Chapter title: Fluorescent proteins: Crystallisation, structural determination and non-natural amino acid incorporation.

Rochelle D. Ahmed[1], Husam Sabah Auhim[2], Harley L. Worthy[3] and D. Dafydd Jones[1].

1. School of Biosciences, Molecular Biosciences Division, Cardiff University, Sir Martin Evans Building, Cardiff, CF10 3AX, UK.

2. Department of Biology, College of Science, University of Baghdad, Baghdad, Iraq.

3. Biosciences, University of Exeter, Geoffrey Pope Building, Stocker Road, Exeter, UK.

Correspondence:

RDA: AhmedRD@cardiff.ac.uk

HSA: husam.auhim@sc.uobaghdad.edu.iq

HLW: H.Worthy@exeter.ac.uk

DDJ: JonesDD@cardiff.ac.uk

Abstract

Fluorescent proteins have revolutionised cell biology and cell imaging through their use as genetically encoded tags. Structural biology has been pivotal in understanding how their unique fluorescence properties manifest through the formation of the chromophore, and how the spectral properties are tuned through interaction networks. This knowledge has in turn led to the engineering of new variants with new and improved properties. Here we describe the process by which fluorescent protein structures are determined, starting from recombinant protein production through to structure determination by molecular replacement. We also describe how to incorporate and determine the structures of proteins containing non-natural amino acids. Recent advances in protein engineering have led to reprogramming of the genetic code to allow incorporation of new chemistry at designed residues positions, with fluorescent proteins being at the forefront of structural studies in this area. The impact of such new chemistry on protein structure is still limited; the accumulation of more protein structures will undoubtedly improve our understanding and ability to engineer proteins with new chemical functionality.

Key words

Fluorescent protein, recombinant protein production, expanded genetic code, non-natural amino acid, X-ray crystallography, molecular replacement, synchrotron light source, protein structure, protein engineering

Running title: Fluorescent protein structure determination

1. Introduction

Fluorescent proteins have revolutionised our understanding of biological processes as they allow researchers to genetically tag a protein of choice with a detectable probe or respond proactively to changes in cellular events [1–3]. Structural biology has been pivotal to our understanding the function of this unique family of proteins and their subsequent engineering to make them more practical tools for use in cell biology. The first fluorescent protein to be fully characterised both genetically and structurally was the Green Fluorescent Protein (GFP) from *Aequorea victoria* [4]. The isolation of the gene encoding GFP together with its productive expression in other organisms was a turning point in modern biology [5]. As no cofactor (other than oxygen) was required for GFP to become fluorescent, it became clear that to understand GFP the molecular structure needed to be determined, which was released in 1996 [6, 7]. There are now over 400 structures of *A. victoria* GFP and its engineered relatives present in the PDB. The importance of GFP as a tool (and its limitations) led a race to find other sources, with red versions in coral anemones [8, 9], orange versions from sea anemone [10] amongst many others now in use (see FPbase www.fpbase.org [11]).

The GFP structure is archetypical of fluorescent protein (FPs) that do not require an additional cofactor. They are comprised of an 11 stranded β-barrel structure with the chromophore present in the centre of the barrel (Figure 1). A helical region comprising residues 55-73 runs through the middle of the barrel, interrupted by the chromophore. The residues that comprise the chromophore are 3 contiguous residues with the motif 65-Xaa-Tyr-Gly-67 (where Xaa can be various amino acids in both natural and engineered versions). While residue 65 (GFP numbering) can be a variety of different residues (*e.g.,* Met in DsRed and Ser in GFP), Gly67 is strictly conserved. In natural FPs, residue 66 is Tyr but can be engineered to contain Trp (*e.g.,* cyan FP [13]) and His (*e.g.*, blue FP [13]) and even non-natural amino acids (*e.g.,* azido-phenylalanine [14]) without hindering maturation. The maturation process involves first cyclisation through nucleophilic attack of the amine of Gly67 on the carbonyl carbon of residue 65 followed by dehydration/O_2-dependent oxidation to

generate the final chromophore shown in Figure 1b-c; the two flat planer ring systems (P and I) are linked by a methylene bridge so bringing the whole system into conjugation that gives GFP its unique fluorescence properties [15]. An additional oxidation event occurs in red FPs that extends the conjugate double bond system [9]. Structural biology has proved pivotal in understanding the mechanism of maturation through the isolation of various intermediates [16–20]. While the chromophore is critical to FP function, the environment around the chromophore can fine tune the fluorescent properties. For example, replacing Thr203 with Tyr converts GFP to a yellow FP [6] while mutation to His instils photo-responsive properties [21]. The chromophore and the core as a whole, is also heavy solvated, with water molecules contributing to H-bond networks that dictate the charge of the ground state chromophore [22–24]. It is important not ignore internal water molecules during structure determination and analysis. In wild-type GFP, the ground state form of the chromophore has the phenolic OH group protonated [neutral] giving rise to a dominant λ_{max} of ~400 nm. On excitation, the phenolic OH deprotonates to generate an anionic chromophore that emits at ~510 nm. The proton ejected from the chromophore is transferred to a local network of residues and water molecules. Mutations that alter this network (*e.g.,* S65T) results in the anionic chromophore dominating the ground state. Glu222 was considered pivotal to controlling the chromophore charged state and high-resolution structures of EGFP have shown how different conformations of Glu222 can dictate the potential H-bond pathways available that in turn dictate the ground state charge [12, 25].

Structural biology of FPs is still an active area and is especially critical to evaluating and understanding protein design/directed evolution efforts. At time of publication, the most recent GFP-related protein release was the structure of a cAMP sensor G-Flamp1 (PDB code 6M63) on the 22nd Sept 2021. While the advent of predictive approaches such as AlphaFold [26] and RoseTTAfold [27] will undoubtedly revolutionise the design process, currently they cannot generate the chromophore of FPs – the most critical part of the protein (Figure 2a). Thus, crystallisation of FPs will still be required in the near future. Furthermore, current

predictive tools cannot easily work with non-natural amino acids as there is little structural information to act as a reference point. Indeed, there are still relatively few protein structures that contain non-natural amino acids in the PDB (Figure 2b). This in turn limits our understanding of how new chemistry can impact the structure and function of a protein.

The incorporation of selenomethionine into proteins by auxotrophic replacement of methionine was a critical development in protein structure determination by X-ray crystallography [28, 29]. Genetic code reprogramming goes beyond classical auxotrophic methods by allowing incorporation of non-natural (or non-canonical) amino acids of choice at defined residues during cellular protein synthesis [30, 31]; we now have precise control of the protein and residue that a non-natural amino acid (nAA) is incorporated into. In brief, a codon (normally the amber UAG stop codon) is reprogrammed to incorporate nAA using an engineered tRNA/aminoacyl tRNA synthase pair from a different organism. Non-natural analogues of tyrosine and pyrrolysine (a natural but very rarely used amino acid) are the two most used systems. New chemistry means new functionality available to a protein, ranging from non-native post-translational modification to new enzyme mechanisms to optogenetic control. One of the first structures of a protein containing a nAA introduced by genetic code expansion were variants of GFP [14, 32] and fluorescent proteins have since acted as excellent model systems for understanding the impact of engineered nAA changes to the structure-function relationship.

In this methods chapter, we will describe how to determine the structure of fluorescent proteins, including those containing non-natural amino acids. The focus will be on versions of GFP but many of the methods will be relevant to other FPs. Figure 3 below provides an overview of the workflow for determining the structure of a fluorescent protein.

2.Materials and equipment

2.1 *E. coli* transformation and protein expression.

The normal route for production of fluorescent proteins for detailed biophysical analysis is recombinant expression in *E. coli*. The fluorescent protein (FP) is normally in an inducible, high expression plasmid system such as arabinose inducible pBAD or IPTG inducible pET vectors. With regards to the nAA incorporation, we have given an example of incorporation of *p*-azido-L-phenylalanine, but similar approaches will apply for different nAAs. All media were sterilised by autoclaving at 121°C before use. Any components added to medium were prepared using ultra-pure water and filter sterilised using a 0.22 µM syringe filter unit.

1. Electrocompetent *E. coli* strains (e.g., DH5α (NEB), Top 10^{TM} (Invitrogen) BL21 DE3(NEB)). *See* **Note 1.**
2. Optional. If incorporating nAA, plasmid vectors encoding a mutant fluorescent protein and/or an engineered orthogonal tRNA and tRNA synthetase (*e.g.*, pDULE [33], pEVOL [34]) for incorporation of nAA derivatives in bacterial cells.
3. Super optimal condition (SOC) medium: 2% tryptone, 0.5% yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 10 mM $MgSO_4$, 20 mM glucose.
4. LB medium: 1% tryptone, 0.5% yeast extract and 0.5 % NaCl.
5. LB Agar: 1% tryptone, 0.5% yeast extract, 1.5% agar and 1% NaCl ,.
6. Auto-induction medium: 1% tryptone, 0.5% yeast extract, 0.5% (v/v) glycerol, 0.05 % glucose, 0.2% lactose, 25 mM Na_2HPO_4, 25 mM KH_2PO_4, 50 mM NH_4Cl, 5 mM $NaSO_4$,2 mM $MgSO_4$, 1 x trace metals (4 µM $CaCl_2$, 2 µM $MnCl_2$, 2 µM $ZnSO_4$, 0.4 µM $CoCl_2$, 0.4 µM $CuCl_2$, µM 0.4 $NiCl_2$, 0.4 µM Na_2MoO_4, 0.4 µM H_3BO_3 and 10 µM $FeCl_3$ in ultra-pure water) and 0.05 % L-arabinose.
7. 100 µg/mL Ampicillin.
8. 25 µg/mL Tetracycline.

9. Optional. Non-natural amino acids (nAAs) such as *p*-azido-L-phenylalanine (Cambridge Bioscience, Cambridge, UK).
10. Electroporation system.
11. Electroporation cuvette.
12. Micropipettes (10 -1000 µL)
13. Sterile pipette tips.
14. Microcentrifuge tubes (sterile, 1.5 mL).
15. Plastic petri dishes.
16. Glass shaker culture flasks (100 mL and 2.5 L).
17. Shaker incubator.

2.2 Protein extraction and purification.

Most of the FPs we work with have a His-tag sequence at either the N- or C-terminal so can be purified by metal affinity chromatography followed by a size exclusion chromatography polishing step. We normally attach the metal affinity columns to a protein chromatography system for imidazole gradient elution, but purification can be achieved using simple wash and elute steps with gravity flow.

1. Lysis buffer: 50 mM Tris-HCl pH 8.0, protease inhibitor (1 tablet per 50 mL)(cOmplete™ EDTA-free Protease Inhibitor Cocktail) *See* **Note 2**.
2. Equilibration buffer: 50 mM Tris-HCl pH 8.0.
3. Elution buffer: 50 mM Tris-HCl pH 8.0, 500 mM imidazole.
4. Cooling centrifuge.
5. Microcentrifuge (refrigerated preferable but not essential).
6. French pressure cell press or other cell disruption system.
7. ÄKTA Prime Plus or ÄKTA Purifier FPLC systems or any available chromatography system.
8. His Trap™ HP column (binding capacity ~200 mg protein, 5ml, GE Healthcare).

9. Protino^{R} Ni-TED 2000 Packed Columns, 1 ml bed volume with 5 mg protein binding capacity (Machery-Nagel, Germany).
10. Size exclusion columns
    a. Hiload^{TM} 16/600 Superdex™ s75 pg (preparative grade, 120 mL bed volume, separation range 3 kDa to 75 kDa, (GE Healthcare)).
    b. Hiload^{TM} 16/600 Superdex™ s200 pg (preparative grade, 120 mL bed volume, separation range 10 kDa to 600 kDa, (GE Healthcare)).
11. 10 kDa molecular weight cut-off spin filters (500 µL and15 mL capacity).

2.3 Protein analysis.

While sample colour should be indicative of functional protein, it is advised to undertake spectroscopic analysis of samples. Access to a plate reader capable of absorbance and fluorescence or a UV-Vis spectrophotometer and fluorimeter is advised.

1. 4x SDS-PAGE sample buffer: 40% Glycerol, 240 mM Tris-HCl pH 6.8, 8% SDS, 0.04% bromophenol blue, 5% beta-mercaptoethanol.
2. SDS-PAGE staining solution: 50% ultra-pure water (v/v), 40% (v/v) methanol, 10% (v/v) acetic acid, 0.1% (w/v) Coomassie blue R250.
3. SDS-PAGE destain solution: 20% (v/v) methanol, 10% (v/v) acetic acid, 70% ultra-pure water (v/v).
4. Prestained protein marker (Broad range 245 -5 kDa).
5. TruPAGE^{TM} Precast (4-20%) gels and 20x running buffer (Sigma).
6. Mini-PROTEIN electrophoresis system.
7. Bio-Rad DC protein assay reagents.
8. A 96 well plate (Nunc, Thermo-Fisher).
9. Microplate reader (e.g. CLARIOstar, BMG LABTECH, UK).
10. Optional. Spectrophotometer (Cary UV-Vis Spectrophotometer, Agilent Technologies).
11. Optional. Fluorimeter (Cary Eclipse Fluorescence Spectrometer, Agilent Technologies).

2.4 Crystallography.

A liquid handling robot will make the process quicker and more reliable, but it is not necessary, especially if crystallisation conditions are known.

1. Concentrated protein sample (~ 20 mg/l).
2. Micropipettes (10 µl) and sterile pipette tips.
3. Swissci Triple Drop Plate, UVP (Molecular Dimensions, Suffolk, UK)
4. Crystallisation screens *e.g.*, PACT *Premier*™ HT-96 and JCSG-plus™ HT-96 broad crystallisation screen (Molecular Dimensions, Suffolk, UK).
5. ClearVue™ sealing sheet (Molecular Dimensions, Suffolk, UK)
6. Automated liquid handling systems Phoenix (Alpha Biotech, UK).
7. Light microscope.
8. Mounted litholoop (Molecular Dimensions).
9. Liquid nitrogen.
10. X-ray source - either synchrotron *e.g.,* Diamond Light Source or in house X-ray generator and detector.
11. The CCP4 package software (Potterton *et al*. 2003) for structure determination.
12. PyMOL programme (Delano 2002) or similar for visualisation and graphical representations of protein structure. (*See* **Note 3**).

3.Methods

3.1 Transformation and protein expression.

Transformation of plasmid DNA into *E. coli* was performed by electroporation. The heat shock method can alternatively be used if you do not have access to an electroporator.

3.1.1 Transformation

1. Warm SOC medium to room temperature.

2. Thaw cells on ice.
3. Add 1-2 µL of plasmid DNA (10 pg - 100 ng) into cells and gently mix.
4. Transfer cells into electroporation cuvette.
5. Electroporate cells with a short pulse of 2500 V (*see* **Note 4**).
6. Aseptically add 200 µL of pre-warmed SOC medium to the cuvette (*see* **Note 5**).
7. Transfer the solution to a 1.5 mL Eppendorf and incubate at 37 °C for 1 hr with shaking (200 RPM).
8. Prepare petri dish with LB agar supplemented with appropriate antibiotics for plasmid selection (*see* **Note 6**).
9. After 1 hr, spread 10 - 150 µL of cells on to the LB agar plates and leave to dry for 5 min. (*see* **Note 7**).
10. Invert the agar plates and incubate at 37 °C overnight.
11. Select colonies and confirm the presence of target plasmid by mini-expression cultures or visualising plate under UV light if inducer is added.

3.1.2 Protein expression

The production of your FP should be obvious, and it turns the cells a distinctive colour (e.g., see Figure 4a). If not, then the protein yield of mature FPs is unlikely to yield enough protein for crystallisation. For example, we achieve yields of up to 50-80 mg/L for native proteins and 30-50 mg/L of culture medium for nAA containing FPs.

1. Select a colony and add to 5 mL LB medium supplemented with the appropriate antibiotic for your chosen vector.
2. Incubate starter culture overnight with shaking (200 RPM) at 37 °C
3. Add appropriate antibiotic, auto induction medium to 2 L culture flask and make up to a 1 L final volume with LB medium (*See* **Note 8**)

4. Add the starter culture to the 1 L solution and incubate overnight at the optimum growing temperature (*see* **Note 9**).

3.3 Protein extraction & purification

The presence of the FP in a fraction should be very clear from the colour of the solution (see Figure 4b as an example). This is useful if purification is done manually using a wash-then-elute procedure. If using a protein chromatography system, UV wavelength of 280 nm can be used to detect total protein. For FP's, additional wavelengths can be monitored corresponding to the protein's absorbance maxima (*e.g.*, 485 nm for WT GFP). If UV light can affect the nAA, (*i.e.,* if using photo-reactive nAAs such as azF) absorbance monitoring should be switched off at 280 nm, but longer wavelengths (i*.e*., the excitation maxima of the FP) can still be used.

3.3.1 Cell lysis by French Press

1. Harvest cells by centrifugation at 5000 *xg* for 30 mins at 4 °C. (*see* **Note 10**).
2. Resuspend cells in 50 mL of lysis buffer.
3. Use French pressure cell press to lyse cells at 1250 psi of pressure (*see* **Note 11**).
4. Centrifuge the lysed cells at 25,000 *xg* for 45 mins (*see* **Note 12**).
5. Remove clarified cell lysate from the insoluble fraction (*see* **Note 13**).

3.3.2 Immobilised metal affinity chromatography (IMAC)

We use ÄKTA FPLC as it is useful to monitor UV wavelengths. However, ProtinoR Ni-TED 2000 Packed Columns can also be used according to the manufacturer's procedures.

1. Assemble the His Trap™ HP column onto the ÄKTA FPLC.
2. Equilibrate the column with 10 column volumes of equilibration buffer.
3. Load clarified cell lysate onto the column at a flow rate of 1 ml/min.

4. Continue to wash with 5 column volumes of equilibration buffer or until the 280 nm absorbance from the FPLC absorbance detector returns to the baseline (*see* **Note 14**).
5. Use the elution buffer to perform an imidazole gradient (increasing the imidazole concentration to 500 mM) (*see* **Note 15**).
6. Collect fractions of protein according to a rise of absorbance on the detector (*see* **Note 16**).
7. Analyse collected fractions by SDS-PAGE and pool those with the greatest purity for size exclusion chromatography (SEC).

3.3.3 Size exclusion chromatography (SEC)

1. Assemble the size exclusion (SE) column onto the ÄKTA FPLC. For monomeric FPs, a HiLoad Superdex s75 16/60 GL (separation range 3-75 kDa) column is used; for oligomeric FPs a HiLoad Superdex s200 16/60 GL (separation range 10-600 kDa) column is used. Equilibrate the column with equilibration buffer (s*ee* **Note 17).**
2. Pool the purest protein fractions as determined by SDS-PAGE from 3.3.2 step 7, and concentrate using a molecular weight cut-off spin filter (MW cut-off 10 kDa) and centrifugation at 4000 x*g* until a maximum volume of 2 mL is reached. (*See* **Note 18**).
3. Load the protein onto the column with a maximum flow rate of 1 mL/min.
4. Collect fractions of protein according to a rise in UV absorbance.
5. Stop collecting fractions when the UV absorbance returns to the baseline.
6. Analyse purity of protein fractions by SDS-PAGE.

3.4 Spectral analysis of purified FPs

3.4.1 UV-visible absorption spectroscopy and calculation of extinction coefficients

Fluorescent proteins absorbance spectra (UV-visible) were recorded using a Cary spectrophotometer in 1 cm pathlength cuvettes (Helma, Müllhein, Germany).

1. Record the absorbance spectrum from 200-800 nm using a maximum of 500 µl of protein sample.
2. Record the absorbance value at λ_{Max} of the fluorescent protein (maximum excitation wavelength (λ_{ex})).
3. Calculate the extinction coefficients of fluorescent proteins by recording absorbance spectra, as above, of protein samples with known concentrations (10-20 µM).
4. Substitute the values of concentration and absorbance at λ_{Max} into the Beer-Lambert law equation (Equation 1).

Equation 1 $A = \varepsilon.c.l$

Here, A is the absorbance value at λ_{Max}, ε is the extinction coefficient (M 1cm 1), c is the concentration (Molar), and l is the pathlength of cuvettes (cm).

3.4.2 Fluorescence spectroscopy

Excitation and emission spectra of fluorescent proteins were recorded on a Cary Eclipse Fluorimeter (Agilent Technologies) using 5 mm x 5 mm QS quartz cuvette.

1. Use a maximum of 400 µl (0.5 - 2 µM) of protein sample for the measurement.
2. Record the spectra at a fixed scan rate of 120 nm/min with a 5 nm slit width and medium scan control.
3. Determine the maximum excitation wavelength (λ_{ex}) for the fluorescent protein from its absorbance spectrum as described in section 3.4.1.
4. Record the emission spectrum (λ_{em}) up to 650 nm using the maximum excitation wavelength (λ_{ex}) determined in section 3.4.2 as a fixed excitation wavelength.

5. Record the excitation spectra by monitoring excitation at a fixed maximum emission wavelength (λ_{em}) for each fluorescent protein over a range of wavelengths down to 350 nm.

3.5 Crystallography

3.5.1 Crystal formation and harvesting

To grow protein crystals, numerous techniques can be used including, sitting, and hanging drop vapour diffusion, and micro-batch. It is advised to try multiple techniques when crystalising protein. In general, all techniques follow the following steps of concentrating protein, incubation, selection of cryoprotectant and harvesting. The key differences are in the type of plates used and how the plates are sealed (either with airtight film or oil). However, we have found the best results using sitting drop vapour diffusion and will describe this process below.

1. Concentrate the purified protein using a molecular weight cut-off spin filter (MW cut-off 10 kDa) and centrifugation at 4000 x*g* until a concentration between 5-20 mg/mL is reached. *(See* **Note 19**).
2. Remove any precipitated protein from the concentrated sample with a final centrifugation step at maximum speed for 5 mins in a microcentrifuge.
3. Determine final protein concentration using preferred method (*see* **Note 20**).
4. Set up Swissci Triple Drop Plate with crystallisation screen using either a liquid handling robot or by hand (*see* **Note 21**).
5. Set up drops with equal volumes of protein and crystallisation buffer (0.2 µL each) (*see* **Note 22**).
6. Seal crystallisation plates with ClearVue™ sheet and store at 25 °C. If using a nAA that is photosensitive, store the plates in the dark (see **Note 23**).

7. Take images of crystals at day 0, 1, 3, 7, and then once a week to monitor crystal growth.
8. Harvest any grown crystals using appropriately sized loop diameter (Fig. 5). *See* **Note 24**.
9. Soak crystals briefly (10-60s) using an appropriate cryoprotectant before plunge freezing into liquid nitrogen. *See* **Note 25**.
10. Store loops in Dewar flask containing liquid nitrogen, until ready to send to the synchrotron.
11. Send crystals in Dewar flask to Diamond Light Source (Harwell, UK). *See* **Note 26**.

3.5.2 Data acquisition and refinement

Both the FP chromophore and the nAA are non-standard protein components. See section 3.5.3 below about how to deal with these components during the structure determination process. Given the common overall structure of FPs and the presence of wild-type proteins in the PDB, the most common process for phasing data is via molecular replacement, which we will focus on here. Please note that if using a synchrotron, steps 2-4 are done automatically during data collection, using in house pipelines (*see* **Note 27**).

1. Perform X-ray diffraction experiments at synchrotron (e.g., Diamond Light Source, Didcot or DESY, EMBL, Hamburg) at 100 K using X-rays of wavelength 0.9795 Å. Detect diffracted X-rays using a Pilatus 6M-F detector or similar. *See* **Note 28.**
2. Integrate and reduce data using the XIA2 package [35]. *See* **Note 29**.
3. Determine space group and unit cell dimensions with POINTLESS [36]. *See* **Note 30.**
4. Use SCALA [36] and TRUNCATE [37] to scale and merge the data, respectively.
5. Determine cell contents using Mathew's test and mature peptide sequence. *See* **Note 31.**

6. Solve phases using molecular replacement with PHASER [38] in basic mode or similar refinement program using a homologous structure *e.g.,* WT GFP (PDB ID: 1EMA), to generate an initial model. *See* **Note 32.**
7. Improve phases using a density modification program such as ACORN [39–42]. *See* **Note 33.**
8. Using COOT [43], manipulate starting model to improve density fit by:
    a. Correcting Ramachandran angles
    b. Adding/removing residues and side chains to fill empty density/lack of density respectively
    c. Adding waters, other solvent molecules, and ligands
    d. Correcting issues highlighted by MolProbity
9. Refine phases using REFMAC [44] to generate updated refinement statistics (R, R_{Free}, and RMSDs of bonds and angles) and updated MolProbity report. There are multiple parameters you can adjust but we find running REFMAC5 with the additional settings to work well in most cases:
    a. Automatic Translation-Liberation-Screw (TLS) parameters
    b. Non-crystallographic symmetry (NCS restraints)
    c. Execute multiple run refinement runs to optimise X-ray/geometry weight
10. Repeat 8 and 9 until R and R_{Free} no longer improve (*see* **Note 34**).
11. Run final structure and maps through PDB deposition validation server (can be done via CCP4i2) (*see* **Note 35**).
12. Deposit structure onto wwPDB. *See* **Note 36**.

3.5.3 Inserting non-standard residues and ligands

Before generating new molecular components for use in COOT and REFMAC5 it is worth checking the PDB for structures that have used the chemical type before and download the

.cif dictionary file directly from there. For example, there are several different forms of the basic green FP type chromophore, mostly with the residue 65 component changed (*e.g.*, CRO from 1EMA for wild-type GFP (TYG), CR2 from 1MYW for Venus (GYG) and CH6 from 4ZIO for mCherry (MYG)). Most extra residues can be imported directly into COOT by using the 3-letter code *e.g.*, HOX. However, we find it's always worth having a local copy of the .cif dictionary file. If a new residue is required, then we have found the following steps work best after initial molecular replacement (Section 3.3.2 Step 8). Please note that when using a model containing non-standard residues, Phaser has a tendency to set atom occupancies to 0 (*i.e.*, as if they are not there). To correct this, either manually edit the PDB file (as text) or delete and re-introduce using COOT.

1. Design ligand/non-standard residue using 3D molecular structure *e.g.*, ChemDraw. Make sure all bond orders (single/double/triple/aromatic) are correct, and hydrogen atoms are included.
2. Give residue a unique 3 letter code (*e.g.*, p-amino-L-Phenylalanine=HOX, [PDB ID: 5BT0], or CQ1 for azide-containing chromophore [PDB ID: 4J88])
3. Generate parameters with AceDRG [45], which creates a .cif dictionary file (e.g., HOX.cif) which can be used inside COOT and REFMAC5.
4. Load .cif dictionary parameter files into COOT.
5. Insert non-standard residues manually via COOT (File > Insert Monomer> “3 letter residue code”.
6. Refine structure with REFMAC5 including .cif dictionary file in the geometry dictionary field.
7. Continue solving Structure as in Section 3.3.2.

4. Notes

1. It is important to have the right *E. coli* strain for your plasmid. For example, pET vectors will need DE3 strains such as BL21, due to the requirement for the host strain to produce T7 RNA

polymerase. Another strain we have used for the expression of FPs are TOP10 (Invitrogen) cells for arabinose induction.

2. Presence of EDTA causes issues with Cation binding columns (e.g., nickel and cobalt affinity columns) so best to avoid in lysis and purification buffers.
3. Any 3D molecular viewer is sufficient, and readers should use software they are most comfortable with.
4. Transformation is unsuccessful if a loud 'popping' sound is heard. This can occur if the concentration of DNA is too high, high levels of salt present (DNA purification not performed correctly) or bubbles in the cuvette solution. Restart transformation with fresh cells.
5. This should be performed quickly to ensure recovery of cells.
6. A suitable chemical inducer can be added to the agar so FP cells can be detected by coloured colonies e.g*.,* GFP colonies can fluoresce green under UV light.
7. We recommend plating two different volumes (e.g., 20 µl and 200 µl) to ensure that at least one plate will have well-spaced colonies.
8. We use a flask that holds double the volume to ensure there is sufficient aeration *e.g.*, 2L flask for 1 L media. Auto induction media is used for expression of our FP's although this will depend on your expression system. If incorporating a nAA into the FPs, make sure that the two required antibiotics are present in the culture media together with the nAA itself. If the nAA cannot be dissolved in water, then try 0.1 M NaOH (the final solvent concentration in the culture is below 0.1% (v/v)) or according to the manufacturer's instructions. If the nAA is photosensitive, grow the cultures in the absence of ambient light (e.g., in the dark or covered in foil).
9. This can range between 20 - 37 C. Expression trials at different temperatures can be performed to determine optimal temperature.
10. The cell pellet should be coloured, with the colour dependent on the FPs expressed.
11. Sonication can also be used; however, we have noticed that this method can lead to misfolding and breakdown of certain FP's and their nAA mutants.
12. We normally use a F21-8x50y rotor (Thermo Scientific) in combination with a Sorvall floor standing centrifuge.
13. The soluble fraction should be coloured. There may be some colouration to the pellets but don't be too concerned unless there is no colour to the solution.

14. To reduce non-specifically bound contaminants, add up to 20 mM imidazole to the binding buffer and 150-500 mM NaCl to binding and elution buffers.
15. Typically, FPs elute at around 150-250 mM imidazole depending on the accessibility of the HIS-tag.
16. If using 280 nm and the maximum excitation wavelength, only collect fractions where both wavelengths show an increase, *i.e.*, not 280 nm increase only as this will contain mostly contaminants and non-functional protein. The fractions containing the FP should be very clear by their colour.
17. The 16/60 GL columns have an ideal sample volume range of 0.5-2 mL. If a larger loading volume is required use the equivalent 26/60 GL columns which have a maximum loading volume of 5 mL. These columns are for preparative use only, if looking to use SEC for analytical purposes *(I.e.*, determining oligomeric states) consider using the Superdex Increase range (sample volumes 0.1-0.5 mL)
18. Different concentrator columns have different maximum speed tolerances. Remember to check the manufacturer's instructions for recommended speed and time. If you can get the volume lower than 2 mL (0.5 mL is preferred) then this will improve the resolution of the chromatography. If you can't get to 2 mL, do separate SEC runs.
19. To maximise crystal hits, use a range of protein concentrations in crystal trials but avoid concentrations where the protein precipitates.
20. The easiest and most accurate approach to determine final protein concentration is absorbance spectroscopy using the known molar extinction coefficient (available from FPBase https://www.fpbase.org). If working with a novel mutant, use a Bradford-type assay with the wild-type FP from which the mutant was derived as the standard.
21. If manually creating crystal dishes, use a multi-channel pipette for better efficiency and reduced evaporation. Normally we screen for crystal formation using sitting drop vapour diffusion across a wide variety of conditions using broad range commercial screens such as PACT *premier*™ HT-96 and JCSG (Molecular Dimensions, Suffolk UK). Typically, these broad screens are sufficient to generate crystals that diffract to high resolution (We have generated crystals that diffract down to ~1.28 Å without further optimisation). Should FP crystals not diffract to the

desired resolution, then screens can be optimised by varying pH (+/- 0.5 pH unit) and precipitant composition (+/- ~ 5%) around hits to improve lattice formation.

22. Varying the ratios of protein to crystallisation buffer (*i.e.*, 1:1, 1:2, 2:1 *etc*) can lead to improved crystal growth. In our setup, solutions for crystallisation are dispensed using an Art Robbins Phoenix robot (Alpha Biotech, UK) or mosquito ® crystal (sptlabtech, UK), however most liquid handling robots can be programmed for setting up crystal dishes.
23. Other temperatures e.g., 4 or 20 ºC can be used to incubate crystal trays. We have found 25 ºC to be the optimal temperature for FP crystal growth.
24. Ideally use a loop that is slightly bigger than the crystal to reduce excess liquid. This should aid in reducing background noise during X-ray diffraction experiments. The crystals should be coloured to match that of your FP. If they are not coloured, then it is likely they are not protein crystals or crystals of your FP.
25. A general rule of thumb is to pick a buffer of a similar pH to the crystallisation condition and a high concentration of glycerol/MPD/low molecular weight PEG (based on precipitant in crystal screen)
26. Please check with your local synchrotron source about how to supply sample. Crystals remain in liquid nitrogen except during transit to Diamond Light Source (Harwell, UK), due to shipping health and safety regulation regulations where a dry shipping Dewar is used (provided by DLS).
27. You or your institution will have to apply for access to centralised facilities, especially if there is no in-house X-ray diffraction equipment suitable for proteins. From our experience, access is normally easily granted (especially is samples are submitted remotely). Staff at these facilities are generally excellent and a valuable source of information and help. We are eternally grateful to the staff at Diamond Light Source.
28. Use screening protocols and auto-centring to determine the centre of the crystal for maximum resolution. To reduce radiation damage, use lower intensity X-rays over more images.
29. You should be able to solve the structure with a resolution of 3 Å, but the higher the resolution the better. Should crystals not diffract to high resolution consider setting up crystal trials using a fine screen around previous hits (See 3.5.1 step 4). At this point you should be able to tell if the data are of good quality (the reports after Xia2, POINTLESS and AIMLESS will highlight any issues with the data).

30. The most common space group we have observed with FPs is $P2_12_12_1$. However, we have also observed others such as $P4_32_12$, $P6_5$ and $C2$.
31. Do not include any primary elements removed during expression and purification *e.g.*, signal sequences and purification tags. This will affect the calculations and could lead to the wrong number of components in the unit cell.
32. Running PHASER in advanced mode can lead to complications when working with non-standard residues without the correct .cif dictionary present. It may also be necessary to correct the atom occupancies of chromophore atoms as they tend to default to zero during molecular replacement.
33. This sharpens the map and improves refinement statistics by averaging out solvent peaks.
34. A gold standard is to have both R and R_{Free} below 0.2 although this is not always possible. However, there shouldn't ever be a more than 0.04 gap between the two as this is a sign that there is too much bias toward the model.
35. Ideally you want to limit Ramachandran outliers (< 1%) and atomic clashes to the bare minimum and that any ligands are associated to the correct protein chain.
36. Before depositing your structure, we would highly recommend using the deposition task on CCP4i2, which will ensure you have all the required information for deposition. For any issues with the deposition, we have found the team at wwPDB are very helpful at resolving problems.

Acknowledgements.

We would like to thank the staff at the Diamond Light Source (Harwell, UK) for the supply of facilities and beam time, especially beamlines I02, I03 and I04 staff, under beamtime code mx18812. This work was supported by BBSRC (BB/H003746/1 and BB/M000249/1), EPSRC (EP/ J015318/1) grants (to DDJ). RDA was supported by KESS 2-Tenovus studentship and HSA. by the Higher Committee for Education Development in Iraq. We would like to thank the Protein Technology Hub, School of Biosciences, Cardiff University for use of facilities to generate protein and analyse protein essential to structural studies. We would also

like to thank to Ben Bax and Magdalena Lipka-Lloyd for help with protein crystallisation. Finally, we are eternally grateful to Pierre Rizkallah for helping on all aspects of protein crystallography over the years and teaching many members of the DDJ group how to determine protein structures – we salute you sir.

References

1. Rodriguez EA, Campbell RE, Lin JY, et al (2017) The Growing and Glowing Toolbox of Fluorescent and Photoactive Proteins. Trends Biochem Sci 42:111–129. https://doi.org/10.1016/j.tibs.2016.09.010

2. Duwe S, Dedecker P (2019) Optimizing the fluorescent protein toolbox and its use. Curr Opin Biotechnol 58:183–191. https://doi.org/10.1016/j.copbio.2019.04.006

3. Shaner NC, Steinbach PA, Tsien RY (2005) A Guide to Choosing Fluorescent Proteins. Nat Methods 2:905–909. https://doi.org/10.1038/nmeth819

4. Tsien RY (1998) The green fluorescent protein. Annu Rev Biochem 67:509–544. https://doi.org/10.1146/annurev.biochem.67.1.509

5. Chalfie M, Tu Y, Euskirchen G, et al (1994) Green Flourescent Protein as a Marker for Gene-Expression. Science (80-) 263:802–805. https://doi.org/10.1126/science.8303295

6. Ormo M, Cubitt AB, Kallio K, et al (1996) Crystal structure of the Aequorea victoria green fluorescent protein. Science (80-) 273:1392–1395

7. Yang F, Moss LG, Phillips Jr. GN (1996) The molecular structure of green fluorescent protein. Nat Biotechnol 14:1246–1251. https://doi.org/10.1038/nbt1096-1246

8. Matz M V, Fradkov AF, Labas YA, et al (1999) Fluorescent proteins from nonbioluminescent Anthozoa species. Nat Biotechnol 17:969–973. https://doi.org/10.1038/13657

9. Wall MA, Socolich M, Ranganathan R (2000) The structural basis for red fluorescence in the tetrameric GFP homolog DsRed. Nat Struct Biol 7:1133–1138

10. Merzlyak EM, Goedhart J, Shcherbo D, et al (2007) Bright monomeric red fluorescent protein with an extended fluorescence lifetime. Nat Methods 4:555–557. https://doi.org/10.1038/nmeth1062

11. Lambert TJ (2019) FPbase: a community-editable fluorescent protein database. Nat Methods 16:277–278. https://doi.org/10.1038/s41592-019-0352-8

12. Arpino JAJ, Rizkallah PJ, Jones DD (2012) Crystal Structure of Enhanced Green Fluorescent Protein to 1.35 angstrom Resolution Reveals Alternative Conformations for Glu222. PLoS One 7:. https://doi.org/10.1371/journal.pone.0047132

13. Heim R, Prasher DC, Tsien RY (1994) Wavelength mutations and posttranslational autoxidation of green fluorescent protein. Proc Natl Acad Sci U S A 91:12501–12504

14. Reddington SC, Rizkallah PJ, Watson PD, et al (2013) Different Photochemical Events of a Genetically Encoded Phenyl Azide Define and Modulate GFP Fluorescence. Angew Chemie-International Ed 52:5974–5977. https://doi.org/10.1002/anie.201301490

15. Craggs TD (2009) Green fluorescent protein: structure, folding and chromophore maturation. Chem Soc Rev 38:2865–2875. https://doi.org/10.1039/b903641p

16. Barondeau DP, Kassmann CJ, Tainer JA, Getzoff ED (2005) Understanding GFP chromophore biosynthesis: controlling backbone cyclization and modifying post-

translational chemistry. Biochemistry 44:1960–1970. https://doi.org/10.1021/bi0479205

17. Barondeau DP, Kassmann CJ, Tainer JA, Getzoff ED (2006) Understanding GFP posttranslational chemistry: structures of designed variants that achieve backbone fragmentation, hydrolysis, and decarboxylation. J Am Chem Soc 128:4685–4693. https://doi.org/10.1021/ja056635l

18. Barondeau DP, Tainer JA, Getzoff ED (2006) Structural evidence for an enolate intermediate in GFP fluorophore biosynthesis. J Am Chem Soc 128:3166–3168. https://doi.org/10.1021/ja0552693

19. Rosenow MA, Huffman HA, Phail ME, Wachter RM (2004) The crystal structure of the Y66L variant of green fluorescent protein supports a cyclization-oxidation-dehydration mechanism for chromophore maturation. Biochemistry 43:4464–4472. https://doi.org/10.1021/bi0361315

20. Auhim HS, Grigorenko BL, Harris TK, et al (2021) Stalling chromophore synthesis of the fluorescent protein Venus reveals the molecular basis of the final oxidation step. Chem Sci 12:7735–7745. https://doi.org/10.1039/d0sc06693a

21. Patterson GH, Lippincott-Schwartz J (2002) A photoactivatable GFP for selective photolabeling of proteins and cells. Science (80-) 297:1873–1877. https://doi.org/10.1126/science.1074952

22. Remington SJ (2011) Green fluorescent protein: A perspective. Protein Sci 20:1509–1519. https://doi.org/10.1002/pro.684

23. van Thor JJ (2009) Photoreactions and Dynamics of the Green Fluorescent Protein. Chem Soc Rev 38:2935–2950. https://doi.org/10.1039/b820275n

24. Worthy HL, Auhim HS, Jamieson WD, et al (2019) Positive functional synergy of structurally integrated artificial protein dimers assembled by Click chemistry. Commun Chem 2: https://doi.org/10.1038/s42004-019-0185-5

25. Royant A, Noirclerc-Savoye M (2011) Stabilizing role of glutamic acid 222 in the structure of Enhanced Green Fluorescent Protein. J Struct Biol 174:385–390. https://doi.org/10.1016/j.jsb.2011.02.004

26. Jumper J, Evans R, Pritzel A, et al (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596:583–589. https://doi.org/10.1038/s41586-021-03819-2

27. Baek M, DiMaio F, Anishchenko I, et al (2021) Accurate prediction of protein structures and interactions using a three-track neural network. Science (80-) 373:871–876. https://doi.org/10.1126/science.abj8754

28. Larsson AM (2009) Preparation and crystallization of selenomethionine protein. In: Bergfors TM (ed) IUL Biotechnology Series 8 (Protein Crystallization, 2nd ed. International University Line, p 135154

29. Hendrickson WA, Horton JR, LeMaster DM (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

30. Liu CC, Schultz PG (2010) Adding New Chemistries to the Genetic Code. In: Kornberg RD, Raetz CRH, Rothman JE, Thorner JW (eds) Annual Review of Biochemistry, Vol 79. pp 413–444

31. Crameri A, Whitehorn EA, Tate E, Stemmer WPC (1996) Improved green fluorescent protein by molecular evolution using DNA shuffling. Nat Biotechnol 14:315–319. https://doi.org/10.1038/nbt0396-315

32. Wang F, Niu W, Guo J, Schultz PG (2012) Unnatural Amino Acid Mutagenesis of Fluorescent Proteins. Angew Chemie-International Ed 51:10132–10135. https://doi.org/10.1002/anie.201204668

33. Hammill JT, Miyake-Stoner S, Hazen JL, et al (2007) Preparation of site-specifically labeled fluorinated proteins for 19F-NMR structural characterization. Nat Protoc 2:2601–2607. https://doi.org/10.1038/nprot.2007.379

34. Plass T, Milles S, Koehler C, et al (2011) Genetically Encoded Copper-Free Click Chemistry. Angew Chemie-International Ed 50:3878–3881. https://doi.org/10.1002/anie.201008178

35. Winter G (2010) xia2: an expert system for macromolecular crystallography data reduction. J Appl Crystallogr 43:186–190. https://doi.org/doi:10.1107/S0021889809045701

36. Evans P (2006) Scaling and assessment of data quality. Acta Crystallogr Sect D 62:72–82. https://doi.org/doi:10.1107/S0907444905036693

37. Collaborative (1994) The CCP4 suite: programs for protein crystallography. Acta Crystallogr Sect D 50:760–763. https://doi.org/doi:10.1107/S0907444994003112

38. McCoy AJ, Grosse-Kunstleve RW, Adams PD, et al (2007) Phaser crystallographic software. J Appl Crystallogr 40:658–674. https://doi.org/10.1107/S0021889807021206

39. Jia-xing Y, Woolfson MM, Wilson KS, Dodson EJ (2005) A modified ACORN to solve protein structures at resolutions of 1.7Å or better. Acta Crystallogr Sect D 61:1465–1475. https://doi.org/10.1107/S090744490502576X

40. Foadi J, Woolfson MM, Dodson EJ, et al (2000) A flexible and efficient procedure for the solution and phase refinement of protein structures. Acta Crystallogr Sect D

56:1137–1147. https://doi.org/10.1107/S090744490000932X

41. Jia-xing Y, Woolfson MM, Wilson KS, Dodson EJ (2002) ACORN — theory and practice. Zeitschrift für Krist - Cryst Mater 217:636–643. https://doi.org/doi:10.1524/zkri.217.12.636.20655

42. Yao J (2002) ACORN in CCP4 and its applications. Acta Crystallogr Sect D 58:1941–1947. https://doi.org/10.1107/S0907444902016621

43. Emsley P, Cowtan K (2004) Coot: model-building tools for molecular graphics. Acta Crystallogr Sect D 60:2126–2132. https://doi.org/doi:10.1107/S0907444904019158

44. Murshudov GN, Vagin AA, Dodson EJ (1997) Refinement of Macromolecular Structures by the Maximum-Likelihood Method. Acta Crystallogr Sect D 53:240–255. https://doi.org/doi:10.1107/S0907444996012255

45. Long F, Nicholls RA, Emsley P, et al (2017) AceDRG: a stereochemical description generator for ligands. Acta Crystallogr Sect D 73:112–122. https://doi.org/doi:10.1107/S2059798317000067