The molecular structure of green fluorescent protein — **Source link** ↗

Fan Yang, Larry G. Moss, George N. Phillips

Institutions: Rice University, Tufts University

Topics: Protein structure and Green fluorescent protein

Related papers:

- Crystal structure of the Aequorea victoria green fluorescent protein.

- The green fluorescent protein

- Green fluorescent protein as a marker for gene expression

- Wavelength mutations and posttranslational autoxidation of green fluorescent protein

- Primary structure of the Aequorea victoria green-fluorescent protein.

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

RICE UNIVERSITY

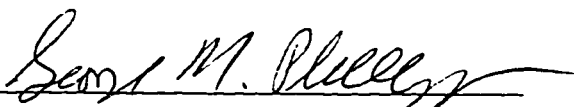The Molecular Structure of Green Fluorescent Protein

by
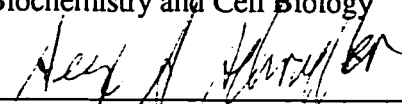
Fan Yang
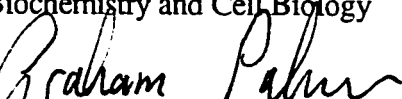
A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE

George N. Phillips, Jr. Professor, Chairman
Biochemistry and Cell Biology

George J. Schroepfer, Professor
Biochemistry and Cell Biology

Graham A. Palmer, Professor
Biochemistry and Cell Biology

Michael C. Gustin, Professor
Biochemistry and Cell Biology

Yin Zhang, Professor
Computational and Applied Mathematics

Houston, Texas

April, 1997

UMI Number: 9727628

UMI

Abstract

The Molecular Structure of Green Fluorescent Protein

by

Fan Yang

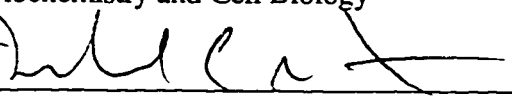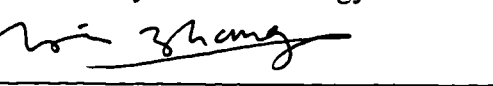The crystal structure of recombinant wild-type green fluorescent protein (GFP) has been solved to a resolution of 1.9 Å by multiwavelength anomalous dispersion phasing methods using selenomethionyl GFP crystals. The protein is in the shape of a cylinder, comprising 11 strands of β-sheet with an α-helix inside and short helical segments on the ends of the cylinder. This motif, with β-structure on the outside and α-helix on the inside, represents a new protein fold. Two protomers pack closely together to form a dimer in the crystal. The fluorophores are protected inside the cylinders, and their structures are consistent with the formation of aromatic systems made up of Tyr66 with oxidation of its $C\alpha$-$C\beta$ bond coupled with cyclization of the neighboring glycine and serine residues. The environment inside the cylinder explains the effects of many existing mutants of GFP and suggests which side chains could be modified to change the spectral properties of GFP. Furthermore, the identification of the dimer contacts may allow mutagenic control of the state of assembly of the protein.

GFP can be reduced by sodium dithionite and as a result, loses its fluorescence. The structure of reduced GFP has been solved which shows that the side chain of

Tyr66 at the fluorophore is the group being reduced. While oxidized, the Cα atom of Tyr66 is not a chiral center since all fluorophore atoms are co-planar. After reduction, the Cα atom returns to a chiral center, but has either an L or D configuration, indicating that the breakage of the resonance system eliminates the fluorescence of GFP and the reduction of Tyr66 is not stereospecific.

Acknowledgments

I would like to express my gratitude to my advisor, Dr. George N. Phillips, Jr., who has provide me with an excellent environment for research in the course of my graduate studies. Without his keen sense in scientific research and his intelligent instruction, it would be impossible for me to solve the structure of the green fluorescent protein (GFP).

I would like to thank Drs. George J. Schroepfer and John S. Olson for not only providing me with the experimental instruments, but also giving me valuable discussion, encouragement, and instructions; Drs. Graham A. Palmer, Michael C. Gustin, and Yin Zhang for their helpful guidance as members of my thesis committee.

I would also like to thank Drs. Michael L. Quillin and Tiansheng Li for teaching me valuable experimental skills including using computer software and instruments; Dr. Larry G. Moss for providing GFP samples and plasmids, and for discussing the project. Special thanks should be given to Dr. Frank Whitby for encouraging me to start the project and helping me to collect data; Dr. Michael B. Berry for helping me to grow GFP crystals and collecting the data that solved the structure of GFP; Ms. Han Xu for helping me to express GFP in *E. coli.*

I am also grateful to Drs. Tod Romo, Eric J. Hnath, Jayashree Soman, Michael Stowell, Michael Soltis, Henry Bellamy, Eric Brucker, Michael Wall, and Craig Ogata for

helping me to do experiments or offering me valuable and enlightening suggestions and discussions.

Finally, I would like to express my gratitude to my parents for supporting and guiding me in all my life.

Table of Contents

List of Abbreviations

ABTS——2,2'-azino-di-(3-ethylbenzthiazoline-6-sulphonic acid

CD——circular dichroism

DTT——DL-dithiothreitol

EDTA——ethylene diamine tetraacetic acid

FRET——fluorescence resonance energy transfer

GFP——green fluorescent protein

HPLC——high-pressure liquid chromatography

MAD——multiwavelength anomalous dispersion

MES——morpholino ethane sulfonic acid

MIR——multiple isomorphous replacement

MPD——2-methyl-2,4-pentanediol

MR——molecular replacement

NMR——nuclear magnetic resonance

PEG——polyethylene glycol

SIR——single isomorphous replacement

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background of bioluminescence

Bioluminescence is a widespread phenomenon that has attracted the interests of scientists for centuries. Bioluminescence was first described in written report by Pliny the Elder in the first century AD who noticed the bright light from some jellyfish (Harvey, 1952). Bioluminescence is much more common in marine habitats than in terrestrial and freshwater species, probably due to the darkness in deep water. No bioluminescence ability has been observed so far for higher plants and higher vertebrates. In the past several decades, with the development of modern biochemistry and molecular biology, large amounts of knowledge about bioluminescence have been accumulated. The origin and evolution, ecological functions, and biochemical mechanism of bioluminescence in numerous organisms such as bacteria, fungi, fireflies, coelenterates, and fish have been studied intensively. In this section, a brief description will be given about bioluminescence.

1.1.1 General process of bioluminescence

Despite the diverse mechanisms in various bioluminescent systems, most of them involve molecular oxygen, an enzyme and its substrate which are generically termed luciferase and luciferin, respectively. Luciferase can catalyze oxidative decarboxylation of luciferin and generate an electronically excited oxyluciferin which may either emit light when returning to the ground state or transfer its energy to an accessory fluorescent protein such as green fluorescent protein (GFP) which has been investigated for a long time and will be studied in this dissertation. The processes can be simplified as:

step one: Luciferin + O$_2$ $\xrightarrow{\text{luciferase}}$ oxyluciferin* + CO$_2$

step two: Oxyluciferin* $\longrightarrow$ oxyluciferin + light

             or

             Oxyluciferin* + GFP $\longrightarrow$ oxyluciferin + GFP*

             GFP* $\longrightarrow$ GFP + light
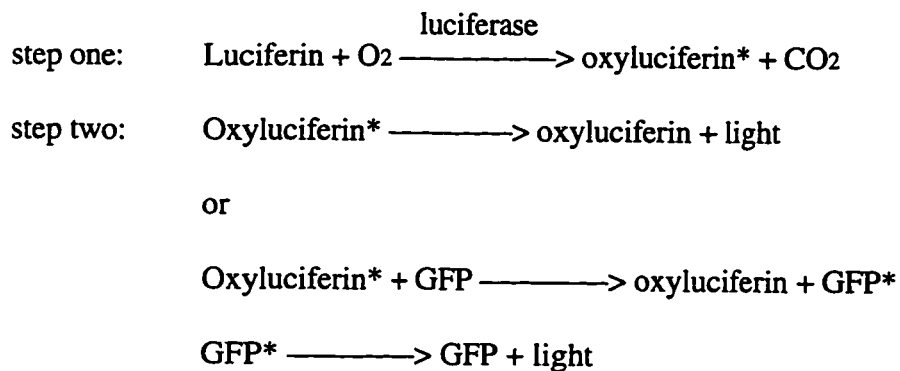
The chemical structures of some luciferins have been determined that revealed the diversity of luciferins (Ward, 1985; Hastings, 1996). Some bioluminescent systems may involve other components such as ATP, Ca^{2+} or Mg^{2+} ions, luciferin binding proteins, accessory fluorescent proteins, etc. (McElroy et al., 1974; Shimomura et al., 1962).

The environment has selective pressure on the colors of bioluminescent species. Statistical results showed that the wavelengths of bioluminescence in marine animals were clustered in the range from 450 nm to 580 nm depending on the environment of the animals (Hastings, 1996). The majority of pelagic and deep sea species have blue emission centered at 470 nm. The coastal marine species are mainly fluorescent green and the

terrestrial and fresh water species are fluorescent around 550 nm to 580 nm. Some organisms may even emit different fluorescent light in different organs of their bodies (Hastings and Morin, 1991).

The bioluminescent colors are determined by several factors. The structure of luciferin is obviously one determinative factor since it is the reagent of oxidative decarboxylation reaction and the carrier of excitation energy. The structure of enzyme luciferase can also change the wavelength of emitted light through different binding interactions with luciferin. Same luciferin can have different bioluminescent colors in different luciferase systems (Seliger and McElroy, 1964). Similarly, mutants of luciferase can change the emission wavelength also (Wood et al., 1989; Wood, 1995; McCapra et al., 1994). Some bioluminescent systems also use biological optical filters or accessory fluorescent proteins such as GFP to control the emission colors (Denton et al., 1985; Hastings and Morin, 1991).

1.1.2 The origin of bioluminescence

The origin of bioluminescence has been a controversial topic but it is much clearer at present. Some early theories proposed that bioluminescence was a side product of electron transport in micro-organisms or oxygen detoxification in primitive anaerobes to remove molecular oxygen (Hastings, J. W., 1968; McElroy and Seliger, 1962). Another attractive hypothesis argued that primitive luciferases might originated from anaerobic hydroxylases that used low levels of molecular oxygen directly as an electron acceptor to metabolize all organic molecules such as aliphatic and polycyclic hydrocarbons as energy

sources (Seliger, 1975). If these hypotheses are correct, bioluminescence could have evolved over hundreds of millions of years without significant ecological functions before the appearance of any photoreceptors. Since the spectral distribution of bioluminescence overlaps with the visual spectral sensitivities of acceptors, it is possible that the appearance of bioluminescence is a relatively recent event, evolutionarily related to the evolution of light detection ability (Harvey, 1952; Seliger, 1982). With the appearance of photoreceptors, bioluminescence obviously was under strong selection pressure in evolution and was very essential for survival of individuals or species. Due to the importance of bioluminescence, other organic molecules such as $FMNH_2$ and long-chain aldehyde were also used as sources of luciferin (Hastings and Nealson, 1977; Hastings, 1978).

It seems unlikely that all the diverse bioluminescent systems originated from a common ancestral biological process such as hydroxylation. A new assumption for the origin of bioluminescence has been proposed based on the diversity of luciferin, luciferase, and bioluminescent mechanisms in different bioluminescent systems (Ward, 1985; Hastings, 1996). The chemical structures of several luciferins from different organisms showed that they were completely unrelated molecules. The molecular weight of luciferases is in a wide range from 21 kDa to 300 kDa in different bioluminescent systems which means they probably originated from different enzymes. The bioluminescent process such as the molecular components and chemical reactions involved in bioluminescence also showed considerable diversity in different systems. However, the luciferases and bioluminescent processes retain the characteristics of the enzyme and

biological process from which the luciferase might evolved. For example, it was proposed that firefly luciferases may be mutants of fatty acyl coenzyme A synthetases and aminoacyl t-RNA synthetases due to the similarity of reaction mechanisms between firefly luciferases and these two classes of enzymes (Rhodes and McElroy, 1958; McElroy et al., 1967). This assumption has been confirmed by the sequence similarity between firefly luciferases and acyl coenzyme A synthetases (Suzuki et al, 1990). Because of these diversities in bioluminescence and the similarities between luciferases and their ancestors, it was suggested that bioluminescence may have originated independently many times through various evolutionary lines (Ward, 1985; Hastings, 1996).

1.1.3 Ecological functions of bioluminescence

Five categories of ecological functions of bioluminescence have been suggested (Buck, 1978): emitter-limited behavior including food getting and defense; illumination; intraspecific communication including courtship and congregation; interspecific adaptation; altruism or group benefits. Several examples are given here.

Food   Bioluminescence can be used solely for the purpose of food predation. Many species of deep-sea fishes use cultures of luminescent bacteria symbiotically living inside the fishes as artificial bait (Herring and Morin, 1978; Bertelsen, 1951). In some cases, the bioluminescent lures are near or even inside the mouth of predators (Bertelsen and Struhsaker, 1977; Buck, 1978). Bioluminescence can also be used for illumination to find prey (Morin et al., 1975).

Defense An example in this category is counterillumination. Some fishes and squids can match the intensity of their bioluminescence with the intensity of ambient light and extinguish completely their bioluminescence in the dark (Lawry, 1974; Young and Roper, 1977; Case et al., 1977). These animals have sophisticated feedback control systems to monitor the intensity of ambient light and their own bioluminescence. This is a strategy of camouflage that conceals fishes from predators. Other defense strategies such as "blink and run" used in *Photoblepharon* have also been reported (Morin et al., 1975). This fish has headlight-like organs filled with symbiotic luminescent bacteria. When in danger, the fish makes a sudden high frequency blinks of light in order to frighten a predator, and abruptly changes its swimming direction at the same time.

Illumination Some species of fireflies can use intensified bioluminescence to illuminate the landing area as they alight to avoid hazards such as a small pool of water (Buck, 1978). Bioluminescence can also be used to assist predation by enabling the predators to see its environment.

Intraspecific communication This category include courtship and congregation. For the purpose of mating, fireflies may recognize the flash pattern of their own species, whether the light is from a male or female (Lloyd, 1978). Evidence has been shown for the correspondence between the wavelength of bioluminescence and the visual spectral sensitivity (Lall et al., 1980; Seliger et al., 1982).

Interspecific adaptation In genus *Photuris*, besides their own species-specific flash pattern for mating, carnivorous *Photuris* females can mimic the flash pattern of other genera to attract foreign males and eat them (Lloyd, 1975; Lloyd, 1980). On the other

hand, the male *Photuris* can mimic the flash pattern of the female's prey to locate and seduce females.

Altruism or group benefits   It has been known for several decades that certain species of fireflies can congregate in hundreds and flash in synchronous and rhythmic style (Buck and Buck, 1968; Buck and Buck, 1978).

1.2 The history of studies of GFP

Some bioluminescent systems use accessory fluorescent proteins as ultimate emitter to change the color of emitted light or increase the quantum yield of bioluminescence, i.e. the amount of bioluminescent light per mole of luciferin molecules oxidized. GFP was the first accessory fluorescent protein found as the ultimate light emitter in bioluminescent systems (Morin and Hastings, 1971a, b). It was found in marine invertebrates such as Pacific Northwest jellyfish *Aequorea victoria* and sea pansy *Renilla reniformis* from the Georgia coastline. In 1974, *Aequorea* GFP was purified, partially characterized, and studied with respect to the energy transfer *in vitro* between aequorin and GFP (Morise et al., 1974). Several isoforms of *Aequorea* GFP were found. The first quantitative study of *in vitro* energy transfer in a natural biological system, sea pansy *Renilla reniformis*, was done in 1976 (Ward and Cormier, 1976). The quantum yield *in vitro* is increased from 5.3% without GFP to 30% after adding GFP to a luciferase solution and the calculated efficiency of energy transfer is nearly 100%. Further studies showed that the energy transfer was radiationless and through highly specific protein-protein interactions between

GFP and luciferase in *Renilla* (Ward and Cormier, 1978; 1979; Hart et al., 1979). In 1979, the first chemical structure model of the fluorophore in *Aequorea* GFP was proposed by isolating the fluorophore and analyzing its components (Shimomura, 1979). In 1980, NMR studies of *Aequorea forskalea* GFP showed that the fluorophore was very tightly bound and the entire protein appeared to show marked conformational inflexibility (Nageswara Rao et al., 1980). It was proposed that *Renilla* and *Aequorea* GFPs had the same fluorophore and the large absorption differences resulted from the unique protein environment (Ward et al., 1980). In 1981, the renaturation of denatured *Aequorea* GFP was achieved the first time (Bokman and Ward, 1981; Ward and Bokman, 1982). The spectral properties of renatured GFP were very similar to those of native GFP, but small spectral differences were found in the absorption and the circular dichroism (CD) spectra, indicating small structural differences between native and renatured GFP. In 1982, it was found that the spectra of *Aequorea* GFP could be perturbed by pH, temperature, ionic strength, and concentration of GFP (Ward et al., 1982). In 1988, native *Aequorea victoria* GFP was crystallized for structural determination (Perozzo et al., 1988). In 1989, a modified chemical structure model of the fluorophore in *Aequorea victoria* GFP was proposed (Ward et al., 1989; Cody et al, 1993). In 1992, the cDNA of *Aequorea victoria* GFP was sequenced and recombinant GFP was available (Prasher et al., 1992). Later, the recombinant GFP was expressed in *E. coli* (Inouye and Tsuji, 1994a). In the same year, it was first shown that GFP could be used as a biological marker for gene expression in both prokaryotic and eukaryotic cells (Chalfie et al., 1994). Since then, extensive studies have been performed on applications of GFP and the preparation of mutants.

1.3 Two typical bioluminescent systems containing GFP

1.3.1 *Renilla* bioluminescence

In sea pansy *Renilla reniformis*, there is a luciferin binding protein that non-covalently binds a molecule of coelenterate-type luciferin (Anderson et al., 1974; Hori et al., 1977). The luciferin binding protein is a single polypeptide chain protein with a molecular weight of 18.5 kDa. The bound luciferin is inaccessible to luciferase and there is no free luciferin *in vivo*. Similar to the calcium binding proteins such as calmodulin, muscle troponin C, and aequorin, *Renilla* luciferin binding protein has two high affinity calcium binding sites and it releases the bound luciferin in the presence of calcium ions. The free luciferin can be oxidized and decarboxylated by molecular oxygen catalyzed by luciferase which is a single polypeptide chain protein with one luciferin binding site and a molecular weight of 35 kDa (Matthews et al., 1977). The luciferase-bound oxyluciferin is in an electronically excited state and emits blue light with broad emission spectrum peaking at 480 nm if no GFP is present. In the presence of GFP, the energy of the excited oxyluciferin is transferred in a radiationless process to GFP which emits green light with a very narrow emission spectrum peaking at 509 nm (Wampler et al., 1971; Wampler et al., 1972; Ward and Cormier, 1976; Ward and Cormier, 1978).

It has been suggested that GFP is not just used for shifting the wavelength of fluorescent light, which can be done easily by changing the fluorophore or its environment inside the primary photoproteins, but that it may improve the overall quantum yield of fluorescence. If the efficiencies of energy transfer and fluorescence of GFP are high, the

jellyfish can produce more light by transferring energy from primary photoproteins to GFP for the same amount of energy input. It has been shown that the quantum efficiencies of primary photoproteins are relatively low (Ward, 1979; Morise et al., 1974). The quantum yield of *Renilla* luciferase without GFP is 5.3% *in vitro* and it was increased to 30% by adding *Renilla* GFP (Ward and Cormier, 1976).

The *Renilla reniformis* GFP is a dimer of two identical single polypeptide chain protein with a molecular weight of 27 kDa for each. Since the radiationless energy transfer is still very efficient at very low concentration of GFP and luciferase, it is very possible that they form a complex before the energy transfer (Ward and Cormier, 1976; Ward and Cormier, 1978). The fact that these proteins are also membrane-bound *in vivo* further facilitates the association and energy transfer of the complex (Anderson and Cormier, 1973). The efficiency of energy transfer has been measured and was determined to be 100% (Ward and Cormier, 1976).

1.3.2 *Aequorea* bioluminescence

In jellyfish *Aequorea victoria* (Figure 1.1), the bioluminescent mechanism is completely different from that in sea pansy *Renilla reniformis*. The generically termed photoprotein aequorin in *Aequorea* is a complex containing several components: apo-aequorin which has the function of luciferase, luciferin, and covalently bound molecular oxygen. Apo-aequorin is a single polypeptide chain protein having a molecular weight of 22.5 kDa and one non-covalent luciferin binding site, one covalent molecular oxygen binding site, and three calcium binding sites (Cormier, 1978; Prasher et al., 1987;
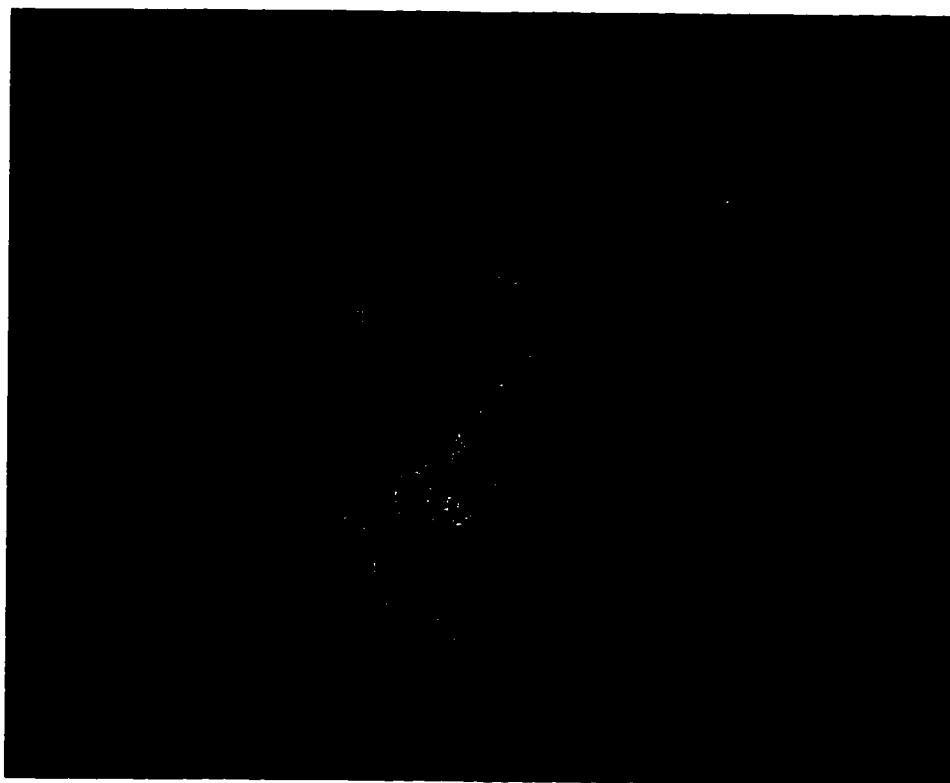
Figure 1.1 Photograph of *Aequorea victoria*. The cells on the edge of its body can express GFP and emit green light.

Kretsinger et al., 1988). The aequorin-bound molecular oxygen is inaccessible to the aequorin-bound luciferin in the absence of calcium ions. The oxidative decarboxylation reaction is triggered by the binding of calcium ions and the aequorin-bound oxyluciferin is in electronically excited state. It can either emit blue light with emission maximum at 469 nm *in vitro* in the absence of GFP or transfer its energy to GFP *in vivo* which emits green light with emission maximum at 508 nm (Morise et al., 1974).

1.4 The spectral characteristics of GFP

The absorption spectra of *Aequorea* GFP and *Renilla* GFP are quite different. *Aequorea* GFP has a major maximum at 395 nm and a minor maximum at 477 nm (Figure 1.2a) (Morise et al., 1974). *Renilla* GFP has a major maximum at 498 nm and a shoulder near 470 nm which was found to be a vibrational structure from the electronic transition of the 498 nm band (Figure 1.2b) (Wampler et al., 1971; Ward and Cormier, 1979). Despite the large differences in absorption spectra, *Aequorea* GFP and *Renilla* GFP have similar emission spectra with a narrow peak at 508 nm and same quantum yield of about 0.8 (Morise et al., 1974; Ward and Cormier, 1978; Ward and Cormier, 1979; Ward, 1979; Hart et al., 1979).

Although numerous results have shown that *Aequorea* GFP and *Renilla* GFP are extremely stable, *Aequorea* GFP can be spectrally affected by temperature, ionic strength, pH, and GFP concentration which is related with its dimerization (Ward et al., 1982). When perturbed by these factors, the absorption spectra showed sharp isosbestic points
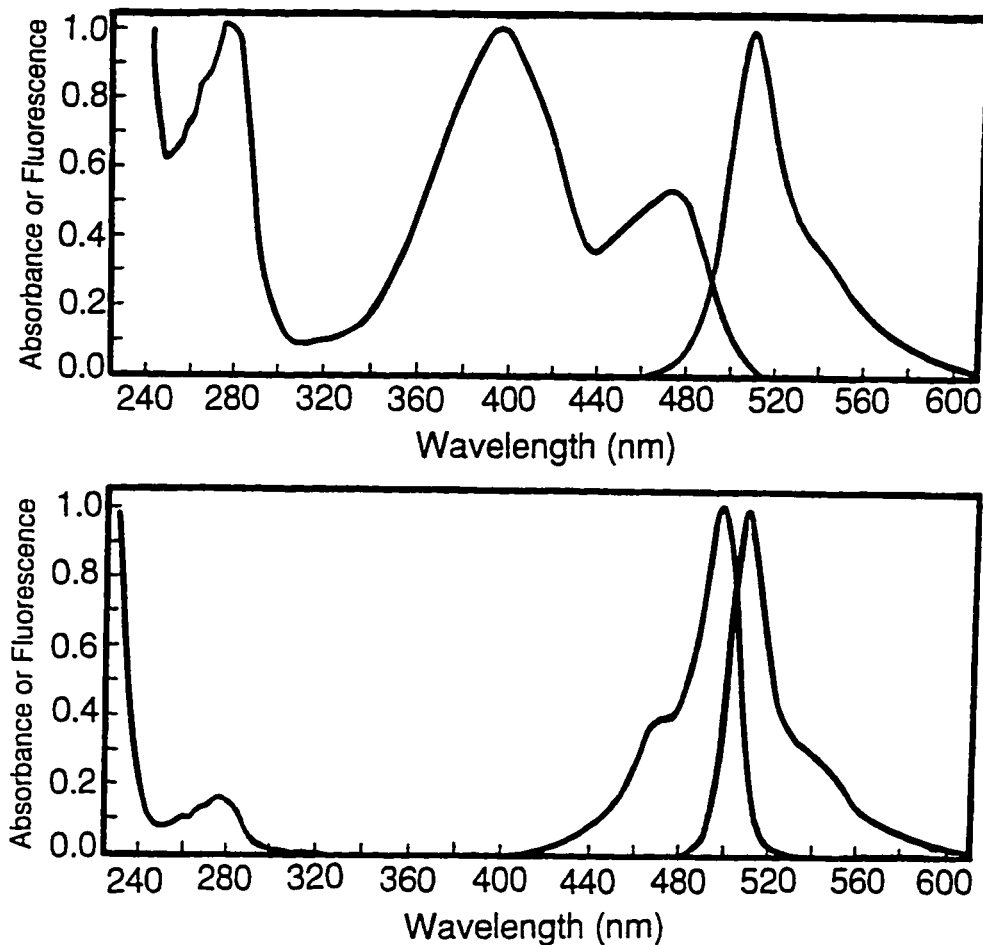
Figure 1.2 Absorption and emission spectra of *Aequorea victoria* GFP and *Renilla reniformis* GFP. a) The absorption spectrum of *Aequorea* GFP has a major maximum at 395 nm and a minor maximum at 477 nm. Its emission spectrum has a maximum at 508 nm. b) The absorption spectrum of *Renilla* GFP has a major maximum at 498 nm and a shoulder near 470 nm. Its emission spectrum has a maximum at 508 nm. Reproduced from Ward et al. (1980).

indicating that the fluorophore existed in one of the two spectral forms absorbing at 397 nm

and 477 nm respectively. This was confirmed by two GFP mutants that each have only

one of the two absorption maximum of native *Aequorea* GFP respectively and have the

other extinguished (Ehrig et al., 1995). *Renilla* GFP is even more stable than *Aequorea*

GFP and all attempts to perturb its spectra without denaturation by changing temperatures,

pH, protein denaturants, proteases, organic solvents, reducing agents, and various ionic

salts, were unsuccessful, suggesting that the fluorophore in *Renilla* GFP is much better

protected than in *Aequorea* GFP.

1.5 The fluorophore and primary structure of GFP

1.5.1 The fluorophore of GFP

Similarity of fluorophores in Aequorea GFP and Renilla GFP Contrary to the

non-covalent bound fluorophores of accessory fluorescent proteins in bioluminescent

bacteria, the fluorophores in *Aequorea* GFP and *Renilla* GFP covalently bind the proteins

as suggested by the results that they could not be separated from the denatured GFP by

dialysis or gel-filtration (Ward et al., 1980). The absorption spectra of fully denatured

Aequorea GFP and *Renilla* GFP were identical in pH range from 6.65 to 10.72 (Ward et

al., 1980). Their spectra also have same isosbestic point at 405 nm, indicating the

existence of two pH-dependent states in denatured GFP. The fluorophore could be

separated from denatured GFP using the non-specific protease pronase (Ward et al., 1980).

The absorption spectra of the yellow, fluorophore-containing peptide separated from

Aequorea GFP and *Renilla* GFP were identical to the spectra of denatured GFP and showed isosbestic point at 405 nm also. Thus, based on the similarities of emission spectra of native GFPs, absorption spectra of denatured GFPs, and absorption spectra of isolated fluorophore, it was suggested that *Aequorea* GFP and *Renilla* GFP have identical or at least very similar fluorophores and the difference in absorption spectra of native GFPs were caused by the non-covalent interactions between the fluorophore and its environment inside proteins (Ward et al., 1980).

Proposed formation of fluorophore The core of *Aequorea* GFP fluorophore was first identified as a substituted 4-p-hydroxy-benzylidene-5-imidazolone based on amino acid component analysis of the fluorophore-containing peptide and the comparison of absorption spectra of the peptide and a synthetic model compound (Shimomura, 1979). However, discrepancies between the model proposed by Shimomura and the components of the GFP fluorophore were found later and a new hexapeptide Phe-Ser-Tyr-Gly-Val-Gln model was proposed and characterized (Ward et al., 1989; Cody et al., 1993). Using the same methods, the fluorophore sequence of *Renilla* GFP was determined to be Phe-Ser-Tyr-Gly-Asp-Arg (SanPietro et al., 1993). Due to this special form of the fluorophore, GFP does not require any cofactors for fluorescence. It is suggested that the fluorophore is formed first by cyclizing the main chain between the Ser carbonyl carbon and the Gly amide nitrogen atoms (Heim et al., 1994; Cubitt et al., 1995) (Figure 1.3). Then, the side chain of Tyr66 is dehydrogenated at its $C\alpha$ and $C\beta$ atoms by forming a double bond. It is interesting that the cyclization reaction is similar to the known deamidation reaction of Asn-Gly sequences in peptide and proteins where the carbonyl carbon atom is also involved in
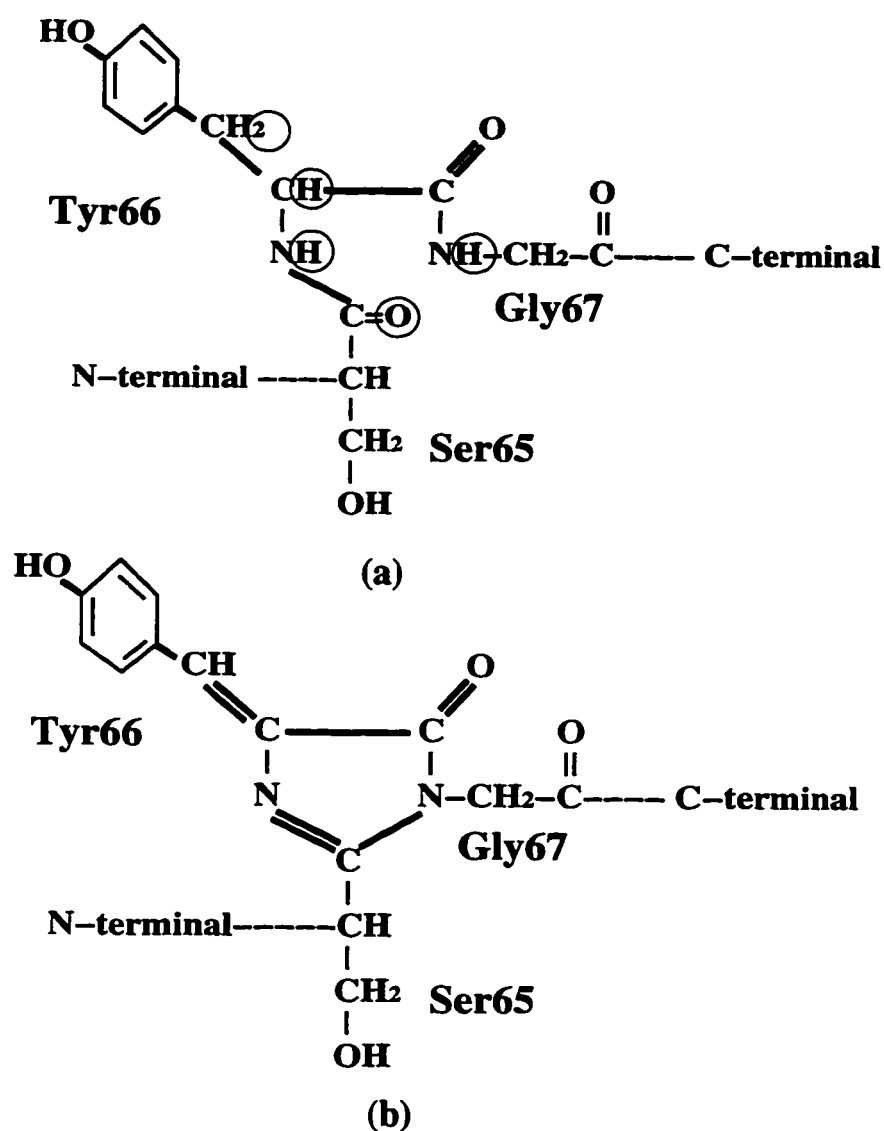
Figure 1.3 The chemical structure of the fluorophore in GFP. a) The amino acid sequence of the fluorophore. b) The chemical structure of the functional fluorophore. The main chain is cyclized by forming a covalent bond between the Ser carbonyl carbon atom and the Gly amide nitrogen atom. The side chain of Tyr66 is dehydrogenated at its Cα and Cβ atoms by forming a double bond.

the cyclization reaction (Wright, 1991). The details of mechanism of the post-translational fluorophore formation still remains unknown. Since GFP has been expressed widely in a wide variety of organisms and the GFP expressed in prokaryotic and eukaryotic cells has the same spectral features as native GFP (Chalfie et al., 1994), the reactions are probably autocatalytic or else require some common cell components.

It has been reported that the oxidation and reduction of Tyr66 is reversible and native GFP could be reduced by sodium dithionite, ferrous sulfate, or 2,2'-azino-di-(3-ethylbenzthiazoline-6-sulphonic acid (ABTS) and lose its fluorescence (Inouye and Tsuji, 1994b). The fluorescence of GFP could be recovered after the sample solution was exposed to air. It was not known at that time whether only the side chain of Tyr66 was reduced or the five-member ring was also broken. On the other hand, irreversible oxidation could be accomplished with hydrogen peroxide and H_2/Pt. It was found that if GFP was reduced by sodium dithionite, the absorption spectrum of reduced GFP was similar to that of native GFP in the range of from 400 nm to 600 nm, but if GFP was oxidized by hydrogen peroxide, the absorption peaks at 397 nm and 476 nm disappeared and could not be recovered.

Structural properties of GFP   The fluorophore is extremely stable in GFP under different denaturing conditions such as high temperature, acidic or alkaline environment, various proteases and several chemical denaturants commonly used (Bokman and Ward, 1981; Ward and Bokman, 1982). For example, at room temperature, the absorption and emission spectra of *Renilla* GFP were unchanged after long incubation in 6 M guanidine, 8 M urea, or 1% sodium dodecylsulfate (Ward and Cormier, 1979). Both *Renilla* and

Aequorea GFPs showed pH stability in a wide range from pH 5.5 to 12 (Ward, 1981; Bokman and Ward, 1981). GFP is also resistant to many proteases such as trypsin, chymotrypsin, papain, etc. The fluorescence remained unchanged after treatment with these proteases for two days (Roth, 1985). Similar to the results of absorption spectra, CD spectra of *Aequorea* GFP and *Renilla* GFP suggested that only the fluorophore in *Aequorea* GFP could be perturbed without denaturing the protein (Ward et al., 1982). When the temperature was increased from 25 °C to 80 °C, the changes in CD spectra at 205 nm which is related to the loss of secondary structure of protein, closely resembled the loss of fluorescence for both *Aequorea* GFP and *Renilla* GFP (Ward, 1981). Thus, the fluorophore is probably buried inside the protein and not accessible to the solvent, and the secondary structure of GFP is essential for protecting the fluorophore (Ward, 1981).

The denaturation at several conditions only disrupted the secondary structure of GFP and the fluorophore remained intact since denatured *Aequorea* GFP could be renatured and had identical emission spectra as those of native GFP at neutral pH and pH 12.2. The absorption spectra and CD spectra of renatured *Aequorea* GFP were also very similar to those of native GFP except in some regions especially around 280 nm in the absorption spectra and 215 nm in CD spectra, indicating small structural changes in renatured GFP (Bokman and Ward, 1981; Ward and Bokman, 1982). Since CD spectra around 215 nm is mainly due to the β structure in proteins, the renatured GFP probably has slightly distinct β strands from native GFP due to the differences in this region in their CD spectra.

1.5.2 Primary structure of GFP

In 1992, the sequence of *Aequorea victoria* GFP was determined from both cDNA and genomic clones (Prasher et al., 1992) (Table 1.1). The recombinant *Aequorea* GFP has 238 amino acids and a molecular weight of 26888 Da. Three GFP genes were found in the *Aequorea* population at Friday Harbor, Washington, which is consistent with the three isoforms of GFPs purified from the jellyfish. On the other hand, the sequence of *Renilla* GFP has not been fully determined yet, but the partially determined sequence of *Renilla* GFP already showed some similarity with that of *Aequorea* GFP (Ward, personal communication). Both *Aequorea* GFP and *Renilla* GFP are acidic globular proteins with similar molecular weight of about 27 kDa. *Renilla* GFP forms a dimer with very strong non-covalent interactions between monomers (Ward and Cormier, 1979), but *Aequorea* GFP only forms a dimer at certain conditions due to the relatively much weaker interactions between monomers.

1.6 Applications of GFP

The molecular cloning of GFP cDNA and the demonstration that GFP can be expressed as a functional transgene in prokaryotic (*E. coli*) and eukaryotic (*Caenorhabditis elegans*) cells have opened exciting new avenues of investigation in cell, developmental and molecular biology (Prasher et al., 1992; Chalfie et al., 1994). Since then, fluorescent GFP has been expressed in yeast (Kahana et al., 1995), slime mold (Moores et al., 1996), plants (Casper and Holt, 1996; Epel et al., 1996), Drosophila (Wang and Hazelrigg, 1994),

Table 1.1 Amino acid sequence of wild-type *Aequorea victoria* GFP*.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	S	K	G	E	E	L	F	T	G	V	V	P	I	L	V

17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
E	L	D	G	D	V	N	G	H	K	F	S	V	S	G	E

33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
G	E	G	D	A	T	Y	G	K	L	T	L	K	F	I	C

49	50	51	52	53	54	55	56	57	58	58	60	61	62	63	64
T	T	G	K	L	P	V	P	W	P	T	L	V	T	T	F

65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
S	**Y**	**G**	V	Q	C	F	S	R	Y	P	D	H	**M**	K	R

81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96
H	D	F	F	K	S	A	**M**	P	E	G	Y	V	Q	E	R

97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112
T	I	F	F	K	D	D	G	N	Y	K	T	R	A	E	V

113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	128
K	F	E	G	D	T	L	V	N	R	I	E	L	K	G	I

129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144
D	F	K	E	D	G	N	I	L	G	H	K	L	E	Y	N

145	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160
Y	N	S	H	N	V	Y	I	**M**	A	D	K	Q	K	N	G

161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176
I	K	V	N	F	K	I	R	H	N	I	E	D	G	S	V

177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192
Q	L	A	D	H	Y	Q	Q	N	T	P	I	G	D	G	P

193	194	195	196	197	198	199	200	201	202	203	204	205	206	207	208
V	L	L	P	D	N	H	Y	L	S	T	Q	S	A	L	S

209	210	211	212	213	214	215	216	217	218	219	220	221	222	223	224
K	D	P	N	E	K	R	D	H	**M**	V	L	L	E	F	V

225	226	227	228	229	230	231	232	233	234	235	236	237	238
T	A	A	G	I	T	H	G	**M**	D	E	L	Y	K

* The segment of the fluorophore and five methionine residues are marked by bold letters.

zebrafish (Amsterdam et al., 1996), and in mammalian cells (Ludin et al., 1996; De Giorgi et al., 1996). There are several advantages of using GFP as a reporter over other fluorescent proteins:

a) The formation of fluorophore is a post-translational modification and is either autocatalytic or require some common cell components. No special co-factors or enzymes are required. Thus, theoretically it can be functional in any cell.

b) The fluorophore of GFP is well-protected by the secondary structure of the protein and is very stable in many denaturing conditions such as high pH, high temperature, proteases and detergents.

c) The fluorescence of GFP does not require any substrate or other molecules. It can emit green fluorescent light by absorbing incident light or accepting energy from an excited luciferin or photoprotein.

d) GFP can tolerate N- and C-terminal fusion to a broad variety of proteins. Many of the proteins have been shown to retain their native function, probably due to the relatively small molecular weight of GFP which reduces its influence on the function of target proteins.

e) GFP is not toxic to transformed cells since it has been expressed in a wide variety of organisms, and cells transformed with GFP vector can grow well.

With these advantages, GFP has wide applications as a reporter for gene expression, protein localization, cell transformation, protein-protein interaction, enzyme activity, etc. in bacterial, plant cells, and mammalian cells. Furthermore, the use of multiple GFP mutants can monitor multiple cellular events such as gene expression

cascades and regulation. Three areas of the applications of GFP are described briefly below.

GFP as a marker for protein location in fusion proteins The traditional method of studying the location and movement of a target protein is to isolate the protein molecules and covalently label them with fluorescent organic compounds or fluorescent proteins, and to introduce the labeled target protein molecules into cells. Although this method can theoretically be applied to any proteins, there are several limitations. The target protein needs to be purified in quantity and covalently labeled without denaturation, and cells should not be perturbed by microinjection of the labeled protein molecules. For fluorescent protein labels, there are some other problems such as special prosthetic group may be required as the fluorophore, different cofactors may be necessary for the formation of the fluorophore or protein, and specific substrates may be needed as the energy source of fluorescence and must be provided outside. The use of GFP as a fluorescent tag in a fusion protein can overcome these difficulties. Protein purification and microinjection are not required and the functions of GFP domain and target protein domain can be retained as reported in many applications. For example, the tubulin-GFP fusion protein in yeast is fluorescent and can be incorporated into microtubules. The function of fluorescence-labeled microtubules can be observed over extended periods using video microscopy (Stearns, 1995). Besides targeting a protein, highly specific intracellular localization including the nucleus, mitochondria (Rizzuto et al., 1996), secretory pathway (Kaether and Gerdes, 1995), plasma membrane (Marshall et al., 1995) and cytoskeleton (Kahana et al., 1995) can be achieved via fusions both to whole proteins and individual targeting

sequences presenting in a subcellular structure. The presence of various GFPs with different absorption and emission maxima allows monitor of localization of two or more targets or cellular events simultaneously by selectively exciting and detecting a specific GFP mutant (Rizzuto et al., 1996).

GFP as a reporter for gene expression GFP can be co-expressed with a target protein as a reporter for gene expression by inserting GFP gene under the control of the same promoter of the target gene. GFP can also be used to estimate the relative or absolute expression level of the target gene if the relationship between GFP fluorescence and its expression level has been established. In one application, GFP was introduced into the expression cassette of a virus vector of potato virus X (PVX) (Baulcombe et al., 1995). Virus infection was indicated by fluorescence on plant leaves. When the coat protein gene was replaced by GFP gene, virus was restricted to the inoculated cells as indicated by fluorescence, confirming the assumption that the coat protein is essential for virus infection from cell to cell.

GFP mutants for fluorescence resonance energy transfer (FRET) Energy can be transferred from an excited fluorophore to a ground-state fluorophore if the distance between them is less than a few nanometers. Also, the peak region in the emission spectrum of the donor must overlap with the peak region in the excitation spectrum of the acceptor for efficient energy transfer (Tsien et al., 1993). In order to test the feasibility of energy transfer between two GFP mutants, fusion proteins containing one donor GFP and one acceptor GFP separated by a cleavable peptide linker have been designed (Heim and Tsien, 1996; Mitra et al., 1996). In two independent experiments, the donor GFPs have

absorption maxima at 381 and 385 nm, and emission maxima at 445 and 450 nm, respectively. The acceptor GFPs have absorption maxima at 479 and 488 nm and emission maxima at 507 and 505 nm, respectively. When excited by ultra-violet light, the spectrum of uncleaved fusion protein showed efficient energy transfer between the two GFP domains as indicated by strong green fluorescent light emitted from the acceptor GFP. After treatment with protease, the fusion protein was separated as individual GFP domains and the energy transfer was disrupted as indicated by the increase of blue fluorescent emission from the donor GFP molecules and disappearance of green fluorescence emission. Control experiments showed that the GFP mutants were resistant to the proteolysis. Disruption of FRET has been used to monitor the activities or inhibition of protease using artificial fluorophores linked by synthetic peptides (Krafft and Wang, 1994; Knight, 1995). The use of GFP to replace synthetic fluorophores has the unique advantage of eliminating some processes such as microinjection. FRET between GFP mutants can also be used to monitor the association of two target proteins by constructing fusion proteins each containing a target protein domain and a donor or acceptor GFP domain. The association of the two fusion proteins may cause FRET if the distance between the GFP domains is small enough.

1.7 Mutagenesis of GFP

Different approaches of mutagenesis have been used to search for mutants with desired properties. Site-directed mutagenesis has been used to produce mutations at some

critical positions such as Ser65, Tyr66 to change the spectral properties of GFP (Heim et al., 1994; Heim et al., 1995). Another approach is to simulate molecular evolution using random mutagenesis. The amino acid sequence subjected to mutagenesis can be a short section of several or tens of residues flanking the fluorophore (Delagrave et al., 1995; Cormack et al., 1996) or the whole sequence of the protein (Heim et al., 1994, Ehrig et al., 1995; Heim and Tsien, 1996; Crameri et al., 1996). Hundreds of thousands of colonies can be screened manually or automatically using high throughput screening technique such as digital imaging spectroscopy or fluorescence-activated cell sorting which also are important applications of GFP (Delagrave et al., 1995; Crameri et al., 1996). A combination of these two approaches, i.e. the cDNA of a GFP mutant is subjected to random mutagenesis to further improve its properties has also been used (Heim and Tsien, 1996). Currently known mutants are listed in Table 1.2.

Ser65 mutants   Apparently, the residues forming the fluorophore or flanking it in primary structure can be critical for altering GFP spectral properties. Random mutagenesis targeted on six residues starting from Phe64 to Gln69 (FSYGVQ) showed that the two -fluorophore residues Tyr66 and Gly67 were highly conserved for fluorescence while the type of other four residues were less stringent (Delagrave et al., 1995). Several absorption maximum red-shifted mutants have been sequenced which typically have Ser65 replaced by Thr, Gly, or Ala together with some other mutations (Heim et al., 1994; Delagrave et al., 1995; Cormack, 1996). No charged or aromatic amino acids was found at this site without the disappearance of fluorescence (Delagrave et al., 1995). The absorption spectra of these mutants are similar to each other and to that of *Renilla* GFP, i.e. there is only one

Table 1.2 GFP mutants.

Mutant	Absorption /Excitation maximum (nm)	Emission maximum (nm)	Comments	References
Aequorea GFP	398 (480)*	509	native protein	Morise et al., 1974
Renilla GFP	499	507	native protein	Wampler et al., 1971
S65T/T203H	512	524	red-shifted excitation	Ormo et al., 1996
S65T/T203Y	513	525	red-shifted excitation	as above
F64L/S65G/S72A/ T203Y	513	525	improved folding	as above
S65G/V68L /S72A/T203Y	513	527	improved folding	as above
S65G/S72A/T203W	502	512	improved folding	as above
F64G/S65C /V68M/Q69N			red-shifted excitation	Delagrave et al., 1996
F64G/V68L/Q69L			red-shifted excitation	as above
F64L/S65L/V68A/			red-shifted excitation	as above
F64M/S65G/Q69L	490	505	red-shifted excitation	as above
F64V/S65A /V68M/Q69L			red-shifted excitation	as above
F64L/S65C /V68T/Q69V			red-shifted excitation	as above
E222G	481	506	single excitation maximum	Ehrig et al., 1995
T203I	400	512	single excitation maximum	as above
Y66W	458	480	weak fluorescence	Heim et al., 1994
Y66H	382	448	medium fluorescence	as above
I167T	471 (396)	502 (507)	increased fluorescence	as above
I167V	471 (396)	502 (507)	increased fluorescence	as above
S202F/T203I	398	511	single excitation maximum	as above
Y66F			no detectable fluorescence	as above
F64M/Y66H/V68I	385	450	blue-shifted fluorescence	Mitra et al., 1996
S65A	471	503	single excitation maximum	Heim et al., 1994
S65C	479	507	single excitation maximum	as above
S65T	489	511	single excitation maximum	as above
F64L/S65T	488	507	efficient folding, increased whole cell fluorescence	Cormack et al., 1996
S65A/V68L/S72A	481	507	as above	as above
S65G/S72A	501	511	as above	as above
S72P/F99S/K126E /N146Y/M153T /V163A/E234D			increased whole cell fluorescence	Crameri et al., 1996
F99S/M153T/V163A /Q184R			increased whole cell fluorescence	as above
F99S/M153T/V163A			as above	as above
Y66W/N146I/M153T /V163A/N212K	433 (453)	475 (501)	brighter than Y66W	Heim & Tsien, 1996
Y66H/Y145F	381	445	brighter than Y66H	as above
Y66W/I123V/Y145H /H148R/M153T /V163A/N212K	432 (453)	480	brighter than Y66W	as above
S65T/M153A/K238E	504 (396)	514	longest excitation maximum	as above

* Values in parentheses are lower-amplitude peaks.

absorption maximum around 490 nm. Despite of the dramatic changes in absorption spectra, the changes in emission maximum of the mutants are small compared with that of native *Aequorea* GFP and *Renilla* GFP. One advantage to shift major absorption maximum to about 490 nm is that the argon laser emission at 488 nm can be used to excite the GFP mutants which results in much more intense fluorescence. Also, the existence of the single absorption maximum makes it much easier to discriminate different GFP mutants by selectively exciting GFP mutants as shown to be feasible (Delagrave et al., 1995; Heim and Tsien, 1996).

Tyr66 mutants   Residue Tyr66 has been be replaced by His or tryptophan which resulted in blue-shifted and weakly fluorescent mutants (Heim et al., 1994). Y66H mutant has the absorption maximum at 397 nm shifted to 382 nm and emission maximum at 508 nm shifted to to 448 nm, respectively. However, Y66W mutant has an absorption maximum at 458 nm and emission maximum at 480 nm, and its fluorescence is much weaker than that of Y66H. The replacement of Tyr66 by Phe results in no detectable fluorescence. In order to improve the brightness of Tyr66 single mutant, the cDNA of Y66H and Y66W were subjected to random mutagenesis and screened for increased fluorescence. Three multiple mutants with extinction coefficients comparable to the wild-type GFP have been found (Heim and Tsien, 1996). Most of the mutations occurred in the range from residue 145 to 163, plus residue Aln212.

Mutant with efficient and fast folding   In order to increase the whole cell fluorescence, GFP mutants were produced by recursive cycles of DNA shuffling and screened for the brightest *E. coli* colonies (Crameri et al., 1996). After three cycles of

DNA shuffling, a GFP triple mutant, (F99S/M153T/V163A), was found that increased whole cell fluorescence by over 40-fold greater than wild-type GFP in prokaryotic cells (*E. coli*) and eukaryotic cells (Chinese Hamster Ovary). The mutations M153T and V163A were also found in another independent study which was aimed to increase extinction coefficient and quantum yield of Y66W mutant by introducing additional mutations (Heim and Tsien, 1996). The factors that can cause such an increase of fluorescence are: increased protein expression, more efficient protein folding, and increased absorbance or quantum yield. The increase in quantum yield, if any, could not result in the 40-fold increase in whole cell fluorescence signal since the quantum yield of native GFP was already 80%. The expression levels of mutant and wild-type GFP were similar as studied by SDS-PAGE analyses, and therefore it could not be the reason for the increase in fluorescence either. Instead, significant improvement in protein folding of the mutant compared with wild-type GFP was observed which contributed to the increase in fluorescence. It was reported that large percentage of wild-type GFP molecules were not functional, had no absorption and fluorescent emission bands, probably due to incorrect protein folding which prevented the cyclization and oxidization of the fluorophore, i.e. no formation of the fluorophore (Heim et al., 1994). The incorrectly folded GFP molecules aggregated and were contained in the inclusion bodies of bacteria. On the contrary, the triple mutant has much more efficient folding and only a small percentage of apo-GFP was found in inclusion bodies. Thus, the triple mutations probably improve the folding and maturation of apo-GFP instead of improving other properties. No information about extinction coefficient of this mutant is available.

The mutant also showed more rapid fluorophore formation. Under the experimental conditions, the wild-type GFP is difficult to detect until 1-2 hours after induction. However, the triple mutant could be detected within eight minutes. This feature is very important in immediate real-time monitoring of cellular events.

Mutants with single excitation peak   Mutations far away from the fluorophore in primary structure also have a dramatic influence on spectral properties of GFP. The ratio of the two absorption maximum at 397 nm and 477 nm of native GFP can be perturbed by temperature, ionic strength, pH, and GFP concentration. All these perturbations have isosbestic points which are in a relatively narrow range from 420 nm to 431 nm, indicating the existence of two spectral forms of the fluorophore. This assumption was supported by two GFP mutants selected from colonies after random mutagenesis that have only one of the excitation peaks present and the other completely extinguished (Ehrig et al., 1995). The single mutant E222G has a single excitation maximum at 481 nm and emission maximum at 506 nm, while the single mutant T203I has single excitation maximum at 400 nm and emission maximum at 512 nm. The single mutant I167T, which has the major excitation maximum at 471 nm and minor excitation maximum at 396 nm, and emission maximum at 502 nm, has the similar spectral properties as E222G mutant (Heim et al., 1994).

1.8 The importance of solving the structure of GFP

Although significant progress has been made toward a better understanding of the mechanism and spectral characteristics of GFP fluorescence, there are still many properties

of GFP that can not be understood without detailed information regarding its molecular structure. For example, some residues such as Thr203 and Glu222 have significant influence on the spectral properties of GFP, but in the primary structure they are remote to the fluorophore. The explanation for the spectral changes caused by these mutations can only be given based on the molecular structure of GFP. Besides, it is not clear that whether or not there are any unidentified residues which might be also critical to the spectral properties of GFP. Without knowledge of the molecular structure of GFP, random mutagenesis is the only method to screen for such critical residues. The structure of GFP will show all the residues that are in the vicinity of the fluorophore and how they interact with it. Also, site-directed mutagenesis will be more helpful if all critical residues can be identified. The fluorophore in GFP has been suggested to be well-protected by the secondary structure of the protein. However, there is no clue about how the fluorophore is protected. Reversible oxidation and reduction of the side chain of Tyr66 has been proposed. However, no structural information is available about this intermediate state. In order to answer questions like these, have a better understanding about the fluorescent mechanism of GFP, and improve its properties by supporting studies such as mutagenesis, we chose to attempt to determine the molecular structure of wild-type *Aequorea* GFP.

Chapter 2

Basic Theory and Practice of
Macromolecular X-ray Crystallography

A brief description of macromolecular X-ray crystallography, methods of crystal

growth, Cryo-temperature experiment, and MAD data collection, will be given in this

chapter to facilitate the discussion later. The details of the theory and its application on

macromolecular structural determination have been well described in textbooks (Blundell

and Johnson, 1976; McRee, 1993; Ladd and Palmer, 1993; Rhodes, 1993; Drenth, 1994;

Gluster et al., 1994). The experimental methods have been summarized and discussed and

in publications due to their importance in solving molecular structures.

2.1 Elementary considerations of crystallography

A crystal can be built up by infinitely packing identical blocks known as unit cell in

three dimensions (Figure 2.1). Each unit cell may contain one or more symmetry-related

(a)   (b)
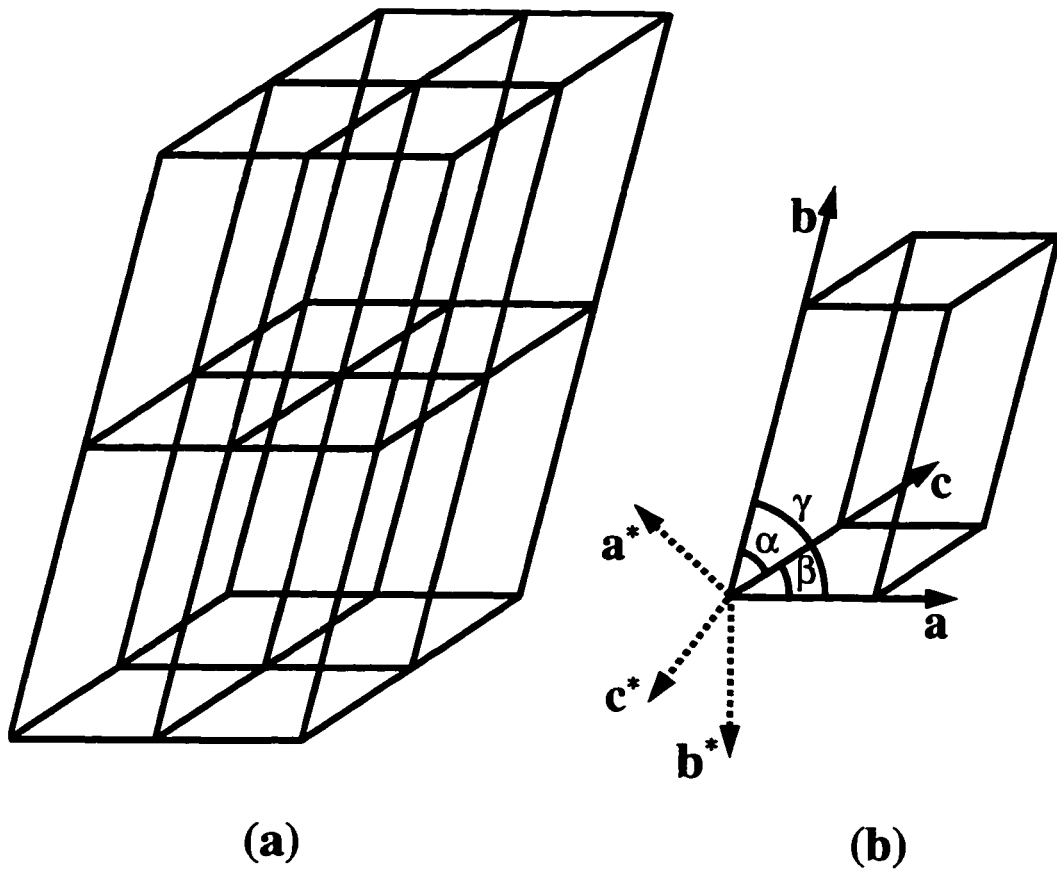
Figure 2.1 Crystal lattice and unit cell. a) A crystal can be built up by infinitely

packing the unit cell (shown in 2.1b) in three dimension space. Only translation operation

is required here. b) A unit cell is the element of a crystal. In real space, it is defined by

three vectors, a, b, and c. The angle between pairs of the vectors are traditionally defined

as α, β, and γ. These three vectors also define the axes in reciprocal space, a*, b*, and

c*, which define the reciprocal lattice corresponding to the crystal lattice in real space.

Note that each of the three vectors in the reciprocal space is perpendicular to a plane defined

by two vectors in the real space (equation 2.1).

lattice points, which are known as asymmetric units. All lattice points are equivalent to each other, i.e. the environment of one lattice point is exactly the same as that of others. Each asymmetric unit contains same molecular components. A unit cell that contains only one lattice point is a primitive unit cell. Otherwise, it is a non-primitive unit cell. In many cases, a non-primitive unit cell is chosen in order to show the symmetry of the lattice. The definition of a unit cell for a given lattice is also arbitrary, but preference is given to those who have short vectors and show the highest symmetry of the lattice. For a non-primitive unit cell, an operation that can superimpose any asymmetric unit to another in the same unit cell is named crystallographic symmetry operation. If one asymmetric unit contains two or more identical molecules, the rotation and/or translation operations that can superimpose one molecule to another in the same asymmetric unit is called non-crystallographic symmetry operations.

Due to the limitation of symmetry operations, the number of types of lattices is finite. For three-dimensional lattices, there are 7 crystal systems, 14 Bravais lattices, 32 classes of symmetry, and 230 space groups. The crystal systems are defined based on the highest rotation symmetry of the lattice. The Bravais lattices are defined based on the type of the seven crystal systems, i.e. whether the lattice is primitive, body centered, face centered, etc. The classes of symmetry are also named point groups. Each point group, or class of symmetry, contains a complete set of symmetry operations so that a combination of any of the two operations is still an operation of the group, and no translation operation is included. The space group is defined based on point groups with translation operations

added to these symmetry operations. The GFP crystals are belong to tetragonal crystal class, primitive lattice, point group 422, and space group P41212.

The three non-parallel edges of a unit cell form three vectors, defined as **a**, **b**, and **c** by convention (bold letters stand for vectors) (Figure 2.1). The three angles between **b** and **c**, **c** and **a**, **a** and **b**, are defined as α, β and γ, respectively. This system defines the framework for subsequent structural determination. An important concept in X-ray crystallography is reciprocal lattice which is built by a set of new unit cell vectors, **a***, **b***, and **c***, corresponding to a set of given unit cell vectors, **a**, **b**, and **c**, in real space:

$$\mathbf{a^*} = \frac{\mathbf{b} \times \mathbf{c}}{V}$$

$$\mathbf{b^*} = \frac{\mathbf{c} \times \mathbf{a}}{V}$$

$$\mathbf{c^*} = \frac{\mathbf{a} \times \mathbf{b}}{V} \tag{2.1}$$

where V is the volume of the unit cell. Similar to the real space lattice that can be constructed by translation along axes **a**, **b**, and **c**, a reciprocal lattice can also be constructed based on **a***, **b***, and **c*** in the space known as reciprocal space. Each reciprocal lattice point is associated with an integral index (h, k, l) which is the coordinates of the lattice point relative to its origin.

2.2 X-ray diffraction by a crystal

X-rays can be diffracted by electrons in crystals and the diffraction patterns depend on the symmetries and orientations of crystals. The diffraction condition can be determined by Bragg's Law (Figure 2.2). The lattice points in real space form crystal planes and the diffraction of incident X-rays can be treated as reflected by these planes. Since X-rays reflected by different planes have different path length, the glancing angle θ must be chosen so that constructive interference occurs:

$$2 \, d \sin\theta = n \, \lambda \qquad\qquad (2.2)$$

where d is the distance between crystal planes, λ is the wavelength of incident X-rays, and n is an integer. This is called Bragg's Law. Note that Bragg's Law only determines the direction of diffracted X-rays, but gives no information about the intensities.

The diffraction condition of X-rays can also be expressed geometrically using the concept of reciprocal lattice and the so called Ewald sphere (Figure 2.3). Draw a sphere with a radius of $1/\lambda$, where λ is the wavelength of the incident X-rays which pass through the center of the sphere. Take the cross point of incident X-rays and the sphere surface as the origin of the reciprocal lattice, i.e. one reciprocal lattice point is on this cross point and has index (0, 0, 0). If a reciprocal lattice point is rotated onto the sphere, X-rays will be diffracted in the direction from the center of the sphere to that lattice point. By the definition of reciprocal lattice, each vector from one reciprocal lattice point to another lattice point in reciprocal space is perpendicular to a set of crystal planes in real space. Thus, the method of Ewald sphere is actually a geometric representation of Bragg's Law.
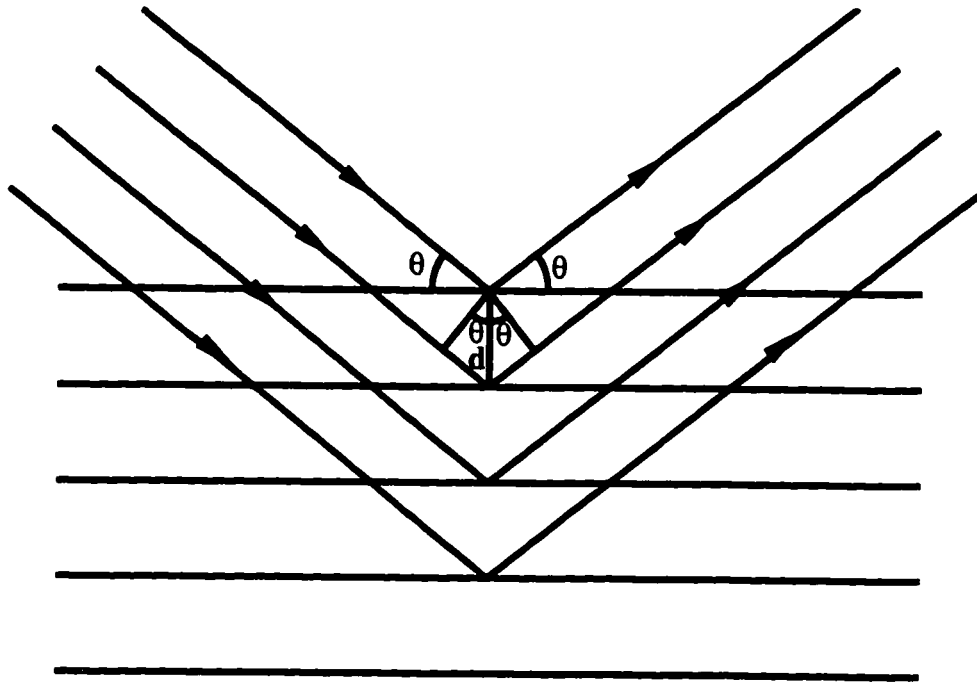
Figure 2.2 Bragg's Law. The crystal planes are defined by the crystal lattice points. The distance between planes is d. Diffraction of incident X-rays can be regarded as reflection from the planes. The incident X-rays have a glancing angle of θ and the diffracted X-rays also have the same glancing angle. The constructive interference occurs when Bragg's law is satisfied, i.e. $2 d \sin\theta = n \lambda$, where λ is the wavelength of X-rays and n is an integer starting from 1. The physical meaning of Bragg's law is that, the difference in optical path between X-rays diffracted from different crystal planes must be the integral times of the wavelength of X-rays.
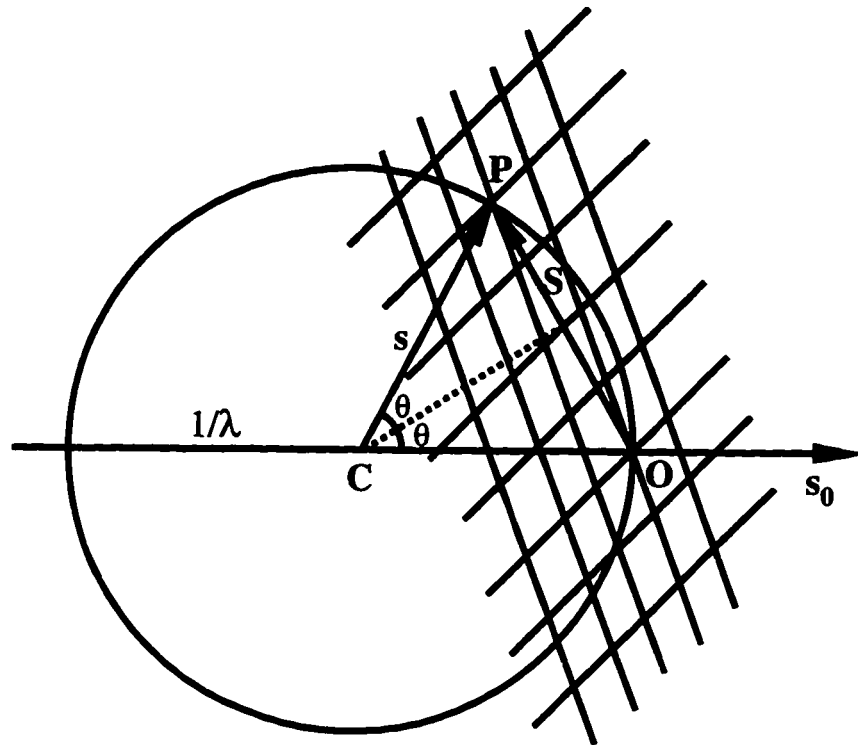
Figure 2.3 The Ewald sphere. Radius of the Ewald sphere is $1/\lambda$ where λ is the wavelength of incident X-rays. The incident X-rays (vector s_0) pass the center of the sphere and has a cross-point (O) on the sphere which is the origin of the reciprocal lattice. When a lattice point (P) is cutting the sphere by rotating the reciprocal lattice around its origin, incident X-rays are diffracted along the direction from the center of the sphere to that lattice point (vector s). The vector S from the origin of the reciprocal lattice to lattice point P is perpendicular to the crystal planes diffracting X-rays in Figure 2.2. The angle θ here is equal to the angle θ in Figure 2.2. The Ewald sphere is a geometrical expression of Bragg's law.

The intensities of diffractions depend on type and positions of all atoms in unit cell.

They are the square of amplitude of structure factors $F_{(h, k, l)}$ given by Fourier transform:

$$I_{(h, k, l)} = |F_{(h, k, l)}|^2$$

$$F_{(h, k, l)} = \int \rho_{(r)}\, e^{2\pi i(x \cdot h + y \cdot k + z \cdot l)}\, dr$$

$$= F_{(h, k, l)}\, e^{i\alpha(h, k, l)} \tag{2.3}$$

where $\rho_{(r)}$ is the function of electron density at position r which has the coordinate of (x, y,

z), and vector $F_{(h, k, l)}$ is named structure factor which has two components, amplitude $F_{(h,}$

$_{k, l)}$ and phase angle $\alpha(h, k, l)$. The integral is over the volume of whole unit cell. If all

atoms have no absorption of X-rays, i.e. the scattering of X-rays is elastic, the amplitude

of diffraction $F_{(-h, -k, -l)}$ is equal to that of $F_{(h, k, l)}$, but the phase of $F_{(-h, -k, -l)}$ is negated

compared with that of $F_{(h, k, l)}$. This is known as Friedel's Law and the diffraction spots

$F_{(h, k, l)}$ and $F_{(-h, -k, -l)}$ are termed Friedel pairs or Bijvoet pairs. If some atoms in crystal

have absorption of X-rays, it is the case of anomalous dispersion and the Friedel's Law is

broken. Anomalous dispersion is a very important phenomenon for solving

macromolecular structures in X-ray crystallography. By analyzing anomalous dispersion

signals, e.g. the differences between $I_{(h, k, l)}$ and $I_{(-h, -k, -l)}$ at same or different

wavelengths, phase information can be obtained for solving molecular structures (see

section 2.3.3).

The inverse Fourier transform is:

$$\rho_{(r)} = \frac{1}{V} \sum \left(F_{(h, k, l)} \times e^{-2\pi i(x \cdot h + y \cdot k + z \cdot l)} \right)$$

$$= \frac{1}{V} \Sigma \ (\ F_{(h, k, l)} \ X \ e^{i\alpha(h, k, l) \ -2\pi i(x \cdot h + y \cdot k + z \cdot l)} \ ) \qquad (2.4)$$

where the summation is over all reciprocal lattice points (h, k, l), V is the volume of unit cell, and $\rho_{(r)}$ is the electron density function at position r. The intensity of diffractions can be measured in experiment, which gives the amplitude of structure factors according to equation 2.3. However, the phase angles of all structure factors $F_{(h, k, l)}$ can not be obtained in experiment and hence the electron density function $\rho_{(r)}$ can not be calculated directly from equation 2.4. This is called the phase problem in macromolecular X-ray crystallography which is one of the two major problems in solving protein structures besides crystallization.

2.3 The solutions of the phase problem

2.3.1 Direct methods

For small molecules, direct methods, which use the statistical relationships between phase angles and amplitudes of structure factors to predict the phase angles, have been developed and routinely used for solving structures. However, it is difficult to apply direct methods to macromolecular X-ray crystallography because of the large number of atoms in protein molecules and poor diffraction resolutions of protein crystals relative to small molecule crystals. The largest number of atoms in one asymmetric unit that can be solved by direct methods so far is about 600 (Miller et al., 1994). However, direct methods are

still useful in solving heavy atom sites in isomorphous replacement methods as described below.

2.3.2 Multiple isomorphous replacement (MIR)

Like direct methods, isomorphous replacement (IR) was also initially developed for small molecule X-ray crystallography. It was shown that this method could also be used for solving the phase problem in protein crystallography by Perutz and his colleague (Green et al., 1954). The basic idea of isomorphous replacement is to introduce one or several heavy atoms, which have strong diffraction or absorption ability to incident X-rays, into the protein crystals without disturbing the crystal packing, i.e. the heavy atom soaked crystals are isomorphous derivatives of native crystals. The changes in structure factor amplitudes of such isomorphous derivative crystals are caused only by the heavy atoms and hence they can be used for determining coordinates of heavy atoms by direct methods or Patterson methods. Then, the phase angles of structure factors can be estimated from the triangles formed by three vectors of each diffraction (h, k, l) (Figure 2.4):

$$F_{PH(h, k, l)} = F_{P(h, k, l)} + F_{H(h, k, l)} \tag{2.5}$$

where the F's are the structure factors of protein molecules (P), heavy atoms (H), and the protein-heavy atom complex (PH). In equation 2.5, the vector of $F_{H(h, k, l)}$ is known by calculation based on the coordinates of heavy atoms, and the amplitudes of $F_{PH(h, k, l)}$ and
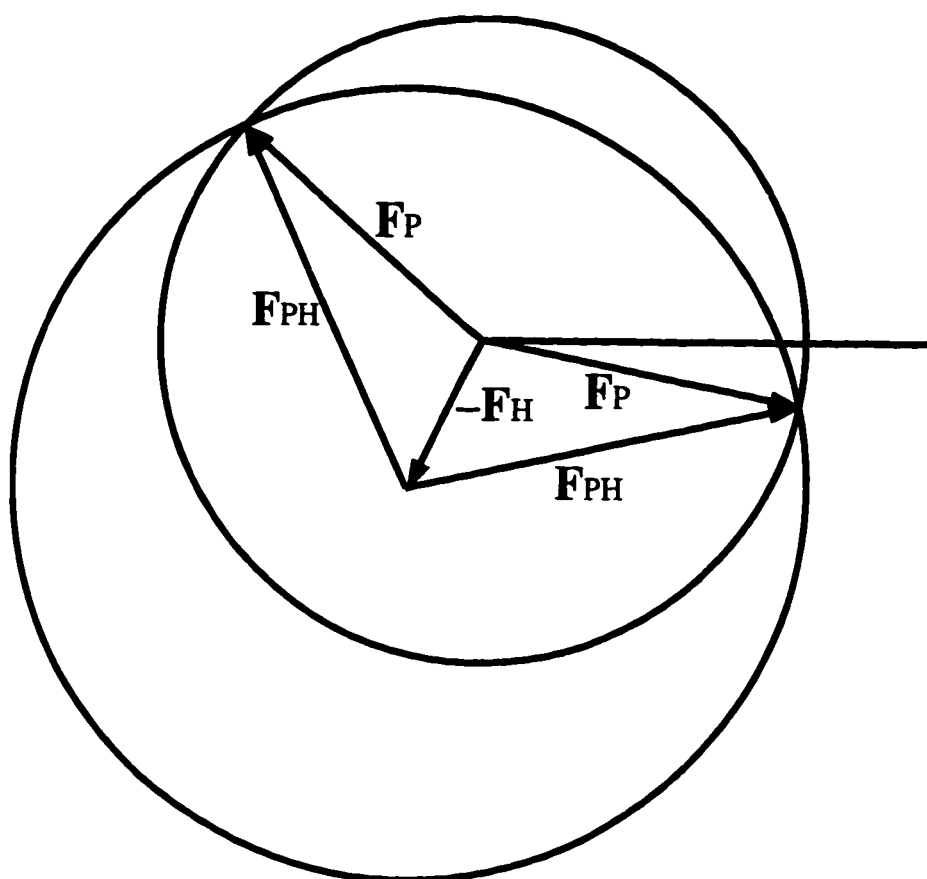
Figure 2.4 Determination of protein phase angle by isomorphous replacement methods. Vectors Fp, FH, and FPH are the structure factors of protein, heavy atom, and protein-heavy complex, and they have the vectorial relationship: FPH = Fp + FH . The structure factor FH can be calculated based on the coordinates of heavy atoms obtained previously. Since only the intensity of a diffraction spot can be measured, only the amplitudes of the structure factors FPH and Fp are known. Therefore, the phase angle of the protein structure factor Fp can not be calculated based on the vectorial relationship above and have to be determined by the cross-points as shown in the figure. The radii of the two circles are the amplitude of FPH and Fp respectively. Note that there are two cross-points in the figure.

$F_{P(h, k, l)}$ have been determined by experiments. Since the phase angles of $F_{PH(h, k, l)}$ and $F_{P(h, k, l)}$ are unknown, there are two possible phase angles for $F_{P(h, k, l)}$ so that normally two or more heavy atom derivatives are required to solve the ambiguity. This method is named multiple isomorphous replacement (MIR) and is the traditional method of solving macromolecular structures.

Depending on the type of the heavy atoms and wavelength of X-rays, two kinds of diffraction information can be obtained for solving structures. If heavy atoms have no significant absorption at the wavelength of X-rays used, they can be treated as normal scattering centers like nitrogen, oxygen, or carbon atoms, and multiple heavy atom derivatives are normally required to solve the unknown structure. However, many heavy atoms have significant absorption to incident X-rays such as the copper $K\alpha$ radiation at 1.5418 Å. In such cases, there is anomalous dispersion and heavy atoms should be treated as anomalous scattering center. Friedel's law is not valid if anomalous dispersion presents and the ambiguity of phase angles can be eliminated by considering the two triangles defined by the three vectors of (h, k, l) and its Friedel mate (-h, -k, -l) (Figure 2.5). The absorption of X-rays by heavy atoms can provide important structural information and sometimes only one heavy atom derivative is required for solving structures.

2.3.3 Multiwavelength anomalous dispersion methods (MAD)

Although the phenomenon of anomalous dispersion has been known long time ago and been used for solving macromolecular structures, MAD methods were developed rapidly only in recent years (Karle, 1989; Hendrickson et al, 1985). In MAD methods,
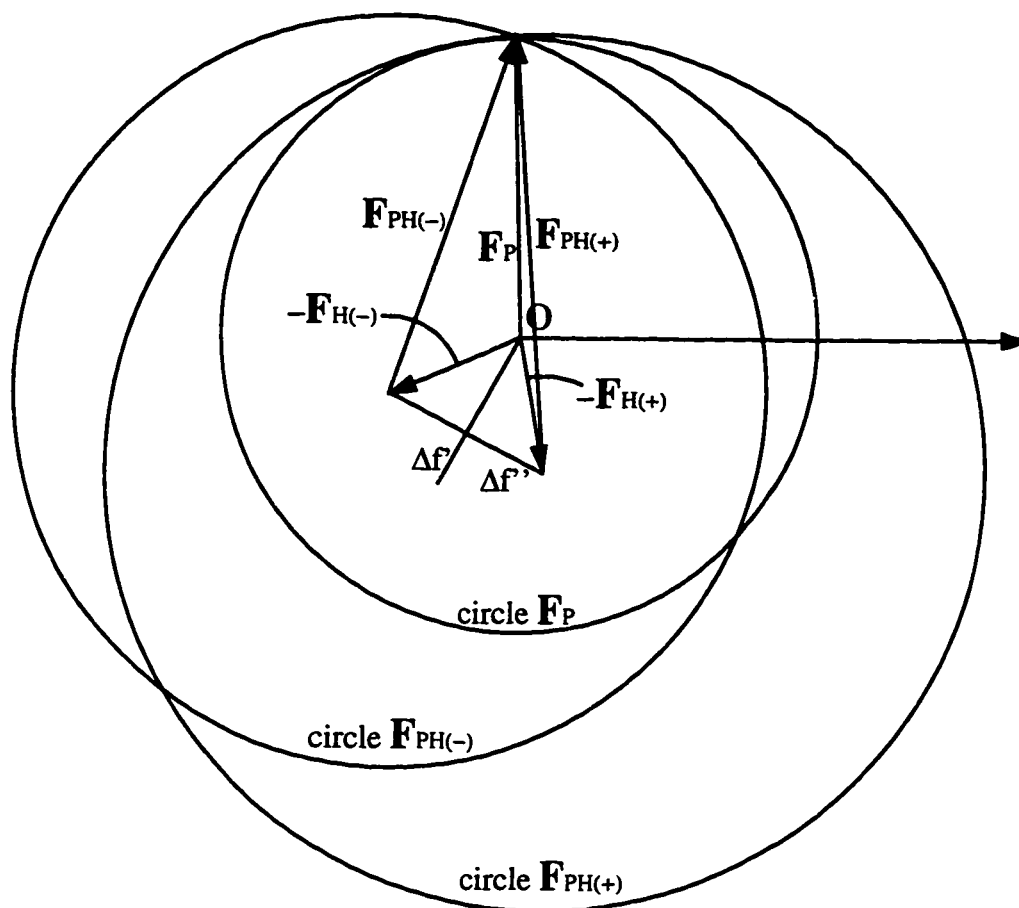
Figure 2.5   Determination of protein phase angle by single isomorphous replacement with information of anomalous dispersion. Vectors F_P is the structure factor of protein. Vectors $F_{H(+)}$ and $F_{H(-)}$ are the structure factors of Friedel mates of heavy atoms. Vectors $F_{PH(+)}$ and $F_{PH(-)}$ are the structure factors of Friedel mates of protein-heavy complex. The radii of the three circles are the amplitudes of the structure factors F_P, $F_{PH(+)}$, and $F_{PH(-)}$. Similar to the multiple isomorphous replacement methods described in Figure 2.4, the two Friedel mates are equivalent to two heavy atom derivatives.

only one heavy atom derivative is required and diffraction data are measured at several wavelengths of X-rays which are selected based on the absorption edge of the heavy atoms. The use of multiple wavelengths is equivalent to the preparation of multiple heavy atom derivatives in MIR. The phase information from MIR and MAD can be combined in order to improved the accuracy of phase angles. Since krypton and selenium atoms have much less electrons compared with traditional heavy atoms like platinum, mercury, etc., they do not have significant contributions to the amplitudes of structure factors. Therefore, MAD methods must be used for such derivatives and the wavelength can be tuned to their absorption edges around 0.96 to 0.97 Å at synchrotron sources. In the past several years, several protein structures has been solved by selenomethionyl crystals and this method was proved to be very successful since replacement of native methionine by selenomethionine is almost guaranteed (Kahn et al., 1985; Hendrickson et al., 1990; Hendrickson, 1991; Weis et al., 1991; Leahy et al., 1992; Leahy, 1994; Burling et al., 1996).

Because of the special theory of MAD methods, the data analyzing methods are quite different from those in MIR methods. The special formula of MAD methods used for data analysis can be derived in a vary simple way as shown below. By the definition of anomalous scattering parameters Δf and $\Delta f''$ which are the real and imaginary parts of anomalous dispersion, the contributions of anomalous dispersion are (Figure 2.6):

$$\text{Fano} = \frac{\textbf{Fa} \cdot \Delta f + i \cdot \textbf{Fa} \cdot \Delta f''}{f_o} \tag{2.6}$$

where **Fa** is the structure factor from normal scattering of all heavy atoms. The structure factor **F** from normal and anomalous scattering of all atoms can be expressed as:

$$F = Ft + Fano$$

$$= Ft + \frac{Fa \cdot \Delta f' + i \cdot Fa \cdot \Delta f''}{f_o} \tag{2.7}$$

where **Ft** is the structure factor from normal scattering of all atoms. The intensities of structure factors **F** can be derived from equation 2.7 as (Figure 2.6):

$$I = |F|^2 = |Ft|^2 + |Fano|^2 - 2 \cdot |Ft| \cdot |Fano| \cdot \cos\omega \tag{2.8}$$

where ω is the angle between vector **Ft** and **Fano**. The factor $\cos\omega$ can be expressed as:

$$\cos\omega = \cos(180° - \Delta\phi - \gamma)$$

$$= -\cos\Delta\phi \cdot \cos\gamma + \sin\Delta\phi \cdot \sin\gamma \tag{2.9}$$

where $\Delta\phi$ is the phase angle between vectors **Ft** and **Fa** for each diffraction (h, k, l) and:

$$\cos\gamma = \frac{|Fa| \quad \Delta f''}{|Fano| \quad f_o}$$

$$\sin\gamma = \frac{|Fa| \quad \Delta f'}{|Fano| \quad f_o} \tag{2.10}$$

The length of vector **Fano** is:

$$|Fano| = (\Delta f''^2 + \Delta f^2)^{1/2} |Fa| / f_o$$

Replace $\cos\gamma$ in equation 2.9 with 2.10, we have:

$$|F|^2 = |Ft|^2 + \frac{(\Delta f^2 + \Delta f''^2)}{f_o^2} |Fa|^2 + \frac{2 \Delta f' |Ft| |Fa| \cos\Delta\phi}{f_o} - \frac{2 \Delta f'' |Ft| |Fa| \sin\Delta\phi}{f_o}$$
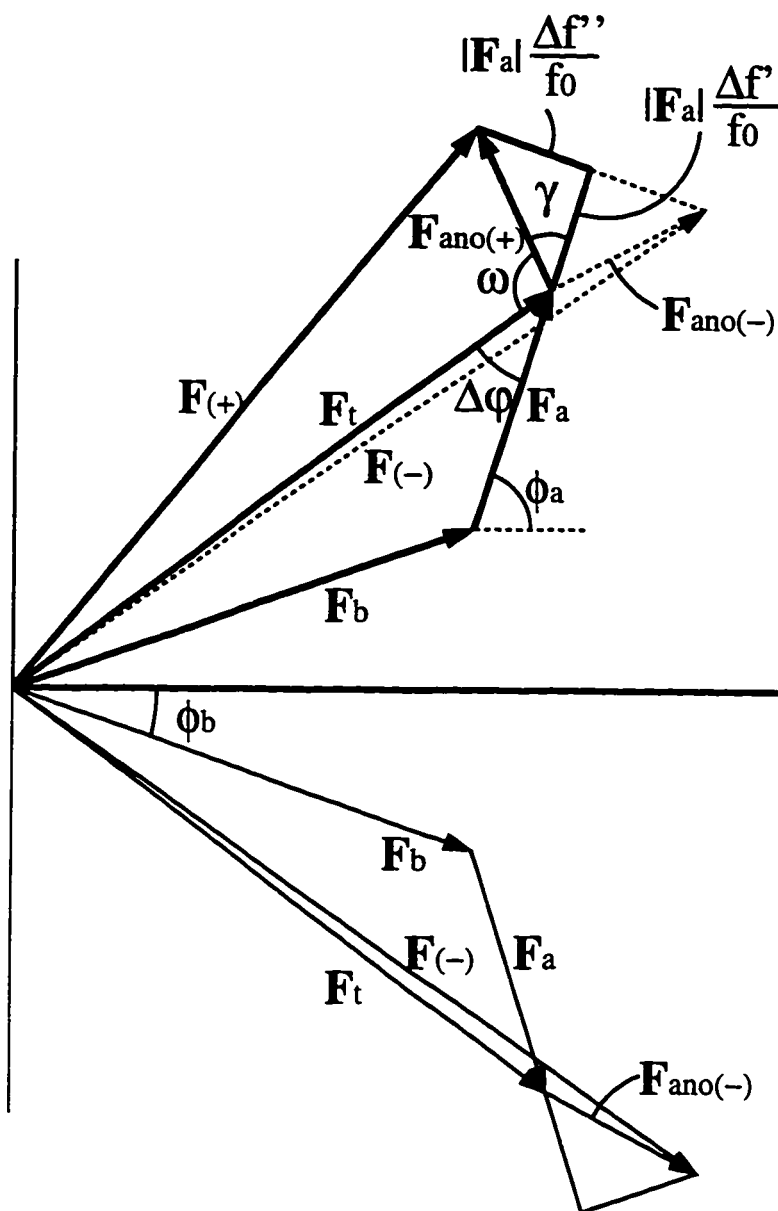
Figure 2.6 Vector representation of multiwavelength anomalous dispersion. The meanings of the vectors in the figure are given in the text. Note that there are three important vectorial relationships: $F = Ft + Fano$, $Fano = Fa \cdot \Delta f' / fo + i \cdot Fa \, \Delta f'' / fo$, and $Ft = Fa + Fb$. The Friedel mates of all vectors with opposite sign of phase angles are also given on the lower half of the figure.

$$= |Ft|^2 + w_1 |Fa|^2 + w_2 |Ft| |Fa| \cos\Delta\varphi - w_3 |Ft| |Fa| \sin\Delta\varphi \qquad (2.11)$$

where

$$w_1 = \frac{\Delta f^2 + \Delta f'^2}{fo^2}$$

$$w_2 = \frac{2 \Delta f}{fo}$$

$$w_3 = \frac{2 \Delta f''}{fo}$$

For the Bijvoet mates |F+| and |F-|, the last term in equation 2.11 has opposite sign. In summary, the intensities for any Bijvoet pairs are:

$$|F\pm|^2 = |Ft|^2 + w_1 |Fa|^2 + w_2 |Ft| |Fa| \cos\Delta\varphi \pm w_3 |Ft| |Fa| \sin\Delta\varphi \qquad (2.12)$$

Equation 2.12 can be used for data analysis. For each wavelength used, there are two heavy atom anomalous dispersion parameters, Δf and $\Delta f''$, which can be obtained in experiment or refined against the MAD data. The amplitude of $F\pm$ can be measured in MAD experiment. Thus, there are only three quantities left in equation 2.12 to be determined, i.e. |Ft|, |Fa|, and $\Delta\varphi$. For each (h, k, l), the data should be measured at least at two wavelengths, or even four wavelengths, so that there are four to eight equations to solve three unknowns if both Bijvoet mates are measured. After solving |Ft|, |Fa|, and $\Delta\varphi$ for each (h, k, l) with Δf and $\Delta f''$ refined in the process, |Fa|, the amplitude of heavy atom structure factor, can be used to locate the positions of heavy atoms by traditional methods such as direct methods or Patterson methods. With the coordinates of heavy atoms

determined, the phase angle of vector **Fa** can be calculated and vector **Fa** can be constructed by combining experimental amplitude or calculated amplitude with calculated phase angle. According to the definition of vector **Ft**, its phase angle is the sum of $\Delta\varphi$ and phase angle of **Fa**, which are all known now (Figure 2.6). Thus, the initial electron density maps can be calculated from vectors **Ft**'s.

2.3.4 Molecular replacement (MR)

Molecular replacement is the most commonly used method for solving macromolecular structures since most of the structures solved are those of mutants of native proteins or proteins from other organisms or species, for which only minor structural changes occur compared to the existing structures. The idea of molecular replacement is to use an existing structures as the initial model, and translate and/or rotate it as a rigid body in real space to search for the correct position and orientation for current crystal form. The initial model may be a variant or a homologous molecule of that under study. After the real space search, the model is refined against current diffraction data to obtain modified structures. This method is only applicable when structural information is available to build an initial model accurate enough for the real space search.

2.4 Methods in macromolecular crystallization

One of the major obstacles in X-ray crystallography is the crystallization of macromolecules. Due to the complication of intermolecular interactions among

macromolecules, solvent-macromolecules, solvent-precipitants, and precipitants-macromolecules, etc., it is almost impossible to predict crystallization conditions for a specific macromolecule. In spite of some general theories about crystallization, it is still a trial-and-error process and typically needs to screen a number of conditions before obtaining the first crystal (for a review, see McPherson, 1990; Weber, 1991; Berry, 1995). After that, the crystallization condition is refined by screening in finer grids around the initial conditions to obtain larger and better crystals suitable for X-ray diffraction experiment. Apparently, to grow high quality crystals is the first important step in obtaining high resolution molecular structures in X-ray crystallography. Some basic theories and methods in this area are summarized below.

2.4.1 General description of crystal growth

A crystal is a three dimensional lattice in which molecules are packed according to the crystallographic symmetries and non-crystallographic symmetries, if present. The resolution of molecular structures refined against diffraction data depends on the quality of crystals, i.e. the internal order of crystals. At least two factors contribute to the quality of crystals: the order of protein molecules packed in crystals and the structural uniformity of protein molecules. Protein molecules must be well ordered in crystals in order to have coherent interference for X-rays and diffract X-rays to high resolution. Defects that decrease the degree of internal order of crystals may be caused by incorporating protein contaminants, vacancies, linear dislocations, planar dislocations, etc. On the other hand, the structures of macromolecules are much more flexible than those of small molecules.

The uniformity of macromolecular structures decreases with increase of their molecular weight which reduces the order of atoms in crystals. Thus, protein crystals diffract to much lower resolution compared with small molecule crystals and effort should be made to improve the quality of macromolecular crystals.

Protein crystals normally grow in supersaturated solution in which protein concentration exceeds its solubility. A supersaturated solution is thermodynamically metastable and protein molecules tend to associate with each other and precipitate in order to decrease the concentration of dissolved protein molecules until an equilibrium is reached. The association may be either ordered which results in crystals or crystalline material, or random which results in amorphous material. Solutions can be supersaturated by the so called precipitants which can change the interactions between protein molecules and bulk solvent water, or change the surface charges of protein molecules resulting in new interactions between protein molecules. By changing the properties of protein molecules or solutions, precipitants can promote the growth of crystals or crystalline material. However, since the forces that pack protein molecules in crystals are very weak, the structural differences between molecules in crystal and in solution are negligible.

The process of crystal growth can be divided into three stages: nucleation, crystal growth, and cessation of growth. At the first stage in supersaturated protein solution, nuclei are formed spontaneously from fluctuation at proper conditions. The nucleation rate highly depends on the supersaturation level in solution. If the supersaturation level is too high, numerous nuclei will appear and grow to many small crystals instead of a few large ones due to the depletion of protein molecules in solution. However, if the nucleation rate

is too low, it may take a long time to form a few nuclei for crystal growth. The size of nuclei must exceed a critical point in order to be thermodynamically stable for subsequent crystal growth. Otherwise, nuclei may spontaneously dissolve in solution. The critical size of nuclei is a function of supersaturation level and temperature. The critical size decreases with an increase of the supersaturation level.

At the second stage, protein molecules continue to precipitate onto the nuclei which provide molecular surfaces for crystal growth. Crystal growth rate at this stage depends on several factors such as supersaturation level, concentration of protein, solubility, and diffusion rate of protein molecules, etc. A high growth rate may result in more crystal packing defects which should be avoided. Since the supersaturation level required for nucleation is typically higher than that for crystal growth, the suitable supersaturation level for nucleation may cause crystal packing defects in the early stage of crystal growth. With the decrease of concentration of dissolved protein molecules in solution, supersaturation level and crystal growth rate also decrease.

In the third stage, crystal growth ceases at certain condition due to several reasons such as the poisoning of the crystal growth surfaces by contaminants, the depletion of protein molecules in solution, etc. It has been frequently observed that crystals cease to grow beyond a certain size even when excess protein molecules still present in solution. A possible explanation for this phenomenon is that the surfaces of crystals are poisoned by defects caused by protein contaminants since the concentration of contaminants increases as protein molecules being crystallized are depleted from solution. Another possible reason for growth cessation is the introduction of crystal strain or accumulation of defects in

crystals. The defects on surfaces cause the addition of new protein molecules energetically unfavorable and crystal growth ceases. It has been reported that some protein crystals halved at full size will grow to their original size, suggesting that the defects causing the cessation of crystal growth are accumulated inside crystals (Berry, 1995). Final crystal size may depend on initial supersaturation level, initial protein concentration, temperature, and purity of protein samples, etc.

The purity of protein samples is very important. Protein contaminants can poison the growing surface and cause cessation of the growth of crystals as described above. If a protein contaminant molecule is incorporated inside the crystal, a defect will occur which may propagate during the process of crystal growth. If a number of such defects present in a crystal, the resolution and quality of the diffraction data will be significantly reduced. The homologous protein molecules such as mutants of the crystallized protein may be even more harmful than completely different protein molecules since they may precipitate on the crystal growth surface more easily than other protein contaminants. These homologous protein molecules may cause dislocation or other flaws that poison growth surface. In the present case, such impurity in protein samples resulted in crystals with very irregular and irregular shapes and almost random crystal surfaces. In many cases, if a protein can not be crystallized or crystallization is irreproducible, the purity of samples is probably the reason.

2.4.2 Strategies in protein crystallization

Many factors can influence protein crystallization such as precipitant type and its concentration, ionic strength, pH and buffer, temperature, and additives. Crystallization

occurs only in some small regions in the parameter space. When a new protein is to be crystallized, the total number of conditions in the parameter space is too large to screen every one. Another problem in protein crystallization is that the amount of protein samples available may be limited. One may have to work with small quantity of samples which limits the number of conditions to be screened. Therefore, different strategies have been proposed to find crystal growth conditions in an efficient and economical way in terms of the number of conditions screened and protein samples required.

Shotgun approach This is a random screening method requiring minimal intelligence and experience. Various crystallization parameters are randomly combined to generate a number of conditions for trial. It is assumed that, by trying enough conditions randomly, one of it may yield some crystals. The conditions successfully used before for crystallizing proteins can be included in the list (Jancarik and Kim, 1991). One disadvantage of this method is that it may require large amount of proteins which are not available in many cases. Also, if all the trials fail, there are not many clues about what should be done next except trying more conditions.

Footprint screening This is a method that crystallization conditions are systematically screened (Stura et al., 1991; McPherson 1992). Although the screening is sparse and not complete in order to reduce the total number of trials, the results give information about the solubility of the sample at different conditions. These screening efforts may not yield crystals, but they may give clues about which conditions may be promising for further crystallization trials and eliminates the region in the parameter space

that are obviously useless. After the initial footprint screening, finer screening can be done in the promising region and parameters can be combined for a better results.

Incomplete factorial experiments This method tries to estimate the correlation between polymorphism of precipitates and various parameters (Carter and Carter, 1979). Different parameters are randomly combined but balanced. The polymorphism of precipitates at these conditions are evaluated by subjectively scoring each condition. For example, a score of 1, 7, and 10 can be assigned to conditions at which the precipitates are amorphous material, crystalline material, and crystals. The impact of each parameter can be estimated by fitting the equations:

$$Q_n = \Sigma \ (\ \beta_i \cdot F_i \ ) \tag{2.13}$$

where F_i is the index of the ith parameter which normally has two values, "1" if presence and "0" if absence, and β_i is the correlation coefficient of the ith parameter. Q_n is the score of the nth crystallization trial. A large positive value of β_i suggests that the ith parameter may promote crystallization due to the contribution of this parameter to the score. However, this model does not include the interactions between different parameters. The equation of the model involving the first order interactions is:

$$Q_n = \Sigma \ (\ \beta_i \cdot F_i \ ) + \Sigma \ (\ C_{ij} \cdot F_i \cdot F_j \ ) \tag{2.14}$$

where C_{ij} is the interaction coefficient and the second summation is over all the ith and jth parameters. This model has not been tested yet.

Screening at the pI The isoelectric point, pI, is the pH value at which protein molecules have no net charge. The neutral protein molecules have low solubility and tend to associate with each other. Supersaturation is easy to reach at the pI of a specific protein.

However, this does not necessarily mean that the precipitates are crystals or crystalline material instead of amorphous material. In many cases, protein crystals are found at pH values far away from their pI. In the present case, GFP was crystallized at pH 6.8 while its pI was estimated about 4.7.

Seeding As described above, the supersaturation level required for spontaneous nucleation is relatively higher than that for crystal growth. In some cases, high supersaturation level causes high growth rate at the early stage of crystal growth which reduces the quality of crystals. For some proteins, it may be difficult to control nucleation rate which results in too many small crystals growing in solution. In order to overcome these problems related to nucleation, seeding can be used to skip the stage of nucleation so that conditions can be optimized for crystal growth (Stura and Wilson, 1990). The seeds can be made from crystals, crystalline material, or even crystals of homologous proteins such as proteins from other species. Different seeding methods have been developed for various purposes. Analytical seeding is used for searching crystallization conditions at which spontaneous nucleation can not take place, but crystal growth can proceed if crystal seeds are introduced. In this method, crystallization solutions are seeded with microcrystals by either mixing seeding solution with it or dipping a fiber or hair into seeding solution and streaking it through the crystallization solution. The number of seeds introduced in this method is not important and not controlled. Another application of seeding is to grow a few large crystals with good quality suitable for X-ray diffraction experiment at an optimized condition for crystal growth. For this purpose, limited number of microcrystal seeds or even a single macrocrystal seed can be introduced so that only a

few or one crystal will grow in solution. In such cases, supersaturation level is not high

enough for spontaneous nucleation. Before seeded into crystallization solution, the

macrocrystal seed is normally soaked in washing solution to dissolve its surface and

remove any contaminants which may hamper further crystal growth.

2.4.3 Precipitating agents

Several classes of precipitants and additives have been used successfully for

crystallization and are described briefly below.

Salts Salts are one class of the most commonly used precipitants which decrease

the solubility of protein molecules by competing with them for solvent water molecules in

high ionic strength solutions. The solutions are supersaturated and protein molecules start

to precipitate which is known as "salt out". This effect is related to ionic strength S of

solutions defined as:

$$S = 0.5 \, \Sigma \, (\, C_i \bullet Z_i^2 \,) \tag{2.15}$$

where C_i and Z_i are the concentration and charge of the ith ion. Equation 2.15 shows that

ions with higher charges have larger influence on ionic strength than ions with lower

charges. Some commonly used salts are: $(NH_4)_2SO_4$, $MgSO_4$, Na_2SO_4, Li_2SO_4, $LiCl$,

KH_2PO_4, $CaCl_2$, etc. Different salts may have completely different interactions with

protein molecules and it is frequently necessary to screen a broad range of salts. In some

cases, salt ions may stabilize the lattice structure by electrostatic interactions with two

protein molecules. One disadvantage of using salts for crystallization is that they may

interact strongly with some heavy atom compounds which causes problems in preparing

heavy atom derivatives. Another disadvantage is that salts increase the mean electron densities of crystals which increase the background scattering of X-rays and decrease the signal-to-noise ratio of diffraction data.

Polymers The most successfully used polymers is polyethylene glycol (PEG). The molecular weight of PEG may range from 200 Da to over 1,000,000 Da, containing from several to thousands of monomers. PEG can compete with protein molecules for solvent water molecules and supersaturate protein solutions. Successful crystallization conditions show that low ionic strength is preferred for using PEG. Several advantages of using PEG as precipitant can be listed. First, PEG does not increase background scattering of X-rays due to its similar electron densities as water molecules compared with salts. Second, it does not react with heavy atom compounds which makes it easy for preparing heavy atom derivatives for MIR. Third, PEG at low molecular weight is a good cryo-protectant when crystals need to be frozen. Finally, the range of PEG concentration resulting in successful crystal growth is fairly narrow, from about 4% to 14% which reduces the conditions to be screened. Due to all these advantages, PEG is a class of excellent precipitants for crystallization and is frequently used.

Non-volatile organic compounds 2-Methyl-2,4-pentanediol (MPD) is a typical compound in this class. MPD may share many useful properties of polymers and organic solvents in crystallization. It is a good cryo-protectant, does not react with heavy atom compounds, and has lower electron densities compared with salts. A high concentration of MPD (> 40% v/v) is typically required. In the present case, the wild-type *Aequorea victoria* GFP was crystallized at 58% MPD with a variation of about ±3%.

Organic solvents   Some hydrophobic organic solvents can be used if they do not denature protein molecules and the ionic strength of solution is as low as possible. Commonly used organic solvent as precipitants are ethanol, methanol, acetone, isopropanol, etc. Organic solvents have similar influence on crystallization as other organic precipitants. To prevent protein denaturation, a temperature at or below 0 °C may be necessary and organic solvents should be added very slowly into protein samples. Compared with salts, organic solvents are similar to MPD and PEG that can reduce the dielectric constant of solutions so that protein molecules are less screened for electrostatic interactions, resulting in relatively stronger electrostatic interactions among protein molecules.

Additives   In some cases, small amount of additives are critical for crystal growth. All precipitants discussed above can be used as additives with other precipitants. Some metal ions such as cadmium, cobalt, and manganese can also be used as additives to promote or improve crystallization.

2.4.4 Crystal growth methods

Several crystallization methods have been developed as described below. The selection of the methods depends on the specific problem that we are facing at, such as the amount of protein samples available, the purpose of crystal growth setup, i.e. whether it is for screening new crystal growth conditions or for producing large crystals for data collection, etc.

Hanging-drop vapor diffusion and sitting-drop vapor diffusion   These two crystallization methods are the most commonly used. In the hanging-drop vapor diffusion method, a well in a Linbro plate contains about 1 ml of mother liquor as the reservoir solution. A drop of protein sample is mixed with the mother liquor in a siliconized cover slide so that the concentration of precipitant in the drop is lower than that of mother liquor and the solution is not saturated. The well is sealed by the slide using silicone vacuum grease or immersion oil with the drop hanging over the mother liquor (Figure 2.7a). Water molecules in the sample drop will diffuse into the reservoir through evaporation due to the high concentration of precipitant in reservoir until equilibrium is reached. In this process, the hanging drop is supersaturated and protein molecules start to associate and precipitate. Typically, an 1:1 ratio of mixing of the sample solution and mother liquor is used. If it is necessary to dilute the drop and slow down the equilibrium process in order to improve the quality of crystals, de-ionized water can be added to the sample drop. The distance between the hanging drop and the surface of the reservoir solution can also influence the equilibrium rate. The hanging-drop vapor diffusion method is suitable for analytical screening for crystal growth conditions since it requires only a very small volume of protein sample (~2 µl) in each drop. Also, it is easy to set up and a number of screening conditions can be tested in a short time. However, it may not be a good method for growing a number of large crystals since the maximal volume of drops is limited. A drop may fall off the cover slide if its volume is too large.

The sitting-drop vapor diffusion method is similar to the hanging-drop vapor diffusion method. Instead of hanging sample drops over the mother liquor, they are

(a)   **(b)**
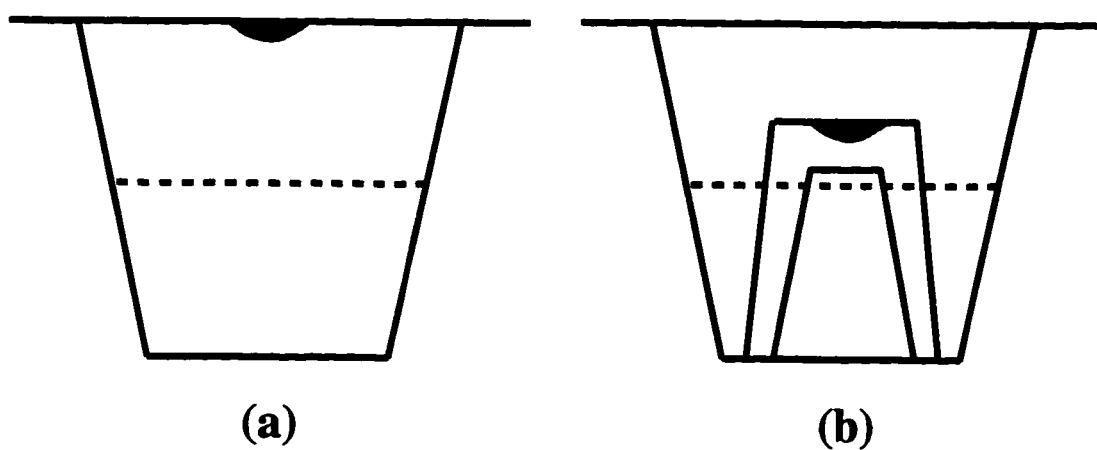
Figure 2.7 Vapor diffusion crystal growth setup. a) Hanging drop vapor diffusion method. Protein sample is hanging on the cover slide over the well solution. b) Sitting drop vapor diffusion method. Protein sample is contained in the dent of a plastic bridge which stands in the well solution.

contained in small dents of plastic bridges standing inside the reservoir (Figure 2.7b). Due to this design, the volume of drop in each setup can be much larger than that in hanging drop setup, and it is easy to pick a crystal from a dent. This method is suitable for producing a number of large crystals. In both vapor diffusion methods, the pH values of drop and the mother liquor are same if volatile precipitants such as ammonium sulfate and some organic solvents are used.

Batch method In this method, protein sample solution is mixed with precipitant solution directly and supersaturation is accomplished immediately. The protein molecules start to aggregate after mixing and in some cases, crystals appear within 20 hours. The batch method is suitable for mass production of crystals if the growth condition is known. Since the batch method typically requires large amount of proteins, it is not efficient and economical, and not suitable for screening crystallization conditions for a new protein.

Free interface diffusion As described above, the supersaturation level required for nucleation is not optimal for crystal growth. The free interface diffusion method is design to solve this problem by obtaining high supersaturation level at the nucleation stage and decreasing the degree of supersaturation level after it. Thus, spontaneous nucleation can still occur and crystals can grow at a proper rate with good quality. In this method, a protein solution is layered over a precipitant solution (Figure 2.8a). The protein molecules at the liquid-liquid interface experience high degree of supersaturation and ideally, will start to nucleate. As the molecules diffuse between the two solutions, the degree of supersaturation decreases which favors crystal growth. In some applications, the bottom solution is frozen first before applying the top solution above it to ensure a sharp

(a) **(b)**
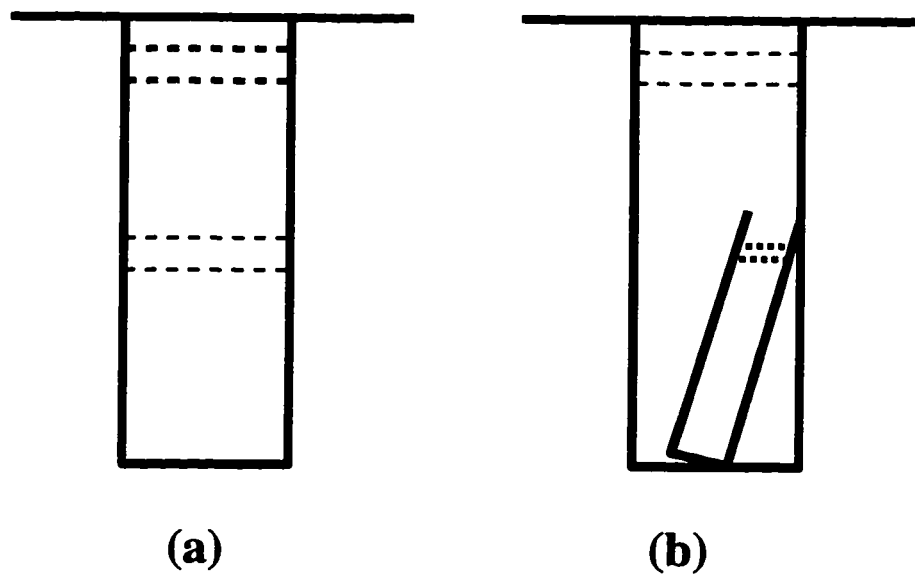
Figure 2.8 Free interface diffusion setup. a) The protein sample is layered over the precipitant solution. b) The protein sample is contained in the small tube which is soaked in precipitant solution. In both setups, the protein sample and precipitant are free to diffuse into each other.

interface. Equilibrium will be reached when protein solution and precipitant solution are mixed completely. In some cases, the protein solution is contained in small capillary tubes and submerged in precipitant solution (Figure 2.8b).

Dialysis This method is similar to the free interface diffusion method. A protein solution is sealed in a vessel such as a dialysis button by dialysis membrane which allows the passage of water, salts, and precipitant molecules, but not macromolecules like protein molecules. In some cases, the membrane also serves as heterogeneous nucleation site. Buttons are submerged in mother liquor to reach equilibrium gradually through diffusion of solvent, salt and precipitant molecules. The protein sample is supersaturated at equilibrium. One advantage of this method is that the protein sample can be used repeatedly if crystallization fails at one condition by transferring the button to another crystallization solution, assuming that the protein molecules are not denatured and the precipitation is reversible. However, it has been suggested that the path, through which the protein sample is brought to supersaturation and returns to equilibrium by precipitation, is critical for crystallization. Thus, it may be necessary to dialyze the sample solution against washing solution before dialyzing it against the next crystallization solution.

2.5 Methods in cryocrystallography

X-ray radiation can generate free radicals which can diffuse in crystals and react with protein molecules which is known as radiation damage. Radiation damage can significantly reduce resolution and intensity of diffraction spots by reducing the internal

order of crystals. Therefore, MAD data were usually collected at a cryo-temperature in order to reduce radiation damage caused by intense synchrotron beam. By freezing crystals, diffusion of the free radicals is significantly localized which reduces the extent of reactions in crystals. The duration of diffraction of crystals irradiated by X-rays is extended greatly. Several data sets can be collected on one crystal which is extremely important for solving phase problem since the problem of non-isomorphism is avoided. Another advantage of freezing crystals at cryo-temperatures is that the thermal movement of protein molecules is reduced which maintains the internal order of crystals and stabilizes protein conformations.

Since water molecules can form ice and expand its volume during freezing, special treatment is necessary to reduce this effect and protect integrity of crystals. This is achieved by flash freezing crystals and using a cryo-protectant. By freezing crystals as quickly as possible, water molecules have little time to rearrange and expand volume. Cryo-protectant can reduce the formation of ice and the expansion pressure from solvent molecules surrounding crystals. Some commonly used cryo-protectants are: glycerol, ethylene glycol, MPD, low molecular weight PEG, propylene glycol, sucrose, glucose, methanol, isopropanol, etc. It is a great advantage to grow crystals using precipitants which are also cryo-protectant such as MPD, low molecular weight PEG, methanol, etc. Otherwise, suitable cryo-protectant and its concentration need to be screened so that the soaking in cryo-protectant solution does not increase the mosaicity of crystals or decrease the resolution and quality of diffraction data.. The concentration of cryo-protectants is typically from 20% to 25% and the soaking time can be from a few seconds to several

days. The concentration of cryo-protectants should be increased gradually in order to minimize its impact on crystals.

Flash freezing can be achieved by two methods. One method is to first pick a crystal in a loop glued on a crystal mounting device. The cold stream of nitrogen gas is blocked at the nozzle of the cryo-system and the crystal mounting device is quickly put on the magnetic base of a goniometer head. The block is then removed immediately to let the nitrogen stream freeze the crystal. This method is a little tricky and requires some skill. It needs to be practiced several times before mounting a good crystal for experiment. Otherwise, the crystal may be poorly frozen and damaged. The other method is more reliable but a little complicated. The cryo-loop is glued on a metal string such as copper wire which can be soldered on a 1/8" inch brass pin to have good heat conductivity. The brass pin has two notches to be clamped easily by tweezers. A brass block is cut to half and a 1/8" inch hole is drilled on the interface. The two pieces of brass blocks are soldered at the tip of a metal tong which can be used to hold the brass pin. The hole should be deep enough to hold part of the pin with the cryo-loop covered inside. A stopper made by a metal ring is soldered on the brass pin so that when the brass pin is put on the goniometer head, the cryo-loop is centered inside the cold stream. For flash freezing a crystal, a tweezers is used to hold a brass pin and pick a crystal in the cryo-loop at the top of the pin. The brass pin is dipped into a small volume of liquid nitrogen as fast as possible. The tip of a metal tong is submerged in liquid nitrogen until it is cooled down to liquid nitrogen temperature. The brass pin is clamped carefully inside the hole of the brass blocks of the tong and put on the goniometer head. The tong is removed quickly to let the cold stream

blow on the crystal. The brass pin should be able to fit into the hole tightly so that it does not fall off during transfer, but leave no gap between the two half brass pieces so that room temperature air can not be leaked into the hole and warm up the crystal. Finally, in order to uniformly freezing a crystal, its size should not be too large, and crystals with size from 0.1 mm to 0.5 mm are typically used. Crystals not uniformly frozen typically have large mosaicity which reduces the quality of diffraction data.

The material for cryo-loop is important and several conditions should be satisfied. First, loop materials should not have strong absorption or diffraction to X-rays. A shadow or increased diffraction background will reduce the quality of data. Second, the fiber for loops should be as thin as possible. A thick loop may hold more mother liquor with the crystal which may slow down the freezing process and damage the crystal. Third, the materials for loops should be strong enough to tolerate some bump when picking a crystal from solution and hold the crystal steady in cold stream. Finally, the shape of the loop and its stem should be smooth. A bulky shape may cause turbulence and moisture in air may form ice on the crystal. Several types of materials have been used for loops such as nylon, rayon from cosmetic puff, glass fiber from glass wool. There are several methods to make a cryo-loop depending on the size and material used. Here, the method for making glass fiber loop is described. Glass fiber is a good material in consideration of low level of absorption and diffraction of X-rays. However, since glass fiber is very fragile, it can be broken easily when the loop size is smaller than 0.2 mm or when it is bumped onto something hard. An easy method has been designed to make glass fiber loops. First, some micro glass tubing were made by melting glass tubes in a flame and pulling both ends

of it. One micro glass tubing is cut to about 2 to 3 cm and both ends of a glass fiber are inserted from one end of the glass tubing. The two ends of the glass fiber are pulled from the other end of the glass tubing until the size of the glass fiber loop is proper. A small volume of extra fast setting epoxy is applied at the stem of the loop to fix the loop. The glass tubing is cut to about 1 cm and glued on a brass pin or other crystal mounting device.

2.6 Methods for the preparation of heavy atom derivatives

There are several methods of preparing heavy atom derivatives for isomorphous replacement. Three of them have been tried for GFP crystals. The first method is to soak crystals in buffer solutions containing heavy atom compounds. It is expected that, at suitable conditions such as proper concentration of the heavy atom compound, pH, temperature, and soaking time, etc., the heavy atom compounds may bind protein molecules at only a few specific sites with relatively high occupancies. In order to be isomorphous to the native crystals, the binding of the heavy atom compounds or the changes in the ionic strength of the buffer should not disturb the crystal packing of protein molecules and the protein structure itself. The second method is to use high pressure noble gases such as xenon or krypton. Under the pressure of about 100 to over 200 pound per square inch (psi), the noble gas molecules may be squeezed into some cavities in the protein molecules. Xenon has been used successfully as the heavy atom for several proteins such as myoglobin (Vitali et al., 1991). The third method is to replace the native methionine residues, occurring naturally or mutated into protein sequence, by

selenomethionine residues in recombinant proteins. By expressing recombinant proteins in

an *E. coli* strain unable to produce native methionine residues, nearly 100% replacement by

selenomethionine residues can be achieved in this method. However, since the selenium

atom is relatively light containing only 34 electrons, selenomethionyl protein crystals are

used only for multiwavelength anomalous dispersion methods (MAD).

2.7 Methods in data collection

The X-ray diffraction data of macromolecular crystals are typically collected on an

area detector such as multiwire area detector, image plate area detector, or charge coupled

device (CCD) area detector. Hundreds of diffractions can be collected simultaneously on

an area detector. In traditional diffraction data collection, oscillation frames are collected

continuously over a range from $50°$ to $150°$ depending on the symmetry of the crystal. In

this way, true Friedel pairs, i.e. the original (h, k, l) and its centrosymmetrically related

partner (-h, -k, -l), may not be measured together, or measured with relatively long time

interval, which may cause significant error in anomalous dispersion signals due to factors

such as radiation damage. Since anomalous dispersion signals may be very weak, special

data collection strategies are required for collecting MAD data. In order to measure true

Friedel pairs immediately, the crystal can be rotated $180°$ relative to the previous orientation

which is known as inverse-beam method. With this inverse-beam geometry, the true

Friedel mate of a diffraction spot that appears on the upper-left part of a frame, will appear

on the lower-left part of the inverse-beam frame. The diffraction patterns of the two frames

are related by a pseudo horizontal 2-fold axis across the direct-beam point. In data process according to equation 2.12, only the intensities of true Friedel mates are compared and experimental errors can be reduced.

Data collection parameters must also be selected carefully. Important parameters include the distance from crystal to detector, oscillation angle and exposure time of each frame, the number of frames required, and the 2θ angle of detector if it can be adjusted. Balance must be maintained among these parameters. A small distance from crystal to detector is suitable for collecting high resolution data and results in relatively intense diffractions since diffraction intensity is inverse-proportional to the square of the distance. But the physical separations among spots on detector are also small and the overlapped spots will reduce the quality of the diffraction data. Similarly, a long exposure time of each frame can increase the signal-to-noise ratio, but it also increases radiation damage if the crystal is not frozen and increases the time of data collection. The selection of oscillation angle of each frame is related to the type of detector used, the feature of the data processing software, the total angle to be scanned, and other limitations on the experiment. A small oscillation angle will results in many partial spots collected on each frame so that intensities on several frames need to be summed up to obtain the intensity of a spot. If the detector has relative large background noise, this may increase the noise level in collected data. On the other hand, if the oscillation angle is too large, the noise from the defects of crystal may be increased which also reduces the quality of data. In short, data collection parameters depend on the current experimental conditions and limitations of experimental time,

capability of instruments, etc., and should be selected carefully to obtain data with possibly the highest quality.

Chapter 3

Materials and Methods

3.1 Overview of experimental procedures

The determination of the structure of GFP required expression of GFP,

purification, crystallization, and diffraction analysis. The experimental procedure started

with protein expression in *Escherichia coli*. *E. coli* strain BL21(DE3)pLysS and the

methionine auxotroph strain B834(DE3)pLysS were used for expression of wild-type and

selenomethionyl-substituted GFP, respectively. Competent cells were prepared and

transformed with plasmid pTu#58 bearing the wild-type *Aequorea victoria* GFP gene,

selected by ampicillin and chloramphenicol resistance. BL21(DE3)pLysS cells were grown

in LB medium and B834(DE3)pLysS cells were grown in M9 minimal medium supplied

with selenomethionine acid, the other nineteen amino acids, and other components. After

the expression of GFP was induced by isopropyl β-D-thiogalactopyranoside (IPTG), cells

were harvested and lysated by ultrasonication. The crude cell extract was purified by one

DEAE anion exchange column, one hydrophobic interaction column, and two cycles of

HPLC anion exchange column and HPLC gel filtration column. The purity of GFP

samples was monitored by the ratio of absorbance A_{397}/A_{280} which was typically larger than 1.0. After purification, various crystal growth conditions were tried and the best GFP crystal growing conditions were found. Traditional heavy atom derivatives for multiple isomorphous replacement (MIR) methods were prepared by soaking GFP crystals in mother liquor containing heavy atom compounds. After a number of trials, MIR methods failed due to non-isomorphism between native and derivative crystals. As a substitution, multiwavelength anomalous dispersion (MAD) methods were successfully applied by collecting data on selenomethionyl GFP crystals at synchrotron source. Diffraction data were collected with inverse-beam geometry at four wavelengths at Brookhaven National Laboratories, beamline X4A. Data were processed using the program DENZO to obtain amplitude of structure factors and program MADSYS to obtain phase angles. The initial structural model was built based on the MAD phased experimental electron density maps using the program O. The final structure was refined against a native data set collected on an R-Axis IIC area detector at room temperature using program X-PLOR.

3.2 Protein expression and purification

3.2.1 Preparation of competent cells of B834(DE3)pLysS

E. coli strain BL21(DE3)pLysS competent cells were used for expressing wild-type GFP and methionine-auxotroph B834(DE3)pLysS cells (Novagen, Madison, WI) were used for expressing selenomethionyl GFP. The BL21(DE3)pLysS competent cells were dispensed into 20 µl in each Eppendorf tube using sterilized pipet tips. The tubes were

frozen in liquid nitrogen first and stored at -70 °C. Since the B834(DE3)pLysS cells were supplied as a glycerol stock and stored at -70 °C, the B834(DE3)pLysS competent cells were made for plasmid transformation from glycerol stock as described below.

A piece of ice was scratched from the glycerol stock and added into 5 ml of LB solution. A volume of one liter of LB solution contains 10 g of tryptone, 5 g of yeast extract, and 10 g of NaCl. The pH was adjusted to 7.5 with 1 N NaOH and the solution was sterilized using high pressure steam. The solution was incubated overnight at 37 °C. A volume of 1 ml of this cell solution was added to 50 ml of LB solution which was incubated at 37 °C until the absorbance at 600 nm is about 0.6 in a cuvette of 1 cm path length. Cells were spun down at 6000 rpm for 10 minutes at 4 °C. The pellet was resuspended in 20 ml of autoclaved 50 mM CaCl2 solution. After standing for 30 minutes in an ice-bath, the cells were pelleted again at 6000 rpm for 10 minutes at 4 °C. The pellet was resuspended in 2 ml of autoclaved solution containing 50 mM CaCl2 and 15% glycerol, which was dispensed into 0.1 ml in each Eppendorf tube. The tubes were frozen in liquid nitrogen and stored at -70 °C.

3.2.2 Transformation of competent cells with pTu#58 plasmid

Plasmid pTu#58 was supplied by the colaborator Dr. Larry G. Moss at Tufts University School of Medicine and the New England Medical Center. A volume of 1 µl of 0.3 mg/ml plasmid solution was added into 20 µl of BL21(DE3)pLysS competent cells solution or 100 µl of B834(DE3)pLysS competent cells solution. After 30 minutes of ice-bath, Eppendorf tubes were heated to 42 °C in water-bath for one minute. A volume of 0.6

ml LB solution was added into each tube and heated at 37 °C in water-bath for 45 minutes. A volume of 0.2 ml of this transformed cell solution was evenly plated on three LB plates which were incubated at 37 °C overnight and the plates were stored in freezer. A volume of one liter of the solution for the LB plate contains 10 g of tryptone, 5 g of yeast extract, 10 g of NaCl, 15 g of agar, 1 ml of 50 mg/ml ampicillin, and 1 ml of 25 mg/ml chloramphenicol. The pH was adjusted to 7.5 with 1 N NaOH. The ampicillin and chloramphenicol were added into the solution after the solution was sterilized using high pressure steam.

3.2.3 Expression of wild-type GFP in BL21(DE3)pLysS cells

The growth medium contains the same component as the solution of LB plate except agar and prepared in the same method. A volume of 2 ml of growth medium was inoculated with a single clone of the transformed BL21(DE3)pLysS cells. The solution was shaken in water-bath at 33 °C for about 10 hours. A volume of 1 ml of this cell solution was added to 60 ml of the growth media which was shaken in water-bath at 33 °C for another 10 hours. Finally, a volume of 5 ml of this cell solution was added to 500 ml of growth media in a 2-liter buffled flask. A total of twelve flasks were shaken at 33 °C and the absorbance was monitored at 600 nm. After the absorbance reached 0.6 in a quartz cuvette with 1 cm path length, 1 ml of 200 mM IPTG solution was added to each flask. After an induction of about 3 to 4 hours, cells were harvested and stored at -70 °C before protein purification.

3.2.4 Expression of selenomethionyl GFP in B834(DE3)pLysS cells

The transformed B834(DE3)pLysS cells were grown in M9 minimum medium. One liter of the growth medium contains 10 g M9 medium salt, 1% (w/w) glucose, 2 mM $MgSO_4$, 0.1 mM $CaCl_2$, 0.05 mM $FeSO_4$, 2 mg thiamine, 2 mg D-biotin, 0.05 g amino acids except methionine, tryptophan, and tyrosine, 0.05 g tryptophan and tyrosine in 1 ml of 1 N HCl solution, 0.033 g selenomethionine, 50 mg ampicillin, 25 mg chloramphenicol.

A volume of 5 ml of the growth medium was inoculated with a single clone of the transformed cells and shaken in water-bath at 33 °C for about 15 hours. A volume of 1 ml of this cell solution was added to 60 ml of the growth medium which was shaken at 33 °C for about 15 hours. A volume of 5 ml of the cell solution was added to each of the twelve 2 liter flasks containing 500 ml of growth medium and shaken at 33 °C for about 10 hours until the absorbance at 600 nm reached 0.6 in a quartz cuvette with 1 cm path length. Then, cells were induced with 0.8 mM IPTG and harvested 10 hours after induction. A total of 17 g cells were harvested from 6 liter of growth medium and were stored at -70 °C before protein purification. For a control, it was found that the transformed B834(DE3)pLysS cells could not grow without selenomethionine acid, suggesting that selenomethionine acid was essential for B834(DE3)pLysS cells and had been used for protein expression.

3.2.5 Protein purification

The buffers used for protein purification are listed below.

A) 10 mM NaH_2PO_4, 2 mM EDTA, pH 7.0

B) 0.22 M NaCl, 10 mM NaH2PO4, 2 mM EDTA, pH 7.0

C) 0.1 M NaH2PO4, 2 mM EDTA, pH 7.0

D) 0.2 mM EDTA, pH 7.0

E) 1 M NaCl, 10 mM NaH2PO4, 2 mM EDTA, pH 7.0

For purification of selenomethionyl GFP, all buffers were degassed with nitrogen gas to prevent oxidation of selenomethionine, and contained 2 mM DTT except buffer B which contained 0.2 mM DTT. All columns were run at room temperature.

Cells were thawed and resuspended in buffer A followed by ultrasonication to lyse cells. A concentration of 1 mM phenylmethylsulfonyl fluoride was used to inhibit proteases like trypsin and chymotrypsin. Cells stood in ice-bath and were lysed for 20 minutes using the pulse mode of the 450 Sonifier (Branson Ultrasonics Corporation). DNA was digested by 40 μl of 20 mg/ml DNase activated by 40 mM MgSO4. The sample solution was shaken at 4 °C for about one hour to complete the digestion. The cell debris was spun down at 12,000 rpm for 50 minutes and the supernatant was loaded in the anion exchange column (Sigma, St. Louis, MO, CL-6B) pre-equilibrated with buffer A. The column was eluted with buffer A for about five hours to wash away unbound molecules before eluting with buffer B to wash GFP off the column. The sample fractions were loaded directly in the hydrophobic interaction column (Sigma, CL-4B) which was pre-equilibrated with buffer C. After eluting the column with buffer C for about five hours, buffer D was used to elute proteins off the column. Then, samples were purified by two HPLC columns, an anion exchange column (Bio-Red, Hercules, CA, DEAE-5PW) and a gel filtration column (Bio-Red, Bio-Sil Sec 125). The sample fractions from the

hydrophobic interaction column could be loaded directly in the HPLC anion exchange column pre-equilibrated with buffer A. Samples were eluted off the column at a flow rate of 1 ml/min and a linear gradient of NaCl concentration at a rate of 0.02 M/min by mixing buffers A and E. The sample fractions could be loaded directly in the gel filtration column which was pre-equilibrated with buffer C. The flow rate of the gel filtration column is 1 ml/min. The sample fractions from gel filtration column was transferred to buffer A by several cycles of concentration and dilution with buffer A, and could be loaded in the anion exchange column again if further purification was necessary. The purity of GFP samples was monitored by the ratio of absorbance at 397 nm and 278 nm. A ratio larger than 1.0 was normally obtained and acceptable. Also, the chromatography of both the anion exchange column and the gel filtration column showed no detectable contaminants in sample solutions. Protein samples were concentrated to about 20 mg/ml and stored in 50 mM morpholino ethane sulfonic acid (MES), 0.1% sodium azide, pH 6.8.

3.3 Crystallization of GFP and Preparation of heavy atom derivatives

The footprint screening method was used to screen pH value, precipitant type and concentration for crystallization of GFP. Crystal screen kits I and II (Hampton Research) were also used for initial trial. The hanging-drop vapor diffusion method was used for all screening. A volume of 2 μl of 20 mg/ml GFP sample solution was mixed with 2 μl of mother liquor and 2 μl of de-ionized water. For initial screening, precipitates PEG 400,

PEG 2000, PEG 4000, PEG 10000, (NH4)2SO4, phosphate, MPD, and citrate were used at various concentrations, and the pH was screened at 5.0, 7.0, and 8.5. The first GFP crystal was found in 60% MPD, 50 mM MES, 0.1% NaN3, pH 7.0. This initial condition was optimized by finer screening to 58% MPD, 50 mM MES, 0.1% NaN3, pH 6.8 at which GFP crystals grew larger and more rapid. In order to grow GFP crystals large enough for X-ray experiments, the sitting-drop vapor diffusion method was used. A volume of 10 µl of 20 mg/ml GFP sample solution was mixed with 10 µl of mother liquor in the dent of a plastic bridge. Usually it took four days for crystals to appear and over two weeks to reach full size which was large enough for data collection. The selenomethionyl GFP crystals were grown at the same condition and same growth rate as wild-type GFP.

Various heavy atom compounds were tried for MIR methods at different conditions such as concentration of the heavy atom compound, soaking time of crystals, concentration and type of buffer used, etc. Fresh heavy atom solutions were always prepared before crystal soaking. Crystals with size of 0.3 mm to 0.8 mm were selected and transferred directly from the mother liquor to heavy atom solutions. GFP crystals showed high sensitivity to ionic strength and buffer concentration that caused changes in mosaicity and unit cell parameters after soaking in heavy atom solutions. In order to solve this non-isomorphous problem, crystals were crosslinked by 0.0156% glutaraldehyde in 58% MPD, 50 mM MES, pH 6.8 solution for 12 hours before they were transferred to heavy atom solutions. Crystals must be crosslinked at exactly the same degree in order to be isomorphous. Also, the concentration of glutaraldehyde and crosslinking time were tested in order to avoid over-crosslinking crystals which resulted in high crystal mosaicity.

Native and derivative data sets collect on crosslinked crystals were compared with each other.

3.4 Data collection and processing

3.4.1 Data collection

Preparation of cryo-temperature experiments   Both flash freezing methods were tried as described before (section 2.5).   GFP crystals were found to be sensitive to temperature changes and they could be destroyed completely when cooled from room temperature to 4 °C. The method of using a cryo-nitrogen stream to flash freezing crystals was found to be unreliable due to the sensitivity of GFP crystals to temperature. The method of submerging crystals in liquid nitrogen was used and accessory tools were made as described in section 2.5. It was found that crystals could be more easily lifted and picked in the loop if the loop was oriented with its side facing the experimenter under a microscope. Since MPD is a good cryo-protectant, no attempt was made to freeze crystals at other conditions.

Data collection on an R-Axis IIC area detector   Native and heavy atom derivative data sets were collected on a Rigaku R-Axis IIC image plate area detector at room temperature in our laboratory. GFP crystals were sealed in 1.0 mm quartz capillary tubes. The oscillation angle was either $0.9°$ or $1.0°$ and the distance from crystal to detector was 70 or 80 mm. The number of final data frames for a specific crystal was determined by the quality of crystals and the partial data processed during data collection. Data collection

would be stopped if the data showed no binding of heavy atoms or the intensity differences compared with the native data was too high.

MAD data collection at Stanford Synchrotron Radiation Laboratories Single crystal multiwavelength anomalous dispersion data were collected at a cryo-temperature at beamline 1-5 at Stanford Synchrotron Radiation Laboratories. A selenomethionyl GFP crystal was mounted on a glass fiber loop and flash frozen by dipping it into liquid nitrogen. Data were collected at 12450 ev (0.99582 Å), 12652 ev (0.97992 Å), 12658.5 ev (0.97942 Å), and 13000 ev (0.95369 Å) on a CCD area detector. The oscillation angle was either $0.5°$ and the distance from crystal to detector was 160 mm. At each wavelength, 200 continuous frames were collected, i.e. inverse-beam geometry was not used and the wavelength was not changed until $100°$ was scanned.

MAD data collection at Brookhaven National Laboratories Single crystal multiwavelength anomalous dispersion data were collected at a cryo-temperature at Brookhaven National Laboratories beamline X4A. Four wavelengths were selected at 12550 ev (0.9879 Å), 12659 ev (0.9794 Å), 12662 ev (0.9792 Å), and 12800 ev (0.9686 Å) (Figure 3.1). The first wavelength λ_1 was on low energy side far away from the absorption edge of selenium atom so that it had weak absorption of X-rays. The second wavelength λ_2 was at the inflection point of the absorption edge at which selenium atom had largest change in real part of its scattering factor $\Delta f'$. The third wavelength λ_3 was at the peak of the absorption edge at which selenium atom had largest change in imaginary part of its scattering factor $\Delta f''$. The fourth wavelength λ_4 was on high energy side far away from the absorption edge at which selenium atom has relatively weak absorption to
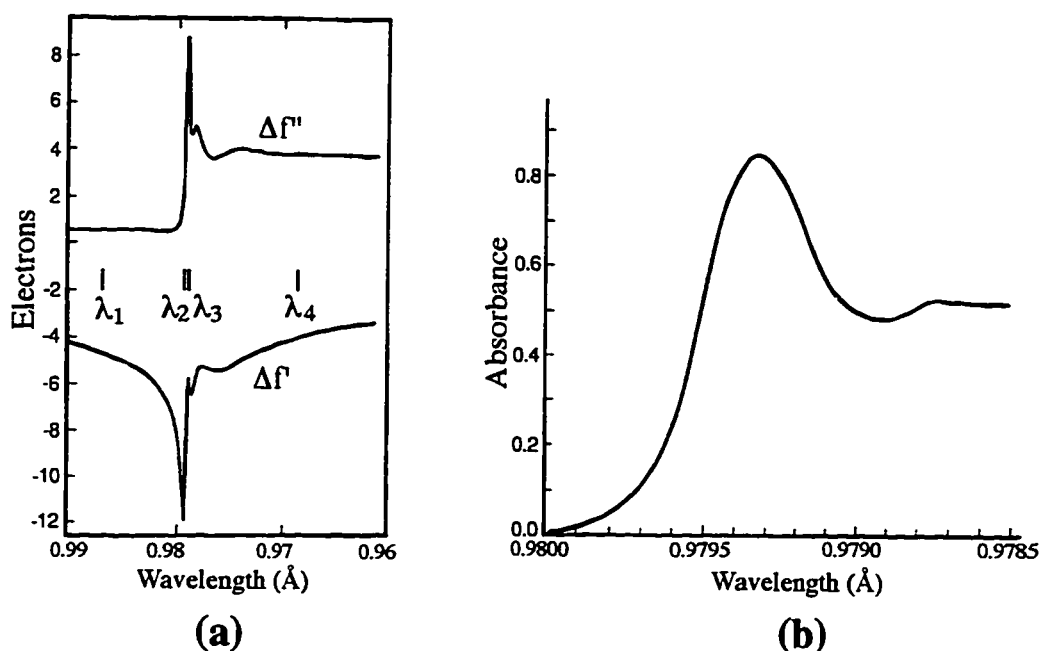
Figure 3.1 Anomalous dispersion of selenium atom. a) The imaginary and real parts of anomalous dispersion of selenium atom in the unit of electrons are shown on the upper and lower curves. Four wavelengths traditionally selected for multiwavelength anomalous dispersion experiment are marked in the figure. Note that λ_2 is at the peak of $\Delta f'$ curve and λ_3 is at the peak of $\Delta f''$ curve. b) The X-ray absorption spectrum of the selenomethionyl GFP crystal used for the MAD data that solved the structure of GFP. The Y axis is at arbitrary scale.

X-rays. A selenomethionyl GFP crystal was mounted on a nylon loop and flash frozen by dipping it into liquid nitrogen. The monochromator was calibrated to the absorption peak of the selenium reference at 12659 ev. A crystal scan showed that the inflection and peak points were at 12659 and 12662 ev. Data were collected on Fuji image plates and scanned manually on an off-line scanner. Since the instrument can load four plates each time, four frames covering a wedge of 4° with 1° oscillation for each frame were collected at the first wavelength. The crystal was rotated 180° relative to the starting position of the wedge and collected a 4° wedge opposite to the previous one. Then, the wavelength was changed and the above processes were repeated. After frames at all four wavelengths were collected, the crystal was rotated 4° forward relative to the starting angle of current wedge and collected data for the next wedge. A total of about 84° were scanned on one side and about 700 frames were collected. Although frozen at a cryo-temperature, the crystal still showed a black core and green edge at the end of data collection, indicating that radiation damage on the crystal was obvious but limited to the region irradiated. Despite the black core in the crystal, the data at the end of experiment were still good enough as shown by the diffraction images and the results of data processing.

3.4.2 Data processing

Several programs were used for data processing. The data collected on an R-Axis detector were processed using DENZO and SCALEPACK (Otwinowski Z., 1993). MAD data were processed using DENZO, SCALEPACK, and MADSYS (Yang et al., 1990). HEAVY and SHELX96 were used to locate heavy atom sites for MIR data and selenium

atom sites for selenomethionyl GFP data collected on an R-Axis detector (Terwilliger et al., 1987; Sheldrick et al., 1993). CCP4 software package was used to scale data sets, refining heavy atom sites, and solvent flattening and non-crystallographic symmetry averaging (CCP4, 1994). Program MAMA was used to generate masks for non-crystallographic symmetry averaging (Kleywegt and Jones, 1994).

Program DENZO was typically run twice for each data set. The results of the first DENZO job were processed using SCALEPACK which gave the estimated mosaicity and useful resolution range of the data set. These parameters were used in the second DENZO job. The input parameter "error positional" for DENZO was modified so that the indices χ^2 of x and y coordinators of diffraction spots were close to 1. The integration box size for diffraction spots should also be inspected and chosen carefully based on the output file of DENZO. Program SCALEPACK was typically run repeatedly as suggested by "The HKL Manual". The individual error parameters of each resolution shell "estimated error" were decreased together with a global error parameter "error scale factor" so that they had rational small values and the χ^2 of intensities in each resolution shell were close to 1. The input parameter "rejection probability" was increased to reject more diffractions if the quality of the data was low. The rejected diffractions were written to a file which was read into in the next SCALEPACK job to exclude these diffractions from calculation.

After being processed by DENZO and SCALEPACK, the MAD data were continued to be processed using MADSYS which gave the values of Ft, Fa and Δφ for each diffraction. The positions of selenium atoms were determined by calculating Patterson maps using Fa which were the structure factors of selenium atoms as normal scattering

centers. The first set of trial sites were given by running subroutine HASSP in HEAVY. Additional trial sites could be found by inputting sites previously found and searching for cross peaks in Patterson maps and finally, five selenium atom sites were found using HASSP and refined using HEAVY. Similarly, SHELXS-96 also found six selenium atoms which were equivalent to the results from HASSP with one extra atom. The heavy atom Fourier map calculated using the amplitude of **Fa** and the partial phase angles calculated for the known heavy atom sites, would reveal new atom site in addition to the known sites. One additional heavy atom was added to the known atom list in each cycle and refined by HEAVY. The heavy atom phases were updated and a new Fourier map was calculated. This process was repeated until no more outstanding site was found or the parameters of any new atom added were refined to low occupancies and high temperature factors, indicating that it should be deleted from the heavy atom list.

For the SIR data, programs HEAVY and SHELXS-96 were used to find selenium atom sites. The methods used here were similar to those used in MAD methods and coefficients $|Fa|^2$ were replaced by $||F_{native}| - |F_{derivative}||^2$. Similar to the results of MAD data, eight selenium atoms were found from SIR data, confirming that the fifth methionine residue was probably disordered in both molecules in one asymmetric unit. Due to the changes in unit cell parameters caused by different temperatures for data collection and other factors such as errors in data, the coordinates of selenium atoms were slightly different from those determined from MAD data.

3.4.3 Determination of non-crystallographic symmetry

The eight selenium atoms were divided into two groups related by the non-crystallographic symmetry of GFP dimer by comparing distances among the atoms. Supposing the distances between atoms A and B, C and D are the same, it is highly possible that atoms A and B belong to one group and atoms C and D belong to the other. This is also true if the distances among a group of three or more atoms are similar to those of another group of atoms. However, a pair of non-crystallographic symmetry related atoms can be switched between the two groups without changing the intra-group distances, i.e. the atoms in one group may not belong to the same protein monomer. Correct assignment of selenium atoms to each GFP monomers could only be done when the initial model was built into the electron density maps. This ambiguity has no influence on non-crystallographic symmetry calculation which can be done by programs KABSCH or X-PLOR. After dividing the eight selenium atoms into two groups, the non-crystallographic symmetry was calculated and the two groups of atoms were superimposed with each other by the calculated non-crystallographic symmetry operation.

3.4.4 Determination and modification of phase angles

The coordinates and temperature factors of the eight selenium atoms were further refined using MADSYS and the phase angles of Ft could be determined. Initial MAD phased electron density maps were calculated using structure factors Ft. Due to the errors caused in data collection and processing, the initial electron density map typically has many noise peaks. A noise screening method known as "solvent flattening" can be used to

modify the electron density maps and improve the accuracy of experimental phase angles. Since there is a dimer in one asymmetric unit, the electron density maps of the two monomers can be averaged to further improve the accuracy of phase angles which is known as non-crystallographic symmetry averaging.

Program DM in CCP4 package was used for solvent flattening and non-crystallographic symmetry averaging. Each density modification method requires its own masks. Solvent flattening mask was determined automatically by setting the percentage of solvent content in unit cell and defining how the solvent mask would be calculated. For a given percentage of solvent, this percentage of grid points with lowest averaged electron density in a unit cell were treated as solvent region and the electron densities at these grid points were reset to the value of solvent density. Also, the electron densities of the grid points with unusual high average values were reset to the maximal value of allowed electron density. After solvent flattening, the modified electron density map was reverse-Fourier transformed and new phase angles were calculated. The new phases were combined with the original phases based on probability distribution functions and used for calculating a new electron density map.

A mask was also required for the non-crystallographic symmetry averaging. This mask is equivalent to the envelope of the protein molecule. Theoretically, it is possible to calculate the molecular envelope based on the electron density map of the protein. However, such program was not available in the laboratory. Program MAMA was used to produce a sphere mask for initial averaging. After the initial model was built up, the molecular envelope was built by calculating the surface of the structural model using

MAMA. The non-crystallographic symmetry operation can be refined in DM for averaging, but the results should be monitored by the changes in pseudo-R$_{free}$. In the present case, the refinement of non-crystallographic symmetry did not improved the pseudo-R$_{free}$, probably because the symmetry operation has been accurately determined from the positions of selenium atom sites.

Masks should be visually inspected before using them for solvent flattening or non-crystallographic symmetry averaging. Program "mama2ccp4" in CCP4 suite was used to transform MAMA masks to CCP4 masks for DM, and the CCP4 masks could be transformed to CCP4 maps using program "mapmask" which could be viewed by program CHAIN (Sack, 1988). It is also helpful to display masks and the electron density maps together to check that masks covered the correct region of the density maps.

3.5 Building of the initial model and structure refinement

The structural model was built into electron density maps using program O (Jones et al., 1991). An accessory program BONES was used to generate the traces of all potential atoms and make assumption for main chain and side chain atoms. It was suggested to edit the traces manually based on electron density maps to facilitate the building of structural model. Since the editing might be time consuming, it could be omitted in some clear regions where the trace of main chain atoms could be easily followed. However, in some unclear regions such as the ends of the barrel, editing of the traces was

necessary and guesses for trace must be made in order to connect two segments. The four selenium sites were used as the starting sites in the primary sequence. By inspecting the feature of the electron density of the side chains around each methionine residues, the residue numbers could be assigned to each of the four methionine residues. The sequence was traced from these sites to both directions until clues for building the structural model were run out. Other residues with clear electron densities for their side chains were also used for sequence recognition and building of segments into the electron density map. One rule for recognizing electron density maps is that hydrophilic residues should normally be on the surface, contacting solvent molecules, and hydrophobic residues should be hide inside the barrel.

The final model has about 90% of the residues built into the density maps before starting structure refinement. The commands Lego_Auto_MC and Lego_Auto_SC, Refi_zone etc. in program O were used to build and refine the model to a rational geometry. Many regions needed manual adjustments to fit the model into the electron density maps. In order to obtain a good geometry of the model, command Refi_zone in program O was used to adjust the region manually modified. The final model fitted the electron density very well.

Program X-PLOR (Brunger, 1992) was used to refine the structures using parameter sets tophcsdx.pro and parhcsdx.pro (Engh and Huber, 1991). Some topology command and empirical energy parameters were assumed to maintain a rational geometry for the fluorophore. Two programs, PROCHECK and WHATIF were used to check potential errors in structures (Laskowski et al., 1993; Vriend, 1990).

The space group of GFP crystal could be either P4₁2₁2 or its enantiomorph

P4₃2₁2. This ambiguity can only be solved after initial electron density maps were

calculated and inspecting the chirality of helices or residues. The chirality of helices in the

model was right-handed if the space group was assumed to be P4₁2₁2, and left-handed if it

was P4₃2₁2.

3.6 Reduction of GFP crystals

3.6.1 Preparation of reduced GFP crystals

Wild-type GFP was purified from *E. coli* and crystals were grown as described

above. Since GFP crystals were very sensitive to the concentration of MPD and ionic

strength, mother liquor was deoxygenated by a new method instead of bubbling solution

with nitrogen gas which would dehydrate the solution. A volume of 4 ml mother liquor

(58% MPD, 50 mM MES, 0.1% sodium azide, pH 6.8) was added into a sealed test tube

and air was purged with nitrogen gas. After the test tube was shaken for twenty minutes,

air was purged with nitrogen gas again. This process was repeated three times to

deoxygenate the solution without dehydrating it. In the second test tube, 0.01 g of sodium

dithionite powder was added in and purged with nitrogen gas. A volume of 3 ml

deoxygenated mother liquor in the first test tube was transferred by a syringe into the

second one to dissolve the sodium dithionite powder. The final concentration of sodium

dithionite was about 20 mM. Finally, in a glove box filled with nitrogen gas, several GFP

crystals were soaked in the 20 mM sodium dithionite solution in a well of Linbro plate (ICN Biomedicals, Inc.) sealed with immersion oil.

3.6.2 Data collection and processing

After two weeks of reduction, the crystals were significantly bleached and only barely greenish. A reduced GFP crystal was sealed in an 1 mm quartz capillary tube in the glove box. Diffraction data were collected on an R-AXIS IIC area detector with CuKα radiation at room temperature. The distance between the crystal and detector was 70 mm and the oscillation angle was set to 1 degree per frame. The data were processed using DENZO and SCALEPACK.

3.6.3 Structure refinement

Wild-type GFP structure was used as initial model and structures were refined using X-PLOR using parameter sets tophcsdx.pro and parhcsdx.pro. All structures were refined under same conditions except resolution range of data so that the reduced structures were comparable with the wild-type structures. Besides refining single structural model for one GFP molecule in crystal, two independent structural models can also be refined simultaneously for one GFP molecule which will be referred as dual-model refinement. There were no interactions between the two equivalent models and all empirical energy terms were properly weighted so that the sum of empirical energies was similar to that in the single-mode refinement. For example, the bond and angle energy terms should have a weighting factor of 0.5. Non-crystallographic symmetry restraints were applied on all

models requiring similar conformations except the region from residue 63 to 69 for the equivalent models. Therefore, the fluorophore can have two independent conformations in the equivalent models.

Another important parameter is the improper energy term for the four main chain atoms of residue Tyr66, Cα, Cβ, N, and C. In normal L amino acids, these atoms form a tetrahedron and Cα is a chiral center. As part of fluorophore in GFP, they should be co-planar with other fluorophore atoms and Cα is not a chiral center. In the reduced fluorophore, they were expected to form a tetrahedron again. Therefore, the improper energy term should not be too large so that the experimental data can still determine the conformation, or too small so that the noise in the experimental data can not alter the conformation. By using parameter sets tophcsdx.pro and parhcsdx.pro (Engh and Huber, 1991), it was found that a value of 10 and 30 were proper for single-model and dual-model respectively since it is just large enough to keep the atoms in a plane in the wild-type fluorophore. Simulated annealing refinements were used by heating the structure to 500 K and cooling down to 0 K followed by Powell minimization and refinement of temperature factors.

3.7 Determination of the molecular weight and molar absorbance of GFP

3.7.1 Determination of the molecular weight of GFP using HPLC gel filtration column chromatography

Besides used for protein purification, HPLC gel filtration column can also be used for estimating molecular weight of proteins by the elution time through the column. Small molecules are more retarded by the pores of column matrix and lagged behind large molecules. Therefore, small molecules will be eluted after large molecules. This method can be used to estimate the molecular weight of GFP and the degree of its dimerization at different conditions such as different ionic strength.

Protein standards were used to establish the relationship between molecular weight and column elution time. The mixture of the protein standards (Sigma Chemical Company) contains:

a) Vitamin B12   MW   1,350 Da

b) Equine myoglobin   MW   17,000 Da

c) Chicken ovalbumin   MW   44,000 Da

d) Bovine gamma globulin MW 158,000 Da

e) Thyroglobulin   MW 670,000 Da

An HPLC gel filtration column (Bio-Sil Sec 125 from Bio-Red) was pre-equilibrated with one of the following buffers:

A) 0.05 M NaH_2PO_4, 0.05 M Na_2HPO_4, 0.15 M NaCl, 0.05 M NaN_3, pH 6.8

B) 0.05 M NaH_2PO_4, 0.05 M Na_2HPO_4, pH 6.9.

C) 0.01 M Tris, pH 6.9

Buffer A was recommended in the instructions of Sigma Chemical Company for running protein standard. Protein standard and GFP samples were loaded into and eluted through

the column separately using one of the buffers. A linear relationship between the elution time and the logarithm of molecular weight could be established using the protein standard. The apparent molecular weight of GFP could be estimated based on its elution time by interpolating this relationship.

3.7.2 Determination of the molar absorbance of GFP

The molar absorbance of GFP can be determined either by protein assay or by calculation of absorption of tyrosine and tryptophan residues in GFP. Bicinchoninic acid protein assay kit (No. BCA-1) was bought from Sigma Co. Bovine serum albumin was used as protein standard and the procedure (Procedure No. TPRO-562) provided with the assay kit was followed. The relationship between the absorbance of the processed protein standard and concentration of the standard was established. The absorbance of a GFP sample with unknown concentration was measured at 395 nm before it was processed according to the assay procedure. The absorbance of the processed GFP sample comparable with that of the standard was measured. The concentration of the GFP sample could be estimated by interpolating the absorbance-concentration relationship of the standard. Then, the molar absorbance of GFP could be calculated according to Beer's law.

The other method to determine the extinction coefficient of GFP is through the ratio of extinction coefficient at 395 nm and 280 nm. Residues tryptophan and tyrosine have strong absorption at 280 nm and their extinction coefficients at 280 nm have been measured and published (Handbook of Biochemistry, 1970). Thus, the extinction coefficient of the protein could be estimated as the sum of extinction coefficients from all tryptophan and

tyrosine residues. For a pure GFP sample, the ratio of extinction coefficients at 395 nm and 280 nm is equal to the ratio of absorbance at these wavelength which can be measured. Therefore, extinction coefficient of GFP at 395 nm can be easily obtained.

Chapter 4

Results and Structure Analysis

4.1 Solution of molecular structure of GFP

4.1.1 Preparation of heavy atom derivatives

The GFP crystals typically have a shape of bipyramid, indicating a possible four-fold axis (Figure 4.1). The crystals are also greenish under daylight and have intense green fluorescence under UV light, indicating the functional fluorophore inside the protein.

In order to solve the structure of GFP, heavy atom derivatives are required. Attempts of using multiple isomorphous replacement methods to solve the GFP structure failed due to the problem of non-isomorphous replacement and existence of multiple mutants of GFP. It was found that the internal order of GFP crystals were very sensitive to the changes in solution. For example, two diffraction data sets were taken on one GFP crystal soaked in 50 mM MES and 10 mM MES, respectively. Although the concentration of precipitant MPD might also vary a little, the diffraction differences were from 25% to 50% in different resolution ranges while the typical noise level was about 5%. The changes in unit cell parameters were very small in this case. Obviously, the concentration
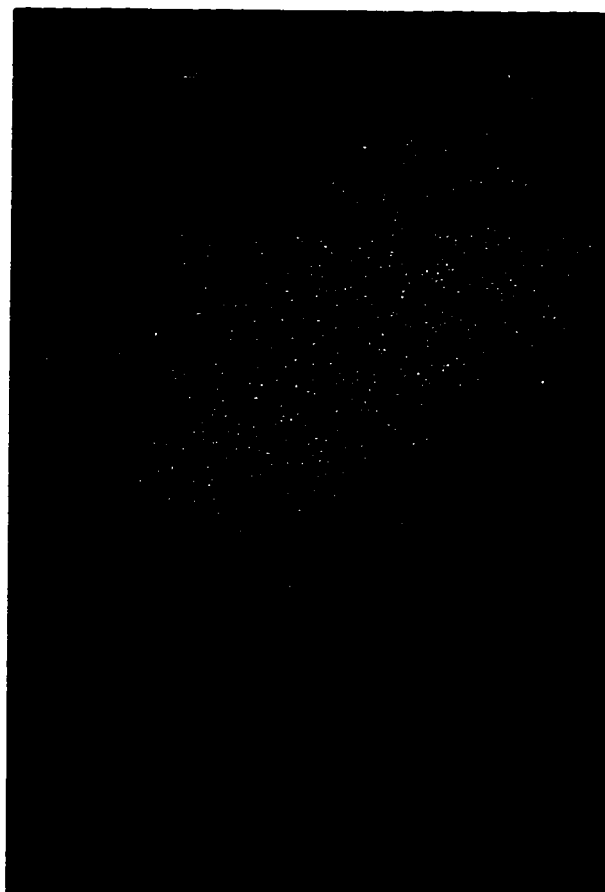
Figure 4.1   Photomicrograph of GFP crystals.   The crystals grew as square bipyramids and showed four-fold symmetry.   They were greenish and had bright fluorescence when illuminated with UV-light, indicating that the GFP molecules were functional in the crystals.

of precipitant and/or ionic strength changed the internal order of GFP crystals such as the orientation of the GFP molecules. When preparing heavy atom derivatives, the ionic strength of solutions was changed by the heavy atom compounds which resulted in significant non-isomorphous replacement, even the unit cell parameters were changed significantly in some cases.

The problem of non-isomorphous replacement was further exaggerated by the existence of multiple mutants in different batches. It was found that crystals grew from different samples had different shapes. Some of them had very irregular and complicated shapes like those of a trident or a football. They also had different hardness, growth rate, and unit cell parameters. The bipyramid-shaped crystals were relatively soft and fragile, but the football-shaped crystals were very hard and solid like sand. The a axis of unit cell varied from 89.0 Å to 89.4 Å, and c axis varied from 119.8 Å to 120.2 Å. Multiple peaks also appeared in the spectra of anion exchange HPLC columns in protein purification.

The homogeneity of samples was quantitatively checked by N-terminal amino acid sequence analysis and matrix-assisted laser desorption ionization mass spectrometry performed by the University of Texas Health Science Center analytical chemistry service. The correct N-terminal sequence and molecular weight of GFP, coded by the known pTu#58 gene sequence, including the original inadvertent Q80R PCR error during the cloning of the gene for GFP (Chalfie et al, 1994), the cyclization between Ser65 and Gly67, and the oxidation of side chain of Tyr66, are ASKGE and 26835.5 Da. However, the N-terminal amino acids of two GFP samples were ASKGE and AVPGE and their molecular weights were 26058 (±1) Da and 26303 (±3) Da, respectively. Supposing that

the N-terminal sequence is ASKGE and the last four and six residues are cleaved at the C-terminus from the wild-type GFP, the calculated molecular weights are 26300.9 Da and 26054.6 Da, respectively. Since the results of molecular weights are more reliable than that of amino acid sequence analysis, it is probably that the two mutants have the correct N-terminus but a cleaved C-terminus. These are consistent with the results of deletion mapping of GFP that suggest the last six residues at C-terminus are not critical for the function of GFP (Dopf and Horiagon, 1996). More samples were measured later which gave various molecular weights. The diffraction differences in data taken on crystals grown from different batches of samples could be as large as 10%, much higher than the normal noise level.

However, several batches of GFP samples expressed in our lab using the pTu#58 plasmid provided by the collaborator all have the correct N-terminal sequence and molecular weight. Thus, the GFP mutants found before were probably produced by unexpected proteolysis during the process of cell lysis or protein purification. Since the mutants with different C-termini were physically and chemically similar to each other, it was very difficult to separate them. It is not surprising that crystals grown from the mixture of GFP mutants have different shapes due to the defects on crystal growth surfaces caused by inhomogeneity.

Crosslinking using glutaraldehyde was tried to overcome the non-isomorphous problem. Different concentration of glutaraldehyde and soaking time were tried. Although higher concentration of glutaraldehyde or longer soaking time could make crystals more resistant to the changes in solution, they also resulted in larger mosaicity and lower

diffraction capability. In order to reduce noise from non-isomorphism, native and heavy atom derivative crystals must be crosslinked at same conditions. Since MAD methods were chosen to solve the structure, crosslinking method for MIR was not fully tested.

Another method to overcome non-isomorphous replacement is to produce selenomethionyl protein and collect MAD data to solve phasing problem (section 2.6). There are five methionine residues in the wild-type *Aequorea* GFP. The calculated molecular weights of the wild-type and selenomethionyl GFP are 26835.5 and 27070.0 Da, respectively. Their molecular weights determined by mass spectrometry are 26836.1 ($\pm$0.9) and 27069.3 ($\pm$1.4) Da, respectively. The differences of 0.6 and -0.7 Da are very small, less than the error of measurement. The error limits do not allow accurate determination of the degree of oxidation of the dehydrotyrosine, however. These results showed that the starting material for the crystallization was completely functional GFP and 100% of methionine residues were replaced by selenomethionine residues.

4.1.2 Data collection

Data collected at Stanford Synchrotron Radiation Laboratories   The MAD data collected at Stanford Synchrotron Radiation Laboratories were not used for solving GFP structure due to inaccuracy of the data. The data were collected at a cryo-temperature which resulted in the increase of mosaicity from 0.3° at room temperature to about 0.8° and the decrease of unit cell from about 89 Å to 87 Å in the a and b axes. The change in the c axis was relatively small which was in the order of tenth of angstrom. Since the mosaicity was much larger than the oscillation angle 0.5°, which was small in order to reduce overlapping

of diffraction spots due to the large mosaicity, all diffraction spots were partially measured which might be a reason for the low quality of the data. Another reason could be the data collection methods. The wavelength was changed after a wedge of $100°$ was scanned continuously. This may cause significant noise compared with anomalous signal in the intensities of Friedel pairs. Although the data showed the anomalous signal, the noise level was too high to determine the positions of the five selenium atoms. It has been suggested that changing wavelengths for each frame and using inverse beam geometry, which measures true Friedel pairs in two sequential frames, can reduce noise in the MAD data.

Data collected at Brookhaven National Laboratories   Probably due to the proper strategy of data collection, the MAD data collected at Brookhaven National Laboratories were very accurate. The R_{sym} of MAD data collected at four wavelengths were about 5% (Table 4.1). The signal to noise ratios of anomalous dispersion were estimated by comparing Bijvoet differences of centric and acentric data using the program MADSYS (Yang et al., 1990) (Table 4.2). The conditions of centric diffractions of space group P41212 and its enantiomorph P43212 are that the absolute values of h and k are equal, or at least one of the three indices, h, k, and l, is zero. Bijvoet differences of centric and acentric data at each wavelengths are calculated as the diagonal elements in the table. The dispersive differences, which are the differences in intensities of the same diffraction at two wavelengths, are the off-diagonal elements in the table. Since the centric data have no Bijvoet differences, the intensity differences of centric Bijvoet pairs can be regarded as the noise level of anomalous signal. In each resolution range, the differences of centric data at four wavelengths are similar to each other, indicating the same quality of data at different

wavelengths. At the first wavelength, acentric data have small Bijvoet differences which are similar to the noise level in centric data in all resolution ranges, indicating very weak absorption of X-rays since the wavelength is remote to the absorption edge of the selenium atom. At all other wavelengths, acentric data have significant Bijvoet differences, which could be even two-fold higher than the noise level, indicating significant absorption of X-rays by selenium atoms. For dispersive differences, the data collected at the second and fourth wavelengths have the largest differences in all resolution ranges, and those collected at the first and fourth wavelengths have the smallest. These results are consistent with the $\Delta f'$ and $\Delta f''$ values refined (Table 4.2).

Table 4.1 Results of multiwavelength anomalous dispersion data.

	λ_1	λ_2	λ_3	λ_4
Energy (ev)	12550	12659	12662	12800
Wavelength (Å)	0.9879	0.9794	0.9792	0.9686
Unit cell (Å) a/b	87.143	87.150	87.147	87.150
c	119.856	119.853	119.854	119.858
R_{sym} (%)	5.2	5.4	5.4	5.3
Signal/noise ratio	13.9	14.1	13.6	14.3
Completeness (%)	98.5	98.8	98.8	99.0

4.1.3 Determination of selenium atom sites

Two programs, HEAVY and SHELXS-96, were used to locate the sites of selenium atoms (Terwilliger et al., 1987; Sheldrick et al., 1993). SHELXS-96 works

Table 4.2 Anomalous diffraction differences and scattering factors for selenomethionyl GFP*.

Wavelength (Å)	30 > d > 3.4 (Å)				3.4 > d > 2.7 (Å)				2.7 > d > 2.2 (Å)				Scattering factors (e)	
	0.9879	0.9794	0.9792	0.9686	0.9879	0.9794	0.9792	0.9686	0.9879	0.9794	0.9792	0.9686	Df'	Df"
0.9879	0.026 (0.026)	0.042	0.030	0.020	0.037 (0.035)	0.046	0.037	0.032	0.064 (0.052)	0.071	0.068	0.064	-4.0	1.1
0.9794		0.050 (0.029)	0.026	0.045		0.058 (0.039)	0.035	0.049		0.088 (0.060)	0.065	0.077	-10.5	3.9
0.9792			0.067 (0.031)	0.032			0.074 (0.042)	0.040			0.101 (0.062)	0.071	-7.9	5.5
0.9686				0.049 (0.027)				0.059 (0.039)				0.089 (0.064)	-3.4	3.9

* Following the format used by Yang et al. (1990). Bijvoet differences are given in diagonal elements with centric values in parentheses, and dispersive differences are given in off-diagonal elements. Scattering factors were chosen to minimize the cumulative errors in Bijvoet and dispersive terms.

excellently in finding initial solutions of heavy atom sites. It is easy to use and most of the input parameters have default values, but it can not refine heavy atom parameters which may be done using other programs such as SHELXL-96. Compared with other programs, HEAVY is especially good at refining heavy atom parameters since it can discriminate false heavy atom sites from true ones. For a false site refined by HEAVY, its occupancy will be small and the temperature factor will be large. Probably, the best strategy is to use SHELXS-96 for searching initial trial sites and use HEAVY for refinement.

A maximum of 5 or 10 selenium atoms were expected to be found in one asymmetric unit depending on whether there was a monomer or dimer. The Matthews number V_M is defined as the volume of one unit cell divided by the total mass of protein molecules in one unit cell. The average value of V_M is 2.15 $Å^3$/Da for all protein crystals obtained so far. For GFP crystals, there are eight asymmetric units in one unit cell with a volume of about 950,000 $Å^3$. The V_M will be 2.2 $Å^3$/Da if there is a dimer in one asymmetric unit which is highly possible.

Using the MAD data collected at Brookhaven National Laboratories, a total of eight selenium atoms with rational parameters were found by calculating Patterson maps using IFaI (Table 4.3). The root-mean-square (RMS) of the distances between the two groups of selenium atoms superimposed was about 0.2 Å, showing that their coordinates refined by HEAVY were very accurate. Since there are five selenomethionine residues in one GFP molecule, the existence of eight selenium atoms indicates that there are two GFP molecules in one asymmetric unit, i.e., GFP molecules forms a dimer in the crystal. Based on the

amino acid sequence of GFP, the missing selenomethionine residue is probably Met233

since it is only five residues away from the C-terminus and is probably disordered.

The positions of selenium atoms were also determined independently by SIR

methods. Methionyl and selenomethionyl data sets were collected at room temperature on

an R-Axis IIC area detector with CuKα radiation (Table 4.4). The differences in unit cell

parameters were very small, only about 0.1% of the length of the axes, indicating very

good isomorphous replacement for the selenomethionyl crystals. The heavy atom

derivatives prepared by soaking crystals in heavy atom solutions showed significant non-

isomorphous problem which has up to 2% changes in unit cell parameters due to the

sensitivity of GFP crystals to the changes in solutions.

Table 4.3 The fractional coordinates of selenium atoms.

atom	group 1*			group 2			distance (Å)#
	x‡	y	z	x	y	z	
se1	-0.0704	0.7936	0.0030	0.3935	0.6817	0.0489	0.219
se2	0.0939	0.6827	0.1008	0.2214	0.7432	-0.0576	0.293
se3	-0.0053	0.7346	-0.1460	0.2910	0.5301	0.1406	0.101
se4	-0.0131	0.9478	-0.0501	0.3839	0.7614	0.1694	0.137

* A total of eight selenium atoms were found. They were assigned to two groups. A pair of corresponding atoms in the two groups belong to different GFP monomers, but atoms in the same group may not belong to the same monomer.

The distances between pairs of selenium atoms in the two groups were calculated after the two groups of atoms were fitted to each other by the non-crystallographic symmetry operation.

‡ The fractional coordinates of selenium atoms are given here.

Table 4.4 Results of SIR data.

	Rsym (%)	complete-ness (%)	signal/noise ratio	a/b-axis (Å)	c-axis (Å)
methionyl GFP	9.6	91.0	9.4	88.85	120.54
selenomethionyl GFP	5.9	96.3	13.6	88.78	120.65

The R_{iso} which is the percentage of differences between the methionyl and selenomethionyl data sets is 10.6% for a total of about 14820 common diffractions scaled in the two data sets. Since selenium is relatively light compared with platinum, uranium, mercury, etc., the R_{iso} is expected to be smaller than values typical of other heavy atom derivatives. Despite the relatively small R_{iso}, the diffraction signals of selenium atoms were strong enough for determining the same positions of selenium atoms as given by the MAD data.

4.1.4 The initial electron density map

For solvent flattening, the percentage of solvent in crystal needs to be determined first. Theoretically, the percentage of solvent content $V_{solvent}$ is given by

$$V_{solvent} = 1 - \frac{1.23 \text{ Å}^3 / \text{Da}}{V_M \text{ in Å}^3 / \text{Da}}$$

where V_M is the Matthews number and 1.23 Å³/Da is the specific volume of protein molecules (Drenth, 1994). The $V_{solvent}$ of GFP is 44% given a value of 2.2 Å³/Da for the V_M of GFP. Different values of solvent content were tried and evaluated by DM which calculated a pseudo R_{free} for the difference between calculated and measured structure

factors. Rfree can serve as a good index for a proper percentage of solvent content. It was 48% at 20% solvent content and decreased to a minimum of 39% at 40% solvent content which was very close to the theoretical value.

After solvent flattening and non-crystallographic symmetry averaging that improved accuracy of phase angles, electron density map was calculated which showed eleven β strands forming a barrel and some residues in the barrel. Due to the high quality of experimental data, it was very clear in the map that the electron densities of side chain atoms branched off from those of main chain atoms. In some regions, the electron density clearly showed the hydroxyl oxygen atoms of the tyrosine residues and distinguished them from phenylalanine residues. However, the densities of several loops, especially those on the ends of the barrel, were unclear and the trace was difficult to follow. Electron densities near or on the dimer interface were much clearer than those on the opposite side of the interface which means the atoms near the dimer interface are more stable in their positions. This phenomenon may be caused by thermal movement of GFP molecules. Each GFP molecule has thermal movement as a rigid body which is restricted by dimer interactions in GFP crystals. Thus, the dimer interface probably serves as the hinge for the movement of monomers so that atoms near or on the interface have smaller displacement than those far away from the interface.

4.1.5 Structure refinement

The initial crystallographic R and Rfree of the model is about 45%. After several cycles of simulated annealing refinement, a number of cycles of POWELL minimization,

and adding of water molecules, the final R and Rfree are 21.3% and 26.2%, respectively. Some residues such as Asp, Asn, Glu, Gln, and Lys, have multiple conformations or disordered side chains.

In the structural model, five segments, Phe114 to Thr118, Asp155 to Asn159, Glu172 to Val176, Ile188 to Val193, and Ser208 to Asp216, looped away from the body of the β-barrel and might not be correct due to the weak electron density in these regions. The MAD phased experimental electron densities at these segments were also relatively weak. In order to check the correctness of the structural model at these segments, omit difference (Fobs-Fcalc) electron density maps were calculated and positive densities appeared at these segments, indicating that the structural model was correct.

4.2 The structure of wild-type *Aequorea* GFP

4.2.1 The tertiary structure of GFP

The molecular structure of GFP is a regular β-barrel with eleven strands on the outside of cylinder (Figure 4.2). All strands are anti-parallel except the conjunction between strands 1 and 6. The cylinder has a diameter of about 30 Å and a length of about 40 Å. The volume of the barrel is about 30600 Å^3. Inspection of the density within the cylinders reveals the modified tyrosine side chain as part of an irregular α-helical segment (Figure 4.3). Small sections of α-helix also form caps on the ends of the cylinders. This motif, with a single α-helix inside a very uniform cylinder of β-sheet structure, represents
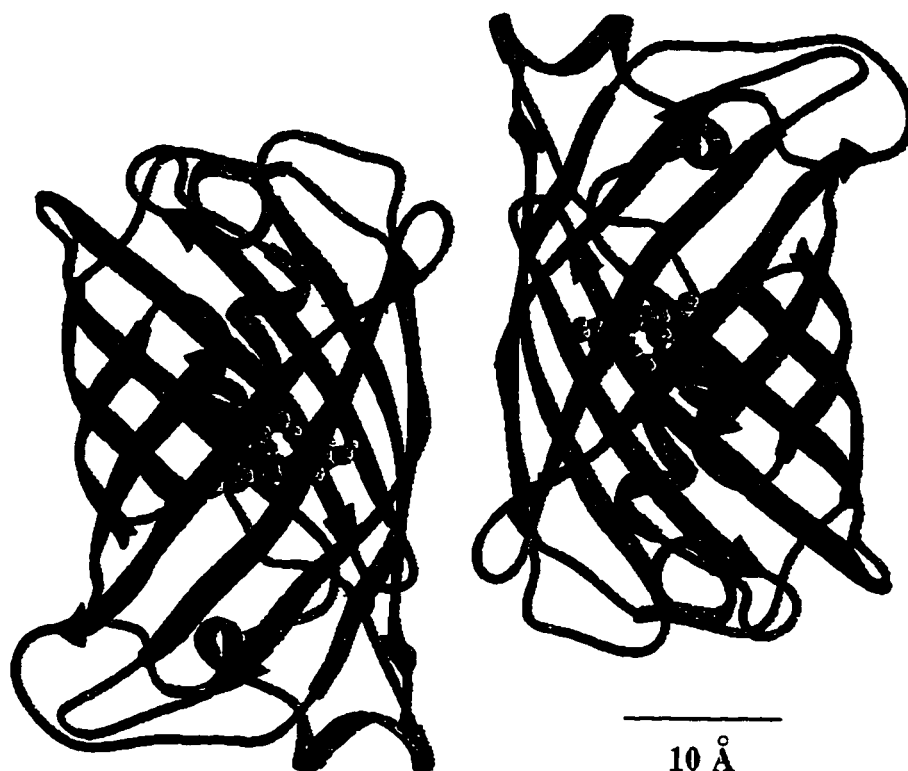
Figure 4.2 The overall shape of the protein and its association into dimers. Eleven strands of β-sheet form the walls of a cylinder. Short segments of α-helices cap the top and bottom of the 'β-can' and also provide a scaffold for the fluorophore which is near the geometric center of the can. This folding motif, with β-sheet outside and α-helix inside, represents a new class of proteins. Two monomers are associated into a dimer in the crystal and in solution at low ionic strengths. This view is directly down the two-fold axis of the non-crystallographic symmetry.

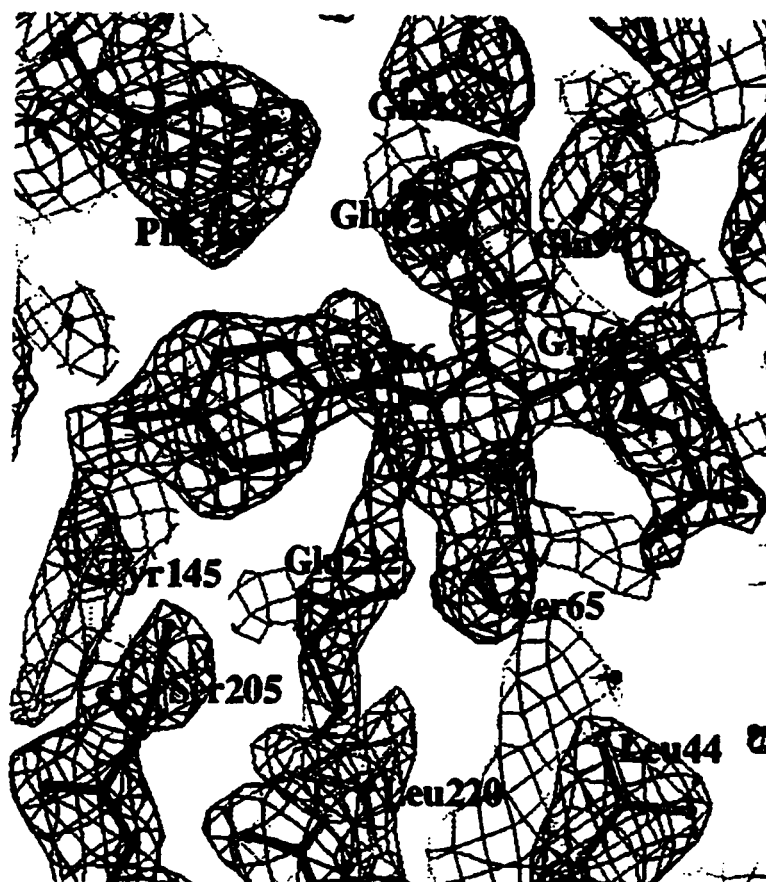Figure 4.3 Model of the fluorophore and its environment superposed on the MAD-phased electron density map at 2.2 Å resolution. The clear definition throughout the map allowed the chain to be traced and side chains to be well placed. The density for Ser65, Tyr66, and Gly67 is quite consistent with the dehydrotyrosine-imidazolidone structure proposed for the fluorophore. Many of the side chains adjacent to the fluorophore are labeled.
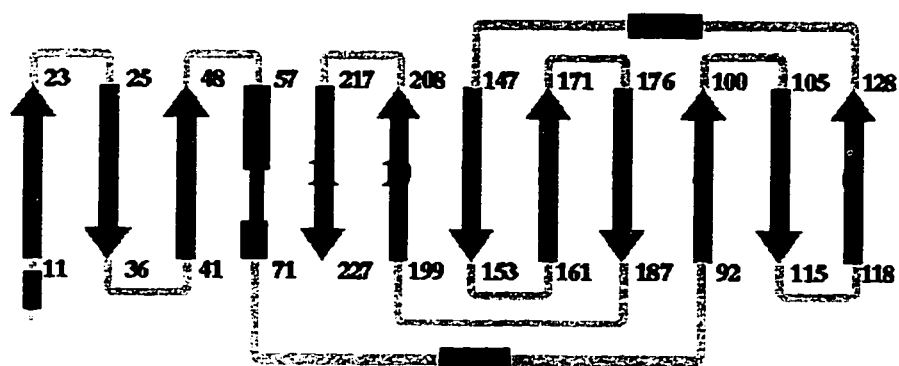
Figure 4.4 A topology diagram of the folding pattern in GFP. The β-sheet strands are shown in medium grey, α-helices in dark grey, and connecting loops in light grey. The positions in the sequence that begin and end each major secondary structure element are also given. The anti-parallel strands (except for the interactions between strands 1 and 6) make a tightly formed barrel.

a new protein class, as it is not similar to any other known protein structure. Within each β-barrel, there are three sheet motifs comprised of three anti-parallel strands each. They are strands 1, 2, 3 for the first sheet motif, strands 4, 5, 6 for the second one, and 7, 8, 9 for the third one (Figure 4.4). It is possible that the three sheet motifs are folded first in apo-GFP.

4.2.2 The fluorophore and its environment

Conformation of the fluorophore   The conformation of the fluorophore was determined carefully. In the structure refinement using X-PLOR, conformations are determined by two factors. One is the structural restraints that define the expected geometry for a model. The other is the experimental data against which a model is refined. If the structural restraints are too strong, the refined model will be biased to the expected geometry. If the structural restraints are too weak, the model may be deformed by noise in the experimental data. Thus, a proper balance must be kept between the two factors. In the current case, the geometry of the fluorophore was to be determined and no experimental information was available except some knowledge about the conformations of the fluorophore. In nature, the $C\alpha$ atom of Tyr66 is a chiral center before the oxidation of its side chain. After the oxidization, all fluorophore atoms are co-planar and the $C\alpha$ atom is no longer a chiral center. The planarity of a chiral center can be estimated by calculating the tetrahedral volume defined by $C\alpha$, $C\beta$, C, and N atoms in each residue. In the single-model refinement for the wild-type GFP while the improper energy term for the $C\alpha$ atom of Tyr66 was set to 10 (section 3.6), the tetrahedral volumes of Tyr66 are 0.04 and 0.02 Å^3

for the dimer. These values are very small compared with the average volume of 0.43 Å^3 for normal L-amino acids which have a standard deviation less than 0.03 Å^3 (Figure 4.5). Even if the improper energy term is set to zero, the fluorophore is still quite flat. Although the bond-length between the carbonyl carbon atom of Ser65 and nitrogen atom of Gly67 is not restrained, it is about 1.43 Å, indicating a well formed covalent bond. Finally, the lack of high peaks in difference electron density (Fo-Fc) maps in this region showed that the crystal contained fully cyclized GFP molecules.

The environment of the fluorophore The fluorophore is highly protected by the β-barrel, located on the central helix within a couple of Ångstrom of the geometric center of the cylinder. The fluorophore is surrounded by a surprising number of polar or charged residues in the immediate environment that definitely have a significant influence on the spectra of GFP. A table of all atoms that come in contact with the fluorophore and their distances to the fluorophore is provided (Table 4.5).

The orientation of the side chain of Ser65 is almost perpendicular to the fluorophore plane (Figure 4.6). The hydroxyl oxygen atom of Ser65 closely interacts with the side chain of Glu222 with a water molecule hydrogen-binding with them (Figure 4.7). The smallest distance between the side chains is 2.7 Å. Four hydrophobic residues, Leu42, Leu44, Val68, and Leu220, are also near the side chain of Ser65 and may interact with it by van der Waals' interactions.

The side chain of Tyr66 has close interaction with His148 (Figure 4.7). The nearest distance between them is 3.3 Å. Like the side chain of Ser65, a water molecule is hydrogen-bound to the hydroxyl oxygen atom of Tyr66 separated by a distance of 2.6 Å.
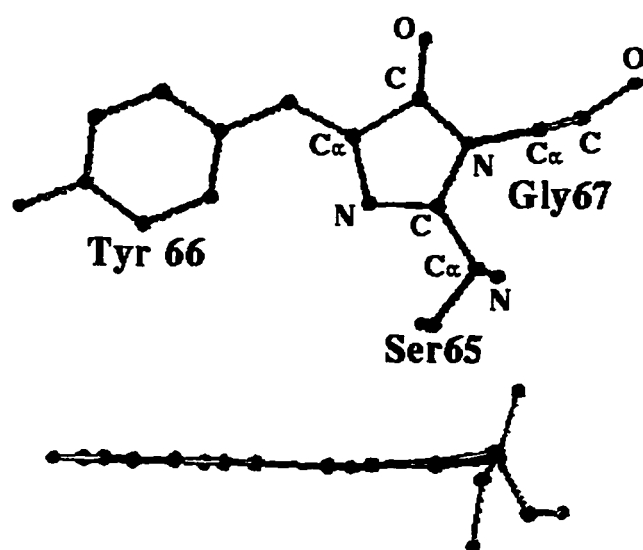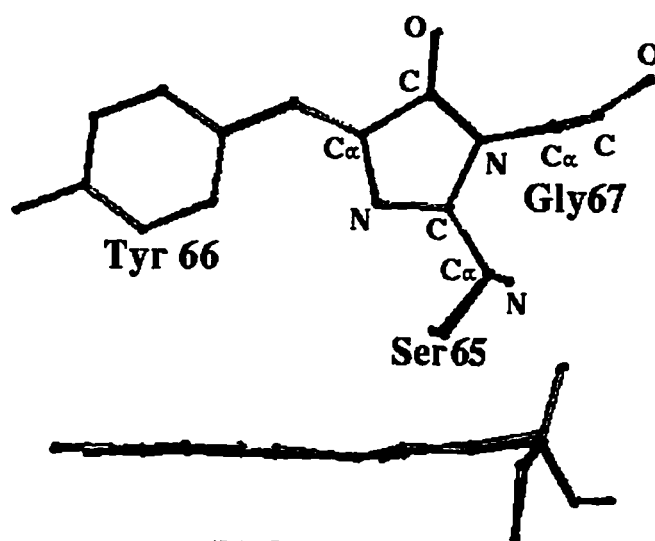
(a) Single-model

(b) Dual-model

Figure 4.5 The fluorophore in wild-type GFP. a) Top view and side view of the fluorophore in single-model refinement. b) Top view and side view of the fluorophore in dual-model refinement (section 3.6). All fluorophore atoms are co-planar.
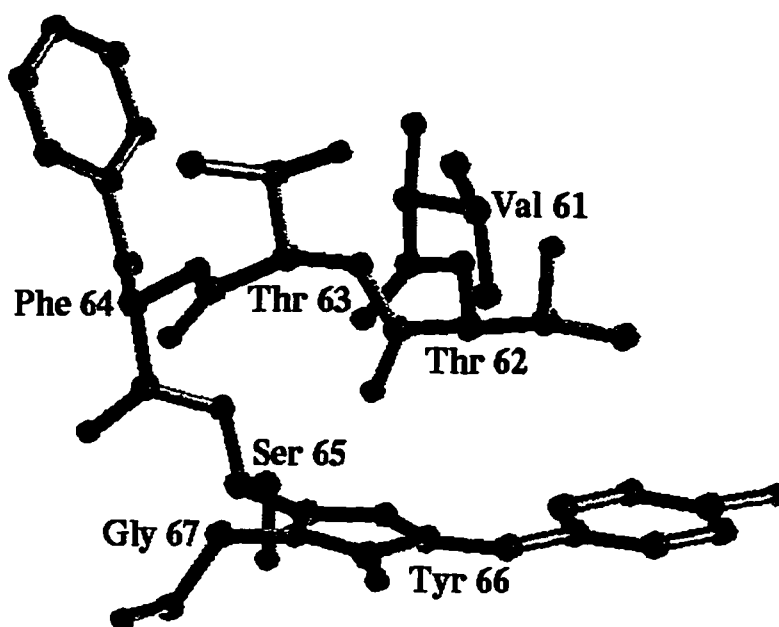
Figure 4.6 The fluorophore and its N-terminal residues. The hydroxyl group of Ser65 is oriented nearly perpendicular to the fluorophore plane. On the N-terminal side of the fluorophore, the carbonyl oxygen of residues Val61, Thr62, and Thr63 are oriented in such way that they are all pointed to the fluorophore with distances as small as 3.0 Å.
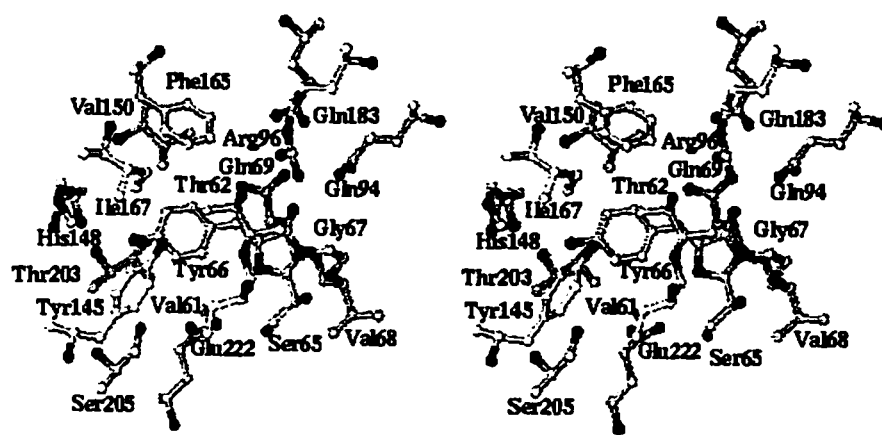
Figure 4.7 Stereo view of the fluorophore and its environment. Important residues

such as Glu222, Arg96, His148, Gln94, etc., are shown. Glu222 has close contact with

Ser65. His148, Gln94 and Arg96 can be seen on opposite ends of the fluorophore and

probably stabilize resonant forms of the fluorophore. Charged, polar, and non-polar side

chains all contact the fluorophore in some way.

Table 4.5 List of amino acid side chains with close contacts (less than 5 Å) to the fluorophore*.

| protein | | fluorophore | | distance |
residue	atom	residue	atom	(Å)
Arg96	NH2	Tyr66	O	2.7
Glu222	OE1	Ser65	OG	2.7
Gln94	NE2	Tyr66	O	3.0
His148	ND1	Tyr66	OH	3.3
Gln69	CD	Tyr66	O	3.4
Leu42	CD1	Ser65	OG	3.4
Thr203	CG2	Tyr66	CE2	3.5
Val150	CG2	Tyr66	CE1	3.6
Phe165	CE1	Tyr66	CD1	3.6
Thr62	CG2	Tyr66	CG	3.6
Ile167	CD1	Tyr66	OH	3.7
Leu220	CD2	Ser65	CB	3.7
Tyr145	CE2	Tyr66	OH	3.7
Ser205	OG	Tyr66	CE2	4.0
Val68	CG2	Ser65	OG	4.0
Leu44	CD1	Ser65	OG	4.3
Phe64	CB	Ser65	CA	4.3
Val61	CG1	Tyr66	CE2	4.5
Gln183	NE2	Tyr66	O	4.8

* The fluorophore is defined as the 7 atoms of the phenol of Tyr66, the 6 atoms of the imidazolidone, the side chain of Ser65, and the bridging methylene between the rings. The amino acid side chains listed here would be expected to have the most direct effects on fluorescence and perhaps fluorophore formation. The atom names are taken from the Brookhaven Protein Data Bank nomenclature.

There are at least four polar residues, Thr62, Tyr145, Thr203, and Ser205, that interact with Tyr66. The side chain of Phe165 is close to Tyr66 and fills the space around the fluorophore. Three other hydrophobic residues, Val61, Val150, and Ile167, are also close to Tyr66. The carbonyl oxygen atom of Tyr66 interacts with a group of polar and charged residues, Gln69, Gln94, Arg96, and Gln183 with distances from 2.7 Å to 4.8 Å. The closest residue is Arg96 with a distance of 2.7 Å. Besides His148 and Glu222, Arg96 is another important and charged residue in the environment of the fluorophore.

The peptide segment at the N-terminal end of the fluorophore interacts with the fluorophore in an unusual way. The carbonyl oxygen of residues Val61, Thr62, and Thr63 are oriented in such way that they all point to the fluorophore with distances as small as 3.0 Å (Figure 4.6).

The space on both side of the fluorophore plane is occupied. On one side, the side chain of Thr62 is just above the fluorophore and there are no water molecules on this side. On the other side, there is no side chain near the fluorophore and three water molecules are found to fill the space. If mutants are designed to replace the water molecules, residues Thr203 and Gln69 are the best ones to be mutated to residues such as histidine, phenylalanine, etc. However, since Gln69 forms two hydrogen bonds with Gln183, a mutation at this site may have some side effects.

Near the five-member ring of the fluorophore, residues Tyr92, Aln121, and Val112 are near the carbonyl oxygen atom of Gly67 with the smallest distance of 3.7 Å. Although the carbonyl group of Gly67 is not part of the fluorophore like the side chain of Ser65, the changes in its interaction with environment should have significant influences on the spectra of GFP.

The central helix does not fill the volume inside the barrel and a total of about 15 water molecules are found in the cavities of the barrel (Figure 4.8). For example, about six water molecules are hydrogen-bound and form a line along the center helix. Four water molecules are near the fluorophore and on one side of its plane. They have hydrogen bonds with the carbonyl oxygen atom of Gly67, hydroxyl oxygen atoms of Ser65 and Tyr66, side chains of Glu222, Thr203 and Gln69. These water molecules are also found
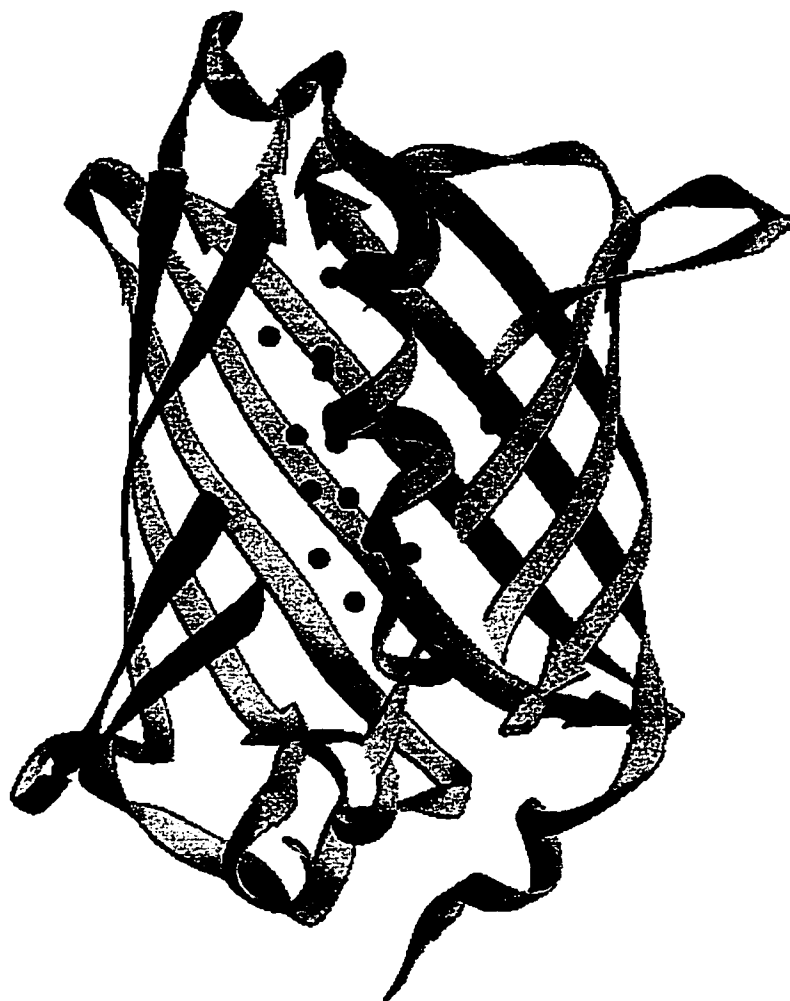
Figure 4.8 The water molecules inside the β-barrel. Water molecules are found inside β-barrel near the central helix. Some of them are hydrogen-bound to the fluorophore or charged residues that interact with the fluorophore.

in native and S65T GFP structures arranged in a similar pattern. They definitely interact with the fluorophore and may change the spectral properties if they are replaced by protein atoms.

4.2.3 Dimer interactions

In each asymmetric unit of a unit cell, there is a dimer of GFP molecules (Figure 4.2). The non-crystallographic symmetry is maintained by extensive contacts and thus is likely to be the source of the dimerization seen in solution studies. The dimer contacts are fairly tight and consist of a core of hydrophobic side chains from Ala206, Leu207, Leu221, and Phe223 from each of the two monomers and a wealth of hydrophilic contacts (Figure 4.9) (Table 4.6). Contacts with other crystallographic molecules are not extensive, and the salt-dependence of this dimer interface and/or the loose contacts with neighboring molecules may explain the difficulties with isomorphism in initial heavy atom phasing studies.

Residue Tyr39 is involved in the dimer interactions. It may interact with residues Ser208, Pro211, and Val219 in the other monomer, and Phe223 in its own monomer. It had two conformations in the initial model built from MAD phased electron density maps. The occupancies of the two conformations are refined to 55% (conformation A) vs. 33% (conformation B) in one monomer and 79% vs. 0% in the other (Figure 4.10). As a control, the structure was refined against another wild-type GFP data set collected at similar conditions. The occupancies are 48%, 32% in one monomer and 76%, 18% in the other,

indicating that the occupancies of the residue are different in the two monomers. It is

possible that the conformation of Tyr39 is partially affected by dimer interactions.

Table 4.6 Dimer interactions*.

monomer		monomer		Distance (Å)
Residue	Atom	Residue	Atom	
Ala206	CB	Phe223	CD1	3.5
Leu221	CD2	Phe223	CZ	4.4
Leu207	O	Phe223	CB	4.7
Tyr145	O	Ser147	OG	2.8
Gln204	OE1	Leu207	O	2.8
Tyr39	OH	Ser208	OG	2.9
Glu142	OE2	Asn149	OD1	3.1
Asn144	CB	Ser147	OG	3.2
Asn144	CA	Gln204	NE2	3.3
Asn146	OD1	Ser147	N	3.3
Tyr145	N	Gln204	NE2	3.4
Tyr39	CZ	Lys209	O	3.5
Tyr39	OH	Val219	CG1	3.5
Gln204	NE2	Ala206	CA	3.5
Tyr143	O	Gln204	OE1	3.6
Asn146	OD1	Arg168	NH1	3.6
Tyr39	CD1	Pro211	CD	3.7
Ser208	CB	Phe223	CD2	3.7
Asn146	CB	Asn146	OD1	3.7
Thr38	O	Pro211	CG	4.0
Asn170	ND2	Ser147	O	4.0
Tyr39	CE1	Asp210	CA	4.1
Arg73	NH1	Pro211	CB	4.1
Arg168	NH1	Asn170	ND2	4.2
Asn144	ND2	Ser202	OG	4.4
Gln204	NE2	Ser205	O	4.4
Tyr145	O	Asn146	C	4.6
Tyr143	O	Ser202	OG	4.8
Asn144	ND2	Asn149	ND2	4.9

* The first three lines are the residues of the hydrophobic core.
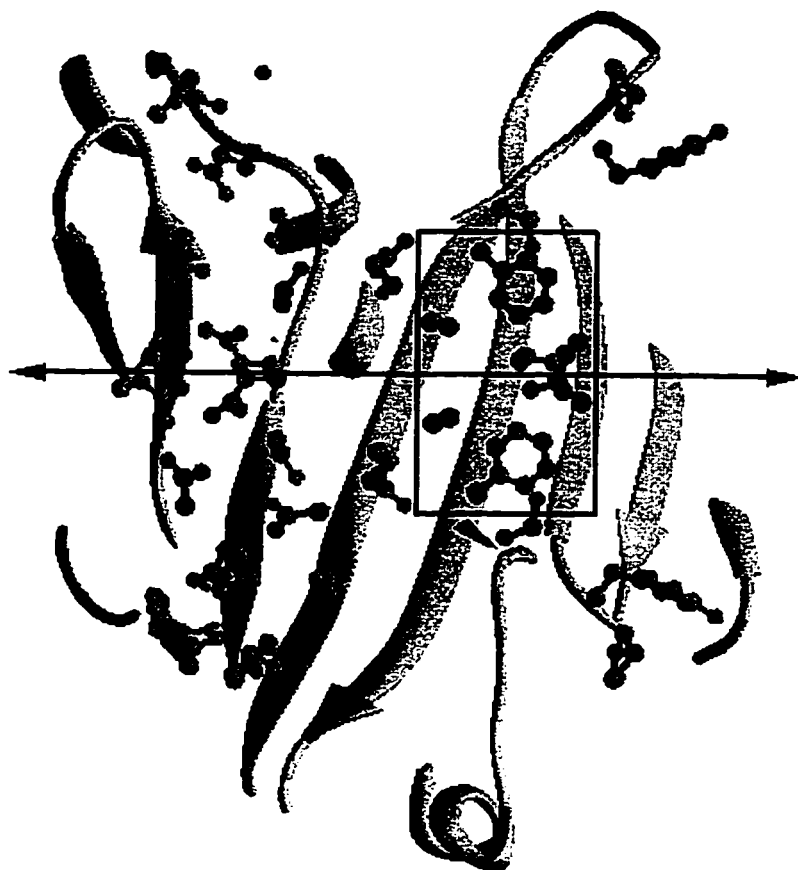
Figure 4.9 The dimer contact region. The two polypeptide chains associate over a broad area, with a small hydrophobic patch (in the box) and numerous hydrophilic contacts. The two-fold symmetry axis is in the plane of the page, and is marked by the arrow.

Figure 4.10   Stereo view of the conformations of Tyr39.   The monomers are shown by different grey levels.   The conformation A has the side chain of Tyr39 pointing to the upper left, and conformation B has the side chain pointing to the lower left.

4.3 Structure of reduced wild-type *Aequorea* GFP

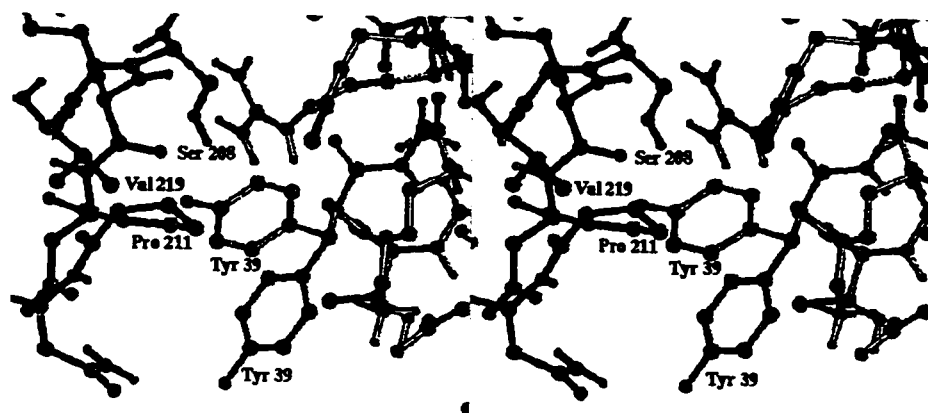It has been reported that GFP can be denatured by several chemical reducing or oxidizing reagents and lose its fluorescence (Inouye and Tsuji, 1994b). Oxidation and denaturation of GFP by hydrogen peroxide is irreversible as the fluorescence is eliminated and can not be recovered. On the contrary, GFP can be reduced reversibly by sodium dithionite, ferrous sulfate, or ABTS, i.e. fluorescence is eliminated after reduction and can be recovered by exposing sample solutions to molecular oxygen. The absorption spectrum of GFP reduced by sodium dithionite is similar to that of native GFP in the range from 400 nm to 600 nm. Since one of the two steps in the formation of the fluorophore is oxidization of the side chain of Tyr66, the reduction may result in a structure similar to the intermediate state after cyclization of the fluorophore but before oxidation of Tyr66. Thus, a structure of reduced GFP will lead to a better understanding about the mechanism of the formation of the fluorophore.

After two weeks of reduction by sodium dithionite, the wild-type *Aequorea victoria* GFP crystal diffracted to only 2.6 Å instead of 1.9 Å in the wild-type state. The R_{sym} and completeness are 6.1% and 97.8%, respectively. The crystallographic R and R_{free} are 23% and 34% for the single-model refinement, and 19% and 28% for the dual-model refinement in which two independent structural models are simultaneously refined as described in section 3.6.

Theoretically, GFP will not be fluorescent if the resonance system is broken by the reduction of the side chain of Tyr66. Reduction of GFP in solution resulted in loss of

greenish color at low sample concentration (Inouye and Tsuji, 1994b) or formation of white precipitate at higher concentration of GFP and sodium dithionite. Although crystals have been reduced in 20 mM sodium dithionite for two weeks, they were still slightly greenish and should be a mixture of the wild-type and reduced GFP. The relative large R and Rfree in the single-model refinement might be caused by the structural differences between the wild-type and reduced GFP molecules, or multiple configurations of reduced GFP.

It is expected that the reduction occurs at the $C\alpha$-$C\beta$ bond of Tyr66 which returns to a single bond. The $C\alpha$ atom becomes a chiral center again and the resonance system is broken after reduction. The planarity of a chiral center can be estimated by calculating the tetrahedral volume defined by $C\alpha$, $C\beta$, C, and N atoms in each residue as described in section 4.2. In order to apply proper structural restraints to the fluorophore in reduced GFP, the value of the improper energy term for the $C\alpha$ atom of Tyr66 was estimated by obtaining a flat or near-flat fluorophore in the wild-type GFP structure. It was set to 10 in the single-model refinement and 30 in the dual-model refinement.

In the structure refinement of reduced GFP, Tyr66 was initially L-amino acid with tetrahedral volumes of 0.16 and 0.05 Å^3 for the dimer (Figure 4.11a). After a simulated annealing refinement, it was refined to D-amino acid with tetrahedral volumes of 0.22 and 0.23 Å^3. The distances between the $C\alpha$ atoms of Tyr66 in L and D configurations are 0.32 and 0.24 Å for the dimer. Except the different configurations at Tyr66, the two structures were same in other regions. The phenol plane of Tyr66 are still parallel to the imidazolidone plane after reduction in the single-model refinement. It seems that the
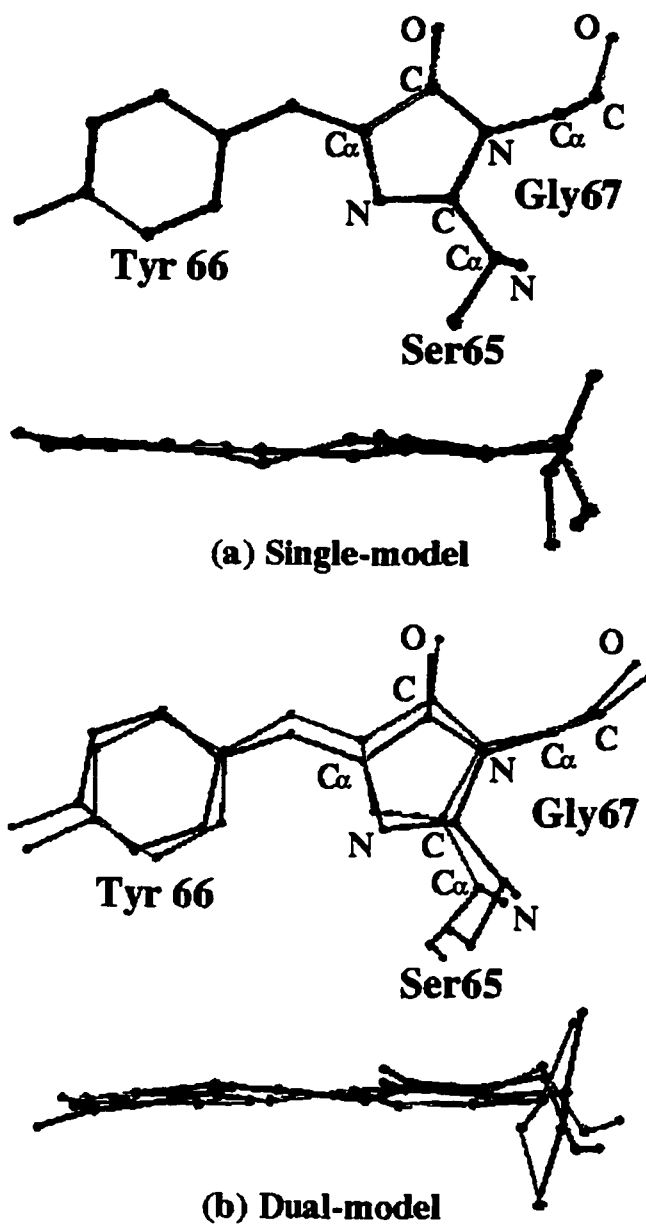
Figure 4.11 The fluorophore in reduced GFP. a) Top view and side view of the reduced fluorophore with L and D configurations in single-model refinements. b) Top view and side view of the reduced fluorophore in dual-model refinements.

reduction only causes a translation of the side chain of Tyr66, but no rotation around its

Cα-Cβ bond or Cβ-Cγ bond. Finally, although the bond length between the carbonyl

carbon atom of Ser65 and amide nitrogen atom of Gly67 was not restrained, it is still about

1.4 Å in the wild-type and reduced structures, indicating a well formed covalent bond. In

another words, the reduction has no detectable influence on the cyclization of the

fluorophore.

Dual-model refinements were used because of the two configurations found for

Tyr66 in reduced GFP structures. The structure of the single-model is nearly at the middle

of the two structures of the dual-model as expected since the single-model should be the

average of the dual-model. The validation of the dual-model refinements is shown by the

decrease in crystallographic Rfree. By replacing the single-model with the dual-model, the

Rfree dropped from 34.7% to 29.4% for reduced GFP, and from 26.2% to 23.7% for the

wild-type GFP (Table 4.7). The crystallographic R factors also decreased with Rfree. The

larger decreases in R and Rfree for reduced GFP than for the wild-type GFP suggests that

molecular conformations are more diverse in the reduced crystal than in the wild-type

crystal, and the dual-model can represent the actual situation better for reduced GFP than

for the wild-type GFP. As expected, Tyr66 was refined to L and D configurations

respectively in the two models, similar to the L and D configurations obtained in the two

single-model refinements (Figure 4.11b). The average tetrahedral volumes of Tyr66 was

0.18 $Å^3$ for the four models in one asymmetric unit, indicating that the Cα atom of Tyr66 is

a chiral center (Table 4.8).

Table 4.7 The R and Rfree of single-model and dual-model refinement of wild-type and reduced GFP.

	R	Rfree	Rdual	Rfreedual
wild-type	21.3%	26.2%	18.2%	23.7%
reduced	22.7%	34.7%	17.7%	29.4%

Table 4.8 The tetrahedral volumes of Tyr66 in wild-type and reduced GFP.

	Volume M1% (Å^3)	Volume M2 (Å^3)	Volume M1' (Å^3)	Volume M2' (Å^3)
wild-type s*	0.04	0.02		
wild-type d	0.08 (L#)	0.04 (L)	0.13 (L)	0.10 (L)
reduced s	0.16 (L)	0.05 (L)		
reduced s	0.22 (D)	0.23 (D)		
reduced d	0.22 (L)	0.17 (L)	0.12 (D)	0.19 (D)

* The letters s and d represent single-model and dual-model refinements.
# The letters L and D represent L and D configurations of amino acids.
% M1 and M2 represent the two monomers in one asymmetric unit.

In summary, the experimental results confirm the theoretical assumption that the reduction occurs at the $C\alpha$-$C\beta$ bond of Tyr66 which breaks the resonance system and eliminates the fluorescence of GFP. A hydrogen atom could be added to the $C\alpha$ atom of Tyr66 from either side of the fluorophore plane resulting in either L or D configurations, indicating that the reduction is not stereospecific.

4.4 The molecular weight and molar absorbance of GFP

4.4.1 Molecular weight of GFP determined by gel filtration column chromatography

With a flow rate of 1 ml/min, the relationship between elution time and molecular

weight for the calibrated column (section 3.7) is:

$$Log(MW) = -0.405\ T + 8.45 \tag{4.1}$$

where T is the elution time using buffer A (Figure 4.12). The differences caused by using

buffer A or B are very small. The elution time of GFP using buffers A, B, and C are 9.8,

9.8, and 8.5 minutes which correspond to apparent molecular weights of 30.5 kDa, 31.3

kDa, and 102 kDa. The calculated molecular weight of GFP monomer according to its

sequence is 26835 Da. Since the ionic strength of buffer A and B is larger than that of

buffer C, it is possible that the wild-type GFP molecules associate with each other at low

ionic strength. However, the interactions between the column material and GFP molecules

may significantly change the elution time of GFP at low ionic strength. Therefore, it is

unreliable to estimate the molecular weight of GFP at low ionic strength using gel filtration

column and the result using buffer C is not accurate.

4.4.2 Molar absorbance of GFP at 397 nm

The linear equation of the amount of protein as a function of absorbance was

established (section 3.7) to be:

$$Y(\mu g) = 88.8\ X\text{-}4.4 \tag{4.2}$$

where Y is the amount of protein in the unit of μg and X is the absorbance (Figure 4.13).
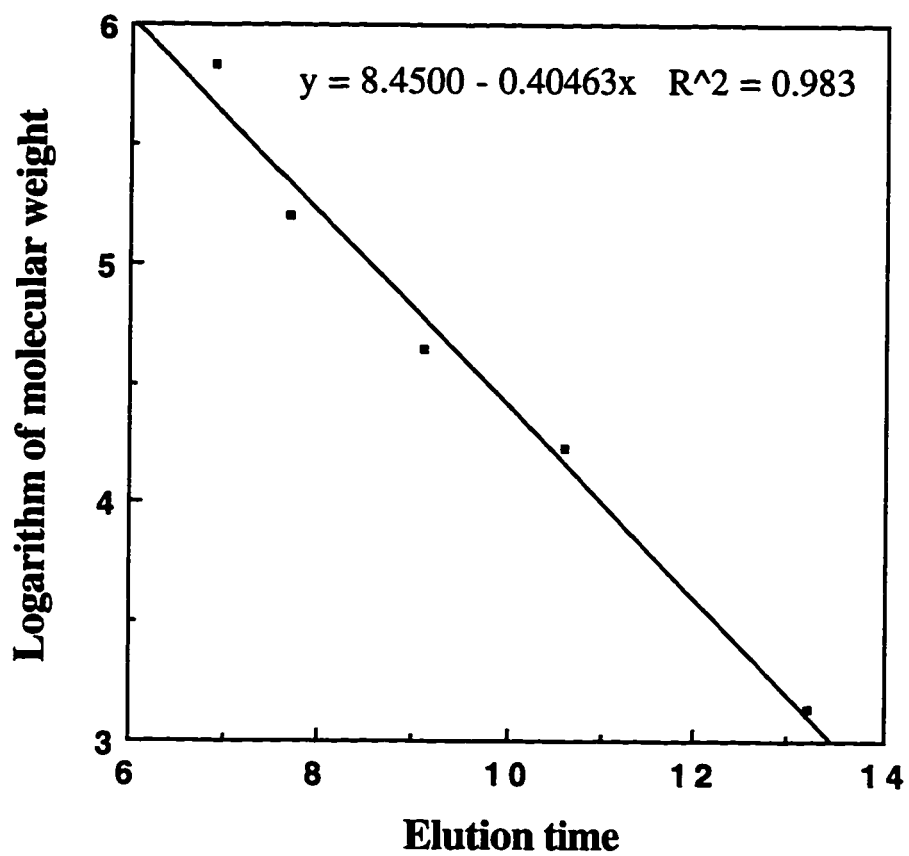
Figure 4.12 Determination of the molecular weight of GFP. The plot shows the elution time of protein standard using buffer A.

For the three measurements using 20 μl, 40 μl, and 60 μl GFP sample, the absorbances were 0.195, 0.328, and 0.451, respectively. These correspond to 0.65 mg/ml, 0.62 mg/ml, and 0.59 mg/ml according to the equation and the average is 0.62 mg/ml. The absorbance of the sample at 397 nm was 0.520 in a cuvette with 1 cm optical path. Thus, the absorbance of GFP at 397 nm is about 0.84 in a 1 cm length cell if the concentration of GFP is 1 mg/ml. This corresponds to a molar absorbance of 22,500 $M^{-1}cm^{-1}$.

Another method to determine the molar absorbance of GFP is to calculate the molar absorbance at 280 nm based on the number of tryptophan and tyrosine residues in GFP. The molar absorbances of tryptophan and tyrosine residues at 280 nm at neutral pH are 5559 $M^{-1}cm^{-1}$ and 1197 $M^{-1}cm^{-1}$, respectively (Handbook of Biochemistry, 1970). Since there are 11 tyrosine and 1 tryptophan residues in GFP, the calculated molar absorbance at 280 nm is about 18726 $M^{-1}cm^{-1}$. Assuming that the ratio of A_{397}/A_{280} is 1.2 for pure GFP sample, the molar absorbance at 397 nm was estimated to be 22471 $M^{-1}cm^{-1}$, close to the experimental result of 22,500 $M^{-1}cm^{-1}$.
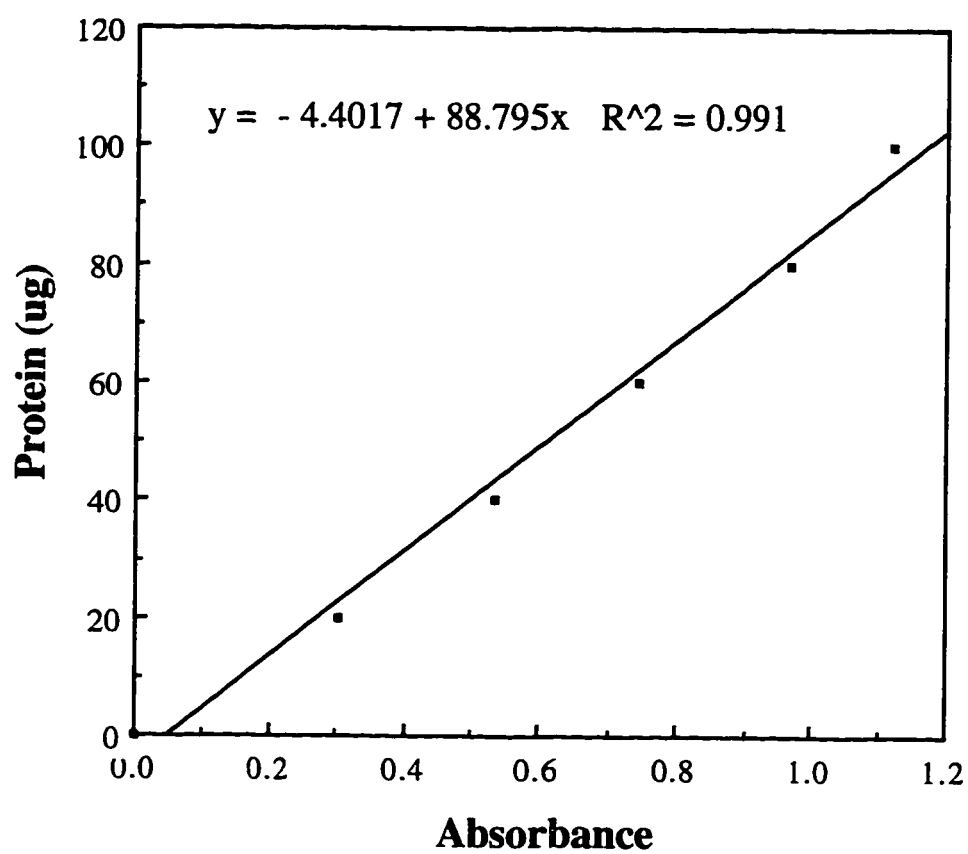
Figure 4.13 Determination of the molar absorbance of GFP. The relationship between amount of protein and its absorbance after chemical processing is plotted.

Chapter 5

Discussion

Being used as a biological marker in a wide range of host cells, GFP is becoming more and more important and the studies about its applications, spectral properties, and fluorescent mechanism is increasing dramatically. The demand for the knowledge of its molecular structure inspired us to work on the crystal structure of the wild-type *Aequorea victoria* GFP. By collecting multiwavelength anomalous dispersion data on a selenomethionyl GFP crystal at four wavelengths and at a cryo-temperature, the high quality MAD data were obtained and processed which led to the solution of the structure (Yang et al., 1996). The structures of GFP mutant S65T and the native *Aequorea victoria* GFP were also solved in other laboratories. Although these structures show high similarity to the structure of the wild-type GFP, several structural differences were observed and will be illustrated. Also, many of the results of mutagenesis and spectroscopy can now be explained by the structure of GFP as discussed in this chapter. Finally, although GFP has been widely used, some of its properties limited its applications in certain conditions.

Some of the future work about improving properties of GFP for various applications will be described.

5.1 Comparison of the structures of wild-type GFP and S65T mutant

The structure of GFP mutant S65T was solved independently and appeared at about the same time as our result (Ormo et al., 1996). The importance of this mutant is due to its altered spectra which have only one excitation maximum at 489 nm and an emission maximum at 511 nm (Heim and Tsien, 1996). Although the changes in emission spectra are small, the single excitation maximum facilitates the detection of its fluorescence from others by selective excitation. Also, the photo bleaching of this mutant is small compared with the wild-type GFP. Another advantage of this mutant is that its excitation wavelength is suitable for using argon laser emission at 488 nm as excitation energy source which will result in intense fluorescence.

The structures of the wild-type GFP and S65T mutant are fitted to each other using main chain atoms from residue 5 to residue 229. The N-terminal and C-terminal residues are excluded from fitting because of either missing coordinates or having large structural differences. The overall root-mean-square (RMS) of distances between pairs of fitted atoms is 0.32 Å. The distances between pairs of Cα atoms of the same residues are plotted as a function of residue number (Figure 5.1). As expected, most of the peaks appear in the regions of loops and structural differences in the regions of β-sheets are small. Some loop

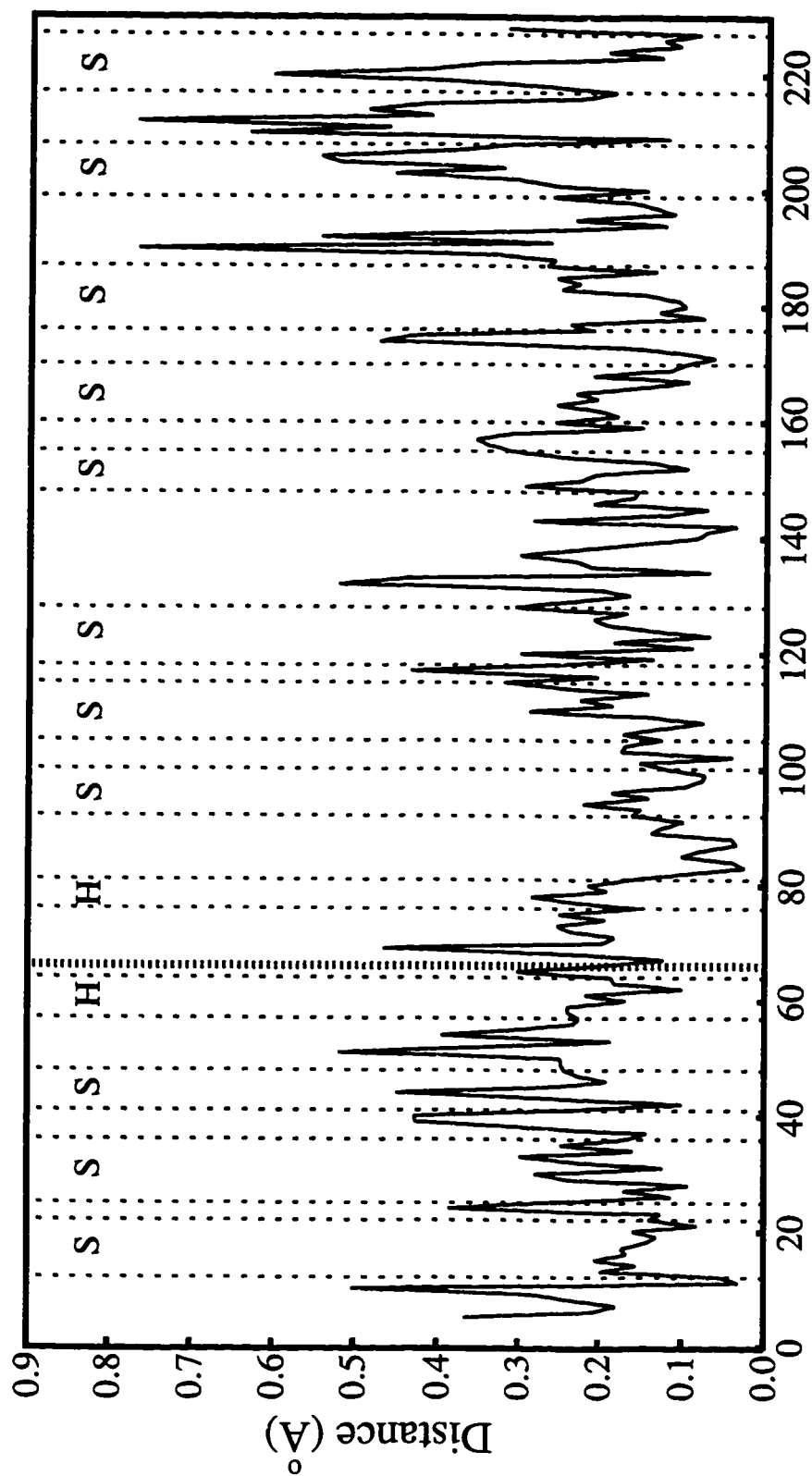Figure 5.1 Comparison of wild-type GFP and S65T mutant. The structures of wild-type GFP

and S65T mutant are fitted to each other using main chain atoms. The distances between pairs of C_a atoms

of the same residues in the two structures are plotted as a function of residue number. The segments of β-

strands (S) and long α-helices (H) are marked. The fluorophore segment is also marked at Tyr66.

regions, which loop away from the β-barrel and where the MAD phased electron densities are weak, have similar conformations in both structures. Thus, even the most flexible regions in GFP are stable in their conformations. In summary, there is no significant global structural differences.

The structural differences around the fluorophore region are small. The $C\alpha$-$C\alpha$ distances in the region are about 0.15 Å to 0.3 Å, close to the average value of the whole protein (Figure 5.1). Despite the mutation at one of the fluorophore residues and the differences in the experimental conditions at which the GFP structures were solved, the conformation of the fluorophore is well conserved which serves as an evidence that the fluorophore is well-protected by the β-barrel and has a relatively invariant environment when the conditions of solution are changed.

One significant structural difference between the wild-type GFP and S65T mutant is at residue 65 (Figure 5.2). In the wild-type GFP, the hydroxyl group of the serine is nearly perpendicular to the fluorophore plane, but in the S65T mutant, it is oriented parallel to the fluorophore plane. The residues that are close to the side chain of Ser65 include Glu222, Leu42, Leu44, Val68, and Leu220. Residue Glu222 was shown to have significant influence on the spectral properties of GFP by mutagenesis (Ehrig et al., 1995). The three leucine residues may also be important to the spectral properties of GFP as they surround the side chain of residue 65. The mutations of these residues may alter the interactions between residue 65 and its environment. Although the side chain of residue 65 is not part of the fluorophore, i.e. the resonance system, its interactions with other residues may change the conformation of the fluorophore and cause changes in spectra.
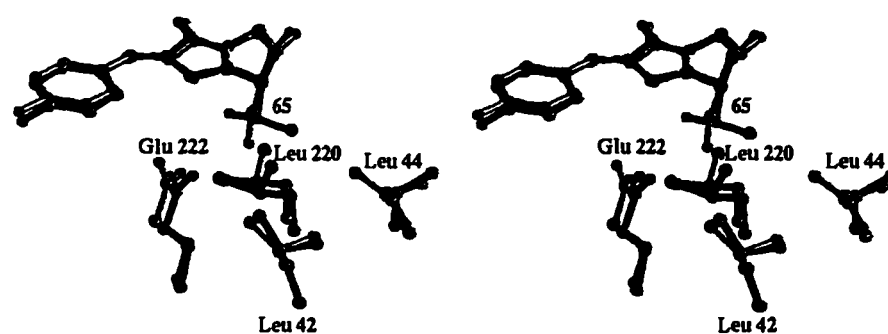
Figure 5.2 Stereo view of the fluorophore and residues near 65 in wild-type GFP and S65T mutant. The structure of S65T mutant (dark) is fitted with the wild-type GFP (light grey). The hydroxyl oxygen atom of serine residue points downward in the wild-type GFP and to the left in S65T mutant.

Although the side chain positions of residues such as histidine, phenylalanine, tyrosine, and tryptophan are very similar in the two structures, there are ambiguities in the conformations of residues such as valine, leucine, threonine, and histidine, such as Val29, Val112, Leu119, Leu194, Thr9, Thr225, and all histidine residues (Figure 5.3). For these residues, the side chains can be rotated 180° and still fit the electron density map well. The distances between the positions of the same atoms in the two structures are about 0.7 Å. Diffraction data with high resolution and high quality are required to solve the ambiguities for residues with a short branched side chain if only one conformation exists in the protein. The electrostatic interactions or hydrogen-binding of the side chain of such residues can also be helpful to determine a possible conformation. The ambiguity of histidine residues can only be solved from their environment instead of collecting high resolution data since diffraction data can not distinguish nitrogen atoms from carbon atoms. The interactions between the side chain of histidine and its environment are the only clue to deduce a possible conformation in X-ray crystallography. Other techniques such as NMR or neutron diffraction may be useful in resolving this ambiguity.

5.2 Comparison of the structures of wild-type and native GFP

Native GFPs purified from jellyfish *Aequorea* and other species have been studied for a long time before recombinant GFP was available in 1992. Although the sequences of the isoforms have not been fully determined yet, at least nine isoforms of native GFP have been identified recently (Cutler, 1995). The spectral properties of these isoforms and the
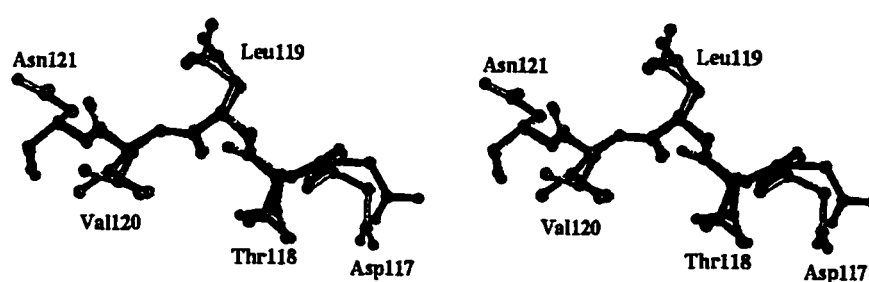
Figure 5.3 Stereo view of some residues with structural ambiguities. Residues in the wild-type GFP (light grey) and S65T mutant (dark) with different conformations are shown here. The four residues, Thr118, Leu119, Val120, and Asn121, have the structural ambiguities that their side chains can be rotated 180° and still fit electron density well.

wild-type GFP are very similar to each other. Crystals of one isoform of native *Aequorea victoria* GFP have been obtained for X-ray crystallography (Perozzo et al., 1988). Using the structure of wild-type GFP as the initial model, the structure of native GFP was solved easily using molecular replacement methods as a collaborative effort.

Similar to comparing the wild-type GFP with S65T mutant, the native and wild-type GFP structures were fitted using residues from 6 to 227. The sequence differences between the two GFPs are: R80Q, F100Y, T108S, L141M, and E172K. Despite these mutations, the RMS of distances between pairs of the same main chain atoms in the two structures is 0.2 Å, smaller than the RMS comparing the wild-type GFP with S65T mutant which is 0.3 Å. All large structural differences appear outside the β-sheet regions and the fitting in the fluorophore region is better than that for the wild-type GFP and the S65T mutant (Figure 5.4 and 5.1). One of the reasons for the larger structural differences between the wild-type GFP and the S65T mutant is that the structures were refined using different conditions such as the refinement program, topology and parameter files used, etc. But the native GFP structure was refined in the same way as the wild-type structure. Since the native structure was derived from the wild-type structure, there is no conformational difference in residues such as valine, leucine, threonine, and histidine. However, this does not mean that there is no ambiguity in their conformations.

Although there is only one GFP monomer in one asymmetric unit in native GFP crystals, it is interesting that the crystal packing of native GFP results in the same protein-protein interactions as the dimer in the wild-type GFP crystals which does not exist in S65T mutant crystals. It is possible that the GFP molecules also dimerize in the same way

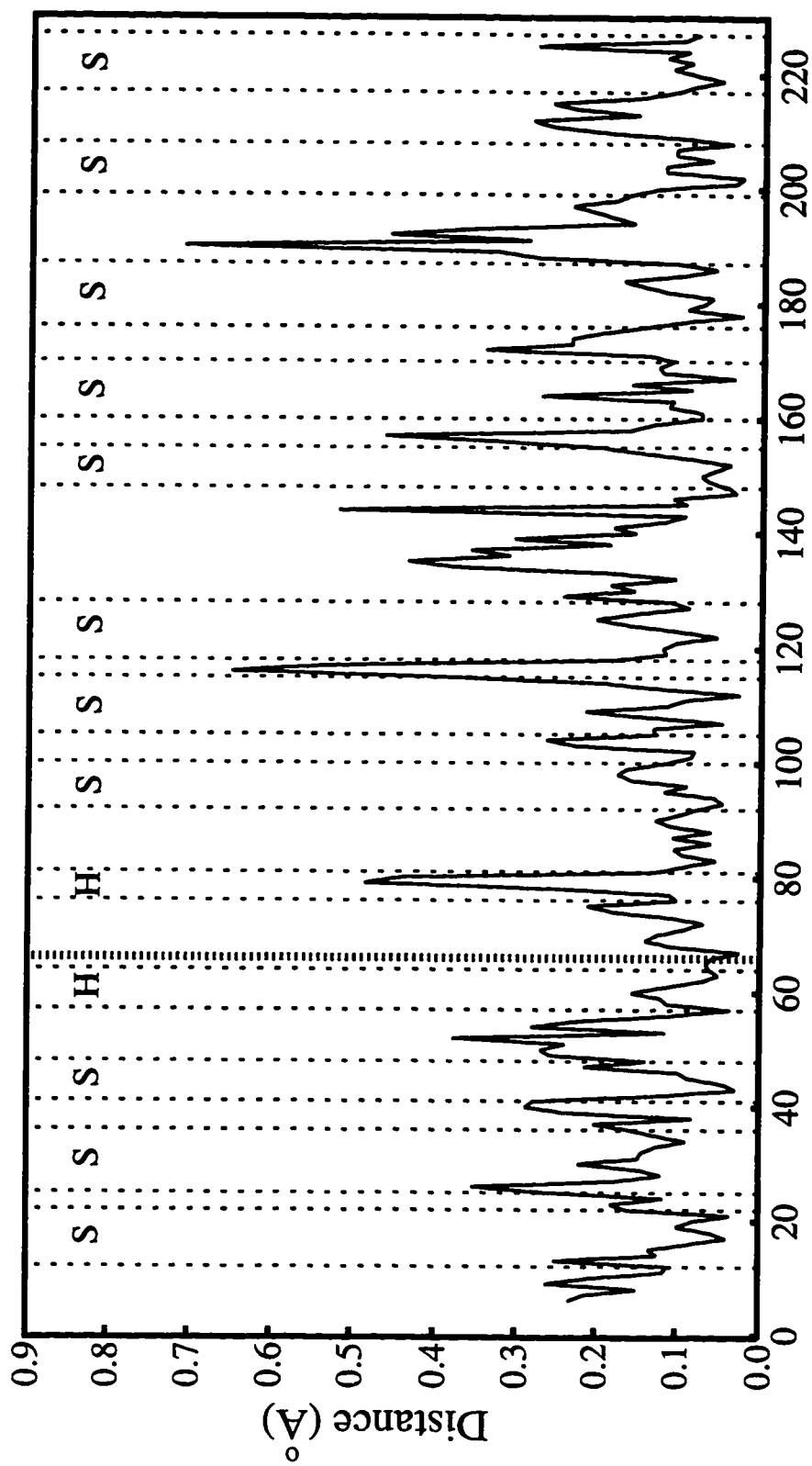Figure 5.4 Comparison of wild-type and native GFP. The structures of wild-type and native

GFPs are fitted to each other using main chain atoms. The distances between pairs of C_a atoms of the

same residues in the two structures are plotted as a function of residue number. The segments of β-

strands (S) and long α-helices (H) are marked. The fluorophore segment is also marked at Tyr66.

in solution as that in crystal. Dimerization may cause problems in some applications of GFP (section 5.4).

On the protein-protein interface, residue Tyr39 has two conformations as observed in the wild-type GFP crystals. The occupancies of conformations A and B are 33% vs. 45%, while they are 53% vs. 33%, and 79% vs. 0% in the wild-type structure. In the structure of the S65T mutant, Tyr39 has only conformation B. It is not known whether this is the only conformation or the major one of two conformations. Note that there is only one GFP molecule in one asymmetric unit and no dimer interactions in S65T mutant crystals. Tyr39 is on the surface of the β-barrel and involved in the dimer interactions in the wild-type and native structure. Since dimerization changes the molar absorbance of GFP so that it is not twice that of monomers, and may influence fluorescence resonance energy transfer between GFP and other photoproteins, this residue may be important to the function and application of GFP.

Small differences in the spectra and the structure of fluorophore of the wild-type and native GFP have been found. The wild-type GFP has slightly higher fluorescence per mole of protein than native GFP, suggesting that the formation of fluorophore is more complete in the wild-type GFP (Heim et al., 1994). In the view of the structure, an oxidized fluorophore should be planar in theory and reduction of the fluorophore will result in either L or D configurations for Tyr66 as shown in the wild-type and reduced GFP structures (Yang et al., 1996; Yang and Phillips, 1997). Comparing the wild-type and native GFP structures, the fluorophore in the wild-type GFP is more planar than that in native GFP if structures are refined using the same conditions, which means the oxidation

of fluorophores in the wild-type GFP molecules is more complete. Thus, the structural and spectral results are consistent with each other. Furthermore, both single-model and dual-model refinements showed that Tyr66 has only the L-amino acid configuration in native GFP. Even if the D-amino acid configuration was used as the initial model for Tyr66, it would be refined to the L configuration. Therefore, the un-oxidized Tyr66 residues in native GFP were not reduced from oxidized tyrosine residues which would have either L or D configurations.

5.3 Structural explanation of properties of GFP

5.3.1 Structural basis of stability, oxidation, and quantum yield of GFP

The remarkable cylindrical fold of the protein seems ideally suited for the function of the protein. The strands of β-sheet are tightly fitted to each other like staves in a barrel, and form a regular pattern of hydrogen bonds. Together with the short α-helices and loops on the ends, the barrel structure forms a single compact domain and does not have obvious clefts for easy access of diffusable ligands to the fluorophore. This fold, taken with the observation that the fluorophore is near the geometric center of the molecule explains the observed protection of the fluorophore from collisional quenching by oxygen ($K_{bm} <$ 0.004 $M^{-1}s^{-1}$) (Nageswara Rao et al., 1980), and increase of the quantum yield. In *Renilla reniformis*, the quantum yield of its luciferase is 5.3% while that of GFP is 30% (Ward and Cormier, 1976). Perhaps more seriously, photochemical damage by the formation of singlet oxygen through intersystem crossing is reduced by the structure. The tightly

constructed β-barrel would appear to serve this role well, as well as provide overall stability and resistance to unfolding by heat and denaturants.

The mechanism of formation of the fluorophore from ordinary protein structure is consistent with a non-enzymatic cyclization mechanism like that of Asn-Gly deamidation (Wright, 1991) followed by oxidation of the tyrosine to dehydrotyrosine, as previously suggested. The role of molecular oxygen in this mechanism and in GFP fluorescence is paradoxical, however. Molecular oxygen is the only cofactor known so far necessary for the formation of the fluorophore in GFP. It is required to oxidize the side chain of Tyr66 to form an extended aromatic system which was proposed to be the second step in the process of fluorophore formation. But oxygen must also be excluded from regular interactions with the fluorophore or else collisional quenching of the fluorescence or damaging photochemistry will occur. The low bimolecular quenching rate suggests that the protein's design sacrifices efficient fluorophore formation for stability and higher quantum yields once fully formed.

5.3.2 Minimal size of GFP required for fluorescence

The minimal size of GFP required for fluorescence has been studied by deleting residues from both ends of GFP (Dopf and Horiagon, 1996). A total of ten truncated GFP mutants were created by the polymerase chain reaction (PCR). Absorption and emission spectra were measured for monitoring formation and fluorescence of the fluorophore in the mutants. It was found that no characteristic absorption or emission of GFP could be detected if residues 2-8 or 226-238 were deleted. Probably the main chain atoms in these

mutants were not cyclized since characteristic absorption spectrum of the wild-type GFP is the result of cyclized fluorophore (Inouye and Tsuji, 1994b). The deletion of residues 233-238 has no influence on the spectra. Although only ten segments were tested in this work and a finer deletion map is necessary for more accurate deletion mapping, it was estimated that amino acid residues from 2 to 232 of the total 238 residues were required for the characteristic emission and absorption spectra of the wild-type GFP.

The N- and C-termini truncation studies (Dopf and Horiagon, 1996) and the fluorescent fusion products (Cubitt et al., 1995; Moores et al., 1996; Olsen et al., 1995) are now understandable given the structure of the protein. At the C-terminus, residues after Thr230 were disordered and could not be determined by experiment. At the N-terminus, Ala1 and Ser2 was only partially disordered and the position of Lys3 was well defined in the structure. Therefore, the amino acid segment that is structurally stable and can be determined by X-ray crystallography is quite consistent with that required for fluorescence of the protein determined by deletion mapping. It is obvious that, those and only those residues, required for the formation of the fluorophore and fluorescence, have well defined positions. The C-terminal residues do not form a stave of the barrel since they loop back outside the cylinder and residues after Thr230 were disordered. Therefore, it should not be critical to have them deleted or further addition of more amino acids as a fusion protein. The role of the N-terminus is a little less clear, as the first strand in the barrel does not begin until amino acid 10 or 11. Thus, barrel formation does not require the N-terminal region. The N-terminal segment, is however, an integral part of the 'cap', on one end of the protein, and may be essential in folding events or in protecting the fluorophore from

accession of solvent molecules. Again, extensions at the N-terminus would not disrupt the motif structure of the protein.

5.3.3 Explanation of mutagenesis results

Residues Ser65 and Glu222   It is obvious that mutation at Ser65 which is part of the fluorophore will cause significant spectral changes. The GFP structure provides the basis for the explanation of the behavior of the Ser65 mutant series. It was shown that Ser65 had close interaction with Glu222 and was surrounded by three leucine residues. Glu222 has been mutated to alanine which caused a single excitation maximum at 481 nm and an emission maximum at 506 nm (Ehrig et al., 1995). S65T has also a single excitation maximum at 490 nm. Based on the structure, it is probably not a coincidence that these two mutants have similar excitation spectra since they interact strongly with each other. Comparing the structures of S65T and wild-type GFP, the angle between the C-O bond in the side chain is larger than $120°$ (Figure 5.2). Therefore, the interaction between Glu222 and Thr66 is quite different from that in the wild-type GFP, and it has a similar effect on the excitation spectrum as that of the mutation of Glu222.

Residue Thr203   Mutation of Thr203 to non-polar residue isoleucine results in a single excitation maximum at 400 nm (Ehrig et al., 1995). Compared to the E222G mutant, which eliminates the excitation maximum at 395 nm in the wild-type GFP spectrum, T203I eliminates the other excitation maximum at 477 nm. In the wild-type GFP structure, Thr203 has close electrostatic and van der Waals' interactions with the phenol group of Tyr66 (Figure 5.5). The S65T structure reveals a new conformation for Thr203
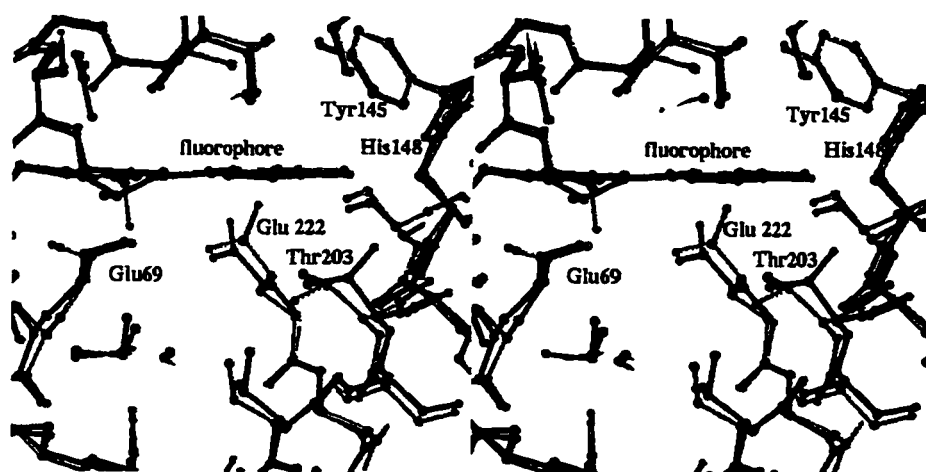
Figure 5.5 Stereo view of the conformation of Thr203. Thr203 has different conformations in the wild-type GFP (light grey) and S65T mutant (dark). The hydroxyl oxygen atom points to the upper right in S65T mutant.

suggesting it may have multiple conformations.

Residues Tyr145, Met153, and Val163   In the three multiple mutants screened

for improving fluorescence of Y66H and Y66W single mutants, mutations occurred at

Ile123, Tyr145, Aln146, His148, Met153, Val163, and Aln212 (Heim and Tsien, 1996).

Residues Tyr145 and His148 have close interactions with the fluorophore (Table 4.5).

Residue Aln146 may influence the fluorophore indirectly through Tyr145 and His148.

Residues Aln212 and Ile123 are remote to the fluorophore and probably have no influence

on GFP spectra. However, the role of Met153 and Val163 is difficult to determine.

Residue Val163 is about 7 Å away from the fluorophore and Met153 is on a loop and even

farther away from the fluorophore. These two mutations were also found in another

independent experiment in which a triple mutant (F99S/M153T/V163A) was found with an

increased whole cell fluorescence signal (Crameri et al., 1996). Residue Phe99 is on the

surface of the β-barrel and should have no contribution to the function of GFP. At present,

the mechanism how residues Met153 and Val163 influence the properties of GFP is not

known.

Residues Gln94, Arg96, and His148   The location of certain amino acid side

chains in the vicinity of the fluorophore begins to explain the fluorescence of the protein.

At least two resonance forms of the fluorophore can be drawn, one with a partial negative

charge on the phenol oxygen of Tyr66, and one with the charge on the carbonyl oxygen of

the imidizolidone ring (Figure 5.6). Interestingly, basic residues appear to form hydrogen

bonds with each of these oxygen atoms, His148 with Tyr66 and Gln94 and Arg96 with the

imidizolidone. The pH dependence of the excitation bands at 395 nm and 475 nm (Ward et
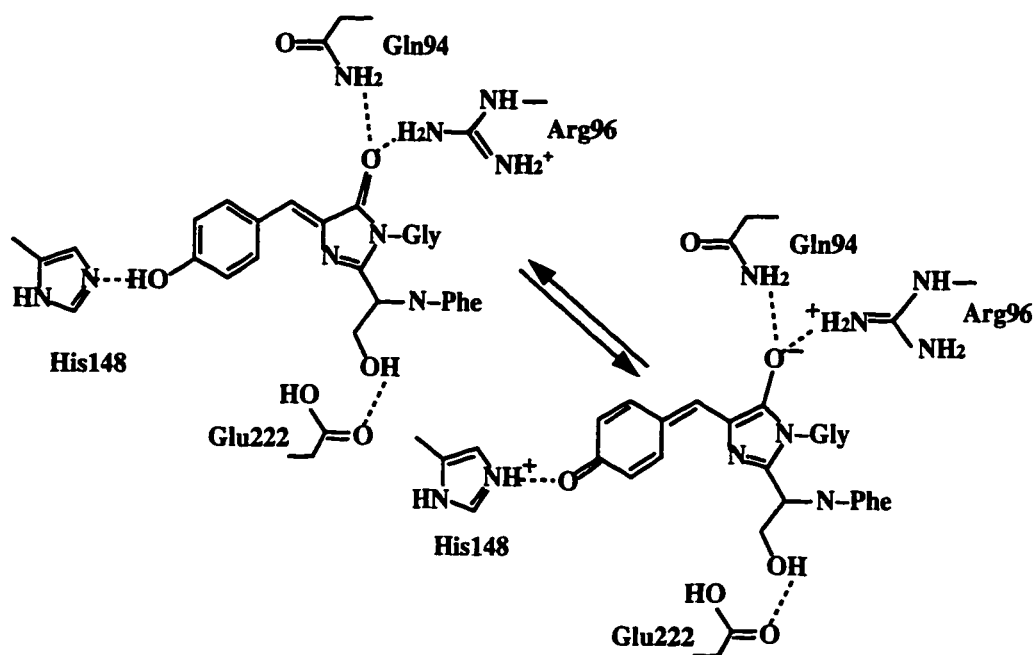
Figure 5.6 Schematic diagram of the resonance forms of the fluorophore with nearby basic amino acids, His148, Gln94, and Arg96 and the acid, Glu222. The bases appear to stabilize anionic oxygen atoms at opposite ends of the fluorophore and the acid forms a hydrogen bond with the hydroxyl of Ser65.

al., 1982) is almost certainly due to His148 whose Nδ atom is 3.3 Å from the Tyr66 hydroxyl oxygen atom of the fluorophore, although NMR pKa measurements or mutagenesis studies are needed for confirmation. These bases presumably act to stabilize and possibly further delocalize the charge on the fluorophore. Most of the other polar residues in the pocket form an apparent hydrogen-bonding network on the side of Tyr66 that requires abstraction of protons in the oxidation process. It is tempting to speculate that these residues help abstract the protons.

5.3.4 Explanation of GFP dimerization

It has been found that the GFP molecules tends to dimerize and the molar absorbance of GFP dimers is not twice as that of monomers. Therefore, the absorption spectra of GFP vary as a function of GFP concentration in solution and there is deviation from Beer's law (Morise et al., 1974; Ward et al., 1982). This result can now be explained based on the structure. Among the residues on the dimer interface and those in the immediate environment of the fluorophore, some are adjacent to each other, i.e. one residue may be involved in dimer interactions, and its neighbor is involved in fluorophore interactions. The formation of dimer may change the conformations of the residues which interact with the fluorophore and change the spectra of GFP. There are at least three such groups which are listed in Table 5.1. Among these three groups, group 1 and 3 are important since residues His148 and Glu222 interact strongly with fluorophore as the structure shows. Phe223 may also be important since it is the neighbor of Glu222 and is the core of the hydrophobic patch in dimer interactions.

Table 5.1 Residues on the dimer interface and in fluorophore environment.

group 1		group 2		group 3	
dimer interaction	fluorophore interaction	dimer interaction	fluorophore interaction	dimer interaction	fluorophore interaction
Asn144	Tyr145	Ser202	Thr203	Leu221	Leu220
Asn146	His148	Gln204	Ser205	Phe223	Glu222
Ser147	Val150	Ala206			
Asn149					
Tyr151					

The dimer in one asymmetric unit in the crystal is likely to be the same one formed in solution, since the ionic strength of the crystallization buffer is low, and we see dimers at low (<100 mM) ionic strengths in solution. Thus, it is not surprising to us to see the large number of hydrophilic dimer contacts. The smaller hydrophobic patch could conceivably be involved in physiological interactions with aequorin, as there would be a natural advantage to close proximity for efficient energy transfer. It is not known at present whether dimers form in physiological circumstances, or what the effect of dimerization is on energy transfer.

On the other hand, contradictory results have been presented in another laboratory. Their results show that *Aequorea* GFP forms reversible dimers by hydrophobic interactions at elevated protein and salt concentrations (Cutler, 1995), while our results show that GFP dimers form at low ionic strength. The preliminary results of analytical ultracentrifugation show that the apparent molecular weight of GFP increases with the decrease of ionic strength which supports our observations by X-ray crystallography. More experiments are required to resolve this issue.

5.4 Future work on GFP

based upon the establishment of its structure

Despite wide applications of GFP in biological research, there are still limitations with the use of wild-type GFP as a biological reporter. Mutagenesis of GFP has been widely studied to improve the properties of GFP for various applications since *Aequorea victoria* GFP has been cloned and characterized (Prasher et al., 1992). However, the research work was restricted before the structure of GFP was solved. Critical residues that interact with the fluorophore can not be identified completely using random mutagenesis. Applications also require new properties of GFP. The solution of the molecular structure of GFP opened a new avenue of engineering of GFP. Some of the future work is discussed below.

New mutants with various spectra properties   For the purpose of monitoring two or more cellular events simultaneously such as the expression of several target genes or the locations of several proteins, GFP mutants with different excitation maxima or emission maxima are required to label different targets. A mutant in a mixture of various GFPs can be monitored by selective excitation and detection of its fluorescence at its emission maximum using optical filters. *E. coli* colonies expressing GFP mutants with different spectral features could be well differentiated, showing the feasibility of detection of multiple GFP mutants in living cells or organisms (Delagrave et al., 1995; Heim and Tsien, 1996).

Most of the residues mutated before the structure was solved are around or at the fluorophore except a few mutants generated by random mutagenesis such as Met153, Val163, Glu222, Thr203, etc. With the knowledge of the structure, more critical residues can be replaced by a selected residues according to the positions of the residues in structure, the purpose of the mutation, etc.

Mutants with increased or reduced dimerization It has been found that *Aequorea* GFP molecules form dimers in certain conditions. The dimerization may also cause aggregation of apo-GFP in protein folding process and results in incorrectly folded GFP which are accumulated in inclusion bodies. The importance of protein-protein interactions of GFP molecules has been realized recently. If GFP is used as part of a fusion protein, the dimerization of the fusion protein caused by the GFP domain may influence the function of the target protein.

With the knowledge of GFP structure, the dimer contacts should now be able to be modified to disrupt the formation of dimers such as the hydrophobic patch without affecting stability and folding. Residues near the hydrophobic patch can also be converted to non-polar residues to enhance dimer stability if desired. Control of the dimerization will be important for fluorescence resonance energy transfer studies of protein-protein interactions using GFP, as one would not want to induce association and hence resonance energy transfer between the differently colored GFP proteins by mechanisms other that of the target protein interactions.

Mutants with short protein folding time In the wild-type GFP, the functional fluorophore is formed several hours after the translation of apo-GFP. The slow post-

translational process limits its application as a real time reporter. It is difficult to monitor two cellular events with a time interval shorter than a few hours. If a GFP mutant can form its fluorophore almost immediately after its translation, time-resolved results can be obtained which is very important in some research such as the study of cell growth. A triple mutant (F99S/M153T/V163A) has been found which is several times faster in maturation than the wild-type GFP (Crameri et al., 1996). With the knowledge of the structure to study the mechanism of protein folding mechanism, mutants with much shorter folding time may be found.

Mutants with protease sites   There are five regions in GFP that loop away from the body of the β-barrel. These regions are relatively flexible compared with other regions in GFP as indicated by electron density. Protease cleavage site may be created in these regions by mutagenesis without affecting the function and stability of GFP. It can be used for at least two purposes. One is to monitor the activity of a protease. The disappearance of GFP fluorescence indicates the existence of the protease. GFP mutants with different colors can carry different protease sites to detect different proteases. The other purpose is to remove the GFP molecules expressed with the first cellular event so that it will not interfere in monitoring a later event.

Mutants for energy transfer   GFP mutants have been made with different spectral properties. If the emission maximum of one mutant is close to the excitation maximum of another mutant, energy transfer from the previous one to the later one is possible as proved when such two mutants were linked by a peptide linker (Heim and Tsien, 1996; Mitra et al., 1996). In those studies, energy transfer between the mutants were used to monitor

protease activity. It can also be used to monitor the association of two target proteins, each is linked with a GFP mutant as energy donor or acceptor. The occurrence of energy transfer indicates the dimerization of the target proteins. However, attention must be paid to eliminate the association among GFP mutants.

There are still other properties of GFP that can be improved. For example, the expression level of GFP as a reporter *in vivo* may be very low since it may be expressed together with a target gene or as part of a fusion protein with low transcription or translation level. Brighter GFP mutants with increased quantum yield, folding efficiency, and molar absorbance are required in such cases. The yield of functional fluorophore from native apo-GFP decreases with the increase of temperature although the fluorophore is very stable after its formation. This temperature requirement limits its applications in animal organisms where the body temperature may be high. Mutants with efficient folding at high temperature will be very useful in such situations. In summary, the solution of the GFP structure accelerates the studies of applications and fluorescent mechanism of GFP.

Conclusion

The three-dimensional structure of GFP has provided a physico-chemical basis for many observed features of the protein, including its stability, protection of it fluorophore, behavior of mutants, and dimerization properties. The structure will also allow directed mutation studies to complement random and combinatorial approaches. Furthermore, with the progress of its studies and applications, GFP may serve as a good model for studying the mechanism of fluorescence and protein folding.

References

Amsterdam, A., Lin, S., Moss, L. G. & Hopkins, N. (1996). Requirements for green fluorescent protein detection in transgenic zebrafish embryos. *Gene* **173**, 99-103.

Anderson, J. M., Charbonneau, H. & Cormier, M. J. (1974). Mechanism of calcium induction of *Renilla* bioluminescence. Involvement of a calcium triggered luciferin-binding protein. *Biochemistry* **13**, 1195-1200.

Anderson, J. M. & Cormier, M. J. (1973). in *"Chemiluminescence and Bioluminescence"*. (M. J. Cormier, D. M. Hercules, J. Lee, eds.). Plenum Press, New York, 387-393.

Baulcombe, D. C., Chapman, S. & Cruz, S. S., (1995). Jellyfish green fluorescent protein as a reporter for virus infections. *Plant J.* **7**, 1045-1053.

Berry, M. (1995). Dissertation: Structure and dynamics of *E. coli* adenylate kinase. Chapter 2.

Bertelson, E. (1951). The ceratioid fishes. Ontogeny, taxonomy, and distribution. *Dana Rep. Carlsberg Found.* **39**, 276.

Bertelson, E. & Struhsaker, P. J. (1977). The ceratioid fishes of the genus Thaumatichthys. Osteology, relationships, distribution, and biology. *Galathea Rep.* **14**, 7-40.

Blundell, T. L. & Johnson, L. N. (1976). *Protein crystallography*. Academic Press.

Bokman, S. H. & Ward, W. W., (1981). Renaturation of *Aequorea* green fluorescent protein: physical separation and characterization of the renatured protein. *Biochem. Biophys. Res. Comm.* **101**, 1372-1380.

Brunger, A. (1992) *X-PLOR Version 3.1: A system for X-ray crystallography and NMR*. New Haven: Yale University Press.

Buck, J. B. (1978). Functions and evolution of bioluminescence, in *"Bioluminescence in Action"* (P. Herring, ed.). Academic Press, 419-460.

Buck, J. & Buck, E. (1968). Mechanism of rhythmic synchronous flashing of fireflies. *Science* **159**, 1319-1327.

Buck, J. & Buck, E. (1978). Toward a functional interpretation of synchronous flashing by fireflies. *Am. Nat.* **112**, 471-492.

Burling, F. T., Weis, W. I., Flaherty, K. M. & Brunger, A. T. (1996). Direct observation of protein solvation and discrete disorder with experimental crystallographic phases. *Science* **271**, 72-77.

Carter, C. W., Jr. & Carter, C. W. (1979). Protein crystallization using incomplete factorial experiments. *J. Biol. Chem.* **254**, 12219-12223.

Case, J. F., Warner, J., Barnes, A. T. & Lowenstine, M. (1977). Bioluminescence of lantern fish (Myctophidae) in response to changes in light intensity. *Nature London* **265**, 179-181.

Casper, S. J. & Holt, C. A. (1996). Expression of the green fluorescent protein-encoding gene from a tobacco mosaic virus-based vector. *Gene* **173**, 69-73.

CCP4: Collaborative Computational Project, N. (1994). The CCP4 suite: Programs for protein crystallography. *Acta Cryst.* **D50**, 760-763.

Chalfie, M., Tu, Y., Euskirchen, G., Ward, W. W. & Prasher, D. C., (1994). Green fluorescent protein as a marker for gene expression. *Science* **263**, 802-805.

Cody, C. W., Prasher, D. C., Westler, W. M., Prendergast, F. G., & Ward, W. W. (1993). Chemical structure of the hexapeptide chromophore of the *Aequorea* green-fluorescent protein. *Biochemistry* **32**, 1212-1218.

Cormack, B. P., Valdivia, R. H. & Falkow, S. (1996). FACS-optimized mutants of the green fluorescent protein (GFP). *Gene* **173**, 33-38.

Cormier, M. J. (1978). Comparative biochemistry of animal systems. in "*Bioluminescence in Action*". (P. Herring, ed.). Academic Press, 75-108.

Crameri, A., Whitehorn, E. A., Tate, E. & Stemmer, W. P. C. (1996). Improved green fluorescent protein by molecular evolution using DNA shuffling. *Nature Biotechnology* **14**, 315-319.

Cubitt, A. B., Heim, R., Adams, S. R., Boyd, A. E., Gross, L. A., & Tsien, R. Y. (1995). Understanding, improving and using green fluorescent proteins. *TIBS* **20**, 448-455.

Cutler, M. W. (1995). Dissertation: Characterization and energy transfer mechanism of the green-fluorescent protein from *Aequorea victoria*.

De Giorgi, F., Brini, M., Bastianutto, C., Marsault, R., Montero, M., Pizzo, P., Rossi, R. & Rizzuto, R. (1996). Targeting aequorin and green fluorescent protein to intercellular organelles. *Gene* **173**, 113-117.

Delagrave, S., Hawtin, R. E., Silva, C. M., Yang, M. M. & Youvan, D. C. (1995). Red-shifted excitation mutants of the green fluorescent protein. *Bio/Technology* **13**, 151-154.

Denton, E. J., Herring, P. J., Widder, E. A., Latz, M. F. & Case, J. F. (1985). The roles of filters in the photophores of oceanic animals and their relation to vision in the oceanic environment. *Proc. Roy. Soc. Lond.*, **B225**, 63-97.

Dopf, J. and Horiagon, T. M. (1996). Deletion mapping of the *Aequorea victoria* green fluorescent protein. *Gene* **173**, 39-44.

Drenth, J. (1994). *Principles of Protein X-ray Crystallography*. Springer-Verlag New York, Inc.

Ehrig, T., O'Kane, D. J., Prendergast, F. G. (1995). Green-fluorescent protein mutants with altered fluorescence excitation spectra. *FEBS Lett.* **367**, 163-166.

Engh, R. A. & Huber, R. (1991). Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallog. Sect A*, **47**, 391-400.

Epel, B. L., Padgett, H. S., Heinlein, M. & Beachy, R. N. (1996). Plant virus movement protein dynamics probed with a GFP-protein fusion. *Gene* **173**, 75-79.

Gluster, J. P., Lewis, M. & Rossi, M. (1994). *Crystal Structure Analysis for Chemists and Biologists*. VCH, New York.

Green, D. W., Ingram, V. M. & Perutz, M. F. (1954). *Proc. Roy. Soc. Lodon*, **A225**, 287.

Handbook of Biochemistry, second edition (1970). The Chemical Rubber Co. page B-75

Hart, R. C., Matthews, J. C., Hori, K. & Cormier, M. J., (1979). Bioluminescence: luciferase catalyzed production of non-radiating excited states from luciferin analogues and elucidation of the excited state species involved in energy transfer to *Renilla* green fluorescent protein. *Biochemistry* **18**, 2204-2210.

Harvey, E. N. (1952). *Bioluminescence*. Academic Press.

Hastings, J. W. (1968). Bioluminescence. *Annu. Rev. Biochem.* **37**, 597-630.

Hastings, J. W. (1978). The chemistry and biology of bacterial light emission. *Photochem. Photobiol.* **27**, 397-404.

Hastings, J. W. (1996). Chemistries and colors of bioluminescent reactions: a review. *Gene* **173**, 5-11.

Hastings, J. W. & Morin, J. G. (1991). Bioluminescence. in *"Neural and integrative animal physiology"*. (C. L. Prosser, ed.). Wiley Interscience, 131-170.

Hastings, J. W. & Nealson, K. H. (1977). Bacterial bioluminescence. *Annu. Rev. Microbiol.* **31**, 549-595.

Heim, R., Cubitt, A. B. & Tsien, R. Y. (1995). Improved green fluorescence. *Nature* **373**, 663-664.

Heim, R., Prasher, D. C. & Tsien, R. Y. (1994). Wavelength mutations and posttranslational autoxidation of green fluorescent protein. *Proc. Natl. Acad. Sci. USA* **91**, 12501-12504.

Heim, R. & Tsien, R. Y. (1996). Engineering green fluorescent protein for improved brightness, longer wavelengths and fluorescence resonance energy transfer. *Curr. Biol.* **6**, 178-182.

Hendrickson, W. A. (1991). Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science* **254**, 51-58.

Hendrickson, W. A., Horton, J. R. & LeMaster D. M. (1990). Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three dimensional structure. *EMBO J.* **9**, 1665-1672.

Hendrickson, W. A., Smith, J. L. & Sheriff, S. (1985). Direct phase determination based on anomalous scattering. *Methods Enzymol.* **115**, 41-55.

Herring, P. J. & Morin, J. G. (1978). Bioluminescence in fishes, in *"Bioluminescence in Action"*. (P. J. Herring, ed.). Academic Press, 273-329.

Hori, K. Charbonneau, H., Hart, R. C. & Cormier, M. J. (1977). Structure of native *Renilla reniformis* luciferin. *Proc. Natl. Acad. Sci. U.S.A.* **74**, 4285-4287.

Inouye, S. & Tsuji, F. I. (1994a). *Aequorea* green fluorescent protein: Expression of the gene and fluorescence characteristics of the recombinant protein. *FEBS Lett.* **341**, 277-280.

Inouye, S. & Tsuji, F. I. (1994b). Evidence for redox forms of the *Aequorea* green fluorescent protein. *FEBS Lett.* **351**, 211-214.

Jancarik, J. & Kim, S. H. (1991). Sparse matrix sampling: a screening method for crystallization of proteins. *J. Appl. Cryst.* **24**, 409-411.

Jones, T., Zou, J., Cowan, S. & Kjeldgaard, M. (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr.* **47**, 110-119.

Kaether, C. & Gerdes, H. (1995). Visualization of protein transport along the secretory pathway using green fluorescent protein. *FEBS Lett.* **369**, 267-271.

Kahana, J. & Silver, P. (1996). in *"Current Protocols in Molecular Biology"*. (F. Ausabel, et al., Eds.). Green and Wiley, 9.7.22-9.7-28.

Kahn, R., Fourme, R., Bosshard, R., Chiadmi, M., Risler, J. L., Dideberg, O. & Wery, J. P. (1985). Crystal structure study of *Opsanus tau* parvalbumin by multiwavelength anomalous diffraction. *FEBS Lett.* **179**, 133-137.

Karle, J. (1989). Linear algebraic analyses of structures with one predominant type of anomalous scatterer. *Acta Cryst.* **A45**, 303-307.

Kleywegt, G.J. & Jones, T.A. (1994). Halloween masks and bones. In *"From First Map to Final Model"*. (S. Bailey, R. Hubbard and D. Waller, eds.). SERC Daresbury Laboratory, Warrington, pp. 59-66.

Krafft, G. A. & Wang, G. T. (1994). Synthetic approaches to conditinuous assays of retroviral proteases. *Methods Enzymol.* **241**, 70-86.

Knight, C. G. (1995). Fluorimetric assays of proteolytic enzymes. *Methods Enzymol.* **248**, 18-34.

Kretsinger, R. H., Moncrief, N. D., Goodman, M. & Czelusniak, J. (1988). The homology of calcium-modulated proteins: Their evolutionary and functional relationships. in *"The calcium channel: Structure, function and implications"*. (M. Morad et al., eds.). Springer-Verlag, New York, 17-34.

Ladd, M. F. C. & Palmer, R. A. (1993). *Structure determination by X-ray crystallography.* Plenum Press.

Lall, A. B., Seliger, H. H., Biggley, W. H. & Lloyd, J. E. (1980). Ecology of colors of firefly bioluminescnce. *Science* **210**, 560-562.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structure. *J. Appl. Cryst.* **26**, 283-291.

Lawry, J. V. (1974). Lantern fish compare downwelling light and bioluminescence. *Nature London* **247**, 155-157.

Leahy, D. J., Erickson, H. P., Aukhil, I., Joshi, P. & Hendrickson, W. A. (1994). Crystallization of a fragment of human fibronectin: introduction of methionine by site-directed mutagenesis to allow phasing via selenomethionine. *Proteins.* **19** 48-54.

Leahy, D. J., Hendrickson, W. A., Aukhil, I. & Erickson, H. P. (1992). Structure of a fibronectin type III domain from Tenascin phased by MAD analysis of the selenomethionyl protein. *Science* **258**, 987-991.

Lloyd, J. E. (1975). Aggressive mimicry in Photuris fireflies: Signal repertoires by *femmes fatales*. *Science* **187**, 452-453.

Lloyd, J. E. (1978). Insect bioluminescence, in *"Bioluminescence in Action"*. (P.J. Herring, ed.). Academic Press, 241-272.

Lloyd, J. E. (1980). Male *Photuris* fireflies mimic sexual signals of their females' prey. *Science* **210**, 669-671.

Marshall, J., Molloy, R., Moss, G. W. J., Howe, J. R. & Hughes, T. E. (1995). The jellyfish green fluorescent protein: a new tool for studying ion channel expression and function. *Neuron* **14**, 211-215.

Matthews, J. C., Hori, K. & Cormier, M. J. (1977). Purification and properties of *Renilla reniformis* luciferase. *Biochemistry* **16**, 85-91.

McCapra, F., Gilfoyle, D. J., Young, D. W., Church, N. J. & Spencer, P. (1994). The chemical origin of colour differences in beetle bioluminescence. in *"Bioluminescence and chemiluminescence"*. (A. K. Campbell, L. J. Kricka & P. E. Stanley, eds.). John Wiley, London, 387-391.

McElroy, W. D., DeLuca, M. & Travis, J. (1967). Molecular uniformity in biological catalysis. *Science* **157**, 150-160.

McElroy, W. D. & Seliger, H. H. (1962). Origin and evolution of bioluminescence, in "*Horizons in Biochemistry*". (M. Kasha and B. Pullman, eds.). Academic Press, 91-102.

McElroy, W. D., Seliger, H. H. & DeLuca, M. (1974). Insect bioluminescence, in "*The physiology of insecta*", Vol.2 (M. rockstein, ed.). Academic Press, 411-460.

McPherson, A. (1990). Current approaches to macromolecular crystallization. *Eur. J. Biochem.* **189**, 1-23.

McPherson, A. (1992). Two approaches to the rapid screening of crystallization conditions. *J. Cryst. Growth* **122**, 161-167.

McRee, D. E. (1993). *Practical protein crystallography*. Academic Press.

Miller, R., Gallo, S. M., Khalak, H. G. & Weeks, C. M. (1994). SnB: Crystal structure determination via Shake-and-Bake. *J. Appl. Cryst.* **27**, 613-621.

Mitra, R. D., Silva, C. M. & Youvan, D. C. (1996). Fluorescence resonance energy transfer between blue-emitting and red-shifted excitation derivatives of the green fluorescent protein. *Gene* **173**, 13-17.

Moores, S., Sabry, J. & Spudich, J. (1996). Myosin dynamics in live *Dictyostelium* cells. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 443-446.

Morin, J. G., Harrington, A., Nealson, K., Krieger, N., Baldwin, T. O. & Hastings, J. W. (1975). Light for all reasons: Versatility in the behavioral repertoire of the flashlight fish. *Science* **190**, 74-76.

Morin, J. G. & Hastings, J. W. (1971a). Biochemistry of the bioluminescence of colonial hydroids and other coelenterates. *J. Cell. Physiol.* **77**, 303-312.

Morin, J. G. & Hastings, J. W. (1971b). Energy transfer in a bioluminescent system. *J. Cell. Physiol.* **77**, 313-318.

Morise, H., Shimomura, O., Johnson, F. H., and Winant, J. (1974). Intermolecular energy transfer in the bioluminescent system of *Aequorea*. *Biochemistry* **13**, 2656-2662.

Nageswara Rao, B. D., Kemple, M. D. & Prendergast, F. G. (1980). Proton nuclear magnetic resonance and fluorescence spectroscopic studies of segmental mobility in aequorin and a green fluorescent protein from *Aequorea forskalea*. *Biophys. J.* **32**, 630-632.

Olsen, K., McIntosh, J. & Olmstead, J. (1995). Analysis of MAP4 function in living cells using green fluorescent protein (GFP) chimeras. *J. Cell Biol.* **130**, 639-650.

Ormo, M., Cubitt, A. B., Kallio, K., Gross, L. A., Tsien, R. Y. & Remington, S. J. (1996). Crystal structure of the *Aequorea victoria* green fluorescent protein. *Science* **273**, 1392-1395.

Otwinowski, Z. (1993). Data collection and processing. in "*Proceedings of the CCP4 study weekend*". Warrington, England: Daresbury Laboratory.

Perozzo. M. A., Ward, K. B., Thompson, R. B., & Ward, W. W. (1988). X-ray diffraction and time-resolved fluorescence analyses of *Aequorea* green fluorescent protein crystals. *J. Biol. Chem.* **263**, 7713-7716.

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

Prasher, D. C., McCann, R. O., Longiaru, M. & Cormier, M. J. (1987). Sequence comparisons of complementary DNAs encoding aequorin isotypes. *Biochemistry* **26**, 1326-1332.

Rhodes, G. (1993). *Crystallography made crystal clear: a guide for users of macromolecular models.* Academic Press.

Rhodes, W. C. & McElroy, W. D. (1958). The synthesis and function of luciferyl-adenylate and oxyluciferyl-adenylate. *J. Biol. Chem.* **233**, 1528-1537.

Rizzuto, R., Brini, M., De Giorgi, F., Rossi, R., Heim, R., Tsien, R. Y. & Pozzan, T. (1996). Double labelling of subcellular structures with organelle-targeted GFP mutants *in vivo. Curr. Biol.* **6**, 183-188.

Roth, A. F. (1985) Purification and protease susceptibility of the green fluorescent protein of *Aequorea aequorea* with a note on *Halistaura*. M.S. thesis, Rutgers University, New Brunswick, NJ.

Sack, J. S. (1988). Chain - A crystallographic modeling program. *J. Mol. Graphics* **6**, 244-245.

SanPietro, R. M., Prendergast, F. G. & Ward, W. W. (1993). Sequence of the chromogenic hexapeptide of *Renilla* green fluorescent protein. *Photochem. Photobiol.* **57**, 63S.

Seliger, H. H. (1975). The origin of bioluminescence. *Photochem. Photobiol.* **21**, 355-361.

Seliger, H. H., Lall, A. B., Lloyd, J. E. & Biggley, W. H. (1982). The colors of firefly bioluminescence-I. Optimization model. *Photochem. Photobiol.* **36**, 673-680.

Seliger, H. H. & McElroy, W. D. (1964). The colors of firefly bioluminescence: Enzyme configuration and species specificity. *Proc. Natl. Acad. Sci. USA* **52**, 75-81.

Sheldrick, G., Dauter, Z., Wilson, K., Hope, H. & Sieker, L. (1993). The application of direct methods and Patterson interpretation to high-resolution native protein data. *Acta Cryst.* **D49**, 18-23.

Shimomura, O. (1979). Structure of the chromophore of *Aequorea* green fluorescent protein. *FEBS lett.* **104**, 220-222.

Shimomura, O., Johnson, F. H. & Saiga, Y. (1962). Extraction, purification and properties of aequorin, a bioluminescent protein from the luminous hydromedusan, *Aequorea. J. Cell. Comp. Physiol.* **59**, 223-240.

Stearns, T. (1995). The green revolution. *Curr. Biol.* **5**, 262-264.

Stura, E. A., Nemerow, G. R. & Wilson, I. A. (1991). Strategies in protein crystallization. *J. Cryst. Growth* **110**, 1-12.

Stura, E. A. & Wilson, I. A. (1990). Analytical and production seeding techniques. *Methods: Companion Methods Enzymol.* **1**, 38-49.

Suzuki, H., Kawarabayasi, Y., Kondo, J., Abe, T., Nishikawa, K., Kimura, S., Hashimoto, T. & Yamamoto, T. (1990). Structure and regulation of rat long-chain acyl-CoA synthetase. *J. Biol. Chem.* **265**, 8681-8685.

Terwilliger, T., Kim, S.-H. & Eisenberg, D. (1987). Generalized method of determining heavy-atom positions using the difference Patterson function. *Acta Cryst.* **A43**, 1-5.

Tsien, R. Y., Bacskai, B. J. & Adams, S. R. (1993). FRET for studying intracellular signalling. *Trends. Cell Biol.* **3**, 242-245.

Vitali, J., Robbins, A. H., Almo, S. C. & Tilton, R. F. (1991). Using xenon as a heavy atom for determining phases in sperm whale metmyoglobin. *J. Appl. Cryst.* **24**, 931-935.

Vriend, G. (1990). WHAT IF: a molecular modelling and drug design program. *J. Mol. Graph.* **8**, 52-56.

Wampler, J. E., Hori, K., Lee, J. & Cormier, M. J. (1971). Structured bioluminescence: two emitters during both the *in vitro* and the *in vivo* bioluminescence of *Renilla. Biochemistry* **10**, 2903-2909.

Wampler, J. E., Karkhanis, Y. D., Hori, K. & Cormier, M. J. (1972). Protein-protein interaction between *Renilla* luciferase and a green fluorescent protein, the energy transfer acceptor during *Renilla* bioluminescence. *Fed. Proc.* **31**, 419.

Wang, S. & Hazelrigg, T. (1994). Implications for bcd mRNA localization from spatial distribution of exu protein in Drosophila oogenesis. *Nature* **369**, 400-403.

Ward, W. W. (1979). Energy transfer process in bioluminescence. *Photochem. Photobiol. Rev.* **4**, 1-57.

Ward, W. W. (1981). Properties of the coelenterate green-fluorescent proteins. in *"Bioluminescence and Chemiluminescence: Basic Chemistry and Analytical Applications"*. M.A. DeLuca and W.D. McElroy (eds.). Academic Press, 235-242.

Ward, W. W (1985). General aspects of bioluminescence. in *"Chemi- and Bioluminescence"*. (John G. Burr, eds.). Marcel Dekker, Inc., New York, 321-358.

Ward, W. W. & Bokman, S. H., (1982). Reversible denaturation of *Aequorea* green fluorescent protein: physical separation and characterization of the renatured protein. *Biochemistry* **21**, 4535-4540.

Ward, W. W., Cody, C. W., Hart, R. C. & Cormier, M. J. (1980). Spectrophotometric identity of the energy transfer chromophores in *Renilla* and *Aequorea* green fluorescent proteins. *Photochem. Photobiol.* **31**, 611-615.

Ward, W. W., Cody, C. W., Prasher, D. C., & Prendergast, F. G. (1989).Sequence and chemical structure of the hexapeptide chromophore of *Aequorea* green-fluorescent protein. *Photochem. Photobiol.* **49**, 62S.

Ward, W. W. & Cormier, M. J., (1976). *In vitro* energy transfer in *Renilla* bioluminescence. *J. Phys. Chem.* **80**, 2289-2291.

Ward, W. W. & Cormier, M. J., (1978). Energy transfer via protein-protein interaction in *Renilla* bioluminescence. *Photochem. Photobiol.* **27**, 389-396.

Ward, W. W. & Cormier, M. J., (1979). An energy transfer protein in coelenterate bioluminescence. Characterization of the *Renilla* green-fluorescent protein. *J. Biol. Chem.* **254**, 781-788.

Ward, W. W., Prentice, H. J., Roth, A. F., Cody, C. W. & Reeves, S. C. (1982). Spectral pertubations of the *Aequorea* green-fluorescent protein. *Photochem. Photobiol.* **35**, 803-808.

Weber, P. C. (1991). Physical principles of protein crystallization. *Adv. Prot. Chem.* **41**, 1-36.

Weis, W. I., Kahn, R., Fourme, R., Drickamer, K. & Hendrickson, W. A. (1991). Structure of the calcium-dependent lectin domain from a rat mannose-binding protein determined by MAD phasing. *Science* **254**, 1608-1615.

Wood, K. V. (1995). The chemical mechanism and evolutionary development of beetle bioluminescence. *Photochem. Photobiol.* **62**, 662-673.

Wood, K. V., Lam, Y. A., Seliger, H. H. & McElroy, W. D. (1989). Complementary DNA coding click beetle luciferases can elicit bioluminescence of different colors. *Science* **244**, 700-702.

Wright, H. (1991). Nonenzymatic deamidation of asparaginyl and glutaminyl residues in proteins. *Crit. Rev. Biochem. Mol. Biol.* **26**, 1-52.

Yang, W., Hendrickson, W. A., Crouch, R. & Satow, Y. (1990). Structure of ribonuclease H phased at 2 Å by MAD analysis of the selenomethionyl protein. *Science* **249**, 1398-1405.

Yang, F., Moss, L. G. & Phillips, G. N., Jr. (1996). The molecular structure of green fluorescent protein. *Nature Biotechnology* **14**, 1246-1251.

Yang, F. & Phillips, G. N., Jr. (1997). The molecular structure of reduced green fluorescent protein. in preparation.

Young, R. E. & Roper, C. F. E. (1977). Intensity regulation of bioluminescence during countershading in living midwater animals. *Fish. Bull. U. S.* **75**, 239-252.