across all variants fall into three categories: (i) residues involved in chromophore formation, (ii) residues on/around the lids of the β-barrel, and (iii) centrally located residues with no known function.

Most conserved residues are located on the ends of the barrel in the β-turns between the β-strands. Use of structural analyses, molecular dynamics simulations, and the Anisotropic Network Model led to finding that the conserved residues in the lids undergo less translation than other lid residues, and that some of these residues could potentially play a role as hinges or folding nuclei for fluorescent proteins.[43]

Glycines 31, 33, and 35 are all located on the second β-strand of the FP β-barrel. They are highly conserved amongst the FPs isolated from naturally occurring organisms and amongst the FPs found in the PDB. G33 is 100% conserved across all FP variants while G31 and G35 are 87% and 95% conserved, respectively.[18]

This trio of glycines does not have a direct role in the formation of the chromophore and they do not line the pore that contributes to chromophore formation. G31, G33 and G35 are behind the non-canonical α-helix where the chromophore is formed. Since each glycine is positioned after every other residue, all three side chains point into the core of the protein. This suggested that having larger side chains could possibly crowd the interior and either hinder or completely prevent chromophore formation (Fig. 21).

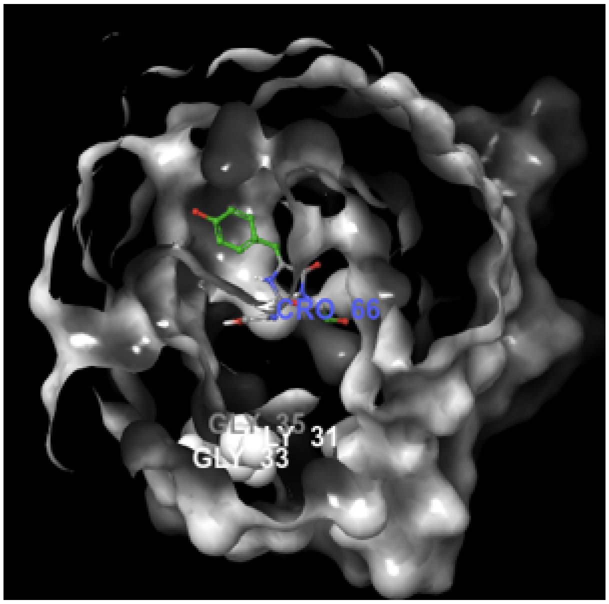


Figure 21. Cross section of avGFP. G31, G33 and G35 are located on β2, adjacent to the imidazolone ring of the chromophore.

The high conservation of the GXGXG motif found in all fluorescent proteins was investigated by performing 100-200 ns molecular dynamics (MD) simulations on immature G31A, G33A, and G35A single point mutant structures. The hydrophobic pocket that G35 is contained in was also investigated by performing 200 ns MD simulations on G35/F71L, G35/F71Y, G35V/F71, and G35V/F71L mutant structures.

Simulation analyses resulted in the following findings:

(1) In order for the amide nitrogen of G67 to attack the carbonyl carbon of S65, they have to be in close proximity to each other. In order to achieve this, immature GFP must adopt the tight-turn conformation. Consequently, glycine is conserved at position 67 in nearly all FPs (only one FP has an alanine at position 67) due to its high flexibility. The tight turn restricts the conformational space and keeps the residues in place for the initial cyclization step of the chromophore formation mechanism.

Although the conserved glycine triad is far from the chromophore forming region, the methyl side chains of the alanine mutants could reduce the size of the cavity located behind the space where the imidazolone ring is formed (see Fig. 21), affecting the tight turn conformation. This led us to monitor the tight turn distance over the course of the MD simulations. The tight turn distance of the alanine mutants were not significantly longer or prone to fluctuation than that of the immature wild type structure.

(2) The crystal structures of immature GFP mutants have shown that the GFP protein matrix creates a dramatic bend at the chromophore region of the central α-helix. This kink prevents interactions that are common in a canonical α-helix from occurring. It was presumed that this is essential to chromophore formation since those hydrogen bonds would have to be broken during chromophore maturation. Enforcing a tight-turn conformation in the chromophore tripeptide sequence causes disruption of the canonical α-helical main chain interactions and formation of a kink in the α-helix. Measurements of distortion in the non-canonical α-helix were taken over the course of each simulation. No significant structural changes were observed in the G3XA mutant simulations.

(3) Glycine residues located in β-strands are typically cross-stranded to aromatic residues that provide a stabilizing effect by stacking the phenyl ring over the glycine backbone, protecting it from bulk solvent. Due to the proximity of G35 and F71, it was presumed that this stabilization effect would occur through an α-helix/β-strand interaction. The G3XA mutations tended to result in more separation between F71 and G35, possibly resulting in the formation of water channels leading directly to the chromophore.

(4) Water migration and water channel prediction analyses showed that the G3XA mutations likely clogged the space around the chromophore, preventing waters from migrating into that area. These mutants were observed to have increased water migration around the

lids of the protein, which was expected due to the high interaction of the lid residues with bulk solvent.

(5) The hydrophobic pocket simulations showed that there is likely a H-π interaction between the side chains of G35 and F71 in order to maintain hydrophobic packing in the pocket while also reducing the amount of clashing in order to maintain the hydrogen bonding across the N-terminus β-strands.

Based on the previous observations, we are fairly confident that glycines 31, 33, and 35 do not influence formation of the chromophore. Due to the literature and mutational studies conducted by Prof. Schneider, which showed that the G3XA mutants are susceptible to misfolding and aggregation, we suspect that the flexibility inherent to the GXGXG motif is crucial in the folding of the FP β-barrel.

Our simulations have shown that mutations of any and all of the glycine triad result in diminished hydrogen bonding interactions amongst the N-terminus sheets. Therefore, the folding and closure of the FP β-barrel will be hindered, if not completely halted.

Preliminary results of the currently run partial structure simulations show that these residues may behave in a zipper-like fashion, which is reasonable in the sense that it would require a high amount of flexibility and smaller side chains in order to easily close the β-barrel. These simulations could then have the potential for us to gain a much deeper understanding in the mysterious nature of these three very unusual residues, and fluorescent proteins as a whole.

References

1. Miyawaki, A. Green fluorescent protein glows gold. *Cell* **2008**, *135*, 987-990.
2. K. Kurose, S. Inouye, Y. Sakaki, and F. I. Tsuji Bioluminescence of the Ca2"-binding photoprotein aequorin after cysteine modification. *Proc. Nati. Acad. Sci. USA* **1989**, *86*, 80-84.
3. Shimomura, O. Structure of the chromophore of Aequorea green fluorescent protein. *FEBS Letters* **1979**, *104*, 220-222.
4. Marc Zimmer GFP: from jellyfish to the Nobel prize and beyond. *Chemical Society Reviews* **2009**, *38*, 2813-2832.
5. Chalfie, M.; Tu, Y.; Euskirchen, G.; Ward, W. W.; Prasher, D. C. Green fluorescent protein as a marker for gene expression. *Science* **1994**, *263*, 802-805.
6. Heim, R.; Prasher, D. C.; Tsien, R. Y. Wavelength mutations and posttranslational autoxidation of green fluorescent protein. *Proc Natl Acad Sci U S A* **1994**, *91*, 12501-12504.
7. Heim, R.; Tsien, R. Y. Engineering green fluorescent protein for improved brightness, longer wavelengths and fluorescence resonance energy transfer. *Curr Biol* **1996**, *6*, 178-182.
8. Miyawaki, A. Green Fluorescent Protein-like Proteins in Reef Anthozoa Animals. *Cell Structure and Function* **2002**, *27*, 343-347.
9. Gross, L. A.; Baird, G. S.; Hoffman, R. C.; Baldridge, K. K.; Tsien, R. Y. The structure of the chromophore within DsRed, a red fluorescent protein from coral. *PNAS* **2000**, *97*, 11990-11995.
10. Matz, M. V.; Fradkov, A. F.; Labas, Y. A.; Savitsky, A. P.; Zaraisky, A. G.; Markelov, M. L.; Lukyanov, S. A. Fluorescent proteins from nonbioluminescent Anthozoa species. *Nature Biotechnology* **1999**, *17*, 969-973.
11. Haddock, S. H. D.; Dunn, C. W. Fluorescent proteins function as a prey attractant: experimental evidence from the hydromedusa Olindias formosus and other marine organisms. *Biology open* **2015**, *4*, 1094-1104.
12. Barykina, N. V.; Doronin, D. A.; Subach, O. M.; Sotskov, V. P.; Plusnin, V. V.; Ivleva, O. A.; Gruzdeva, A. M.; Kunitsyna, T. A.; Ivashkina, O. I.; Lazutkin, A. A.; Malyshev, A. Y.; Smirnov, I. V.; Varizhuk, A. M.; Pozmogova, G. E.; Piatkevich, K. D.; Anokhin, K. V.; Enikolopov, G.; Subach, F. V. NTnC-like genetically encoded calcium indicator with a positive and enhanced response and fast kinetics. *Scientific reports* **2018**, *8*, 15233-19.
13. Sacha N. W. Toussaint Ryan T. Calkins Sumin Lee Brian W. Michel Olefin Metathesis-Based Fluorescent Probes for the Selective Detection of Ethylene in Live Cells. *Journal of the American Chemical Society* **2018**, *140*, 13151-13155.
14. Jasmine N Tutol; Weicheng Peng; Sheel C Dodani Discovery and Characterization of a Naturally Occurring, Turn-On Yellow Fluorescent Protein Sensor for Chloride. *Biochemistry* **2018**, *58*, 31-35.
15. Reifenrath, M.; Boles, E. A superfolder variant of pH-sensitive pHluorin for in vivo pH measurements in the endoplasmic reticulum. *Scientific reports* **2018**, *8*, 11985-8.
16. Ormö, M.; Cubitt, A. B.; Kallio, K.; Gross, L. A.; Tsien, R. Y.; Remington, S. J. Crystal structure of the Aequorea victoria green fluorescent protein. *Science* **1996**, *273*, 1392-1395.
17. Yang, F.; Moss, L. G.; Phillips, G. N. The molecular structure of green fluorescent protein. *Nature Biotechnology* **1996**, *14*, 1246-1251.
18. Ong, W. J. -.; Alvarez, S.; Leroux, I. E.; Shahid, R. S.; Samma, A. A.; Peshkepija, P.; Morgan, A. L.; Mulcahy, S.; Zimmer, M. Function and structure of GFP-like proteins in the protein data bank. *Mol Biosyst* **2011**, *7*, 984-992.
19. Stepanenko, O. V.; Stepanenko, O. V.; Kuznetsova, I. M.; Verkhusha, V. V.; Turoverov, K. K. Beta-barrel scaffold of fluorescent proteins: folding, stability and role in chromophore formation. *International review of cell and molecular biology* **2013**, *302*, 221.
20. Sarkisyan, K. S.; Bolotin, D. A.; Meer, M. V.; Usmanova, D. R.; Mishin, A. S.; Sharonov, G. V.; Ivankov, D. N.; Bozhanova, N. G.; Baranov, M. S.; Soylemez, O.; Bogatyreva, N. S.; Vlasov, P. K.; Egorov, E. S.; Logacheva, M. D.; Kondrashov, A. S.; Chudakov, D. M.;

Putintseva, E. V.; Mamedov, I. Z.; Tawfik, D. S.; Lukyanov, K. A.; Kondrashov, F. A. Local fitness landscape of the green fluorescent protein. *Nature* **2016***, 533*, 397-401.

21. Fu, J. L.; Kanno, T.; Liang, S.; Matzke, A. J. M.; Matzke, M. GFP Loss-of-Function Mutations in Arabidopsis thaliana. *G3 (Bethesda, Md.)* **2015***, 5*, 1849-1855.
22. Barondeau, D. P.; Putnam, C. D.; Kassmann, C. J.; Tainer, J. A.; Getzoff, E. D. Mechanism and energetics of green fluorescent protein chromophore synthesis revealed by trapped intermediate structures. *Proc. Natl. Acad. Sci. U. S. A.* **2003***, 100*, 12111-12116.
23. Grigorenko, B. L.; Krylov, A. I.; Nemukhin, A. V. Molecular Modeling Clarifies the Mechanism of Chromophore Maturation in the Green Fluorescent Protein. *Journal of the American Chemical Society* **2017***, 139*, 10239-10249.
24. Wood, T. I.; Barondeau, D. P.; Hitomi, C.; Kassmann, C. J.; Tainer, J. A.; Getzoff, E. D. Defining the role of arginine 96 in green fluorescent protein fluorophore biosynthesis. *Biochemistry* **2005***, 44*, 16211-16220.
25. Tsien, R. Y. The green fluorescent protein. *Annu Rev Biochem* **1998***, 67*, 509-544.
26. Ganim, Z.; Rief, M. Mechanically switching single-molecule fluorescence of GFP by unfolding and refolding. *PNAS* **2017***, 114*, 11052-11056.
27. Govardhan Reddy; Zhenxing Liu; D. Thirumalai Denaturant-dependent folding of GFP. *Proceedings of the National Academy of Sciences - PNAS* **2012***, 109*, 17832-17838.
28. Jackson, S. E.; Craggs, T. D.; Huang, J. Understanding the folding of GFP using biophysical techniques. *Expert review of proteomics* **2006***, 3*, 545-559.
29. Minor, D. L.; Kim, P. S. Measurement of the β-sheet-forming propensities of amino acids. *Nature* **1994***, 367*, 660-663.
30. Merkel, J. S.; Regan, L. Aromatic rescue of glycine in β sheets. *Folding and Design* **1998***, 3*, 449-456.
31. Dou, J.; Vorobieva, A. A.; Sheffler, W.; Doyle, L. A.; Park, H.; Bick, M. J.; Mao, B.; Foight, G. W.; Lee, M. Y.; Gagnon, L. A.; Carter, L.; Sankaran, B.; Ovchinnikov, S.; Marcos, E.; Huang, P.; Vaughan, J. C.; Stoddard, B. L.; Baker, D. De novo design of a fluorescence-activating β-barrel. *Nature* **2018***, 561*, 485-491.
32. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Research* **2000***, 28*, 235-242.
33. Schrödinger, L. Maestro. **2018***, 3*.
34. Schrödinger, L. The PyMOL Molecular Graphics System. *, 2*.
35. LLC, S. MacroModel. **2017***, 3*.
36. D. E. Shaw Research Desmond Molecular Dynamics System. **2018***, 3*.
37. Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner, R. A. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theory Comput.* **2016***, 12*, 281-296.
38. User Manual Desmond 4.2.
39. Lemay, N. P.; Morgan, A. L.; Archer, E. J.; Dickson, L. A.; Megley, C. M.; Zimmer, M. The Role of the Tight-Turn, Broken Hydrogen Bonding, Glu222 and Arg96 in the Post-translational Green Fluorescent Protein Chromophore Formation. *Chem Phys* **2008***, 348*, 152-160.
40. Li, B.; Shahid, R.; Peshkepija, P.; Zimmer, M. Water Diffusion In And Out Of The β-Barrel Of GFP and The Fast Maturing Fluorescent Protein, TurboGFP. *Chem Phys* **2012***, 392*, 143-148.
41. Pravda, L.; Sehnal, D.; Tou?ek, D.; Navrátilová, V.; Bazgier, V.; Berka, K.; Svobodová Va?eková, R.; Ko?a, J.; Otyepka, M. MOLEonline: a web-based tool for analyzing channels, tunnels and pores (2018 update). *Nucleic Acids Res* **2018***, 46*, W368-W373.
42. Sehnal, D.; Svobodová Vařeková, R.; Berka, K.; Pravda, L.; Navrátilová, V.; Banáš, P.; Ionescu, C.; Otyepka, M.; Koča, J. MOLE 2.0: advanced approach for analysis of biomacromolecular channels. *Journal of cheminformatics* **2013***, 5*, 39.
43. Zimmer, M. H.; Li, B.; Shahid, R.; Peshkepija, P.; Zimmer, M. Structural consequences of

chromophore formation and exploration of conserved lid residues amongst naturally occurring fluorescent proteins. *Chemical Physics* **2014**, *429*, 5-11.