2019

Computational Analysis of the Structural Importance of the Conserved Glycine Residues at Positions 31, 33, and 35 in the Chromophore Formation and Folding of Green Fluorescent Protein

Franceine Welcome
Connecticut College, franceine.welcome@gmail.com

Recommended Citation

Welcome, Franceine, "Computational Analysis of the Structural Importance of the Conserved Glycine Residues at Positions 31, 33, and 35 in the Chromophore Formation and Folding of Green Fluorescent Protein" (2019). *Chemistry Honors Papers*. 27.
https://digitalcommons.conncoll.edu/chemhp/27

Computational Analysis of the Structural Importance of the Conserved Glycine Residues at Positions 31, 33, and 35 in the Chromophore Formation and Folding of Green Fluorescent Protein

Franceine Welcome
Connecticut College Department of Chemistry
New London, CT 06320
Spring 2019

Table of Contents

Acknowledgements

I want to thank my thesis advisor Professor Marc Zimmer for all his guidance, academic and emotional support, not just during my thesis, but for my entire time at Connecticut College. Working in his lab on this thesis has built my confidence in my research abilities, led me to applying for graduate school and taught me a lot about time management. Being a part of the Science leader program has really helped me to navigate undergrad and graduate school. I truly could not have done any of this without Marc's support. Not all superheroes wear capes. I want to thank the other student researchers Sercan Durmus, Justin Nwafor and Christian Salguero for being willing to stay late nights in the lab, being around to bounce ideas off of, and most importantly, for the comedic relief and emotional support. I want to thank Professor Vicki Fontneau and Professor Katherine Launer-Felty for being my external readers and giving me feedback on this work. And finally, a huge thank you to my family and friends for emotional support. Without them, none of this would be possible.

Abstract

Green fluorescent protein, initially cloned and expressed from the bioluminescent jellyfish, *Aequorea victoria* (avGFP), has a wide range of uses in cellular biology, one of which includes uses as a biological marker[1] . Variants of GFP exist, but some residues are highly conserved and necessary for appropriate chromophore formation. Some of these conserved residues are the glycine residues at positions 31, 33 and 35, though it they are not part of the tripeptide that forms the chromophore[2] .

The objective of this honors study and 3 related independent studies was to use computational methods to find out why the glycines at positions 31, 33, and 35 are so highly conserved in all fluorescent proteins and what role they play in chromophore formation, folding and stabilizing the protein. Precyclized immature structures (i.e. with no chromophore) have been used in our simulations because there is been evidence that conserved glycine residues play an important role in protein folding[3] or chromophore formation, which occurs prior to cyclization. Although mutating the glycine at position 35 to a cysteine[3] has been found to be somewhat fluorescent, we chose to make a less aggressive mutations, G31A, G33A, and G35A in order to study because alanine is the most structurally similar to glycine.

Introduction

GFP History

Bioluminescence can be described as the emission of light by living organisms. Bioluminescence reactions produce light via a reaction between an oxyluciferase and its luciferin substrate, which gets oxidized. Organisms are found in many different environments and use bioluminescence to evade predators, attract prey and for communication[1] . The bioluminescent jelly fish *Aequorea victoria* was first reported to produce green bioluminescence in 1955[4] . The bioluminescence from *A. victoria* is the result of the co-activation of the proteins aequorin and green fluorescent protein (GFP)[5]. Aequorin is a photoprotein that consists of apo-aequorin (luciferase) and the tightly bound substrate, coelenterazine (luciferin)[4,5]. When aequorin is activated by three calcium ions, a conformational change takes place that converts it into an oxygenase, oxidizing coelenterazine and emitting blue light[4,5] . The oxidized aequorin complex undergoes radiationlesss energy transfer to GFP, producing the characteristic green light of *A. victoria*[4] . GFP has a major excitation peak at 395nm and a minor peak at 475nm which is caused by different protonation states[6,7]. GFP has an emission maximum of 508nm and a quantum yield of 0.72-0.85[7]. Initially, upon discovering these proteins, scientists were more interested in using aequorin as a calcium sensor to monitor processes in biological systems[8]. In 1979, Shimomura was able to deduce the structure of the of the chromophore of GFP (Figure 1) by comparing it to other synthesized compounds of known structures[4,9].
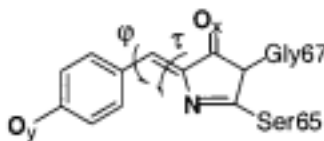

Figure 1: Green fluorescent protein chromophore structure[1].

In 1992, the 238-amino acid residue polypeptide sequence of *A. victoria* GFP was determined by Prasher and colleagues[4,10]. In 1996, the crystal structure of *A. victoria* GFP was solved (Figure 2), revealing its eleven strand, closed bottom beta barrel structure encasing the chromophore[11]. The crystal structure also confirmed that chromophore containing fragment is the result of a cyclized tripeptide formed from amino acid residues 64-69 (Ser-Tyr-Gly)[4,11] . In 1994, GFP was expressed in both *E. coli* and *C. elegans* and showed similar activity as the native protein in *A. victoria*, proving that the chromophore formation occurred via an autocatalytic cyclization and demonstrating the ability of GFP to be used as a genetic luminescent tag[4,12,13] . Together, these discoveries would go on to revolutionize the way that GFP would be used in molecular and cell biology and medicine as a biological marker, and steal some of the spotlight from its bioluminescent partner, aequorin. In 2008, the Nobel Prize in chemistry was awarded to Osamu Shimomura, Martin Chalfie, and Roger Tsien for their work in determining the structure of GFP and its applications in biology to help us to understand many cellular pathways and
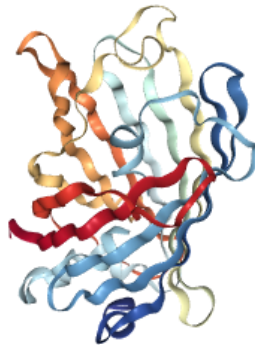
Figure 2: Crystal structure of green fluorescent protein isolated from *A. victoria*[7] . PDB: 1EMB.

proesses[13] . In the past thirty years, GFP has become a frequent and useful tool in biology and has been cited more than 30,000 times. More recently, mutations have been made to GFP to produce GFP-like proteins with other desirable qualities for research and other GFP-like proteins from other organisms have been found and studied. Today, more than 450 entries for GFP-like

proteins currently exist in the Protein Data Bank[2], and fluorescent proteins (FPs) used by many scientists worldwide to track cellular activity.

Fluorescent Proteins in Nature

Fluorescent proteins (FPs) have been discovered in organisms in almost all major phyla including Cnidaria, Ctenophora, Arthopoda, and Chordata[14,15] . Many bioluminescent organisms (about 80% of all known to science) live in the ocean and include fish, jellyfish, squid, bacteria, insects, slam, and sea cacti. Most known fluorescent proteins are found in organisms that are not bioluminescent[4,14,15] . Fluorescent proteins in organisms that are not bioluminescent may have many other functions than to evade predators, attract prey and mates and for communication. Some proposed functions for these FPs include photoprotection of photosynthetic apparatuses, light-induced electron donors, camouflage and as a primitive proton pump[2].

In nature, measurements of fluorescent emission spectra cover the full visible range of light, have different intensities and kinetic profiles[15]. Since most fluorescent organisms live in the ocean and blue light (λ_{max}= ~475nm) travels the farthest in ocean water, most absorption spectra are blue[15]. The second most abundant color of absorption spectra is green and that may be because these species exist in shallow, coastal regions so turbidity from particles scatters blue light so the transmission of longer wavelengths, such as green, are favored[15]. Fluorescence emission in violet, yellow, red and orange occur rarely in the sea and their function and chemistry are poorly understood[14,16] .

FPs derived from corals make up a large portion of the fluorescent proteins isolated from non-bioluminescent organisms. In corals, the function of FPs may be to aid in the photosynthesis of algal endosymbionts and to provide protection from solar radiation[17] . The fact that the

principles of fluorescence in corals are the same, but the molecular implementation is different suggests that fluorescence evolved recently in different phylogenetic groups[17].

Though the GFP from *A. victoria* (*av*GFP) is the most used and studied[1], other fluorescent proteins exist in nature that have been studied and may be more useful for certain studies[18] . An example of this is the similar *Renilla* GFP from the sea pansy that has a 25 percent (GFP-like proteins range from 58-25 percent sequence similarity) sequence similarity to *av*GFP[2,4,18] . On its own, *Renilla* GFP is brighter than *av*GFP, which can be useful for *in vivo* imaging. A *Renilla* GFP – *av*GFP construct has been developed that has an intensity that is brighter than similar firefly luciferase-GFP construct, but the delivery of the coelenterazine substrate must be delivered[18].

Uses of Fluorescent Proteins

I. pH Monitoring

One use of fluorescent proteins is to monitor pH *in vivo*. Many functions of live cells, such as protein structure and steps of metabolism, depend on cellular pH[19] . Enhanced cyan fluorescent protein (ECFP) has been genetically modified to undergo changes to the ECFP lifetime in response to changes smaller than 0.2 in the pH range of 5-7. An advantage of using CFP as opposed to other FPs is that is it suitable for use in a two-color system[19]. Fluorescent tags can be used to monitor pH in cells and in many other types of assays, the two aforementioned FPs are just examples of how they can be used in cells and in other systems.

II. Calcium Indicator

FPs and FRET systems have also been used to determine the presence of calcium in a system. Recalling the consecutive excitation of aequorin and GFP in *A. victoria*, the fluorescence is initiated by the presence of three calcium molecules. In fact, Shimomura paid more attention to aequorin at first due to its ability to indicate the presence of calcium. GFP-aequorin complexes are still used to show increases in cytosolic calcium in a single cell[1]. Calcium plays an important role in the action potential, polarization and subsequent depolarization in neurons and fluorescent probes have been used to detect the influx of calcium ions on the surface of an axon[20] . FPs play an important role in bioimaging. An example of this is the GECO series of Ca^{2+} sensors made up of chimeras of GFP, calmodulin and a peptide derived from myosin light chain kinase[21] . These chimeras are sensitive to calcium concentration and are able to respond very rapidly which is partially the reason that they have gained so much popularity for calcium imaging in olfactory cells and neurons[21] .

A popular type of genetically encoded calcium indicators are GCaMPs. GCaMPs consist of calmodulin (CaM) which interacts with Ca^{2+}/CaM-binding motif M13 from myosin light chain kinase, and a circularly permuted green fluorescent protein[22]. These calcium sensors are particularly useful because interactions between calcium ions and CaM are able to change the protonation state of the FP, resulting in different color emission[6,22]. One disadvantage of GCaMPs is that studies have shown that they are able to impair the health of tissues that are being studied[22].

III. Fluorescence Resonance Energy Transfer

The mechanism through which GFP is able to absorb the blue light emitted from aequorin and emit it at a longer wavelength is through a process called FRET. Fluorescence resonance energy transfer (FRET) is an exchange of energy from a donor fluorophore to an acceptor fluorophore that is nonradiative when the two are within 100Å^4 of each other. The system must also have a donor emission spectrum that overlaps with the acceptors' excitation spectrum. FRET has been used as a highly sensitive method for detecting nitro-aromatic compounds, as calcium ion indicators, and for protease and kinase monitoring[4,23] . In many of these applications, the FRET system works because the change in environment of interest causes a conformational change in the peptide bond that hold the FRET system together, the indication of which is changed fluorescence[4] .

IV. Reporter Gene

GFP was first applied as a reporter gene in *E.coli* and *C. elegans* by Chalfie in 1994[12] , and this is perhaps still the best known application for GFP today. To use GFP as a reporter gene, the DNA sequence must be inserted after a gene of interest that is under the control of an active promotor. GFP will then be coexpressed with the gene of interest. The level of expression directly correlates to the quantitative expression of the gene of interest. Though this technique is popular, it does have some downfalls. Some of the shortcomings of using GFP as a reporter gene include its low sensitivity, slow chromophore formation after translation, and the nonlinearity of the fluorescent signal[4].

FPs may also be fused to a protein of interest, allowing researchers to see where and when a protein is localized in time and space in certain tissues, cells, or organelles[24] . These

fluorescent tags have been useful and popular because fusing them to proteins does not alter its function[1] or disrupt the function of the protein of interest[24] . GFP is commonly used as a fusion tag because its chromophore is able to form via an autocatalytic cyclization *in vivo* and it does not require the presence of a cofactor or a substrate. However, GFP fusion tag activity can be affected by the length of the polypeptide linker[4].

Fluorescent Protein 3-Dimensional Structure

GFP has a 238-amino acid sequence that has a tertiary structure of a can shaped beta barrel[4,11,25] . The wall of the can is composed of eleven beta pleated sheets with an alpha helix running through the middle where the chromophore will form after the posttranslational autocatalytic cyclization. On both ends of the beta barrel are also short alpha helical segments. This tertiary structure has been observed in the crystal structure of all known FPs, though some vary in quaternary structures forming tetramers, dimers and monomers[25] . Further studies have shown that FPs do not have to exist as dimers and dimer presence in crystal structures depends heavily on the growth conditions of the crystal[4] . The height of the GFP beta barrel is 42Å and the diameter is 24Å[4] (Figure 3).
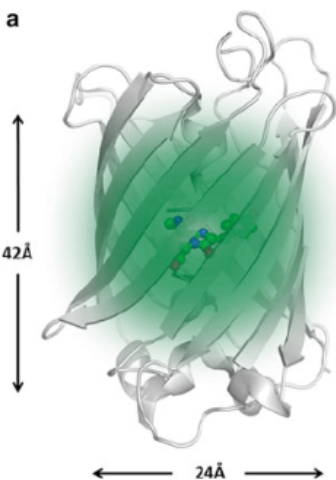
Figure 3: Size of GFP beta barrel[68].

In all fluorescent proteins, the posttranslational chromophore formation requires the presence of oxygen and releases one molecule of water[25] . The beta barrel structure that fluorescent proteins have contributes to FPs resistance to photobleaching and high chemical and thermal stability[26]. The beta barrel also serves to shield the internal helix from the surrounding solvent[24] . The encasing of the chromophore by the beta barrel wall and the alpha helix lid residues protects it from being quenched by oxygen and attack by hydronium ions[4] . Variable residues around the chromophore contribute greatly to the different characteristics of FPs. Unfolding of GFP, which results in the retention of the chromophore but rearrangement of the amino acids surrounding it results in loss of fluorescence, proving that chromophore structure and surrounding amino acids both contribute greatly to fluorescence[24] . Deletion mapping experiments show that residues 2-233 or 7-229 are necessary for chromophore formation and fluorescence in GFP[4] . In GFP, there is also a large ($\sim$135Å)[4] cavity on one side of the chromophore that contains four water molecules, but this is still protected by nearby amino acid residues and is shielded from bulk solvent[4] .

Conserved Residues

Variants of *A. victoria* GFP exist in other organisms and show some structural similarity. In a study of 266 GFP-like proteins[2] , researchers compared and contrasted the 266 crystal structures of fluorescent GFP-like proteins that had been entered into the RCSB Protein Data Bank in 2011. All fluorescent proteins had between 200 and 250 amino acids with an eleven beta strand structure that formed the ß-barrel. All proteins possessed an aromatic residue at the position equivalent to 66 in *av*GFP, and all naturally occurring FPs had a tyrosine in that position, leading researchers to believe that it was also necessary for ancestral function. Since a

tyrosine at this position causes FPs to red shift and play the role of an electron donor that can be activated by light, it is thought that it was the initial function of FPs in nature. The study concluded that all known GFP-like proteins may have emerged from a common ancestor throughout evolution, which is why some residues are so highly conserved and necessary for protein stability and appropriate chromophore formation. Another study found that GFP has evolved at least four times, independently of each other[27]. The lineages don't necessarily correlate with color diversity or non-fluorescence, or organism order making it almost impossible to determine the original common ancestor of GFP.

Although the chromophore forms from amino acid residues 65-67, other amino acids are required for this to occur. In all naturally occurring FPs, the tripeptide sequence that goes on to form the chromophore is X-Tyr-Gly, where the presence of glycine in the third position is essential, suggesting that flexibility is required in that position to provide the necessary kinked conformation[4,28]. The residue at position 65 is variable in fluorescent proteins, but 66 is always an aromatic amino acid (tyrosine in nature), and residue 67 must be a glycine[2]. One study compared precyclized forms of 28 naturally occurring GFP-like proteins in the Protein Data Bank and concluded that in all FPs mature forms were more rigid than immature forms and that the compaction that occurs with maturation may play a role in chromophore formation[2]. It was found that other residues that are highly conserved are those in the lids of the barrel. They thought to be necessary for chromophore formation because they act as hinges and/or folding nuclei[29] (Figure 4). The chromophore is protected from bulk solvents by lid residues and comprise a large portion of the most conserved residues among FPs[2]. The rest are amino acid residues that make up the turns between the ß-sheets (Figure 4). Conservation of lid residues might be linked to some unknown binding function of FPs or folding[2]. Glycine 222 and arginine

96 are highly conserved among FPs because of them important catalytic role they play in the first

cyclization step of the mechanism of chromophore formation[28] (Figure 5).
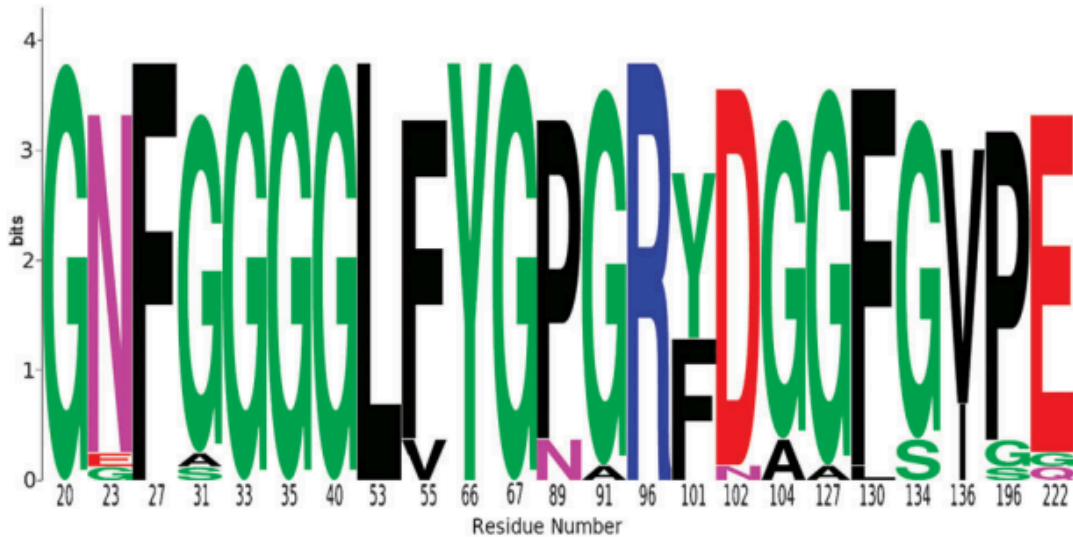
Figure 4: Amino acid diversity among most conserved residues of wild-type GFP-like structure[10] deposited in the PDB. Among the most conserved are glycines at position 31, 33, and 35. The roles of some conserved residues are known, like the catalytic residues at positions 222 and 96, or the chromophore forming residues at positions 66 and 67.

Figure 5: Location of conserved residues in GFP[2]. A large portion of the conserved residues serve as hinges (green dots in structure on left) and in the chromophore. The glycine residues (shown in red) are the only conserved residues with unknown function and the only conserved residues on the beta strands.

Glycine Residues in β-Sheets

Beta pleated sheets usually exist in nature with hydrogen bonding interactions that hold the sheets together. In one study[30], to determine the favorability of one amin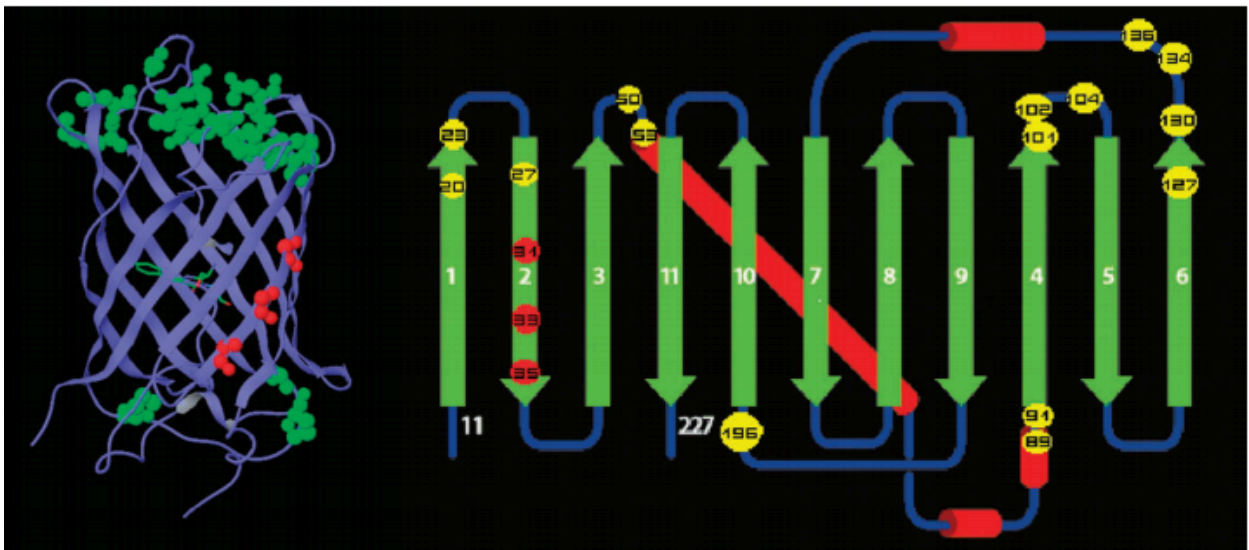o acid occurring in a beta sheet versus another, the difference in Gibbs-Hemholtz free energy of a beta sheet protein with alanine at an important position was compared with that of mutant with each of the 19 other naturally occurring amino acids at this position ($\Delta\Delta G$). A negative $\Delta\Delta G$ value corresponds to an overall decrease in stability of the sheet, while a positive value corresponds to a increase in stability ($\Delta\Delta G = \Delta G_{mutant} - \Delta G_{wt}$). When compared to alanine, the most simple, chiral amino acid, glycine generates a $\Delta G = -1.2$, indicating a decrease in stability when present in beta sheets[30] (Table 1). For this reason, glycine residues are not very frequent in beta sheets, which may be due to their flexibility.

Table 1: Relative stability of amino acid residues occurring in a beta sheet to an alanine residue at the same position[32].

TABLE 1 $\Delta\Delta G$ values for β-sheet formation relative to alanine, thermal melting temperatures (T_m), and relative binding constants to human Fc for the AASS-53Xaa proteins normalized to the binding of AASS-53Thr

Amino acid	$\Delta\Delta G$ (kcal mol^{-1})	T_m(°C)	$K_a/K_a^{AASS-53Thr}$
Thr	1.1	53.7	1.0
Ile	1.0	53.0	1.0
Tyr	0.96	52.5	1.0
Phe	0.86	51.6	1.0
Val	0.82	51.2	1.1
Met	0.72	50.2	1.0
Ser	0.70	50.1	1.1
Trp	0.54	48.7	1.1
Cys	0.52	48.5	1.1
Leu	0.51	48.4	1.0
Arg	0.45	47.9	1.0
Lys	0.27	46.3	1.1
Gln	0.23	45.8	1.2
Glu	0.01	44.0	1.1
Ala	0.00	43.8	1.1
His	−0.02	43.3	1.0
Asn	−0.08	42.9	1.0
Asp	−0.94	35.2	1.1
Gly	−1.2	30.2	1.0
Pro	<−3	<0	0

When they are present, glycine residues can cause kinks[31,32] and they can have a left-hand twist without disrupting the hydrogen bond pattern[31]. The presence of glycine in a beta barrel increases the local curvature, forming a corner in an otherwise circular space and they also alleviate steric hindrance[31]. In GFP, there are four glycine residues in the second strand of the beta barrel. This could be one of the possible functions of these residues.

In parallel ß-sheets, the C – H bond of glycine is too strong to be damaged by many weaker oxidizing radicals such as superoxide and peroxyl[33]. Since glycine is known to be intrinsically destabilizing in ß-sheets (Table 1), this can be counteracted by strand pairing with aromatic residues in both parallel and antiparallel sheets, which is reflected by statistical surveys[34].

Chromophore Formation

GFP chromophore formation is the result of an autocatalytic posttranslational cyclization that does not require any involvement of cofactors of enzymatic components (Figure 6), which is
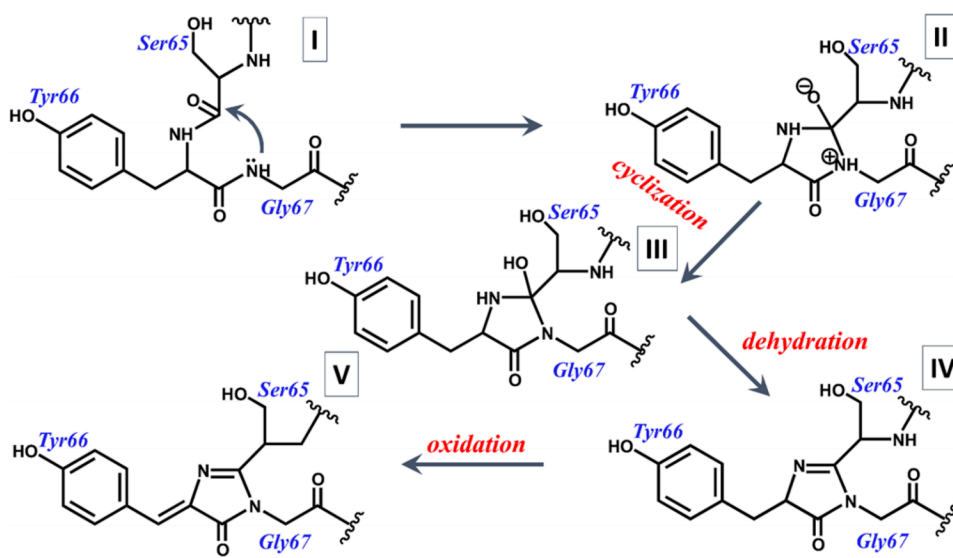

Figure 6: The proposed cyclization-dehydration-oxidation mechanism for GFP's tripeptide chromophore formation[68].

rarely observed outside the FPs. A major requirement for this autocatalytic chromophore formation is that the nitrogen of Gly 67 and carbonyl carbon of Ser 65 must be in close proximity; in immature GFP, the distance between the two is shorter than the sum of their covalent radii[35]. Another requirement for chromophore formation is the 'tight turn' orientation between the amide nitrogen of Gly67 and the carbonyl carbon of Ser 65[36] . This tight turn restricts the conformational space and keeps the residues in place for the cyclization step of the mechanism. GFP mutants that exhibit desirable properties such as thermostability and increased solubility, contain mutations that decrease the distance between these two atoms[37]. Fluorescence is not observed in wild type GFP until 90 minutes to four hours after protein isolation[37] . It is thought that this is because folding occurs rapidly, but chromophore formation and oxidation occur much slower.

Though there has been a considerable amount of effort in deducing the mechanism for chromophore formation, it is not completely understood, although it is confirmed that it requires imidazolinone core ring structure[25] . It is the imidazolinone core ring structure (the outer structure may differ depending on the amino acids that went on to form the chromophore) possesses alternating double and single bonds in the bridge region extend the electron delocalization from the carbonyl of the imidazolinone to the phenolate. This creates a pi conjugated system which provides the efficient visible light absorption that is characteristic of fluorescent proteins. The more extended the conjugated the pi system is, the more red shifted the fluorescence of the protein will be[28] . A study set out to determine the entire mechanism of chromophore formation[38] , and results supported the cyclization, dehydration, oxidation mechanism of chromophore formation using a GFP structure most similar to the *av*GFP. Another study showed that the arginine residue at position 96 plays an important role in chromophore

maturation[37] , that isn't necessarily due to the structure of arginine, but more so to its positive

charge[39] . This was determined by mutating R96 to an alanine, which lacked the large side chain

and the charge of arginine. They did this to see how the loss of charge and open space would

alter the hydrogen bond network around the chromophore and allow for waters to enter this

space. This mutant took considerably longer for chromophore formation. This meant that the

arginine at position 96 was not necessary for chromophore formation, but it plays a significant

role in the speed of chromophore formation in wild type GFP. An R96M mutant was also made

to mimic the steric hindrance of an arginine, without the positive charge. In this mutant,

chromophore formation took up to three months. This showed that the positive charge of

arginine also played a role in chromophore formation. However, when the R96K mutation was

made the chromophore maturation time was similar to that of wild-type GFP. This meant that the

identity of the amino acid at position 96 is not as important as the positive charge it provides,

because lysine and arginine are similar in size and charge.

Some FPs have an additional hydrolysis step in their mechanism in addition to the

cyclization, dehydration, and oxidation mechanism that is best supported for GFP[28,38] . In the

cyclization step, the chromophore forming tripeptide cyclizes. This is then followed by

deprotonation of the alpha carbon of tyrosine 66, that is promoted by the electrostatic catalysis of

arginine 96 and the base catalysis of glutamine 222[28,37] . Glu222 is known to play a stabilizing

role in the presence of excessive x-ray irradiation, and has been proposed to contribute to rigidity

of the chromophore and restricts the chromophore's flexibility, preventing it from nonradiative

deactivation of the excited state[28]. The glycine residue at position 67 allows the alpha helix to

form the necessary kinked conformation, which allows for positioning of the amide nitrogen of

glycine 67 close to the carbonyl carbon of the residue at position 65 for the nucleophilic attack

that follows[28]. A study has suggested that the chromophore formation mechanism may have an addition conjugation trapping mechanism, that serves to lower the energetic cost of the cyclization of the peptide[40] . This part of the mechanism may be necessary because the cyclization step of the mechanism is thermodynamically unfavorable. Cyclization is followed by dehydration, and a water molecule is lost, forming a double bond in the ring structure that will go on to be essential for the conjugation of the system.

FP Folding

The energy landscape theory for protein folding is quantified by the ratio of the transition temperature of the folding to the glass-transition temperature, giving rise to the principle of minimum frustration[41] . The principle of minimum frustration is often depicted as a funnel on a graph of energy, the bottom of which is the native fold of the protein because the native fold is thought to be the most thermodynamically favorable structure. In fluorescent proteins, the most important requirement for proper chromophore formation is correct and complete protein folding. Correct folding consists of bending of the alpha helix in a way that results in the three amino acids that go on to form the chromophore folding inward into barrel and arranging them in a way that is best suited for chromophore synthesis[28], and subsequently forming a more rigid and compact beta-barrel.

However, proper folding occurs less efficiently at temperatures higher than which the *A. victoria* jellyfish experiences in the wild, like the temperature of the human body[42]. This does make FPs useful for studying chaperonin activity when used as a fluorescent tag because fluorescence can be observed when proper folding occurs. GFP is known to fold and unfold

much slower than other globular proteins of comparable size, which is why many FP mutants have been developed that fold faster and more readily[4,25,42] .

A study found that after unfolding then refolding, the refolding of GFP is only partially reversible[42] . A proposed explanation of the inability of GFP to refold is because of GFP does not reach equilibrium within a few days due to the slow kinetics in the intermediate concentration condition of guanidine-HCl that was used to denature the protein in this experiment. At lower concentrations of guanidine-HCl, the fluorescence recovery was not 100%, even though folding was faster. The same irreversibility was observed at temperatures above 50 ˚C, but GFP was still stable. Some proposed explanations for this observation are: (i) since GFP has two cysteine residues that are not bonded in the native fold, there is a possibility that they bond together during refolding, resulting in improper conformations; (ii) that the chromophore may be destroyed if the protein remains unfolded for too long and may not be reformed during refolding; and (iii) there may be a non-fluorescent conformation that is thermodynamically similar to the native fold; i.e. a dip in the wall of the funnel close to the bottom[42] (Figure 7).
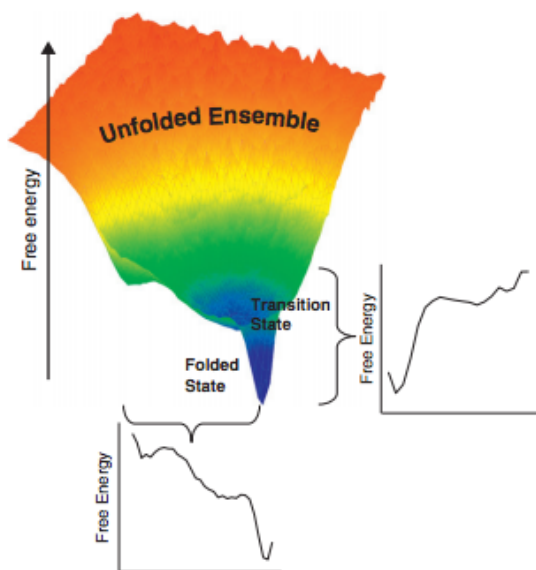

Figure 7: View of protein folding from the top of folding funnel[43].

Another study proposed a four-step process of GFP folding based on the similar cycle3 mutant (a rapidly folding FP with similar structure and characteristics to that of GFP) using acid denaturation[44] . The burst phase corresponded with the collapse of the GFP molecule that is not specific, and tryptophan 57 is shielded from solvent. The very fast phase corresponded with a folding intermediate that has similar characteristics of the molten globule state during folding. The fast phase and parts of the medium phase corresponded with refolding from the intermediate and at these phases the proline configurations are the same as they are in the native fold of GFP. The remaining portion of the medium phase and the slow phase corresponded to the slow proline isomerization in the intermediate state that is known to make up a major portion of the observed kinetics involved with folding of GFP.

Glycine Residues in Fluorescent Proteins

The glycine residue at position 35 is highly conserved in all the GFP-like structures, and all the proteins that were studied and contrasted in the Protein Data Bank, and many others, all have a glycine in this postion[2,4] (Figure 4). The way that this glycine contributes to chromophore maturation is poorly understood, but is necessary for fluorescence. However, studies have shown that the glycine at position 35 is not involved in chromophore formation or the pore that is involved chromophore maturation[2] . The glycine residues at positions 31, 33, and 35 are the only conserved residues located on a beta sheet that are not directly involved in of chromophore formation[40] .

One study found that the glycine residue at position 177 lowers steric hindrance because glycine can adopt a large range of the dihedral angles in a polypeptide chain and that it contributes to the intensity of fluorescence[45] . In another study, researchers were able to engineer

a functionally fluorescent protein that did not contain the glycine residue at position 67 (rather, an alanine)[46]. In many other studies, this exact residue has been deemed essential for fluorescence in GFP, and all other fluorescent proteins known to science. This raises some questions as far as what role the residue at position 67 plays in chromophore formation, in addition to contributing to the tight turn distance and conjugation trapping, and how it interacts with some hydrophobic amino acids in the beta barrel because several hydrophobic residues had to be removed for fluorescence. In another study, insertion of glycine residues between Asn144 and Tyr125 to make space for water molecules near the chromophore resulted in a structure was still functionally fluorescent, but fluorescence intensity was dependent on hydrostatic pressure[47].

The glycine residues in GFP and its beta pleated sheets were used to successfully design a *de novo* FP using computational methods to anticipate structure and function[32]. Designing a protein that only contains beta sheets is more difficult than designing other proteins due to the greater fraction of non-local interactions. If these interactions do not exist, protein folding can be greatly affected. Researchers in this paper found that glycine kinks can be used to reduce the strain in beta sheets.

Overall, the exact role that glycines 31, 33, and 35 play in FPs is poorly understood although they are highly conserved.

Loss of Function Mutations

The G35S GFP mutant generated showed fluorescence in seedlings and very low levels of protein after purification. All other proteins with a mutation to the glycine residues at positions 31, 33, and 35 resulted in a lack of visible fluorescence or detectable protein, despite

normal transcription of the GFP mRNA[3]. This result led researchers to believe that these residues help to facilitate protein folding and stability in GFP.

GFP Local Fitness

Genotypes determine the phenotypes and functions of proteins, which directly correlates with how 'fit' an organism or protein would be to survive and thrive over the course of evolution. The shapes of the maps generated by these studies are effected by factors, including genetic robustness[48]. Traditionally the impact of single point mutations, and their subsequent combined effect are used to generate fitness landscapes. However, researchers in one study used high throughput screening techniques to generate a full local fitness landscape of GFP based on over 50,000 mutants[49]. About 75% of mutations resulted in a protein with no detectable fluorescence, further showing that not only residues that go on to form the chromophore must be conserved for chromophore formation[2]. On average of 3.7 mutations were made per gene sequence. The mutations included single, double and multiple mutations, although 9.4% of single mutations resulted in a fivefold decrease in fluorescence[49]. This study also concluded that mutations that had a strong effect on fluorescence went on to code for amino acids whose side chains were oriented inward toward the chromophore[49]. One of the fitness landscapes generated from this study showed that mutation from a glycine to a cysteine at position 35, though mutations at this position are never tolerated, would result in a protein with some level of detectable fluorescence. These two amino acids are very structurally different; glycine is flexible and non-polar, while cysteine has a sulfur atom and is able to form disulfide bonds with other cysteine residues. The idea that these two residues are interchangeable at this highly conserved place was of interest for further investigation for my biochemistry semester project.

The result of my semester project was that the G35C mutant produced no detectable light a week after purification, and the LCMS data proved that a chromophore was never even formed in this mutant. There seems to be an issue with folding and solubility based on the gel analysis and low yield as compared to wild-type. The protein was successfully purified and the sequence showed that the indended muation and no other was made. Also, it would be worth checking if after a significant amount of time, the chromophore did form because there is a posibility that the glycine at position 35 plays a catalytic role.

Objective

The objective of this honors study and 3 related independent studies was to use computational methods to find out why the glycines at positions 31, 33, and 35 are so highly conserved in all fluorescent proteins and what role they play in chromophore formation, folding and stabilizing the protein.

Precyclized structures have been used in our simulations because there has been evidence that the role conserved glycine residues play are in protein folding[3] or chromophore formation, which occurs prior to cyclization. Although G35S[3] has been found to be somewhat fluorescent, we chose to make a less aggressive mutation, G35A, to study because alanine is the most structurally similar to glycine.

Computational Methods

Molecular Mechanics

Since green fluorescent protein has spectral properties that are extremely similar in both

its solid state and its aqueous phase, and because all FPs in the PDB are structurally similar, it is

ideal for computational studies because it can be assumed that there are very few conformational

changes between these two states[50] . The Schrodinger equation (Figure 8), is useful to determine

the changes of a quantum system over time, but it cannot be solved for a many electron system,

such as a protein[51,52] . When applied to a single hydrogen atom, a particle in three dimensions,

$$\left[-\frac{1}{2}\nabla^2 - \frac{Z}{r} \right] \psi(r) = E\psi(r)$$

Figure 8: Schrodinger equation[40] .

the portion in parenthesis accounts for the kinetic
and potential energy of an electron at some
distance, r, from the nucleus which has a charge
of Z[52] . This does not work for atoms with many
electrons. Instead, molecular mechanics is used

for GFP and many other proteins because it approximates the quantum mechanical energy

surface with a classical mechanical model, creating less computational cost of simulations on a

large system by orders of magnitude[53] . All molecular mechanics methods treat atoms as

particles that are assigned a radius and a net charge and the bonded interactions are springs, with

the calculated length is treated as the equilibrium length.

To investigate the static mechanical properties of GFP, large-scale nonlinear optimization

of the molecular mechanics model is necessary. The equilibrium configuration for a protein, or

any system of atoms, is determined by minimizing the potential energy according to the

molecular mechanics model[53] . To determine the force on each particle, a force field must be

used. In this study the OPLS3 (Optimized Potential for Liquid Simulations) force field[54] was used. The OPLS3 equation (Figure 9) makes it so that the stretching and bend terms of atoms are fit to quantum chemical data, van der Waals terms and core charge sets are obtained from liquid simulations, bond charge corrections are defined and tested, and torsional parameters are fit to quantum data. Some key improvements from earlier OPLS force fields are that the entire force field is improved, there is more use of high quality quantum chemical calculations to yield valence and torsional parameters that are more accurate, improvements to the charge model, and the models have been tested against more large data sets[54].

$$E = \sum_{i<j} [q_i q_j e^2 / r_{ij} + 4\varepsilon_{ij}(\sigma_{ij}^{12}/r_{ij}^{12} - \sigma_{ij}^{6}/r_{ij}^{6})]f_{ij} + \sum_{bonds} K_r(r - r_{eq})^2 + \sum_{angles} K_\theta(\theta - \theta_{eq})^2$$

$$+ \sum_{dihedrals} \left[\frac{V_1}{2}(1 + \cos\varphi) + \frac{V_2}{2}(1 - \cos 2\varphi) + \frac{V_3}{2}(1 + \cos 3\varphi) + \frac{V_4}{2}(1 - \cos 4\varphi) \right]$$

Figure 9: OPLS3 equations[54].

Molecular Dynamics

A molecular dynamics (MD) simulation is a computational method that mimics the actions of atoms at a certain potential energy and quantifies the physical movement of atoms and molecules in an ordered system[55]. Molecular dynamics simulations can be used to characterize structural dynamics information for a specific system. The types of information include, but are not limited to: dielectric properties, system relaxation, elastic and plastic mechanical properties, structural transformations, protein folding, equilibrium conformation, structural transformations, docking of molecules and much more[56]. MD simulations use an algorithm (Figure 10) and high performance computing to mimic the behavior of atoms of a system at a specific time. The
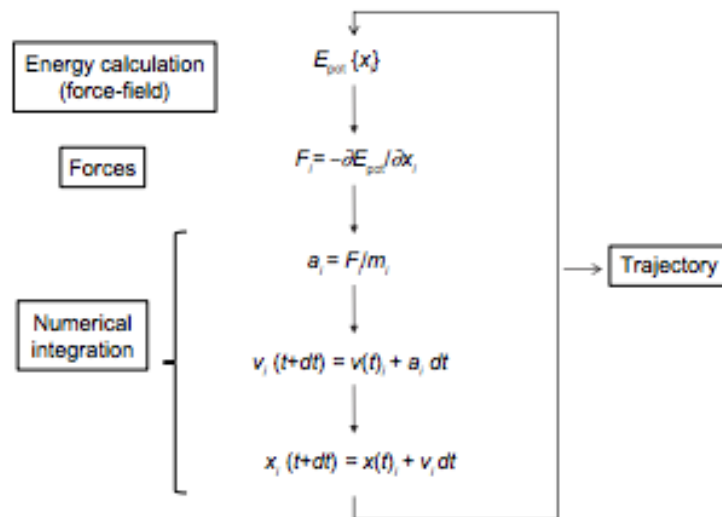
Figure 3 Molecular dynamics basic algorithm.
Notes: The simulation output, the trajectory, is an ordered list of 3N atom coordinates for each simulation time (or snapshot).
Abbreviations: E_{pot}, potential energy; t, simulation time; dt, iteration time; For each spatial coordinate of the N simulated atoms (i): x, atom coordinate; F, forces component; a, acceleration; m, atom mass; v, velocity.

Figure 10: Molecular dynamics simulation alogrithm[57] .

potential energy function is used to calculate the force experienced by a certain atom relative to the positions of other atoms, by using Newton's second law to determine how forces affect the motions of the atoms[56] . These calculations have to be done for small, discrete time steps and is done with a simulated force field to mimic the effects of forces on the movement of atoms, in this case, OPLS3. Ordinary differential equations must be used to solve for the coordinates and the velocity of each atoms via 'time symmetric' integrations methods[56,57]. When the data are plotted against each other, the simulation samples the Botlzmann distribution. When the protein does reach an energy minimum, it does not stay there and usually simulations cannot not go long enough to reach all of the thermodynamically favorable arrangemnts[57,58].

Maestro and Desmond

Desmond is a MD code that was created by D.E. Shaw Research and is used on Linux operating systems. MD simulations can be computationally expensive, especially when dealing with larger systems. For this reason, Desmond can operate with graphics processing unit (GPU) technology, which can run up to 200 times faster than CPU time, allowing the user to simulate for larger time scales[59] . However, we run Desmond on 32 parallel processors. Desmond works well with the OPLS3 force field and is able to calculate energies and forces for it. Desmond is also able to calculate relative free energies of binding and solvation free energies for solvent simulations[59] . In this study the solvent boundary conditions are orthorhombic. Conveniently, all of this quantitative data can be viewed within Maestro, the graphic user interface for Desmond.

A. Assessing the Quality of the MD Simulations

Protein Root-Mean Squared and Root-Mean Fluctuation Graphs

Studying protein behavior during a molecular dynamics simulation can be difficult as data is continually changing per time step. One way to analyze a protein's behavior and to compare the behavior of two proteins is by having a look at the graph generated by comparing a reference structure to all others using the root-mean squared deviation (Figure 11) over the time

$$\text{RMSD}(t) = \left[\frac{1}{M} \sum_{i=1}^{N} m_i |\mathbf{r}_i(t) - \mathbf{r}_i^{\text{ref}}|^2 \right]^{1/2}$$

Figure 11: P-RMSD Formula[44].

29

of the simulation[60] . This graph (Figure 12) can give information about the proteins' structural

conformation throughout the simulation, protein stability and whether or not the simulation has

equilibrated[60,61] . If the graph for this simulation is still fluctuating by more than three Ångstroms

at the end of the simulation, this is a pretty good indicator that the protein is undergoing a large

conformational change, so the simulation should be run for longer. The data obtained from P-

RMSD can also give information about flexible and rigid substructures within a protein during

the simulation[61] . Root-mean squared fluctuations , on the other hand, give information about the

deviation of a reference atom, in this case C^α, and a reference position[61] . The difference between

P-RMSD and P-RMSF is that P-RMSF is averaged over time, giving a value for each particle

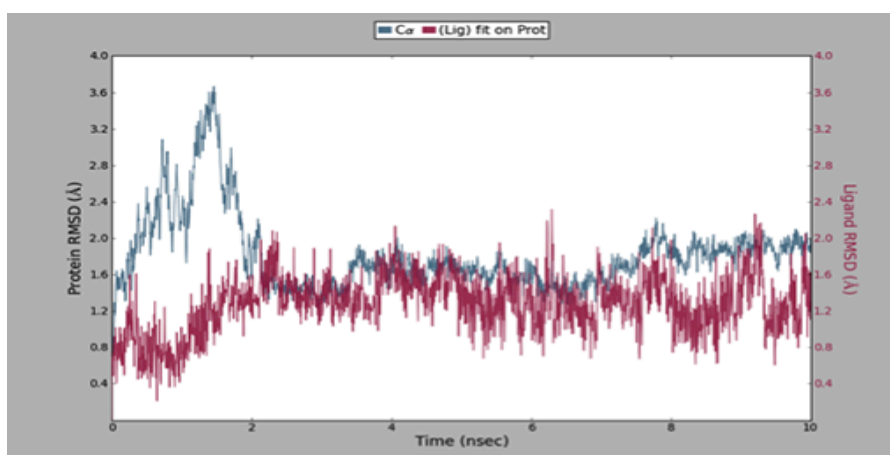and RMSD is the average over each particle per a given time step[61] .

Figure 12: Example graph of P-RMSD. Source: https://www.schrodinger.com/newsletters/introducing-sid-simulation-interactions-diagram

Starting Structure

This past summer, student researchers Justin Nwafor and Christian Salguero prepared the starting structure used for this thesis study. To do this, they began with an immature, precyclized structure of GFP from the Protein Data Bank, 2AWJ. 2AWJ is a precyclized GFP R96M mutant. Once this structure was imported to Maestro, the mutations in 2AWJ were reverted to the original amino acids that are present in wild type avGFP (L64P, T65S, M96R, S99P, T153M, A163V). This was to generate a structure that was a precyclized version of wild type GFP. Next, a 10,000 step torsional sampling conformational search was ran on that structure. The newly generated structure then had another 11,000 step conformational search ran on it with a distance constraint between residue 66 and 96. The structure generated from the second search had a 2,000 step large-scale low-mode sampling conformational search ran on it. Finally, the most stable structure generated from this search was used for a 50nsec MD simulation under default conditions in Desmond.

Last fall, this structure was then used to mutate the glycine at position 35 to an alanine in Maestro, and a 100nsec MD simulation was ran. The simulation and trajectory file output was used to generate data for this work.

Results and Discussion

*In this section, I have used our file names to describe structures.

Preparing the Starting Structure for Molecular Dynamics

The structure 2AWJ-hbond-opt was generated by taking the structure that was the product of the summer research of Justin Nwafor and Christian Salguero, our starting structure, and preparing the protein in Maestro using software defaults. The protein structure was preprocessed to correct the bonding information, cap the termini, create disulfide bonds and to remove the waters, sulfates, and methanol molecules. The hydrogen bond network in the protein was optimized for the protein (hence, 'h-bond-opt'), which may have reoriented hydroxyl and amino terminals of some residues. A model water system was built for this protein with predefined solvent conditions and an orthorhombic box with minimized volume in 0.15M NaCl. After this, a 25,000-step large scale, low mode search was preformed to generate the protein conformation with the lowest energy. A large scale low mode search is a type of conformational search that locates the energy minima for a structure by finding a local energy minima, then using this structure search along low-mode vibrations to generate an even lower energy minima. This searching only continues for the assigned number of steps because this process could theoretically go on forever[62].

Quality of our 2AWJ-G35A 100nsec Molecular Dynamics Simulation

After running a 100 nanosecond MD simulation for 2AWJ-G35A-100nsec-mdsim, the data was exported to Desmond's Simulation Interactions Diagram module. The program generated the P-RMSD and P-RMSF graphs for this structure. The P-RMSF graph (Figure 13A)
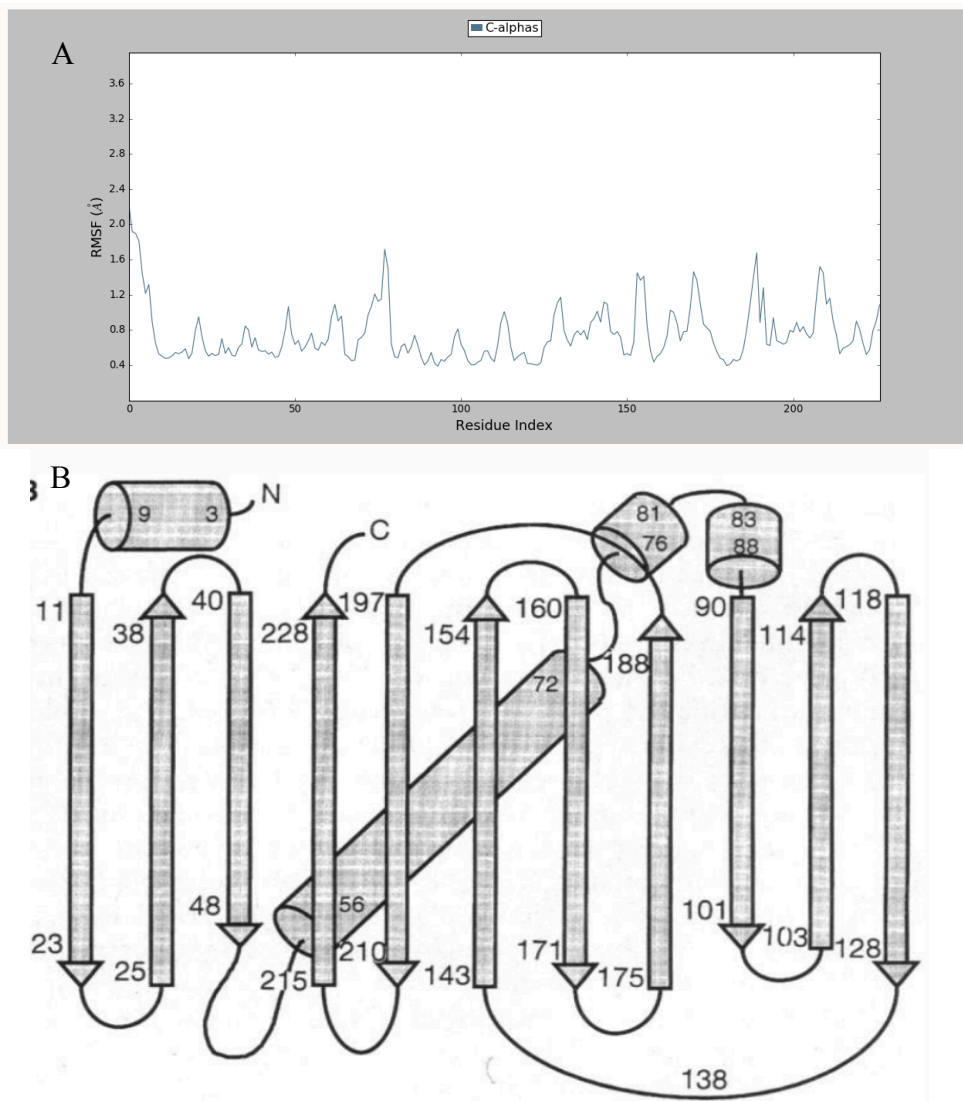
Figure 13: Amino acid residue location and fluctuation. A: P-RMSF graph for 2AWJ-G35A 100nsec MD Simulation. The fluctuations for this structure range from 0.4Å to 2.0Å. Peaks at the alpha and amino terminals are to be expected. B: Amino acid location on secondary elements on the fold of GFP[11].

shows which atoms are fluctuating the most. In this case, the atoms that do fluctuate significantly

are all around 1.6Å with no real outliers which is expected for a protein like GFP, with a distinct

beta barrel shape because the rigid shape does not allow for much movement of alpha carbons.

Figure 13B shows the location of the amino acid residues in respect to the overall fold and

secondary structures of the protein. The N terminus, specifically residue 1, fluctuates the most

over the course of the simulation, with a fluctuation of about 2.2Å. The carboxy terminus does

not fluctuate as much, with a fluctuation of less than 1.2Å. The end of the alpha helix that runs

through the center of the helix forms a small alpha helix that serves as a lid on the end of the

barrel with the termini (residues 76-81). These residues fluctuate from 0.4 to 1.0Å. Residues

156-159 form a loop on the termini end of the barrel, these residues fluctuate from 0.6Å to about

1.6Å. The residues from 187 to 192 form part of a lid loop on the termini capped end of the

barrel and fluctuate from 0.6Å to about 1.6Å also. On the P-RMSF graph, a lot of the larger

fluctuations are from residues that are a part of the lid forming section on the termini capped end

of the barrel. This is to be expected because the beta barrel structure is far more rigid and is

known to have an intricate hydrogen bonding network with the chromophore. Residue 35 does

not have significant fluctuation over the course of the simulation. The P-RMSF noticeably

fluctuates the least in the region that the alpha helix is located. This is to be expected because the

alpha helix has a very specific structure with the tight turn[35] and intricate hydrogen bonding

network which is necessary for chromophore formation. This is a good indication that even when

there is alanine at position 35, the alpha helix structure is still rigid.

    The P-RMSD graph (Figure 14) shows that after about 20nsec of simulation, the protein

undergoes a small change in conformation. After, the graph plateaus for the remainder of the

simulation, giving confidence that the structure is not undergoing a large conformational change.
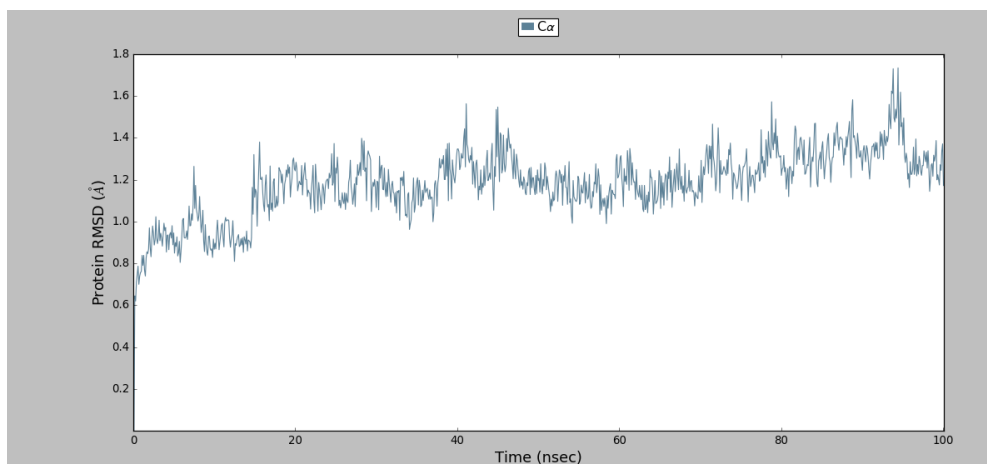

Figure 14: P-RMSD graph for 2AWJ-G35A 100nsec MD Simulation. The RMSD values for this structure range from 0.6Å to about 1.7Å. The protein underwent a conformational change and stabilized at around the 20nsec point in the simulation.

Any fluctuations of this graph are negligible due to the small scale of the y-axis; all changes are under 3Å.

B. Structural Differences

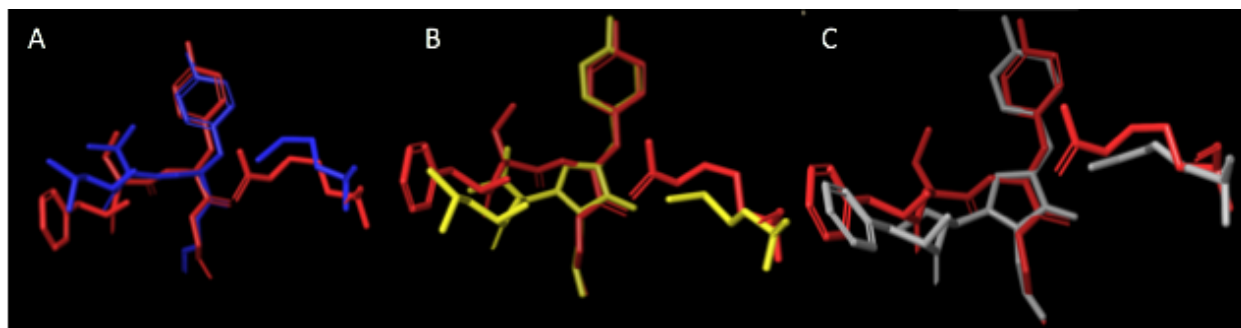*I. Superimposition of 2AWJ-hbond-opt with Similar Fluorescent Proteins in the Protein Data Bank*

Figure 15: (A) Superimposition of precyclized 2AWJ-hbond-opt-25000 steps variant (red) and precyclized 2AWJ-R96M (blue). Image shows that the residues at position 96 hardly overlap, but the backbone where the chromophore would form in GFP overlaps well. (B) Superimposition of precyclized 2AWJ-hbond-opt-25000 steps variant (red) and mature 2AWK-R96M variant. The tyrosine rings and the residues at position 96 overlap well, but the glycine in 2AWJ protrudes where the planar five membered is located in the mature 2AWK. (C) Superimposition of precyclized 2AWJ-hbond-opt-25000 steps variant (red) and cyclized 1EMB WT GFP. This figure shows that the backbone of 2AWJ aligns with where the five membered ring forms in mature GFP, however, the phenylalanine side chains have different orientations in space.

1. 2AWJ-hbond-opt and 2AWJ

2AWJ-hbond-opt and 2AWJ[39] are the two most similar structures as one was derived from the other, so they superimpose the best (Figure 15A). In 2AWJ, the chromophore would be formed via the peptide chain leucine-threonine-tyrosine-glycine if allowed to cyclize. The chromophore of 2AWJ was not allowed to form because the precyclized structure was trapped to help determine the role position 96 plays in chromophore maturation[39] . The corresponding

sequence in 2AWJ-hbond-opt is phenylalanine-serine-tyrosine-glycine. The carbons in the backbone of residues 65-67 of both molecules were superimposed in addition to the nitrogen and the carbonyl oxygen. The side chains for residues 66 and 67 were also superimposed. In 2AWJ-hbond-opt the residue at position 96 is an arginine and in 2AWJ the residue at position 96 is a methionine. For these two residues, the backbone atoms alone were superimposed. For all superimposed atoms, hydrogens were excluded. The remaining two residues that were not superimposed still have similar orientations (Figure 15A), showing that these two molecules are indeed quite similar. The root mean squared value for the superimposition of these two molecules is 0.8625Å. The maximum difference between two atoms is 1.7347Å and that occurs between the carbonyl atoms in the amino acid backbone of the glycine residues.

2. 2AWJ-hbond-opt and 2AWK

  The structure 2AWK[39] is a GFP R96M mutant with a mature chromophore that was used to determine the role that arginine at the 96 position plays in chromophore formation, by mutating it to a methionine. In 2AWK (yellow) the peptide sequence that forms the chromophore is threonine-tyrosine-glycine and the corresponding sequence in 2AWJ-hbond-opt is phenylalanine-serine-tyrosine-glycine. In 2AWJ-hbond-opt, there is an arginine at position 96 and in 2AWK there is a methionine residue. All atoms of the tyrosine residues were superimposed and all of the atoms in the glycine residues were superimposed, including the oxygen atoms, even though one was a part of a carbonyl and the other was a part of an alcohol. The backbone atoms in 2AWJ that correspond with the formation of the lower portion in the five membered ring were also superimposed, excluding the carbonyl oxygen because in 2AWK this oxygen is lost as water during chromophore formation (Figure 15B). The residues at position 96

for these two molecules only had the two carbons and the nitrogen in the backbone superimposed. The carbonyl oxygen atoms were not superimposed. The chromophore region of these two molecules superimpose relatively well, but the residue at position 96 seems to be oriented quite differently. The root mean squared value for these two molecules is 0.5911Å. The maximum difference between two atoms is 1.5486Å and that is between the two carbons in the benzene ring that are one carbon away from the carbon with the alcohol substituent.

3. 2AWJ-hbond-opt and 1EMB

Of all of the GFP variants, 2AWJ-hbond-opt and 1EMB (grey) are the most different of the three. In 1EMB the chromophore is formed from the peptide chain serine-tyrosine-glycine and the corresponding amino acids in 2AWJ-hbond-opt are phenylalanine-serine-tyrosine-glycine. Both structures have an arginine residue at position 96. In these two structures, the all atoms expect hydrogens in the side chain groups of corresponding tyrosine residues were superimposed. The carbons and nitrogen in the amino acid skeleton of tyrosine were also superimposed in the backbone. Again, the oxygen molecules were superimposed even though one was a part of a carbonyl, and the other was a part of an alcohol. The backbone molecules (two carbons and one nitrogen) of the corresponding glycine residues were also superimposed. The atoms that corresponded with the five-membered ring were also overlapped. No other atoms in the chromophore sequence were overlapped because their orientation was much too different. The arginine residues at position 96 only had the nitrogen and the two carbons of the backbone superimposed, but the side chains were in different orientations in space (Figure 15C). No hydrogen atoms for these two proteins were overlapped. The root mean square value for the superimposition of these two proteins was 0.6729Å. The maximum difference between two

atoms was 1.4888Å, and that occurred between the two carbons that are two carbons away from the alcohol substituent on the benzene ring of tyrosine.
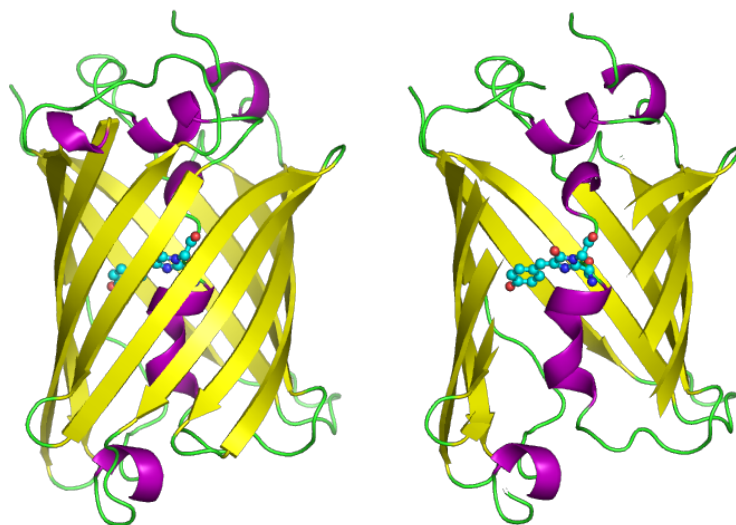

Figure 16: The crystal structure of GFP with emphasis on the alpha helix that runs through the center of the beta barrel. PDB 1GFL[56].

II. Hydrogen Bond Distances For Starter Structure

The alpha helix that runs through the center of GFP (Figure 16) doesn't form many hydrogen bonds, and this lack of hydrogen bond formation is thought to contribute to the low interconversion energy barriers[40] . The hydrogen bonds that do form in the alpha helix that runs down the center of the precyclized structure must be broken during chromophore formation, which can be energetically costly. To account for this, the protein architecture creates a dramatic bend in the alpha helix where the chromophore will form, which removes hydrogen bonds, lowering the energetic cost of peptide cyclization[40], see Figure 18 this called a conjugation trapping mechanism. After chromophore formation, an intricate hydrogen bond network must reform (Figure 17), which play an important role in photophysics, specifically the two protonation states of GFP[1,11]. The atoms that form hydrogen bonds in wild type GFP were
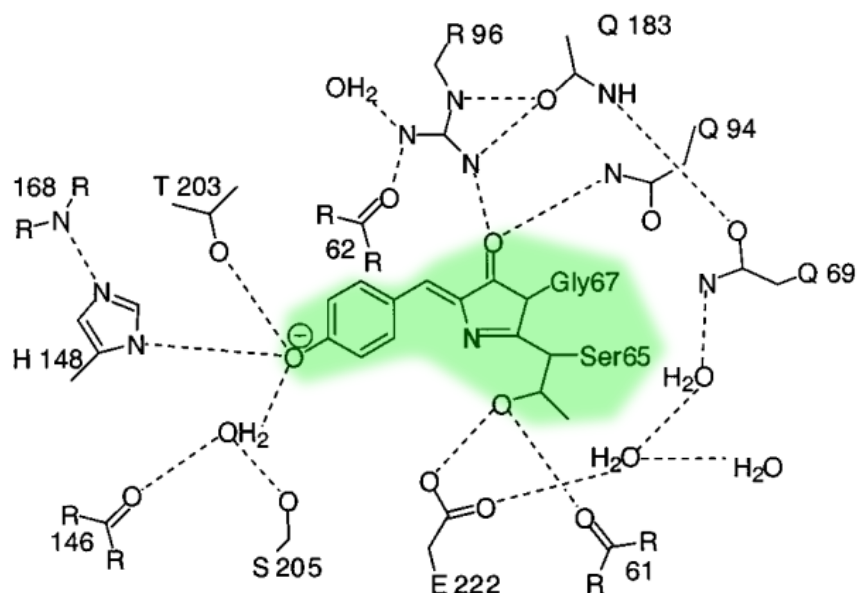
Figure 17: Hydrogen bond network around chromophore in GFP[1,11]. The chromophore (green) forms hydrogen bonds with six surrounding water molecules. Water molecules hydrogen bond with four additional residues, keeping them in close proximity with the chromophore. The negative charge on the chromophore is from the deprotonated alcohol group on the side chain of tyrosine. This negatively charged oxygen forms hydrogen bonds with a water molecule, Thr 203, and His 148. The carbonyl at the stop of the imidazolinone ring is hydrogen bonded to Arg 96 and Gln 94. The oxygen in the backbone of Ser 65 is hydrogen to the backbone of Val 61 and Glu 222.

determined in the process of determining the mechanism of chromophore formation (Figure 18)[40]. Unlike a traditional alpha helix which would form hydrogen bonds with every four residues, the alpha helix in GFP only forms 6 of 24 possible hydrogen bonds. The distances between these atoms was measured for 2AWJ-hbond-opt, 2AWJ, 2AWK, and 1EMB (Table 2).

The hydrogen bonds that are formed by a traditional alpha helix structure are between residues that are *i* and *i+4* in the linear sequence, but end up being on top of each other in the native fold (Figure 18B). These two residues hydrogen bond between the carbonyl peptide in the peptide backbone and the nitrogen in the peptide backbone of another. This bonding results in a native fold that is very stable and in a low energy state. The alpha helix in GFP that runs down the center of the barrel is different in energy due to differences in hydrogen bonding[40] . Two *i* and *i+4* bonds form between the residues at 60 and 64 and between 61 and 65 (Figure 18A). Residues at positions 62 and 66 form hydrogen bonds with the inward facing side chains of
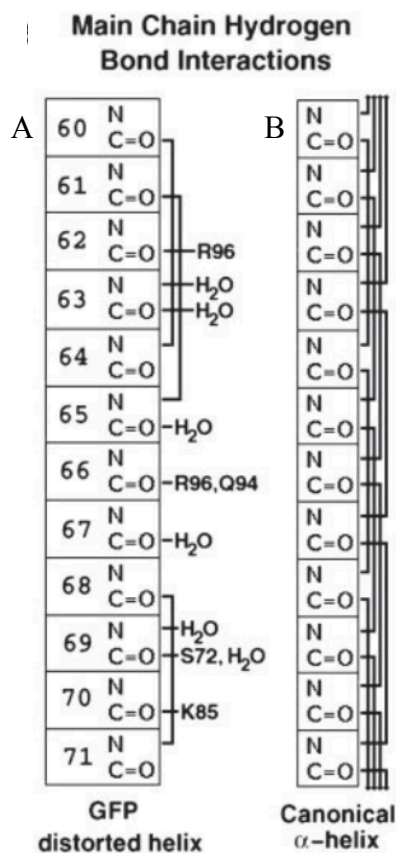
Main Chain Hydrogen Bond Interactions

A 60 N C=O
61 N C=O
62 N C=O —R96
63 N C=O —H$_2$O / H$_2$O
64 N C=O
65 N C=O —H$_2$O
66 N C=O —R96,Q94
67 N C=O —H$_2$O
68 N C=O
69 N C=O —H$_2$O / S72, H$_2$O
70 N C=O —K85
71 N C=O

GFP distorted helix

B N C=O (repeated for each residue position)

Canonical α–helix

Figure 18: Hydrogen bonds present in alpha helix of GFP[40] . There is much less hydrogen bonding in the alpha helix of GFP than compared to a traditional alpha helix structure. A: The hydrogen bonding system along the alpha helix of GFP. Traditional alpha helix *i* and *i+4* hydrogen bonds exist between the peptide backbone of positions 60 and 64 and 61 and 65. B: The *i* and *i+4* hydrogen bonds present in a traditional alpha helix between the carbonyl carbon and the nitrogen in the peptide backbone. These bonds stabilize the helix and play a hydrogen. The differences in the energy states of the two structures due to hydrogen bonding of GFP and a traditional alpha helix plays a major role in the energetics of chromophore[40] . Solid lines represent the location of hydrogen bonds between two atoms.

amino acids in the barrel. Residue 70 forms a hydrogen bonds with the side chain of a turn

residue. An unsual *i* and *i+3* hydrogen forms between the peptide backbone of residues 68 and

71 (Figure 19). This bond helps to bend the alpha helix to transition into a turn at the lid of the

barrel and has an effect on the positioning on the other residues in the alpha helix that go on to

form the chromophore. All hydrogen bonds that are formed by residues on the alpha helix help

give the helix a conformation that is higher in energy than that of a canonical alpha helix (Figure

20) the conjugation trapping mechanism. Since this conformation has higher energy it lessens the

energy necessary for the first step of chromophore formation to begin. Figure 20 also shows that

the alpha helix structure that GFP has lowers the total energetic cost of the chromophore

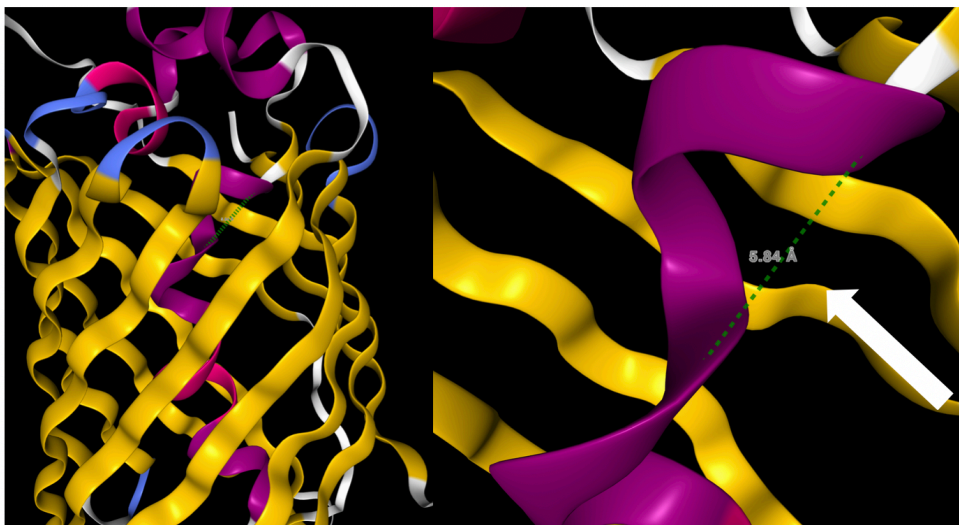formation mechanism, and the mature structure is lower than that of the immature structure.

Figure 19: Location of the hydrogen bond that forms between the carbonyl carbon of the peptide backbone of reside 68 and the nitrogen in the peptide backbone of residue 71. The view of the entire protein (left) with the hydrogen bond location (green). The location of the hydrogen bond effects the distortion of the alpha helix. This view is shown on precyclized GFP (PDB 2AWJ).
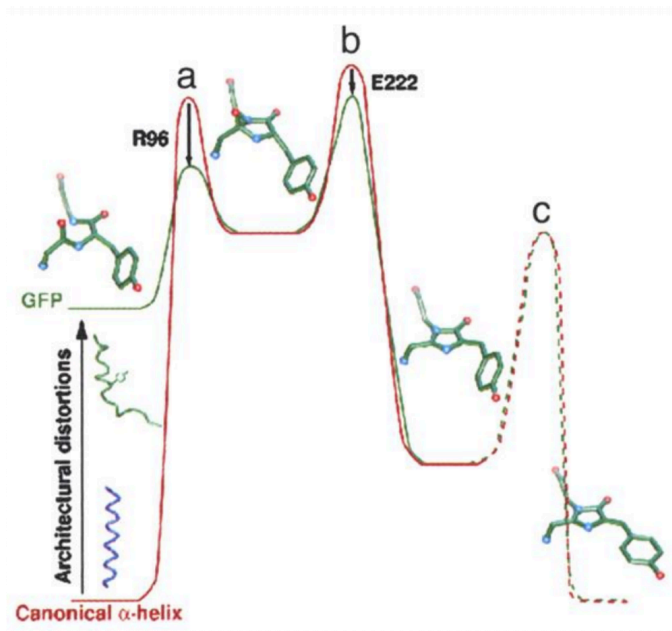

Figure 20: Reaction coordinate diagram for canonical alpha helix with normal hydrogen bonds and the alpha helix of GFP with its architectural distortions undergoing the chromophore formation mechanism. Reaction progress in on the x axis and energy is on the y axis. The path of GFP's alpha helix (green) shows that the reaction is more favorable for GFP than for the canonical alpha helix (red).

Table 2: Hydrogen Bond Distances Along Alpha Helix in Center of Beta Barrel[40]

Residue 1	Residue 2	Distance (Å)			
		2AWJ-hbond-opt‡	2AWJ	2AWK	1EMB
60 (O, carbonyl)	64 N	2.79	2.78	2.92	2.89
61 (O, carbonyl)	65 N	2.83	2.96	2.88	2.76
62 (O, carbonyl)	R 96 (dbond, sb)	4.68, 4.77	(met) 3.60	(met) 3.81	4.86, 2.60
63 N O	H_2O H_2O	-	-	-	-
65 (O, carbonyl)	H_2O	-	-	-	-
66 (O, carbonyl)	R96, Q94 N	6.95, 2.91	M, Q 6.54, 3.12	M, Q 5.24, 2.87	3.03, 3.01
67 (O, carbonyl)	H_2O	-	-	-	-
68 (O, carbonyl)	71 N	3.36	2.99	2.96	3.06
69 (O, carbonyl)	S 72	3.20	2.81	2.78	2.67
70 (O, carbonyl)	K 85	3.34	2.98	2.97	2.86

‡Distances measured with starting structure of 2AWJ-hbond-opt, prior to 100nsec MD simulation. See Section III for distances over course of simulation.

The carbonyl oxygen of the peptide backbone at position 60 forms a hydrogen bond with the nitrogen of the peptide backbone at position 64. The distance between these atoms in 2AWJ-hbond-opt, 2AWJ, 2AWK, and 1EMB all fall within 0.13Å of each other. This data represents no real difference in hydrogen bond distance between these atoms precyclization (2AWJ-hbond-opt and 2AWJ) and post cyclization (2AWK and 1EMB). Another hydrogen bond forms between the peptide backbone carbonyl oxygen at position 61 and the nitrogen in the peptide backbone at

position 65 and the distances are all within 0.2Å for all structures. This again shows that these distances are not very different before and after cyclization. The carbonyl oxygen of the peptide backbone of residue 62 forms a hydrogen bond with one of the nitrogen atoms on the side chain of the arginine that is located at position 96 in 2AWJ-hbond-opt and 1EMB. The oxygen forms a hydrogen bond with the methionine side chain at position 96 for 2AWJ and 2AWK. For the two structures with an arginine, the distance was measured for the nitrogen with the double bond and the one with the single bond. The hydrogen bond mostly forms with the atom that is closer, which is the single bond in mature wild type GFP (more than 2Å closer!). However, in 2AWJ-hbond-opt, both nitrogen atoms are within 0.05Å. For the two structures with a methionine present at position 96 (this was used to determine the catalytic role Arg 96 plays[39]), the bond distances are within 0.2Å of each other, showing that this distance does not change much, if anything it increases, during cyclization. The data does show that the distance between 62 and 96 changes depending on the amino acid present at position 96, even though methionine and arginine are about the same size[63]. The carbonyl oxygen and the nitrogen of the peptide backbone of residue 63 forms hydrogen bonds with two water molecules. The exact molecule could not be determined to use the measure function in Maestro. The carbonyl oxygen at position 65 also hydrogen bonds with a water molecule. The carbonyl oxygen at position 66 forms a hydrogen bond with the side chain of the amino acid at position 96 and the peptide backbone nitrogen at position 94. For the structures with methionine at position 96 (2AWJ and 2AWK), the distances were shorter after cyclization than before cyclization. For the structures with arginine at position 96, the distance between 66 and 96 was much smaller for the structure with the mature chromophore (1EMB). The distance for 66 and 96 was the shortest for mature wild type GFP (1EMB). The carbonyl oxygen of the peptide backbone at position 67 hydrogen bonds

with a single water molecule. The carbonyl oxygen for residue 68 hydrogen bonds with the nitrogen of the backbone of residue 71. The distances for all four structures fell within 0.4Å of each other. The carbonyl oxygen of the backbone of position 69 hydrogen bonds with the side chain of serine, the amino acid present at position 72. For all of the structures, the distance between these two atoms fall within 0.6Å, with the shortest distance being for 1EMB, 2.67Å. The carbonyl oxygen in the backbone at position 71 forms a hydrogen bond with the side chain of the lysine at position 85. The distances for all four structures are within 0.4Å of each other, but again, 1EMB has the smallest distance of 2.86Å.

The hydrogen bond distances for wild type GFP along the alpha helix that gives the energetically favorable distorted conformation are represented by the measured distances for 1EMB (Figures 18, 20). For the idealized *i* and *i+4* between the peptide backbone of 60 and 64 and 61 and 65, the distance for 2AWJ-hbond-opt are within 0.1Å of wild type. This data corresponds with the idealized alpha helix bonding for this portion of the helix for 2AWJ-hbond-opt, 2AWJ, 2AWK, and 1EMB. Since the values for the mutation at position 35 from glycine to an alanine does not affect the hydrogen bonding network for this part of the alpha helix. The peptide backbone of the residues at positions 62 and 66 form hydrogen bonds with the catalytic residue R96, which is not necessary for chromophore formation, but greatly influences the speed[39]. The distance of the single bonded nitrogen (which is closer than the double bonded) in the side chain of R96 to the carbonyl of the backbone of position 62 in 1EMB is 2.0Å less than it is in 2AWJ-hbond-opt, even though both structures have the same residue at each position. This means that the mutation made at position 35 must be affecting the hydrogen bond network (Figure 17) in a way that has moved R96. This could affect the catalytic ability of this residue due to differences in proximity. The difference in distance between R96 and residue 66 for

2AWJ-hbond-opt and 1EMB is more than 4Å. Not only could this larger distance in 2AWJ-hbond-opt affect the ability of R96 to catalyze chromophore formation, but it shows a difference in the shape of the alpha helix that bends at this location. If the alpha helix is not in the necessary conformation, chromophore formation in this mutant is energetically unfavorable. This distance is also important because after cyclization (like in 1EMB) the carbonyl oxygen of the imidazolinone ring forms the hydrogen bond with R96 (Figure 17). Figure 21 proposes other roles of Arg96 in the chromophore of GFP.
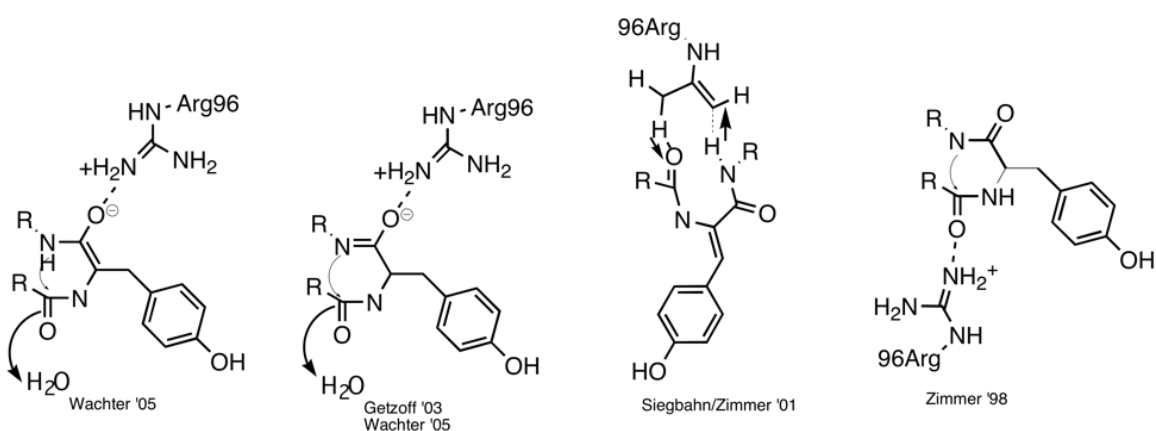

Figure 21: Possible toles for the arginine at position 96 in the posttranslational autocatalytic cyclization reaction that forms the chromophore[35] .

 This same oxygen atom also forms a hydrogen bond with the side chain of Q94, but this distance is within 0.1Å in the two mutants, but this time 2AWJ-hbond-opt is closer than 1EMB. Residues 69 and 70 hydrogen bond with the side chain of 72 and 85 respectively, but the distances for 2AWH-hbond-opt are larger than the three other structures in both cases. This is an indication that the difference in distance is not caused by precyclization (because 2AWJ has comparable distances in both cases to 1EMB), but rather by a structural difference in the conformation of the alpha helix in 2AWJ-hbond-opt that is caused by the alanine at position 35 because that is the only difference between 2AWJ-hbond-opt and 1EMB. The energy of the

conformation will play a large role in whether or not chromophore formation is even possible for this structure.

III. H-Bond Distances Over Course of Simulation

The distances of the atoms known to form hydrogen bonds along the alpha helix that runs down the center of GFP[40] were measured for the initial minimized structure of 2AWJ-hbond-opt (Section II), and showed some variation between precyclized and cyclized structures, with either and arginine or methionine present at position 96. Comparison with another precyclized structure (2AWK) and the wild type structure (1EMB), showed that some differences in distance are not due to the fact that the chromophore has not formed, but rather due to conformational differences in the alpha helix of 2AWJ-hbond-opt. These conformational differences have an effect on the energy state of the helix, and ultimately the reaction coordinate of chromophore formation (Figure 20). Every hydrogen bond distance other than the idealized *i* and *i+4* between the peptide backbone of 60 and 64 and 61 and 65, fell outside the mean length of 2.99 for the idealized O···N bond of an idealized alpha helix[64]. This is a good indication that an idealized alpha helix has not formed in 2AWH-hbond-opt, but the structure is still different than wild type GFP. For this reason, the hydrogen bond distances of the alpha helix of 2AWJ-hbond-opt were monitored over the course of the simulation.
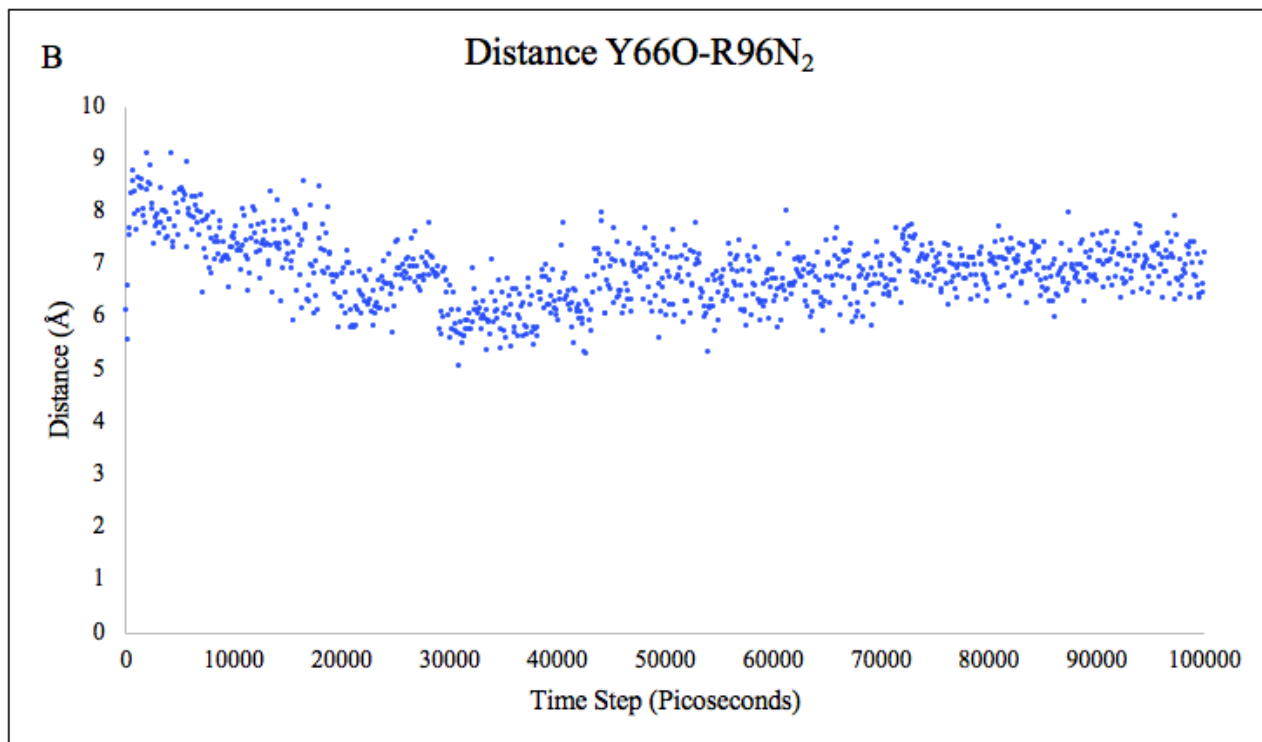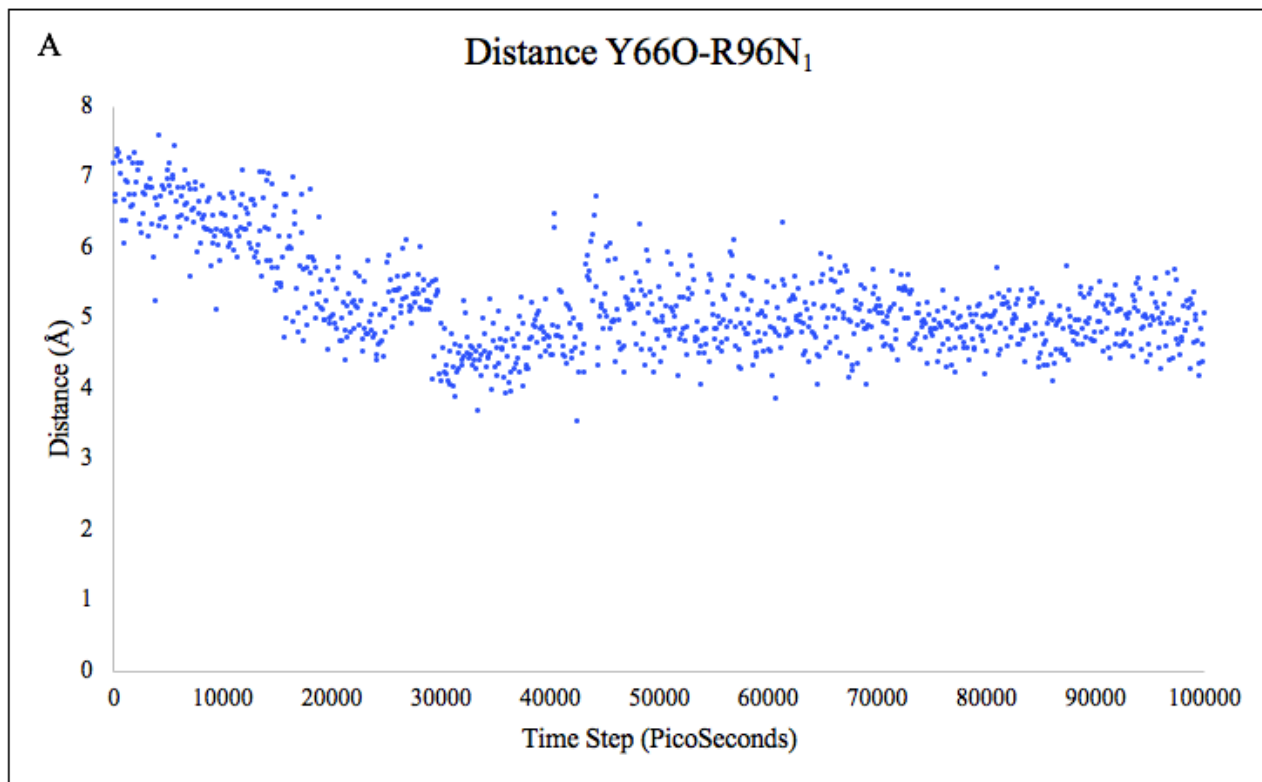
After running a 100nsec MD simulation on 2AWJ-hbond-opt, these distances were monitored over the course of the simulation (Figure 21). Averages of these distances were taken from the data after the first 20nsec of the MD simulation (Table 3), after stabilization, to compare to wild type values and ultimately other structures (Table 2).
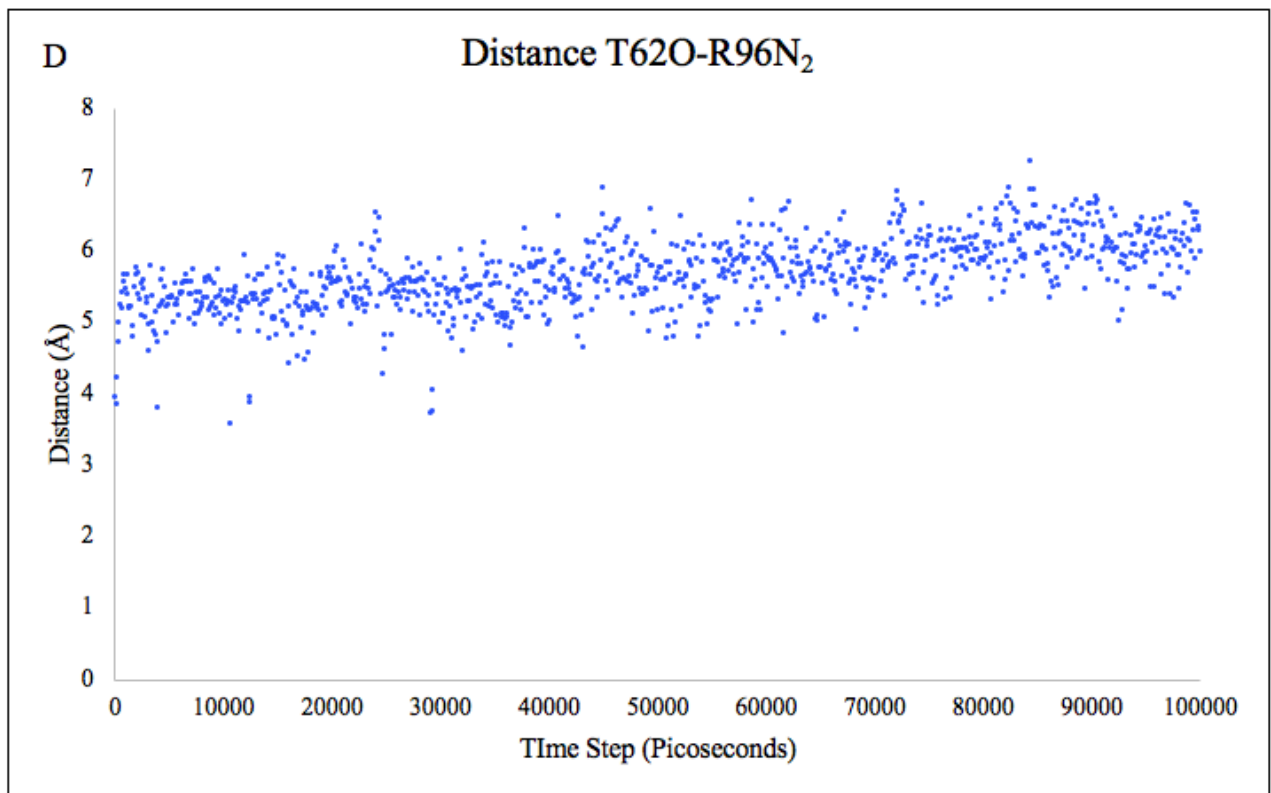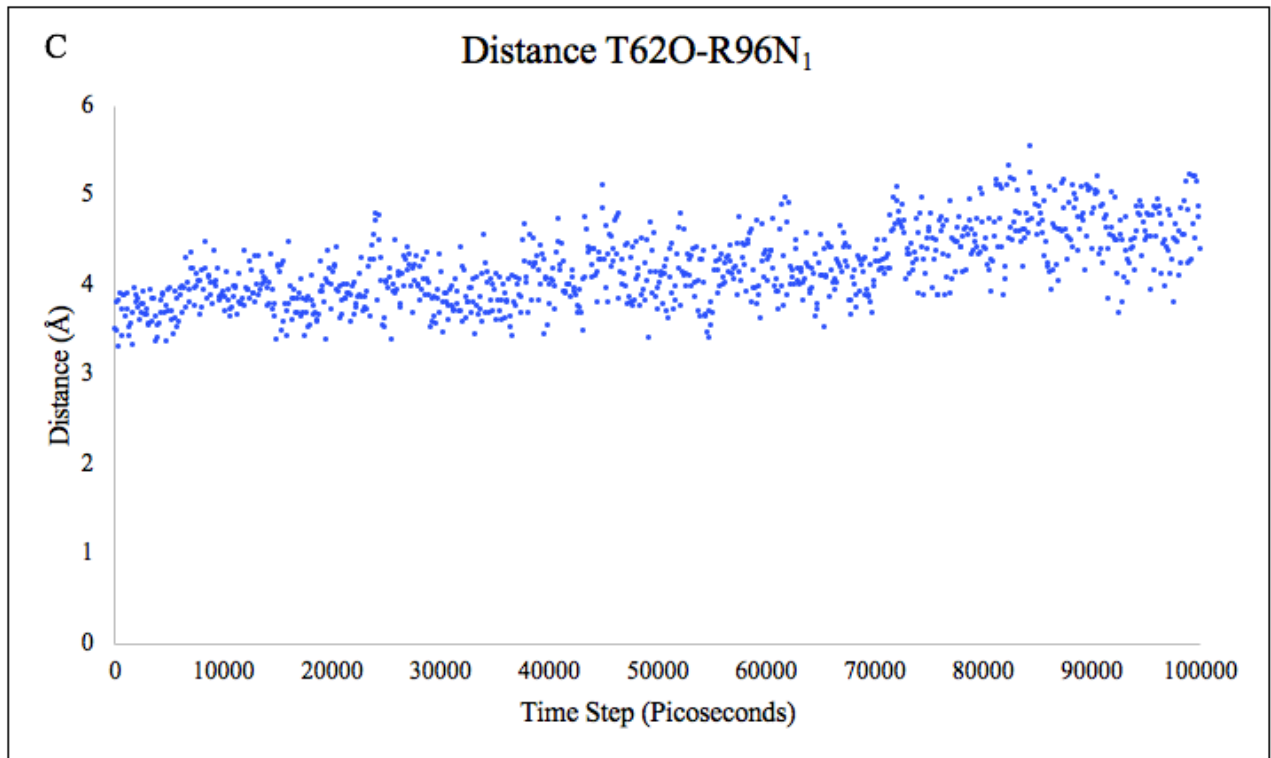
Table 3: Average of H-bond distances of main chain atoms over the course of 100nsec MD simulation for 2AWJ-hbond-opt.
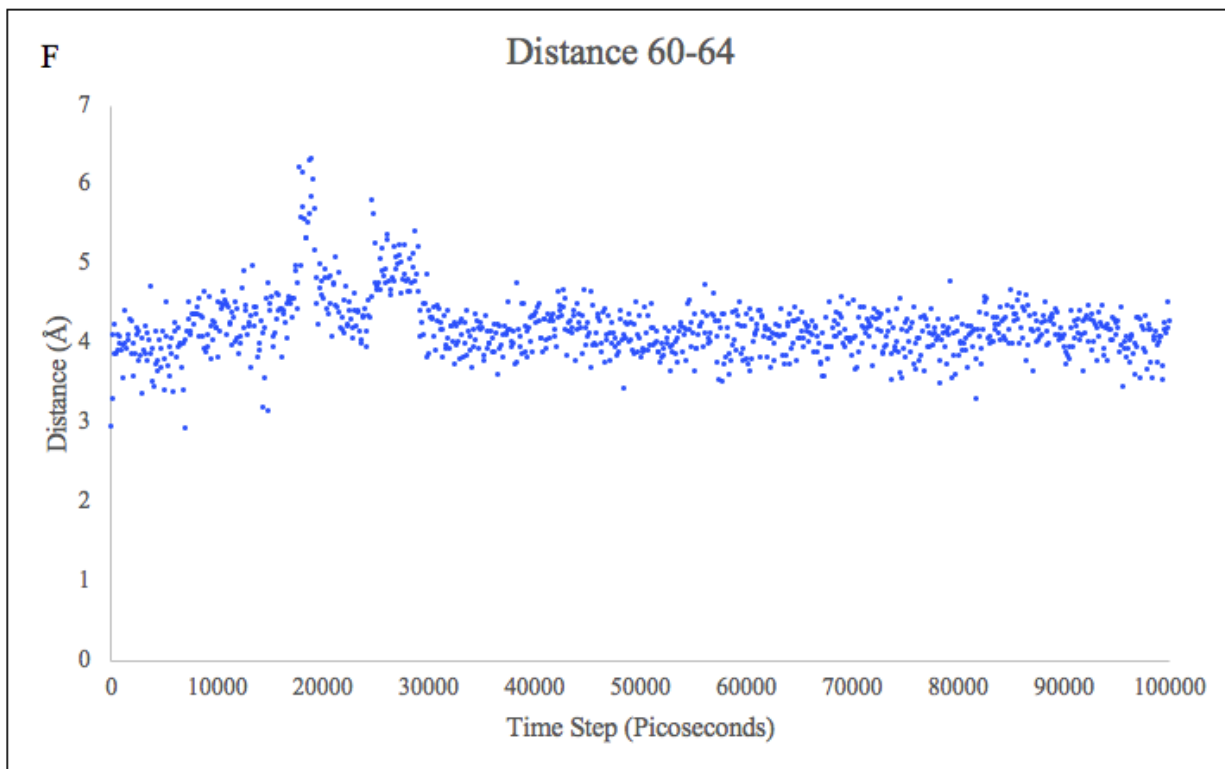
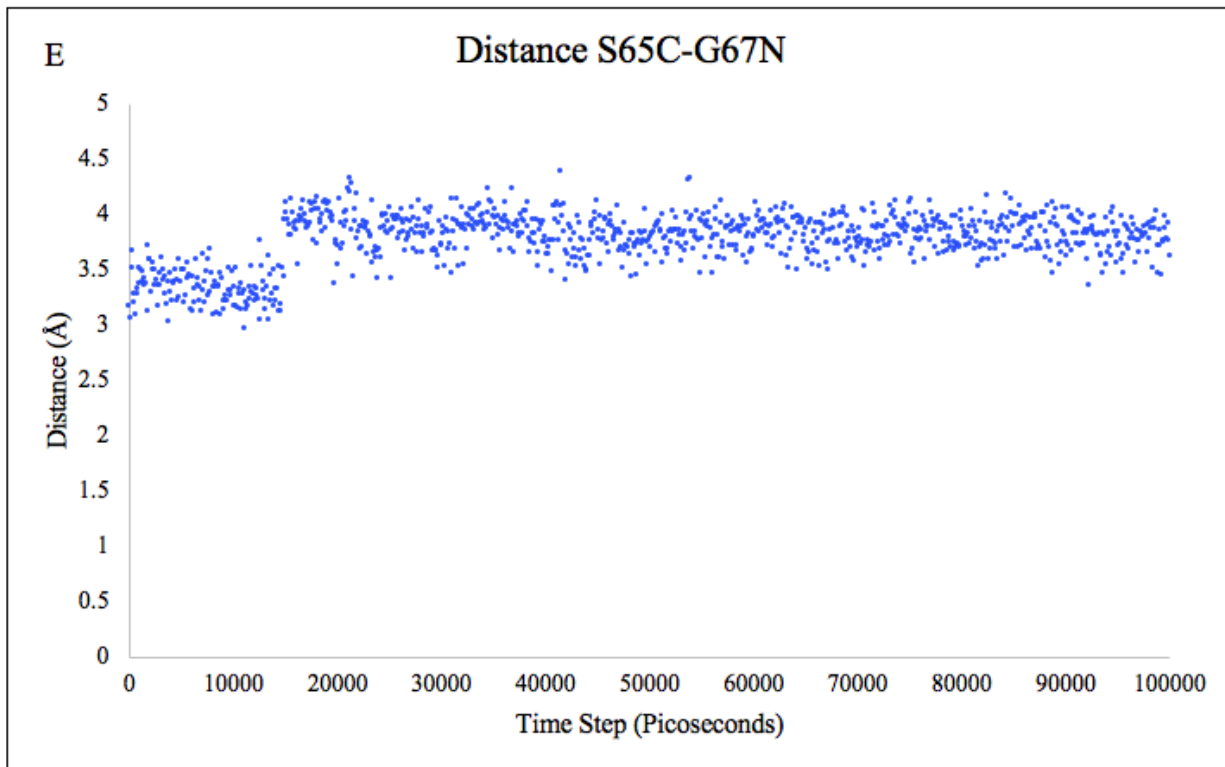| Measurement	Average Distance† (Å) MD sim	H-Bond Distance Starting Structure	Std Deviation (Å)
Y66O-R96N$_1$	4.933	2.91	0.438
Y66O-R96N$_2$	6.697	6.95	0.507
T62O-R96N$_1$	4.259	4.68	0.398
T62O-R96N$_2$	5.779	4.77	0.456
S65C-G67N	3.836	-	0.158
60-64	4.179	2.79	0.315
61-65	3.000	2.83	0.213
66-94	4.713	2.91	1.047
68-71	3.358	3.36	0.307
69-72	4.218	3.2	1.180
70-85	5.238	3.34	1.098

†These averages are after the first 20nsec when the md sim p-rmsd graph stabilizes.

The hydrogen bond distances for the starting structure of 2AWJ-hbond-opt are much smaller than they are over the course of the 100nsec simulation, except between residue 66 and the double bonded nitrogen atom of the side chain of arginine 96, residues 61 and 65, residue 62 and the single bonded nitrogen atom of the arginine at position 96, and residues 68 and 71. The graph of the distance fluctuation over the course of the simulation (Figure 19) for these atoms are also the flattest, showing that they change the least.

Distance Y66O-R96N₁

Distance Y66O-R96N₂

Distance T62O-R96N$_1$

Distance T62O-R96N$_2$

49

Distance S65C-G67N

Distance 60-64

H Distance 61-65

I Distance 66-94

Distance 68-71

Distance 69-72
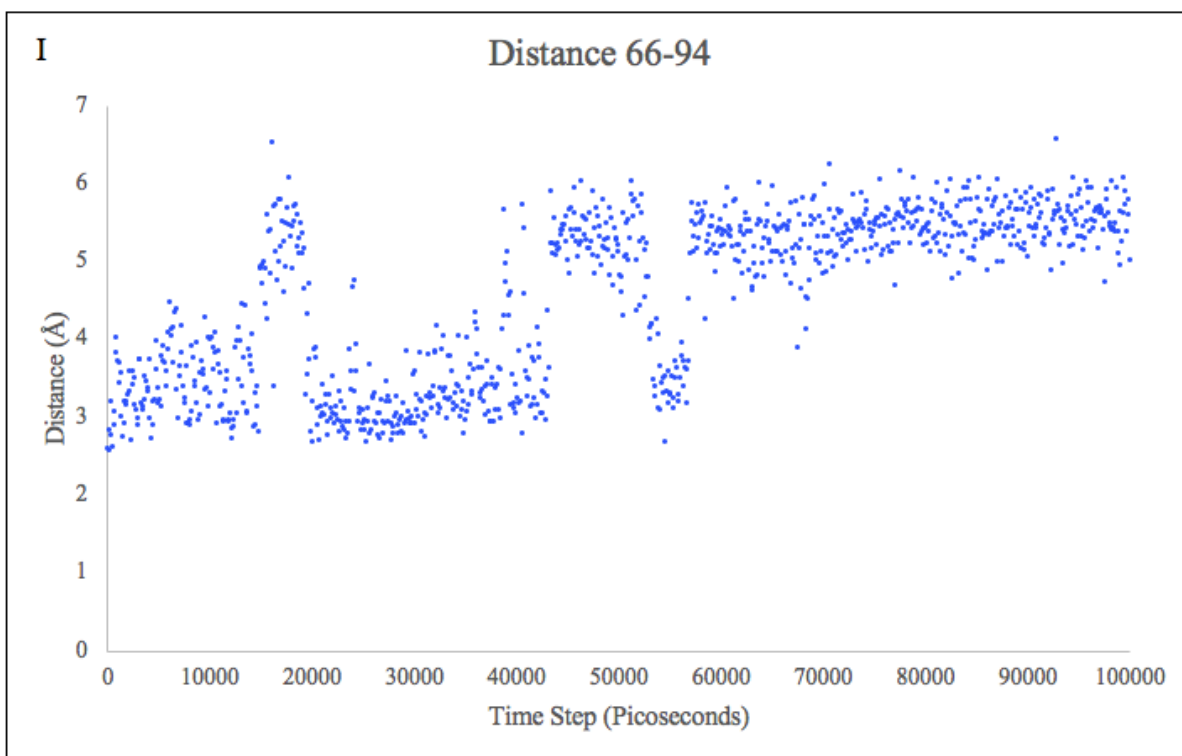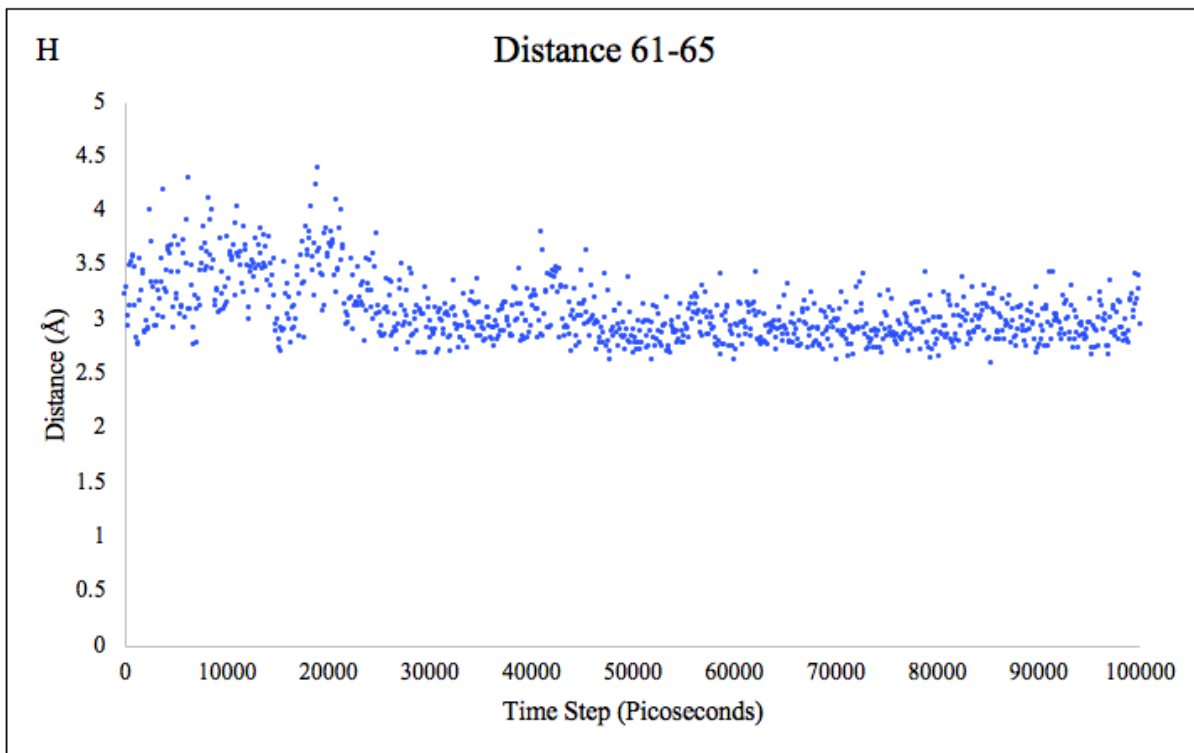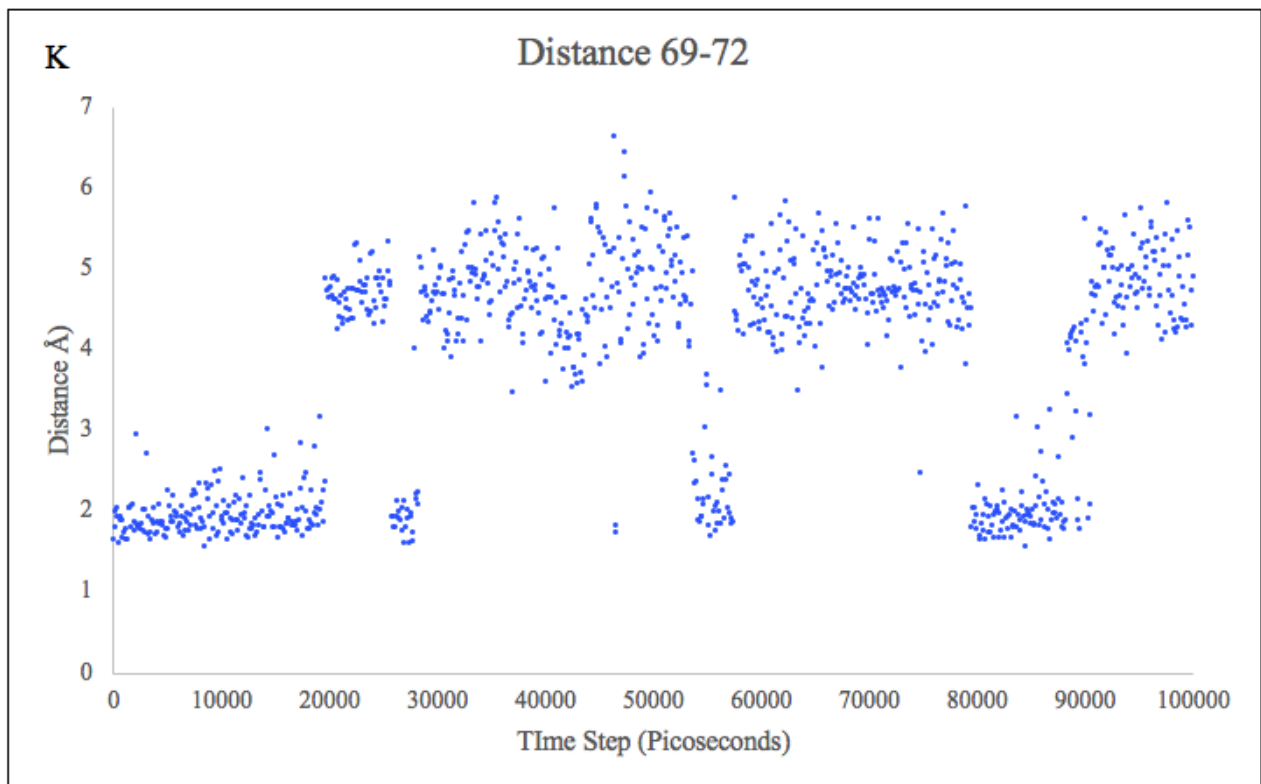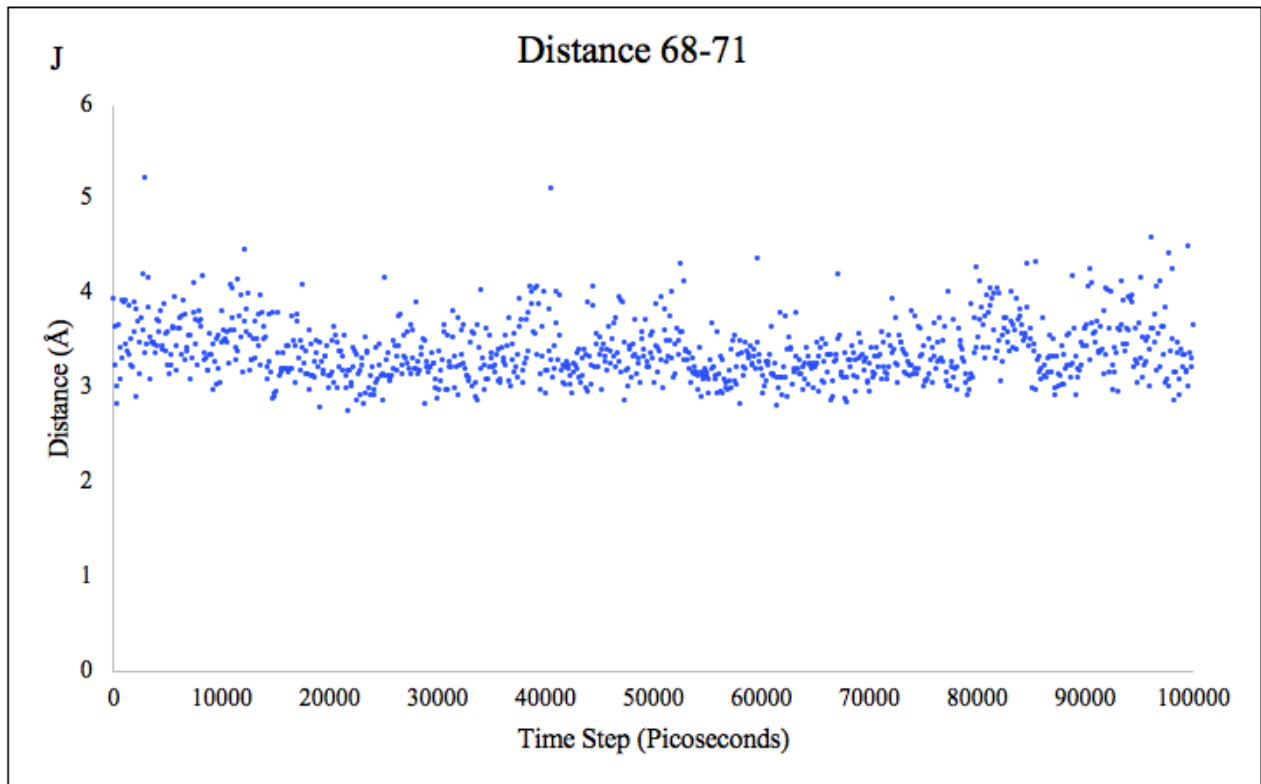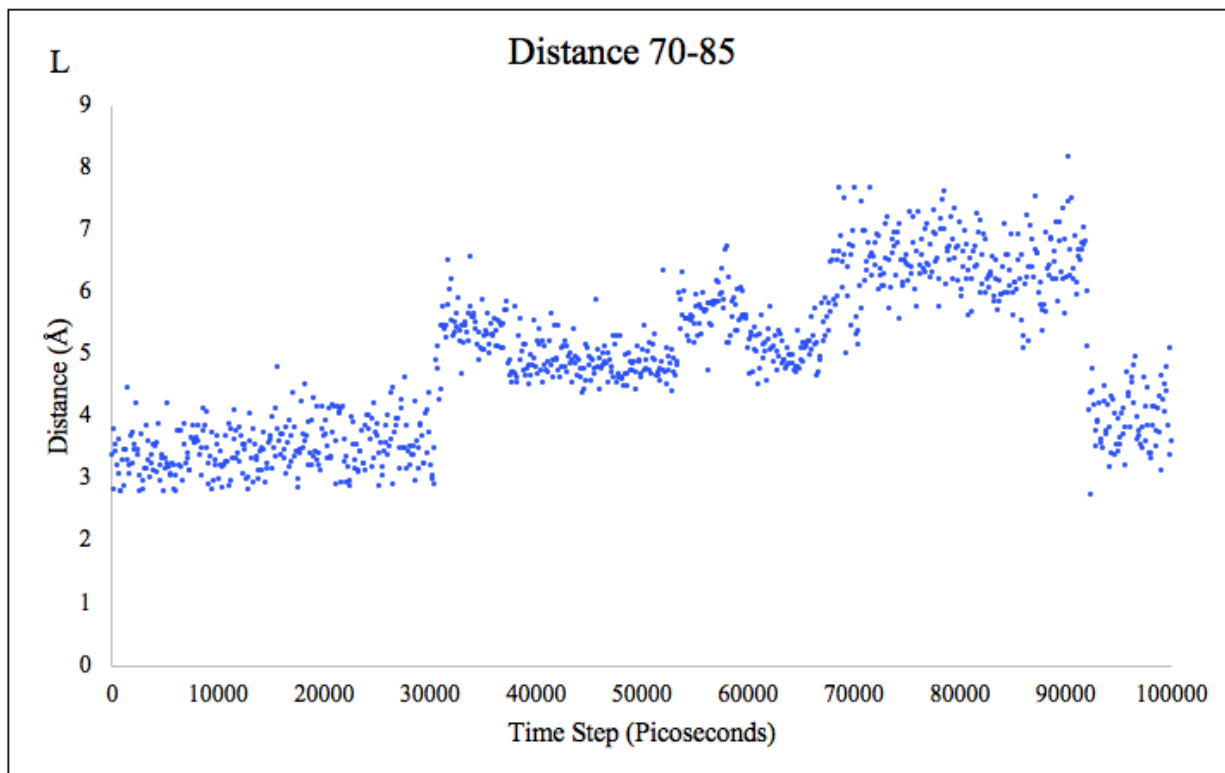
Figure 22: A-L: Hydrogen bond distances if atoms along the alpha helix that runs through the center of GFP that are involved in hydrogen bonding over the course of the 100nsec MD simulation.

The hydrogen bond distance from the carbonyl peptide oxygen of residue 66 and the single bonded nitrogen atom of the side chain of the nitrogen atom of the arginine at position 96 has an average 4.93Å ± 0.44Å, however the distance fluctuates from 7.5Å to 3.5Å (Figure 22A). The hydrogen bond distance between the carbonyl peptide oxygen of residue 66 and the double bonded nitrogen atom of the side chain of the nitrogen atom of the arginine at position 96 has an average of 6.7Å ± 0.51Å, but the bond distance ranges from 9.09Å to 5.07Å over the course of the simulation (Figure 22B). The proximity of these residues plays a role in the ability of R96 to aid in chromophore formation (Figure 21). The average hydrogen bond distance for the carbonyl oxygen of the peptide backbone at position 62 to the single bonded nitrogen atom of the side chain of the arginine atom at position 96 is 4.26Å ± 0.4Å, but the distance fluctuates from 5.55Å to 3.43Å (Figure 22C), which is quite smaller than many other atoms over the course of the

simulation. The hydrogen bond distance for 66 and the double bonded nitrogen of Arg96 has an average of 5.78Å ± 0.46Å, but ranges from 3.58Å to 7.26Å (Figure 22D). The hydrogen bond distance of the achiral carbonyl carbon of the backbone of the residue present at position 65 and the nitrogen present in the backbone at position 67 has an average 3.84Å ± 0.16Å, but the distance ranges from 2.97Å to 4.4Å over the course of the simulation (Figure 22E). There is a 1Å jump for the range of the distances at about 15nsec which corresponds with the change seen in the P-RMSD graph (Figure 14), that shows the protein going through a conformational change at this point. The graph is otherwise relatively flat other than at the aforementioned part. The hydrogen bond distances between residues 60 and 64 has an average of 4.18Å ± 0.32Å, and ranges from 2.92Å to 6.3Å (Figure 22F). The largest fluctuations are from the 20-30nsec region of the MD simulation. The average hydrogen bond distance between residues 61 and 65 is 3.0Å ± 0.21Å, and the distance ranges from 2.6Å to 4.4Å over the course of the simulation (Figure 22H). This graph fluctuates the most for the first 30nsec of the MD simulation, then stabilizes. The hydrogen bond distance between residues 66 and 94 is 4.18Å ± 1.05Å, and the distance between these two residues fluctuates significantly over the course of the simulation, from 2.67Å to 6.53Å (Figure 22I). The structure does not settle at a conformation with about the distance between these two atoms until about 60nsec into the MD simulation. The hydrogen bond distance between residues 68 and 71 has an average of 3.36Å ± 0.31Å, with distances ranging from 2.75Å to 5.22Å over the course of the simulation (Figure 22J). The different conformations the protein exists in over the course of the simulation effects the distance between residues 69 and 72 the most. There is an average distance of 4.22Å ± 1.18Å, with distances ranging from 1.56Å to 6.45Å; this graph never stabilizes (Figure 22K). The hydrogen bond distance from residues 70 to 85 has an average of 5.24Å ± 1.1Å, with distances ranging from 2.75Å to 8.15Å;

the protein conformations over the simulation results in distances that fluctuates between these atoms quite a bit (Figure 22L).

IV. Water Migration Over Course of Simulation[65,66]

The presence of water, and a minimal hydration level is necessary for proper protein function[66]. In GFP, water molecules are able to move in and out of the barrel because the atoms are required for chromophore formation[65,66]. The oxidation step of chromophore formation is the slowest, and water molecules help to speed it up. The mature structure of GFP is much more rigid than the immature structure of GFP, and it is thought to be because of the hydrogen bonds that barrel residues form with the chromophore post cyclization[28]. Most crystal structures have waters within hydrogen bonding distance to Arg96 and Glu222[65]. Simulations have showed that water diffusion in and out of GFP beta barrel decreases after cyclization due to this rigidity, and may be porous during chromophore formation due to a necessity for catalytic water and oxygen molecules[1] .

One study[66] found pores for water diffusion are location between strands 7 and 10 and 7 and 8 in wild type GFP. The pore between strands 7 and 10 opens up at temperatures about 310 K due to fluctuations in the loop between strands 6 and 7, and exhibits the highest RMSD for all nonloop segments[66] . During this fluctuation, the hydrogen bonds change in the protein, and some hydrogen bonded residues are replaced with water molecules[66].

Molecules are able to move in and out of the cavity over the course of the MD simulation, but it is hypothesized that the water molecules that stayed in over the entire course of the simulation were hydrogen bonded to a nearby residue. The water molecules present inside the barrel at the beginning of the MD simulation (Table 4) and at the end of the MD simulation

(Table 5) were monitored to see which ones left and stayed inside and where they came and went from.

Table 4: Water Molecules Inside at Beginning of 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 that runs down the center of the barrel • 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

†Water molecule remains inside the barrel over course of simulation

Table 5: Water Molecules Present at 100ns

Atom #	SPC	Description
53521	16612	• At frame 1001, this water molecule is closest to Lys 101 on the end of the barrel without termini • Water molecule enters GFP at frame 882 through the loops next to Lys 101
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 Ser 147 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 Phe 84 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 Arg 73
38710	11675	• At last frame, water molecule is closest to Tyr 66 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 Asp 36 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

There are 14 water molecules present in the barrel at the beginning of the MD simulation, and six of these waters stay inside over the course of the entire simulation. All of the waters that remain inside the barrel for the course of the simulation begin either within hydrogen bonding distance to a residue in the alpha helix that runs down the center of GFP. This data supports the findings of studies that described an intricate hydrogen bonding network around the chromophore before and after cyclization occurs. Many other water molecules that were present

inside the barrel at the beginning of the simulation were very close to one of two ends of the barrel where they could readily move in and out.

There are eight water molecules present at the end of the simulation. These water molecules are more distributed throughout the barrel. Some are near the alpha helix, some are at the termini capped end of the barrel and others are located around the wall of the can. The tunnels of the structure that allow for water diffusion at the beginning will be examined with MOLEonline in Section VI.

V. MOLEonline

In proteins, pores control the transport of water in and out of structures and tunnels connect and active site to the surface of a protein [67]. MOLEonline[67] is a web-based application that detects both tunnels and pores in protein structures and their geometries. The 2018 version of MOLEonline used in this honors study contains updates that allow for channel visualizing with the original biomolecular structure in the LiteMol Viewer ran by JavaScript, or a file that can be exported and viewed within Maestro, which is what was done. Structure import can be done with a structure directly from the Protein Data Bank, or from a frame of a MD simulation directly from Maestro because the file type is compatible. Calculations were done with channel mode, the system default and the beta structure parameter. Calculations occur with a channel search algorithm that contains seven steps: (i) computation of the Delaunay triangulation/Voronoi diagram for the atomic centers: (ii) building of the molecular surface; (iii) identification of cavities; (iv) identification of possible start points; (v) identification of possible end points; (vi) localization of channels; and (vii) filtering of localized channels[67].

A study of the water diffusion in and out of the beta barrel of GFP and the quick maturing FP, TurboGFP[65] generated simulations that confirmed the existence of a pore that goes from the surface of the beta barrel directly to the chromophore. TurboGFP was found to be more permeable to water than wild type GFP, which may be one of the reasons the chromophore formation in TurboGFP forms faster.

MOLEonline was used to locate the tunnels in 2AWJ-G35A at the beginning of the MD simulation (Figure 23) and at the end of the MD simulation (Figure 24). The channels will then be compared to the places where the waters came and left from during the MD simulation.

There are eight water molecules present at the beginning of the MD simulation that leave the beta barrel before the end of the simulation. Six water molecules stay inside the entire simulation. The paths they used to exit the barrel through were recorded in Table 4 and will be compared with the channels generated using MOLEonline in Figure 24. Water molecules with atom numbers 55639, 55195, and 55952 (all atom numbers are for oxygen of water of interest) all leave the barrel through the termini capped end of the beta barrel (Table 4). The paths they

Figure 23: MOLEonline channel generation for the first frame of 2AWJ-100nsec MD simulation. There are nine tunnels (orange) and no pores for this structure. G53A is modeled using ball and stick to show emphasis on proximity from outside the barrel (left) and the termini capped end (right).

took correspond with the channels depicted in Figure 22. Water molecules with atom numbers 55486, 55189 and 55246 exit the beta barrel through strands 7 and 10, which is known hole in the beta barrel[66] and a pore was generated at this location by MOLEonline. Water molecule 55567 leaves at the end of GFP without the termini which MOLEonline generated a tunnel for (Figure 23). For the water molecule 55273 that exited the barrel between strands 6 and 7, there was no channel for this path. G35A is modeled using the ball and stick to show its location with respect to tunnels where waters can migrate. One channel appears to be close to G35A from the view from the termini capped end (Figure 23, right), but upon view from another angle (Figure 23, left) it is clear that this tunnel is located on the turns on the top of the barrel, while G35A is located on strand 2 (Figure 5). However, it is possible that a water that uses this tunnel may interact.

There are eight water molecules present in the beta barrel at the end of the MD simulation that are not there are the beginning of the simulation. MOLEonline generated five tunnels (Figure 24) for the last frame of the simulation. Water molecules with atom numbers 22708 and 16060 enter the beta barrel though the end of GFP without the termini (Table 5) and the channel generated by MOLEonline corresponds with their path (Figure 24). Water molecules with atom numbers 37009, 19321, 32431, and 23893 enter the barrel through the termini capped end, the
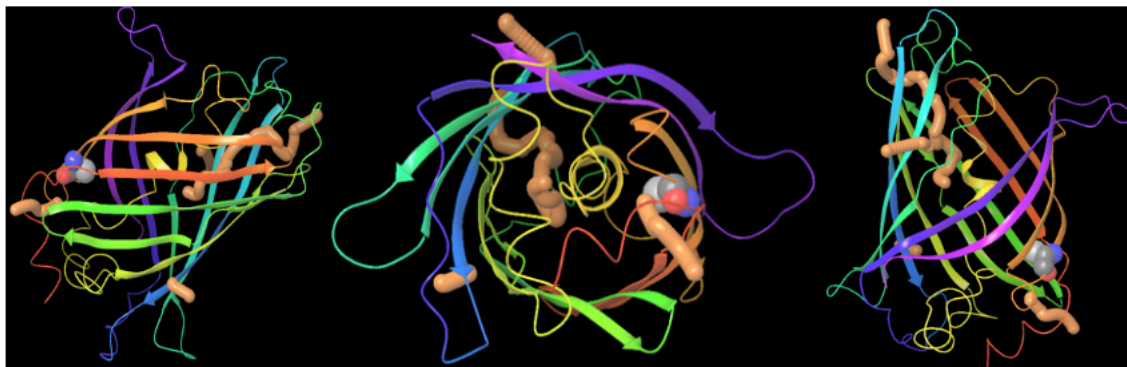

Figure 24: MOLEonline channel generation for the last frame of 2AWJ-100nsec MD simulation. There are five tunnels (orange) for this structure and no pores. G35A is modeled using the ball and stick to show its location in respect to the tunnels. The picture in the middle is an aerial view from the termini capped end of the barrel.

61

paths by which they enter are not channels recognized by MOLEonline. There is only one channel at that end. The water molecule with the atom number 38710 entered the barrel through the hole between strands 7 and 10, which is represented by a pore generated by MOLEonline (Figure 24). The water molecule with the atom number 53521 enters the barrel at the end of the barrel without the termini through the loops, which is represented by a channel generated by MOLEonline (Figure 24). There is a tunnel present in very close proximity to G35A, giving a water molecule that migrates in and out of the barrel using this tunnel opportunity to interact with this residue.

The reason that some of the paths the water molecules took are not generated by MOLEonline is not known. It would be a useful further study to run the simulation for another 100nsec with the last structure from the first 100nsec MD simulation to see of this tunnel is still present in close proximity to G35A and if this a place where water molecules are able to interact with it and other residues involved in the alpha helix hydrogen bonding network.

VI. Centroid Measurements

One study[34] concluded that in beta sheets when glycine and phenylalanine residues are paired across from each other on opposing beta sheets, it results in an overall synergistic increase in protein stability. This pairing occurs quite frequently in naturally occurring beta sheets, both parallel and antiparallel. In GFP, there is a phenylalanine residue at position 71 that is across from the glycine residue at position 35 in the native fold. Since glycine 35 was one of the highly conserved residues in all known FPs with unknown function, the distance from the centroid of phenylalanine to glycine 35 was measured for wild type GFP (Table 6), 2AWJ-G31A, 2AWJ-G33A (Table 7-8), and 2AWJ-G35A (Table 9) over the course of the MD simulation. Since a

direct centroid measurement could not be determined, this was done by measuring the distance to each carbon in the phenyl ring of phenylalanine over the course of the simulation, averaging each distance, then taking an average of the averages.

Table 6: Aromatic Carbons Gly35-Phe71 Wild Type Distances

	Average Distance (Å)	Std Deviation (Å)
Phe71-C1	4.297	0.521
Phe71-C2	4.275	0.434
Phe71-C3	4.468	0.495
Phe71-C4	4.578	0.501
Phe71-C5	4.729	0.452
Phe71-C6	4.529	0.543
Average of the Avgs:	4.4795	

Table 7: Aromatic Carbons Phe71-CAlpha G35 in G31A 2AWJ Mutant 100-150ns[¥]

	Average Distance (Å)	Std Deviation (Å)
Phe71-C1	4.799	0.550
Phe71-C2	5.240	0.7499
Phe71-C3	5.472	0.730
Phe71-C4	5.290	0.537
Phe71-C5	4.779	0.820
Phe71-C6	4.510	0.839

¥Tables 7 and 8 are for the same structure, an additional 50nsec MD simulation had to be ran on the last structure because the P-RMSD graph never stabilized, meaning the protein was undergoing a large conformational change after the first simulation.

Table 8: Aromatic Carbons Phe71-CAlpha G35 in G31A 2AWJ Mutant 150-200ns¥

	Average Distance (Å)	Std. Deviation (Å)
Phe71-C1	4.879	0.597
Phe71-C2	4.999	0.954
Phe71-C3	5.272	0.908
Phe71-C4	5.414	0.565
Phe71-C5	5.212	0.840
Phe71-C6	4.940	0.854
Avg of Avgs	5.067	

¥Tables 7 and 8 are for the same structure, an additional 50nsec MD simulation had to be ran on the last structure because the P-RMSD graph never stabilized, meaning the protein was undergoing a large conformational change after the first simulation.

Table 10: Aromatic Carbons Phe71-CAlpha G35 in G35A 2AWJ Mutant

	Average Distance (Å)	Std Deviation (Å)
Phe71-C1	4.685	0.193
Phe71-C2	4.389	0.235
Phe71-C3	7.023	0.518
Phe71-C4	6.847	0.502
Phe71-C5	5.973	0.503
Phe71-C6	5.184	0.515
Avg. of Avgs	5.683	
Std. Dev.	1.110	
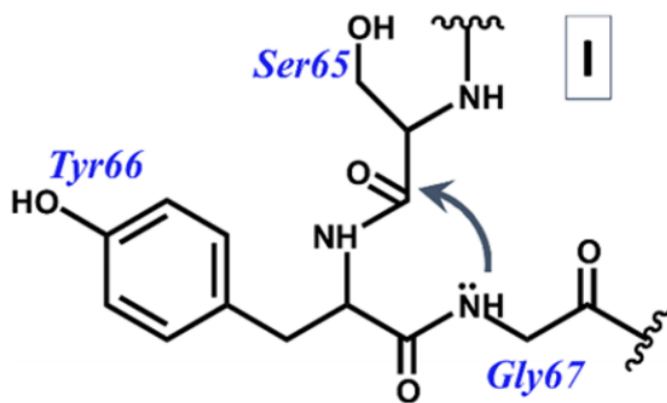
All of the mutants have a centroid distance that is at least 0.5Å greater than that of wild type. The centroid measurement for 2AWJ-G35A mutant is greater than the other structures that had other glycine residues mutated to an alanine. This means that the wild type structure has the largest propensity to form beta sheets. Mutating glycine 35 to an alanine increases the centroid distance and lowers the likely hood of beta sheet formation.

VII. Tight-Turn Distance

In the first step of the mechanism for chromophore formation for GFP[36] , the carbonyl carbon of Ser 65 and nitrogen in the peptide backbone of Gly 67 must be in close proximity (Figure 25). To achieve this orientation, GFP adopts the 'tight-turn' conformation[35,36], in which the distance between the two atoms is shorter than the sum of their covalent radii. In wild type GFP, the distance between the two atoms is 2.87Å[37] . The conformation also keeps these atoms

Figure 25: The first step of the mechanism of the chromophore formation of GFP[39] .

in place for the slow step of the mechanism, autocatalytic cyclization. This tight turn conformation also exists in all of the structures of GFP with an immature chromophore in the Protein Data Bank[35]. One study[37] has shown that GFP mutants that have greater solubility and thermostability, properties that have been linked folding and chromophore formation, have a

deceased distance between the carbonyl carbon of Ser65 and the amide nitrogen of Gly67. For these reasons, the tight turn distance was monitored over the course of the 100nsec MD simulation for 2AWJ-G35A (Figure 22 E, shown again below). The average distance over the course of the simulation was 3.84Å ± 0.16Å. This distance is just under 1Å larger than that of wild type GFP. After about 15nsec of the simulation, 2AWJ-G35A adopts a conformation that results in an increase in the tight turn distance, and it stays about the same for the remainder of the simulation. This data shows that the mutation of glycine to alanine at position 35 increases the tight turn distance, making the first, and subsequent, steps of chromophore formation less likely.
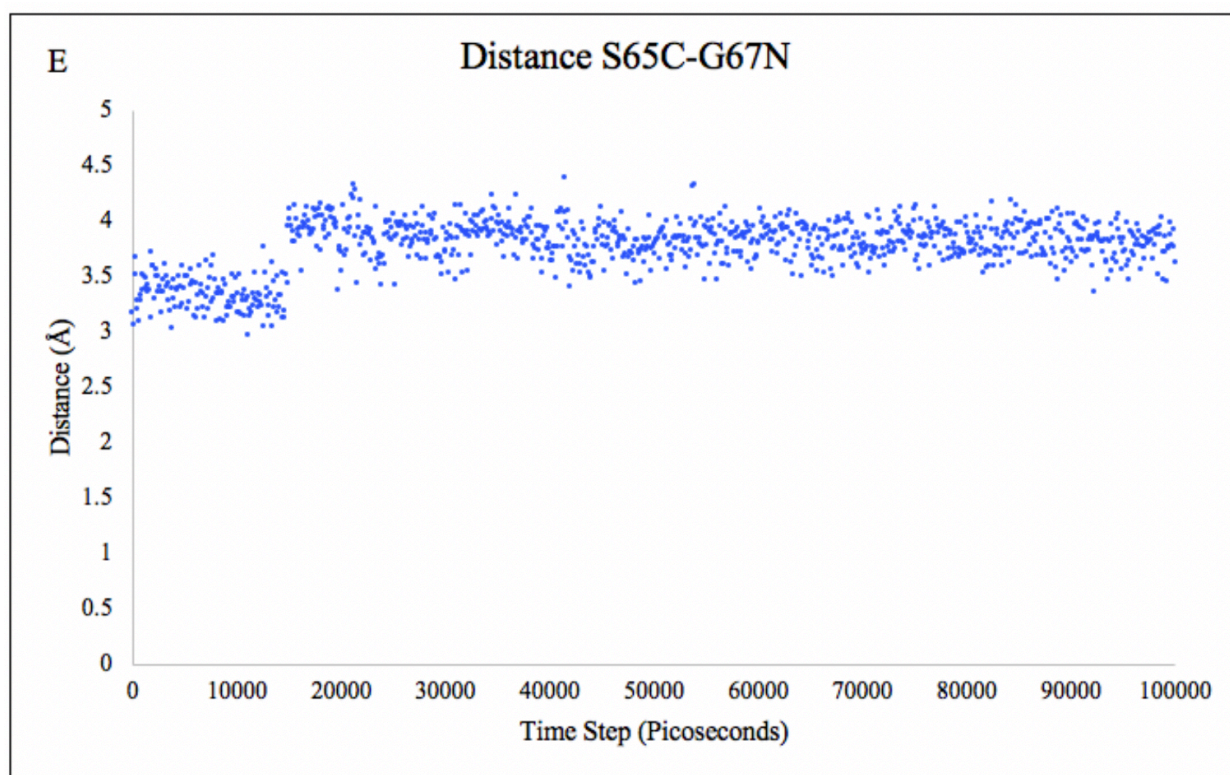

Figure 22E: The distance between the carbonyl carbon of Ser65 and the amide nitrogen of Gly67, the 'tight-turn', over the course of the 100nsec MD simulation for 2AWJ-G35A. The distance has an average of 3.84Å ± 0.16Å.

Conclusion

A 100nsec MD simulation was performed on the precyclized structure, 2AWJ-G35A. The structure underwent a conformational change at around 20nsec; the conformation of the structure did not change much after that. The data from the last 80nsec of the MD simulation was uses to assess its quality, and to compare 2AWJ-G35A to another precyclized mutant, 2AWJ, a cyclized mutant, 2AWK, and wild type GFP, 1EMB. The data produced to compare the structure differences resulted in several findings: (i) the chromophore forming residues of 2AWJ-G35A deviates the most structurally from wild type GFP, and the least from 2AWj; (ii) the alpha helix that runs through the center of GFP in 2AWJ takes on a different conformation than wild type, although the conformation of wild type is the most energetically favorable for chromophore formation; (iii) however, an idealized alpha helix does not form over the course of the simulation though hydrogen bond distances are much larger than that for wild type; (iv) waters migrate in and out of the barrel over the course of the simulation, but six waters stay inside the entire time; (v) there are nine water channels are present at the first frame of the simulation, five in the last frame and one of these five is located in close proximity to G35A; the G35A mutation increases the centroid measurement to Phe71, lowering the likelihood of beta sheet formation; (vi) that the G35A mutation increases the 'tight-turn' distance by just under 1Å from wild type, decreasing the likelihood of a nucleophilic attack by the amide nitrogen of Gly67 to Ser65; (vii) and the centroid measurement increases from that of wild type with the G35A mutation, decreasing the likelihood of beta sheet formation; (viii) the G35A mutation increases the distance from Y66 to R96, limiting its ability to act as a catalyst in the posttranslational autocatalytic chromophore formation, increasing the time of chromophore formation significantly if it does occur at all.

Some further studies that can be done to investigate the glycine at position 35 would be to check what residues Gly35 can potentially hydrogen bond to, and to see if this distance changes over the course of the 2AWJ-G35A MD simulation. And, most importantly, to compare the data generated from G35A with the G31A and G33A mutants. This would give us information about what role they play together, and separately in chromophore formation, folding and structural stability.

References

1. Marc Zimmer Green Fluorescent Protein (GFP): Applications, Structure, and Related Photophysical Behavior. *Chemical Reviews* **2002**, *102*, 759-782.

2. 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.

3. Fu, J. L.; Kanno, T.; Liang, S.; Matzke, A. J. M.; Matzke, M. GFP Loss-of-Function Mutations in Arabidopsis thaliana. *G3 (Bethesda)* **2015**, *5*, 1849-1855.

4. Zimmer, M. Green Fluorescent Protein: Properties, Applications, and Protocols. (Book Review). *Q. Rev. Biol.* **2006**, *81*, 165.

5. Gorokhovatsky, A. Y.; Marchenkov, V. V.; Rudenko, N. V.; Ivashina, T. V.; Ksenzenko, V. N.; Burkhardt, N.; Semisotnov, G. V.; Vinokurov, L. M.; Alakhov, Y. B. Fusion of Aequorea victoria GFP and aequorin provides their Ca 2+-induced interaction that results in red shift of GFP absorption and efficient bioluminescence energy transfer. *Biochem. Biophys. Res. Commun.* **2004**, *320*, 703-711.

6. Kummer, A. D.; Kompa, C.; Lossau, H.; Pöllinger-Dammer, F.; Michel-Beyerle, M. E.; Silva, C. M.; Bylina, E. J.; Coleman, W. J.; Yang, M. M.; Youvan, D. C. Dramatic reduction in fluorescence quantum yield in mutants of Green Fluorescent Protein due to fast internal conversion. *Chemical Physics* **1998**, *237*, 183-193.

7. Kendall, J. M.; Badminton, M. N. Aequorea victoria bioluminescence moves into an exciting new era. *Trends Biotechnol.* **1998**, *16*, 216-224.

8. Alonso, M. T.; Rodríguez-Prados, M.; Navas-Navarro, P.; Rojo-Ruiz, J.; García-Sancho, J. Using aequorin probes to measure Ca2+ in intracellular organelles. *Cell Calcium* **2017**, *64*, 3-11.

9. Shimomura, O. Structure of the chromophore of Aequorea green fluorescent protein. *FEBS Letters* **1979**, *104*, 220-222.

10. Prasher, D. C.; Eckenrode, V. K.; Ward, W. W.; Prendergast, F. G.; Cormier, M. J. Primary structure of the Aequorea victoria green-fluorescent protein. *Gene* **1992**, *111*, 229-233.

11. 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.

12. 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.

13. Coleman, R. Jellyfish, fluorescent proteins, Nobel Prizes and pioneers in histochemistry. *Acta Histochem.* **2010***, 112*, 113-117.

14. Widder, E. A. Bioluminescence in the ocean: origins of biological, chemical, and ecological diversity. *Science* **2010***, 328*, 704-708.

15. Wojdyla, K.; Rogowska-Wrzesinska, A.; Wrzesinski, K.; Roepstorff, P. Mass spectrometry based approach for identification and characterisation of fluorescent proteins from marine organisms. *J Proteomics* **2011***, 75*, 44-55.

16. Widder, E. A.; Latz, M. I.; Case, J. F. MARINE BIOLUMINESCENCE SPECTRA MEASURED WITH AN OPTICAL MULTICHANNEL DETECTION SYSTEM. *Biol. Bull.* **1983***, 165*, 791-810.

17. 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. *Nat. Biotechnol.* **1999***, 17*, 969-973.

18. Wang, Y.; Yu, Y. A.; Shabahang, S.; Wang, G.; Szalay, A. A. Renilla luciferase- Aequorea GFP (Ruc-GFP) fusion protein, a novel dual reporter for real-time imaging of gene expression in cell cultures and in live animals. *Mol. Genet. Genomics* **2002***, 268*, 160-168.

19. Poëa-Guyon, S.; Pasquier, H.; Mérola, F.; Morel, N.; Erard, M. The enhanced cyan fluorescent protein: a sensitive pH sensor for fluorescence lifetime imaging. *Anal Bioanal Chem* **2013***, 405*, 3983-3987.

20. Rothermel, M.; Brunert, D.; Klupp, B. G.; Luebbert, M.; Mettenleiter, T. C.; Hatt, H. Advanced tracing tools: functional neuronal expression of virally encoded fluorescent calcium indicator proteins. *J. Neurovirol.* **2009***, 15*, 458-464.

21. Oscar, B. G.; Liu, W.; Zhao, Y.; Tang, L.; Wang, Y.; Campbell, R. E.; Fang, C. Excited-state structural dynamics of a dual-emission calmodulin-green fluorescent protein sensor for calcium ion imaging. *PNAS* **2014***, 111*, 10191-10196.

22. Yaxiong Yang; Nan Liu; Yuanyuan He; Yuxia Liu; Lin Ge; Linzhi Zou; Sen Song; Wei Xiong; Xiaodong Liu Improved calcium sensor GCaMP-X overcomes the calcium channel perturbations induced by the calmodulin in GCaMP. *Nature Communications* **2018***, 9*, 1-18.

23. Taya, P.; Maiti, B.; Kumar, V.; De, P.; Satapathi, S. Design of a novel FRET based fluorescent chemosensor and their application for highly sensitive detection of nitroaromatics. *Sensors & Actuators: B. Chemical* **2018***, 255*, 2628-2634.

24. Crivat, G., Justin W Imaging proteins inside cells with fluorescent tags. *Trends in Biotechnology* **2011***, 30*, 8-16.

25. Blum, C., Brockinke, A., Budisa, N., Gensch, T., Gu, W., Helms, V., Hoesl, M., Hötzer, B., Jung, G., Luin, S., Meech, S., Merkel, L., Nienhaus, G., Nienhaus, K., Nifosí, R., Schwedler, S., Subramaniam, V., van Thor, J., Tozzini, V., Veettil, S., Wiedenmann, J. One-Photon and Two-Photon Excitation of Fluorescent Proteins. In *Fluorescent Proteins I: From Understanding to Design*; Gregor Jung, Ed.; Springer, Berlin, Heidelberg: pp 3-40.

26. Amos, W. B. The glow spreads throughout biology: Green Fluorescent Protein: Properties, Applications and Protocols, edited by M. Chalfie and S. Kain (Book Review). *Trends Biochem. Sci.* **1999***, 24*, 251.

27. Labas, Y. A.; Gurskaya, N. G.; Yanushevich, Y. G.; Fradkov, A. F.; Lukyanov, K. A.; Lukyanov, S. A.; Matz, M. V. Diversity and Evolution of the Green Fluorescent Protein Family. *Proc. Natl. Acad. Sci. U. S. A.* **2002***, 99*, 4256-4261.

28. 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. *Int Rev Cell Mol Biol* **2013***, 302*, 221-278.

29. Zimmer, M. H.; Li, B.; Shahid, R. S.; Peshkepija, P.; Zimmer, M. Structural Consequences of Chromophore Formation and Exploration of Conserved Lid Residues amongst Naturally Occurring Fluorescent Proteins. *Chem Phys* **2014***, 429*, 5-11.

30. Minor, D. L.; Kim, P. S. Measurement of the β-sheet-forming propensities of amino acids. *Nature* **1994***, 367*, 660-663.

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. Marcos, E.; Chidyausiku, T. M.; McShan, A. C.; Evangelidis, T.; Nerli, S.; Carter, L.; Nivón, L. G.; Davis, A.; Oberdorfer, G.; Tripsianes, K.; Sgourakis, N. G.; Baker, D. De novo design of a non-local β-sheet protein with high stability and accuracy. *Nature Structural & Molecular Biology* **2018***, 25*, 1028-1034.

33. Rauk, A.; Armstrong, D. A. Influence of ß-Sheet Structure on the Susceptibility of Proteins to Backbone Oxidative Damage: Preference for ªC-Centered Radical Formation at Glycine Residues of Antiparallel ß-Sheets. *American Chemical Society* **2000***, 122*, 4185-4192.

34. Merkel, J. S.; Regan, L. Aromatic rescue of glycine in beta sheets. *Fold Des* **1998***, 3*, 449-455.

35. 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.

36. Grigorenko, B. L.; Krylov, A. I.; Nemukhin, A. V. Molecular modeling clarifies the mechanism of chromophore maturation in the green fluorescent protein. *J. Am. Chem. Soc.* **2017**, *139*, 10239-10249.

37. Branchini, B. R.; Nemser, A. R.; Zimmer, M. A Computational Analysis of the Unique Protein-Induced Tight Turn That Results in Posttranslational Chromophore Formation in Green Fluorescent Protein. *Journal of the American Chemical Society* **1998**, *120*, 1-6.

38. 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.

39. 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.

40. 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.

41. Bryngelson, J. D.; Wolynes, P. G. Spin glasses and the statistical mechanics of protein folding. *PNAS* **1987**, *84*, 7524-7528.

42. Fukuda, H.; Arai, M.; Kuwajima, K. Folding of green fluorescent protein and the cycle3 mutant. *Biochemistry* **2000**, *39*, 12025-12032.

43. Lapidus, L. J. Exploring the top of the protein folding funnel by experiment. *Current Opinion in Structural Biology* **2013**, *23*, 30-35.

44. Enoki, S.; Saeki, K.; Maki, K.; Kuwajima, K. Acid Denaturation and Refolding of Green Fluorescent Protein. *Biochemistry* **2004**, *43*, 14238-14248.

45. Kao, T.; Chen, Y.; Pai, C.; Chang, M.; Wang, A. H. -. Structure of a NADPH-dependent blue fluorescent protein revealed the unique role of Gly176 on the fluorescence enhancement. *J. Struct. Biol.* **2011**, *174*, 485-493.

46. Roldán-Salgado, A.; Sánchez-Barreto, C.; Gaytán, P. LanFP10-A, first functional fluorescent protein whose chromophore contains the elusive mutation G67A. *Gene* **2016**, *592*, 281-290.

47. Watanabe, T. M.; Imada, K.; Yoshizawa, K.; Nishiyama, M.; Kato, C.; Abe, F.; Morikawa, T. J.; Kinoshita, M.; Fujita, H.; Yanagida, T. Glycine insertion makes yellow fluorescent protein sensitive to hydrostatic pressure. *PloS one* **2013**, *8*.

48. De Visser, J. Arjan G. M.; Krug, J. Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics* **2014***, 15*, 480.

49. 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.

50. Saeger, J.; Hytönen, V. P.; Klotzsch, E.; Vogel, V. GFP's mechanical intermediate states. *PLoS ONE* **2012***, 7*, e46962.

51. Tong, X. A three-dimensional time-dependent Schrödinger equation solver: an application to hydrogen atoms in an elliptical laser field. *Journal of Physics B: Atomic, Molecular and Optical Physics* **2017***, 50*, 144004.

52. Schleich, W. P.; Greenberger, D. M.; Kobe, D. H.; Scully, M. O. Schrödinger equation revisited. *Proceedings of the National Academy of Sciences of the United States of America* **2013***, 110*, 5374-5379.

53. Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J Comput Chem* **2003***, 24*, 1999-2012.

54. 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.

55. Hospital, A.; Goñi, J. R.; Orozco, M.; Gelpí, J. L. Molecular dynamics simulations: advances and applications. *Advances and Applications in Bioinformatics and Chemistry* **2015***, 2015*, 37-47.

56. Hansson, T.; Oostenbrink, C.; van Gunsteren, W. Molecular dynamics simulations. *Curr. Opin. Struct. Biol.* **2002***, 12*, 190-196.

57. 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. *Journal of chemical theory and computation* **2016***, 12*, 281-296.

58. Lindahl, E. Molecular dynamics simulations. *Methods in molecular biology (Clifton, N.J.)* **2015***, 1215*, 3.

59. DE Shaw Research Desmond Molecular Dynamics System. **2014**, *3.7*.

60. Folie Lecture on MD Simulation Analysis.
http://www.drugdesign.gr/uploads/7/6/0/2/7602318/lecture_mdanalysis.pdf.

61. Devadoss, F.; Paul Raj, V. Analysis and visual summarization of molecular dynamics
simulation. *J Cheminform* **2014**, *6*, 1-2.

62. Kolossváry, I.; Guida, W. C. Low Mode Search. An Efficient, Automated Computational
Method for Conformational Analysis: Application to Cyclic and Acyclic Alkanes and
Cyclic Peptides. *Journal of the American Chemical Society* **1996**, *118*, 5011-5019.

63. Timothy I Wood; David P Barondeau; Chiharu Hitomi; Carey J Kassmann; John A Tainer;
Elizabeth D Getzoff Defining the role of arginine 96 in green fluorescent protein
fluorophore biosynthesis. *Biochemistry* **2005**, *44*, 16211-16220.

64. Kim, S.; Cross, T. A. Uniformity, Ideality, and Hydrogen Bonds in Transmembrane α-
Helices. *Biophysical Journal* **2002**, *83*, 2084-2095.

65. Zimmer, M.; Li, B.; Shahid, R.; Peshkepija, P. Water diffusion in and out of the β-barrel of
GFP and the fast maturing fluorescent protein, TurboGFP. **2012**.

66. Shinobu, A.; Agmon, N. The Hole in the Barrel: Water Exchange at the GFP Chromophore.
The journal of physical chemistry. B **2015**, *119*, 3464-3478.

67. Pravda, L.; Sehnal, D.; Toušek, D.; Navrátilová, V.; Bazgier, V.; Berka, K.; Svobodová
Vařeková, R.; Koča, J.; Otyepka, M. MOLEonline: a web-based tool for analyzing channels,
tunnels and pores (2018 update). *Nucleic Acids Res* **2018**, *46*, W373.

68. Duan, C.; Adam, V.; Byrdin, M.; Bourgeois, D. Structural basis of photoswitching in
fluorescent proteins. *Methods Mol. Biol.* **2014**, *1148*, 177-202.