Green-lighting green fluorescent protein: Faster and more efficient folding by eliminating a *cis-trans* peptide isomerization event

David J. Rosenman,¹ Yao-ming Huang,^{1,2} Ke Xia,³ Keith Fraser,¹ Victoria E. Jones,¹ Colleen M. Lamberson,¹ Patrick Van Roey,⁴ Wilfredo Colón,³ and Christopher Bystroff^{1*}

¹Rensselaer Polytechnic Institute, Biological Sciences, 110 8th St., Troy, New York 12180

²Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, California 94143-2530

³Rensselaer Polytechnic Institute, Chemistry and Chemical Biology, 110 8th St., Troy, New York 12180

⁴Laboratory of Computational and Structural Biology, Wadsworth Center, P.O. Box 509, Albany, New York 12201-0509

Received 26 July 2013; Accepted 6 January 2014

DOI: 10.1002/pro.2421

Published online 9 January 2014 proteinscience.org

Abstract: Wild-type green fluorescent protein (GFP) folds on a time scale of minutes. The slow step in folding is a *cis-trans* peptide bond isomerization. The only conserved *cis*-peptide bond in the native GFP structure, at P89, was remodeled by the insertion of two residues, followed by iterative energy minimization and side chain design. The engineered GFP was synthesized and found to fold faster and more efficiently than its template protein, recovering 50% more of its fluorescence upon refolding. The slow phase of folding is faster and smaller in amplitude, and hysteresis in refolding has been eliminated. The elimination of a previously reported kinetically trapped state in refolding suggests that X-P89 is *trans* in the trapped state. A 2.55 Å resolution crystal structure revealed that the new variant contains only *trans*-peptide bonds, as designed. This is the first instance of a computationally remodeled fluorescent protein that folds faster and more efficiently than wild type.

Keywords: GFP; folding kinetics; protein design; *cis*; *trans* isomerization

Introduction

Green fluorescent protein (GFP) is a monomeric, kinetically stable protein capable of autonomously forming an intrinsic and highly efficient fluorescent chromophore. GFP and its variants have seen ubiquitous use as a reporter for gene expression, as a fusion tag for the spatial localization of proteins in the cell, as fluorophore partners in Förster resonance energy transfer (FRET)-based applications,^{1,2}

and as biosensors of pH, metal ions, peptides, and proteins.³ GFPs numerous contributions to biotechnology were highlighted by the awarding of the 2008 Nobel Prize in Chemistry to three pioneer investigators who isolated the protein, explained the mechanism behind its autofluorescence, and broadened its applications as a bioluminescent tag.⁴

Initially, the effectiveness of GFP as a reporter was limited by the slowness of its folding, with a half life on the order of minutes,⁵ and its associated proclivity to aggregate in the cell when overexpressed at 37°C.⁶ *In vitro* evolution techniques have been used to create GFP variants with substantially improved fluorescence and folding efficiency. These include “enhanced” (EGFP) with two-point mutations, ‘Cycle3’ with 3, ‘folding reporter’ with 5, ‘superfolder’ (sfGFP)

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Christopher Bystroff; Department of Biology, 1W14 Jonsson-Rowland Science Center, Rensselaer Polytechnic Institute, 110 Eighth Street, Troy, NY 12180. E-mail: bystroff@rpi.edu

with 11, and ‘OPT’ (OPT-GFP) with 16.^{7–10} Despite these improvements in folding efficiency, sfGFP, and OPT-GFP still fold slowly when compared with a typical monomeric protein, and both also misfold to yield a long-lived, nonfluorescent “trapped” state.^{11,12}

Cis–*trans* isomerization of X-Pro peptide groups is a rate-determining step in folding for many proteins.^{13,14} Engineered elimination of *cis* X-Pro in ribonuclease T1 both accelerated folding and removed folding intermediates,^{15–17} while different *trans* Pro→Ala mutations were observed to either speed up or slow down folding and unfolding of CRABP.¹⁸ Stopped flow experiments of Cycle3 in the presence of cyclophilin A, a prolyl peptide isomerase, have demonstrated that *cis*–*trans* isomerization accounts for the slowest phase of GFP’s multiphasic refolding kinetics, composing 5–14% of the refolding amplitude.¹⁹ Further evidence was provided by Steiner,²⁰ who replaced all of the prolines in EGFP with the fast-flipping analog (4S)-fluoroproline, resulting in improved folding kinetics. These results focused the source of the slow phase on X-Pro backbone *cis*–*trans* isomerization.

Most homologs of GFP contain one or two prolines in *cis* peptide bonds. In sfGFP, only P89 is *cis*.⁹ This proline, along with P58 and P75, form a triad of highly conserved prolines that flank the distorted internal α -helix of GFP that harbors the chromophore. Mutation of these prolines abolishes the chromophore maturation and eliminates a kinetically trapped, nonfluorescent state.¹² It has been hypothesized that a near-native conformational state, called N_{iso} , exists which undergoes a very slow transition to the native state, and that the presence of N_{iso} in the folding pathway is responsible for proper chromophore maturation.²¹ Within the transition from N_{iso} to the native state, each of the three conserved prolines (the “triumvirate”¹²) slowly isomerizes as part of a system that is believed to lock the beta barrel in place around the central helix and properly position side chains for chromophore maturation.¹²

In this paper we describe the computational design, crystal structure and refolding kinetics of AT-GFP (all *trans* GFP), a new, faster folding variant designed to contain no *cis* peptide bonds. Our findings identify one of the sequence determinants of the slow phase of GFP folding, help to define the structure of the observed trapped state, and elucidate the sequence requirements for chromophore maturation.

Results

We targeted the P89 loop of GFP to investigate the hypothesized connection between slow folding and chromophore maturation. Since point mutations of P89 failed to produce a soluble and fluorescent protein in previous work,¹² we created a longer loop to replace the proline. The designed variant is stable

and fluorescent, having excitation/emission spectra nearly identical to those of sfGFP. The 2.55 Å crystal structure shows that the loop adopts all-*trans* peptide configurations. Refolding studies in GuHCl show a significant partial loss of the slowest phase of refolding, leading to faster and more efficient folding overall. pH refolding studies show a complete loss of the slow phase. AT-GFP does not form a kinetically stable trapped state under conditions that produce such a state in OPT-GFP and sfGFP.

Loop design

The template for loop design was sfGFP (PDB: 2B3P). On to this template, six surface mutations were modeled based on the work of Cabantous,^{10,39} producing OPT-GFP, which was used as the control for experiments reported here. Sequences used in this study are shown in Supporting Information Figure 2.

In order to remove the *cis*-peptide at P89, we replaced the loop with several candidate database-derived loops having all-*trans* backbone omega angles. We then subjected the loops to rounds of side chain design and energy minimization. No loops with reasonable stereochemistry and backbone angles were found in the initial database search. However, when a two-residue insertion was allowed, loops with good stereochemistry were found. Side chains were added using in-house computational protein design software.²⁴ Starting with the two-residue insertion P89AAA, the iterative design/minimization procedure converged in four rounds to a single structure and sequence, 88-ISNGDGFN-96, replacing the wild-type sequence 88-MPEGYVQ-94.

Characterization of fluorescence

Purified AT-GFP has a single, monomeric FPLC peak. No significant differences were found in the fluorescence excitation and emission spectra of AT-GFP as compared to OPT-GFP, but both variants’ spectra differ from published spectra of wild-type *Aequorea victoria* GFP in the ratio of its two excitation maxima. Peak “A” (395 nm) corresponds to the cationic⁴⁰ or neutral⁴¹ protonation state of the chromophore, while Peak “B” (480 nm) corresponds to the zwitterionic⁴⁰ or anionic⁴¹ state. Peak A is dominant in wild-type and Cycle3 GFP,⁸ but missing in both OPT-GFP and AT-GFP. Peak B is shifted slightly, 480 nm in AT-GFP versus 485 nm in OPT-GFP. Both emit at 507 nm (F_{507}).

The concentration of mature chromophore per mole protein of AT-GFP measured by A_{480}/A_{280} was 79% that of OPT-GFP, signaling a decrease in the catalytic efficiency of chromophore maturation in the new variant. The relative fluorescence quantum yield F_{507}/A_{480} was 76% that of OPT-GFP, where the latter was measured as F_{507}/A_{485} . The product of these two numbers, 61% is the yield of fluorescence

Figure 1. The loop connecting the central helix with strand 4 of GFP as it appears in sfGFP (orange), in the designed model (pink), and in the AT-GFP crystal structure (green). Boxed numbers are strand labels. A thin green circle marks the *cis*-peptide bond in sfGFP.

per mole protein AT-GFP, relative to OPT-GFP. (Spectra are shown in Supporting Information Fig. 3.) We noted that the atomic positions of the imidazolidone ring of the chromophore were subtly but significantly shifted between the least-squares superposed AT-GFP and sfGFP crystal structure, and this shift may have resulted in a subtle change in the stabilization of the photo-excited state, but no specific mechanism for this change has been identified.

Crystallography

The structure of AT-GFP was solved to 2.55 Å resolution by molecular replacement using the sfGFP structure (PDB: 2B3P) as the model. The crystals were slightly twinned and highly mosaic, with a 0.63° mean spot width. There are five molecules in the asymmetric unit. Noncrystallographic symmetry (NCS) operators were determined using the rigid body refined monomer coordinates. Noncrystallographic two-fold symmetry is present but there is no point-group symmetry. NCS operators were used to restrain most backbone atoms throughout refinement. After convergence of simulated annealing refinement, three loop regions showed un-modeled density in $2F_o - F_c$ and $F_o - F_c$ difference maps. These three regions, including the designed loop (86–90), a loop that interacts with the modeled loop (188–194) and a hairpin loop that is involved in crystal contacts (156–159), were each remodeled using omit maps and real-space modeling tools. NCS restraints were turned off for those three regions, and the structure was again refined to convergence. The final MolPro-

bity³³ score was 2.17 (93rd percentile). 94.5% of all backbone angles are in Ramachandran favored regions; 0.36% are outliers, none in the designed loop.

The final model for the designed loop has a α helical turn 85-KSAISNG-91 followed by two residues with positive backbone ϕ angles, 92-DG-93. This conformation differs from the computational prediction, as shown in Figure 1. The designed conformation was a “glycine helix cap” motif as defined in the I-sites Library of local structure motifs,⁴² where G91, the first of the two glycines in the loop sequence, was designed to adopt a positive backbone ϕ angle. Instead it adopts a negative ϕ angle while both D92 and G93 adopt positive ϕ angles. Positive ϕ angles (the α_L region of the Ramachandran plot⁴³) are rare, comprising only 6.8% of all positions in globular proteins, predominantly at glycine residues. Consecutive $\alpha_L\alpha_L$ positions are more rare, occurring in only 0.7% of all dipeptide units. However, the sequence DG is one of three dipeptide sequences (along with GG and NG) that most favor $\alpha_L\alpha_L$. The DG sequence occurs in the $\alpha_L\alpha_L$ conformation 11% of the time. Folding of this rare structural motif may account for the observed residual slow phase in the refolding kinetics.

The quality of the fit to the electron density was assessed by looking at $2F_o - F_c$ and $F_o - F_c$ difference maps. In chains A, C, D, and E, all backbone and side chain atoms fit entirely within the 1-sigma contour of the $2F_o - F_c$ density (Fig. 2).

In chain B, the alpha carbon of G91 was well outside of the 1-sigma contour and there was positive and negative $F_o - F_c$ difference density that

Figure 2. $2F_o - F_c$ electron density contoured at 1σ for the designed loop in chain A. All of the peptide bonds are *trans*. The electron density and coordinates are indistinguishable in the other four chains, which were refined with noncrystallographic symmetric restraints.

suggested that a *cis*-peptide bond was present at 90-NG-91. To explore this possibility, the 90-NG-91 peptide bond was remodeled and restrained in the *cis* conformation, and the structure was refined through one round of simulated annealing. The resulting R-free was slightly improved, the $F_o - F_c$ difference density was reduced, and the G91 alpha carbon was now centered in the density. As a control, we also attempted to remodel 90-NG-91 in chain E as a *cis* peptide, but in this case the result was an increase in both R-free and difference density. Since the density fits a *cis* peptide at 90-NG-91 in chain B only, we attempted to model it as such. The implication would be that both *cis* and *trans* conformers are present in the natively folded state, but that only one of the five sites in the asymmetric unit, chain B, showed a preference for housing the *cis* conformer. Coincidentally perhaps, chain B makes contacts with chain D through the side chain of N90, and the side chain of N90 is shifted in the *cis* conformer. The presence of minority *cis* conformers in AT-GFP may explain the residual slow phase present in the folding kinetics. However, we cannot be sure without higher resolution data that this part of chain B is in the correct conformation. The lack of any difference in the refolding kinetics when CyPA is used (see next section) argues that the fraction *cis* is negligible at best. At the end of the day, the kinetics evidence outweighed the weak statistical evidence from the small drop in R-free, and the presence of a residual *cis* peptide in AT-GFP is therefore unsupported. All five chains were modeled as all-*trans* in the final structure.

Aside from the designed loop and the three variable regions mentioned here, the overall structure of AT-GFP differs only slightly from that of sfGFP

(PDB ID: 2B3P) with an overall RMSD < 1.0 Å for backbone and side chain atoms. The first exception to this trend includes N-terminal residues 1–4. This short N-terminal helix is in close contact with the designed loop, and the helix appears disrupted, forming an extended structure which adopts a different conformation in each of the five copies. The N-terminus is involved in crystal contacts in chain D only. Variability was also found in the loop between strands 9 and 10, residues 189–198. The backbone is shifted away from the nearby 88–92 loop by about 1 Å due to direct contacts, including a hydrogen bond between G91 and G193. This loop conformation is consistent across all five copies of the protein, even though one copy (chain B) is involved in a crystal contact. Finally, the β -hairpin between strands 7 and 8, residues 158–161, are involved in crystal contacts in three of the five copies, and there are large (2–3 Å) backbone shifts. The peptide backbone at 160-KN-161 flips 180° in chain A, which forms a salt bridge crystal contact with chain B at this position.

AT-GFP has been submitted to the Protein Data Bank (PDB ID 4LW5). Table I shows a summary of the crystallographic statistics.

Refolding kinetics

GFPs multiphasic refolding trajectory reveals parallel folding pathways with at least three,³⁹ or as many as five,¹⁹ distinct rates ranging in half-life from less one-half second to hundreds of seconds, but some of these phases were only observed using far-UV CD and tryptophan fluorescence. Cycle3 GFP, when studied using green fluorescence only and pH jump refolding, has three kinetic phases, the slowest of which disappears upon treatment with

Table I. Data Collection and Refinement Statistics (Molecular Replacement) for AT-GFP (PDB ID 4LW5)

	AT-GFP
Data collection	
Space group	P2 ₁ 2 ₁ 2 ₁
Cell dimensions	
<i>a</i> , <i>b</i> , <i>c</i> (Å)	95.557, 110.939, 114.706
Resolution (Å)	2.55
<i>R</i> _{sym}	0.119 (0.536) ^a
<i>I</i> /σ <i>I</i>	20.2 (3.13)
Completeness (%)	99.49 (100.0)
Redundancy	6.9 (7.0)
Refinement	
Resolution (Å)	2.55
No. reflections	40237
<i>R</i> _{work} / <i>R</i> _{free}	0.218 / 0.279
No. atoms	
Protein	8988
Water	427
<i>B</i> -factors	
Protein	36.6
Water	32.3
Rms deviations	
Bond lengths (Å)	0.007
Bond angles (°)	1.4

One crystal was used.

^a Values in parentheses are for highest-resolution shell.

the peptide bond isomerase cyclophilin A (CyPA).¹⁹ Refolding of the template protein for this experiment, OPT-GFP, using GuHCl dilution, has a fourth,

Figure 3. Comparison of rates (*k*) and amplitudes of the multiple phases of GuHCl dilution refolding for AT-GFP (closed circles), and OPT-GFP (open circles), with error bars showing the standard deviation over six replicates. Previously published multiphase kinetic measurements are shown for cycle3 GFP (open squares), and cycle3 GFP in the presence of CyPA (closed squares). Amplitudes are normalized to sum to 100%. Labels on the top are used in the text.

Figure 4. Fluorescence recovery versus time for AT-GFP (red) and OPT-GFP (cyan). For each protein, six experimental traces (thin black lines) are shown superposed on the least-squares fit to a four-phase kinetic model (thick lines). The traces were fit individually and the six rates and amplitudes. The curves were calculated using the averaged rates and amplitudes. The sum of the amplitudes was then scaled to the percent recovery of fluorescence. The inset shows a blow-up of the early portion of the same data.

very fast, phase (1.7 s^{-1}) that accounts for 38% of the total amplitude.

As illustrated in Figure 3, GuHCl dilution refolding is faster overall in AT-GFP, which at 300 s has recovered 99% of its final baseline fluorescence, while OPT-GFP has recovered only 93%. The results of the four phase kinetic fit are shown in Figure 4. The very fast, fast, and medium phases were unchanged within experimental error from those of the template protein, but the slow phase is both twice as fast (0.01 versus 0.005 s^{-1}) and has a diminished relative amplitude (24 versus 30%).

Refolding by GuHCl dilution was compared to pH jump refolding for AT-GFP and OPT-GFP. Refolding by pH jump was carried out using manual mixing rather than stopped flow, so that the very fast phase in stopped flow GuHCl refolding becomes a burst phase in manual mixing pH refolding. With that caveat, the rates for pH jump refolding of OPT-GFP are similar to GuHCl refolding, having four phases. However, AT-GFP has no detectable slow phase when measured using manual mixing pH jump. The loss of the slowest phase cannot be attributed to using manual mixing versus using stopped flow, since the increased dead time only affects the early data. It can however, be attributed to the difference in folding conditions, pH 8.0 with 0 M GuHCl (manual mixing pH jump) versus pH 8.0 with 0.55 M GuHCl (stopped flow GuHCl dilution). The presence of 0.55 M guanidine may change the

Figure 5. (a) Hysteresis experiment. Protein starts at 0M GuHCl, pH8. FWD arrow represents titration to pH 5.5 or 1.8 M GuHCl, and incubation for 96 h. REV arrow represents titration to pH 2 or 8 M GuHCl, incubation for 1 h, followed by titration to pH 5.5 or 1.8 M GuHCl, and incubation for 96 h. Blue arrow: magnitude of the hysteresis effect. (b) Results of hysteresis experiment. Blue lines: OPT-GFP. Orange lines: AT-GFP. Solid lines: pH unfolding/refolding. Dashed lines: GuHCl unfolding/refolding. Error bars are from triplicate measurements.

stability of early intermediates of folding, a common observation in many other systems,⁴⁴ leading to a residual slow phase in AT-GFP that is not present in pH jump refolding (Supporting Information Fig. 4).

The faster folding kinetics after replacing P89 with an all-*trans* peptide loop confirms P89 as one of the sources of the slow folding, but removing this source did not completely remove the slow phase when GuHCl dilution was used to refold. The remaining contributors to slow folding could be one or more of the other three highly conserved prolines, converting slowly to the native *trans* state from an equilibrium distribution of isomers in the denatured state. To investigate this possibility, we carried out pH jump refolding in the presence of CyPA. CyPA eliminates the slow phase in OPT-GFP refolding, mirroring the published results for Cycle3 GFP,¹⁹ but has no effect on the kinetics of AT-GFP (Supporting Information Fig. 5). This result seems to exclude other prolines as contributors to the slow phase of folding.

Note that the OPT-GFP and Cycle3 kinetics are not identical. OPT-GFP has a much higher amplitude of the slow phase. We cannot provide an explanation for this, but the 13 surface positions that differ between Cycle3 and OPT-GFP may in some way affect the equilibrium between *cis* and *trans* X-Pro89.

Another possible explanation for the residual slowness in GuHCl refolding is that the designed loop may have replaced one slow step with another one, not as slow as prolyl peptide isomerization, but slower than the medium phase. The new slow step could be the formation of the rare $\alpha_L\alpha_L$ loop. The

absence of this slow phase in pH jump refolding suggests that guanidine is responsible for the presence of a non-native folding intermediate that is otherwise not present. It is possible that guanidine, but not low pH, destabilizes the $\alpha_L\alpha_L$ loop relative to alternative, non-native conformations.

Refolding efficiency and the trapped state

In previous work, it was noted that sfGFP misfolds into a kinetically trapped and nonfluorescent state when refolded from GuHCl, recovering only 39% of the original fluorescence. OPT-GFP, which differs from sfGFP by six surface mutations, gives similar results in our hands. After unfolding into 8.0 M GuHCl, then refolding by dilution into 0.55 M, OPT-GFP recovered 41.4% of the original fluorescence intensity. Under those same conditions, AT-GFP recovered 60.8%. The results for the six-fold replicate experiments are presented in Figure 3. We tentatively conclude that X-Pro89 *cis-trans* isomerization, or perhaps trapping of the *trans* state during folding, is at least partially responsible for misfolding to a long-lived, nonfluorescent state.

There remains an unexplained 39.2% loss of fluorescence in AT-GFP upon refolding, which may be due simply to presence of residual 0.55 M GuHCl. To quantify the misfolded state we carried out “hysteresis” experiments, in which we approached identical midpoint conditions, 1.8 M GuHCl or pH 5.5, from both the folded side and the unfolded side as illustrated in Figure 5(a). The magnitude of the hysteresis effect was the fluorescence recovered by refolding to the midpoint (REV) divided by fluorescence retained after unfolding to the midpoint

(FWD). As shown in Figure 5(b), OPT-GFP loses about half of its fluorescence to a “trapped” nonfluorescent state, whether refolding by pH jump or by GuHCl dilution. AT-GFP, on the other hand, recovers approximately 100% of its midpoint fluorescence in both cases. We may conclude that X-Pro89 is *trans* in the trapped state.

In these experiments, both AT-GFP and OPT-GFP lose quantum yield at pH 5.5, presumably due to partial unfolding of the beta barrel. In 1.8 M GuHCl, interestingly, OPT-GFP loses more fluorescence than AT-GFP does. The reason for this is unknown, but it suggests that the quantum yield of the chromophore in the context of OPT-GFP is more sensitive to GuHCl than it is in the context of AT-GFP.

Kinetic and thermodynamic stability

The rates of unfolding at 20°C were determined by thermal denaturation (40, 60, and 80°C) and were extrapolated assuming Arrhenius behavior, as shown in Supporting Information Figure 6. AT-GFP folds and unfolds faster than its template, and it is thermodynamically less stable by 7.5 kJ/mol. The unfolding half-life of AT-GFP is approximately 5.72 h as compared to 161 h for OPT-GFP.

A comparison of ^1H - ^{15}N HSQC spectra for AT-GFP and OPT-GFP (Supporting Information Figure 7) reveal that many peaks shift between the designed construct and its original template; see Supporting Information Table I. Overall, there is a reduction in the dispersion of these shifts in both the ^{15}N (−1.73%) and ^1H (−2.88%) dimensions in AT-GFP relative to OPT-GFP and there is an upfield shift in the mean and median of the ^1H chemical shift. However, statistical tests of significance (one sample independent *t* test, Levene’s test) fail for both of these observations, ultimately suggesting that the thermodynamic stability of AT-GFP is similar to that of OPT-GFP.

In vivo relative quantum yield

Because *in vivo* applications of GFP are of interest, we investigated the relative efficiency of AT-GFP versus OPT-GFP when expressed in bacteria. Values were measured for fluorescence of a cell suspension per unit cell density (F_{507}/A_{600}), total soluble protein (A_{280}/A_{600}) per unit cell density, and fraction GFP of total soluble protein, for each of the two constructs. The fraction GFP in total soluble protein was measured by densitometry of the gel shown in Supporting Information Figure 1, lanes 3 and 7. All of both GFPs were found in the soluble fraction (data not shown). Combining terms, we get fluorescence per unit GFP for each variant. AT-GFP was found to have 56% of the quantum yield of OPT-GFP *in vivo*. The same number when measured *in vitro* was 61%, as shown in Supporting Information Figure 3. The

difference is within error. We conclude that the *in vitro* measurements for AT-GFP are relevant to *in vivo* applications.

Discussion

GFP is an excellent model system for the investigation of the backbone peptide isomerization events in folding. However, simply mutating the *cis* prolines to residues less prone to isomerize does not always have the desired effect, and in this case an insertion of two residues was necessary to successfully remodel the loop. The result was a GFP variant that folds faster and recovers more of the fluorescence upon refolding. A crystal structure confirms that the *cis*-peptide conformer has been completely removed.

The GuHCl refolding kinetics of the AT-GFP variant has a slow phase of folding that is significantly faster and diminished in amplitude. However, the slow phase has not been altogether eliminated, suggesting that other parts of the molecule besides P89 also contribute to slow refolding. Residual slow unfolding cannot be due to isomerization of other prolines, since CyPA has no effect on the kinetics of AT-GFP. The residual slow phase is probably not the isomerization of the mature chromophore around the hydroxyphenyl side chain, as was previously suggested,¹¹ since pH jump refolding kinetics of AT-GFP showed no residual slow phase. Instead, it is probably the presence of a rare turn in the newly designed region.

In the study by Andrews,¹¹ the misfolded state of sfGFP was characterized by NMR and found to involve the *cis-trans* isomerization of the chromophore around the oxidized Y66 C α -C β bond. Our results suggest, however, that misfolding and kinetic trapping must result from the presence of the non-native *trans* conformation at 88-MP-89. Our modeling experiments and the strict conservation of P89 suggest that the *trans* conformation is not compatible with the native fold. The absence of this trapped state would explain how AT-GFP recovers more of the fluorescent native state when refolded, as compared to OPT-GFP. It must be noted that our experiments and those of Andrews¹¹ lasted 96 h, less than one half-life of OPT-GFP unfolding ($t_{1/2}$ = 161 h), but several half-lives of AT-GFP unfolding ($t_{1/2}$ = 5.72 h). It is possible that the trapped state is not trapped, but simply very long lived. The lifetime of the trapped state may have been shortened in AT-GFP, rather than completely eliminated. In any case, we would still say that the change in the P89 loop is the key to this difference.

This work helps to isolate the key sequential determinants of the chromophore maturation catalytic activity of GFP. The proposed role of P89 as a key structural requirement for chromophore maturation has been disproven, since this construct, lacking P89, glows green. Our results suggest that P89 is

largely responsible for the misfolded “trapped” state, and that its removal leads to higher recovery of fluorescence upon refolding, and to the elimination of hysteresis in refolding.

AT-GFP is a faster folding GFP with improved recovery of fluorescence upon refolding and may improve applications of GFP where refolding and solubility are important. However, the presence of a residual slow phase when refolded from GuHCl, decreased thermostability, and reduced chromophore maturation efficiency means that there is still room for improvement.

Materials and Methods

Loop design

Using the homology modeling module in MOE (Molecular Operating Environment, Version 2010.10; Chemical Computing Group, Montreal, Canada), a preliminary AT-GFP model was created based on coordinates for sfGFP (PDB: 2B3P) where P89 was replaced with a trialanine “placeholder” loop. A database loop search was carried out to identify plausible structures for the two residue insertion, with neighboring regions 85-KSAM-88 and 90-EGYVQ-94 serving as anchoring constraints. Twenty five candidate models were sampled and assigned a GB/VI interaction score.²² The best scoring models were energy minimized using the AMBER99 force field²³ *in vacuo*. All side chains in the preliminary loop models were iteratively redesigned, using DEE-design,²⁴ and energy minimized with distance restraints, using MOE with the AMBER99 force field.²³ The final, lowest energy loop contained the designed sequence 88-ISNGDFIN-96, which replaces 88-MPEGYVQ-94 in OPT-GFP¹⁰ (using 2B3P numbering).

Construction, expression, and purification

The plasmid construct encoding the AT-GFP gene was constructed by replacing MPEGYVQ loop of OPT-GFP with the designed loop sequence using inverse PCR. The template of pCDFDuet-1 plasmid carrying the OPT-GFP gene and primers 5'-ATCGC CGTTGCTAATGGCACTCTTGAAAAAGTCATG-3' and 5'-GGCTTTATTAACGAACGCACTATATCTTT CAAAG-3' [bold: annealing region; italic: linker region that codes for the designed loop sequence] were used. Amplified linear DNA was purified, blunt-end ligated and transformed into Acella strain of *E. coli* (EdgeBio, Gaithersburg, MD) for clone selection. Clones having the designed sequences were confirmed by sequencing (MCLAB, S. San Francisco, CA).

AT-GFP protein was expressed as follows: cell cultures were initiated by 100× dilution of an overnight culture of transformed Acella cells in LB medium with 50 µg/mL streptomycin. After 2.5 h of

growth at 37°C ($OD_{600} = \sim 0.4$), expression was induced with 0.5 mM IPTG and cultures were incubated at 25°C for 16 h. Cells were then pelleted and lysed using freeze/thaw cycles and lysozyme, and the resulting lysates were clarified by centrifugation at 27,000g for 30 min at 4°C. Supernatants containing AT-GFP were purified by ethanol extraction as described previously,²⁵ followed by gel filtration with a Superose 12 10/300 GL column (GE Healthcare). Two successive runs of gel filtration were performed, necessary to successfully remove all contaminants. Purified AT-GFP was dialyzed against the dialysis buffer (150 mM NaCl, 100 mM HEPES-NaOH, pH 8.0) and concentrated using Amicon Ultra-15 concentrators (Millipore, Billerica, MA) to 15–20 mg/mL. SDS-PAGE analysis confirmed that the purity was >95% (Supporting Information Fig. 1).

Fluorescence spectra

The fluorescence spectra of both purified AT-GFP and OPT-GFP were determined using a Fluorolog-3 TAU fluorometer (HORIBA Jobin Yvon, Kyoto, Japan) at room temperature with 5 nm slit and 0.5 s integration. Proteins were diluted to 0.01 µM in the dialysis buffer before the measurements. The emission intensities were collected from 495 to 550 nm with a 0.5 nm increment while excited under 485 nm, and the excitation intensities were recorded from 400 to 500 nm with a 0.5 nm increment while following 508 nm emission. The recorded spectra were normalized by A_{280} .

X-ray crystal structure

Initial conditions for crystallization were determined by the high throughput screen service at the Hauptman-Woodward Institute.²⁶ 100 µm rod-shaped, green crystals were produced by batch crystallization under paraffin oil, containing 15 µM protein in Condition 434 [0.1 M NaCl, 0.1 M HEPES, pH 7.5, PEG 8000 20% (w/v)] in a vibration-free room temperature (22°) incubator. Crystals appeared after one week. For crystal mounting, equal volumes of Condition 434 and 40% PEG 400 were mixed, then added to the drop. Crystals were mounted on cryoloops (Hampton Research, Aliso Viejo, CA) and immediately frozen in liquid N₂.

Data were collected to 2.55 Å resolution at the Brookhaven National Labs X4C beamline using 0.97907 Å X-rays. Images were reduced using the CCP4 iMosflm program.²⁷ Phases were determined by multicopy molecular replacement using CCP4 Molrep.^{28,29} Least-squares refinement was done using CNS,^{30,31} using noncrystallographic symmetry restraints. Remodeling was done, where necessary, using real-space refinement tools in Coot.³² MolProbity³³ was used to identify problem areas and to automatically fix selected side chain rotamers.

Nuclear magnetic resonance

For NMR experiments, both AT-GFP and OPT-GFP were purified using the previously described protocol except that M9 media supplemented with ^{15}N -labeled ammonium chloride was used as the growth medium. ^1H - ^{15}N heteronuclear single quantum coherence (HSQC) spectra for AT-GFP and OPT-GFP were collected using a 600 MHz Bruker Advance II spectrometer equipped with a cryoprobe (NMR core facility, Center for Biotechnology and Interdisciplinary Studies at Rensselaer Polytechnic Institute). Experiments were done at 200 μM protein concentration in a minimal salt buffer (20 mM KH_2PO_4 , 10% D_2O , pH 8.0) at 25°C. Spectra were processed with nmrPipe³⁴ and subsequently analyzed with Sparky (T. D. Goddard and D. G. Kneller, SPARKY 3, University of California, San Francisco, CA). Statistical tests for obtained chemical shifts were performed with IBM SPSS Statistics, Version 21.

Folding kinetics

Purified AT-GFP and OPT-GFP were unfolded by titrating with 0.1 *N* HCl to pH 2.0, then returned to pH 8.0 by addition of 6 *M* guanidine-HCl, 150 mM NaCl, 1 mM DTT, 100 mM HEPES-NaOH, pH 8.0 (denaturing buffer) to a final protein concentration of 10 μM . The denatured protein was incubated at room temperature in denaturing buffer for at least 1 h prior to refolding. Fluorescence recovery upon refolding was measured using a BioLogic SFM-400 Stopped Flow apparatus attached to the Jasco J-815 Spectrometer. The excitation wavelength was set at 480 nm with 10 nm bandwidth and a 495 nm emission filter was used. Refolding was initiated by 11-fold dilution of denatured AT-GFP or OPT-GFP with 150 mM NaCl, 1 mM DTT, 100 mM HEPES-NaOH, pH 8.0 (native buffer). Fluorescence data were collected for 450 s in steps of 0.05 s. Multiphase folding rates and their associated amplitudes were determined by minimizing the squared residual using Excel's Solver function (Microsoft, Version 14.3.8), running in "GRG nonlinear" mode. The time points were logarithmically downsampled to make fitting more efficient. Each trajectory was fit to a sum of exponential decay functions with an optional constant (burst phase). The number of kinetic phases was determined to be the fewest number that produced a fit with an uncorrelated residual ($r < 0.5$) and contained rates that differed by at least one standard deviation. Six replicates were done for all kinetic measurements.

pH jump refolding experiments were carried out as above, except as follows. Purified AT-GFP and OPT-GFP were unfolded by titrating with 0.1 *N* HCl to pH 2.0 and incubating for at least 1 h in the presence of 1 mM DTT. To initiate pH jump experiments 3 μL of 10 μM unfolded protein was added to 3 mL

150 mM NaCl, 1 mM DTT, 100 mM HEPES-NaOH, pH 8.0 in a stirred cuvette while monitoring green fluorescence at excitation/emission wavelength of 485/508 nm in time steps of 5 s. Multiphase fitting was carried out as above, always including a burst phase. To test for the contribution of backbone peptide bond isomerization to folding rates, pH jump experiments were carried out in the presence of 2.1 μM cyclophilin A isomerase (CyPA) (Sigma), following the method of Enoki.¹⁹ Three replicates were performed for each experiment, and each replicate was fit separately.

Unfolding kinetics and thermodynamics

The equilibrium free energy of folding was estimated by measuring the folding and unfolding rates and taking the equilibrium constant to be the ratio of these rates. Determination of the folding rates is described above. Unfolding was measured in triplicate by the loss of green fluorescence at 40, 60, and 80°C, fit to a single exponential decay, using the Excel Solver function (Microsoft, Version 14.3.8). The rate k_u is the derivative of the fit at time zero. Then, a linear fit of $\ln(k_u)$ versus inverse temperature gives k_u at 20°C.

$$\ln k_u = (1/T)(1/R)(\Delta G_u^\ddagger) - \ln v$$

$\Delta G_u^\ddagger$ is the activation barrier of unfolding (sometimes called E_a), R is the gas constant, and v is the prefactor.³⁵ The folding rate k_f was calculated as the amplitude-weighted average of the multiphase folding rates. The folding free energy at 20°C was calculated as $\Delta G_f = -RT \ln(k_f/k_u)$.

Kinetic trapping analysis

The presence of a kinetically trapped misfolded state was probed using the method described by Andrews²¹ except as follows. Both pH refolding and guanidine hydrochloride (GuHCl) dilution experiments were done. Green fluorescence was measured at conditions where the protein was partially unfolded (pH 5.5 or 1.8 *M* GuHCl), starting from fully folded conditions (pH 8.0 or 0 *M* GuHCl) or from fully unfolded conditions (pH 2.0 or 8 *M* GuHCl). Proteins remained at these conditions for at least 96 h before measurement. All buffers contained 1 mM DTT to prevent disulfide bond formation.

Structural bioinformatics

Statistics for loops in proteins was compiled using in-house software on a nonredundant database of globular proteins called PDBselect.³⁶ GFP structural homologs were found using keywords from the Protein Data Bank (PDB) and by sequence search using PSI-BLAST.³⁷ Structures were aligned (sequentially)

using MOE. Nonsequential alignments were done, where necessary, using SCALI_ENREF_38³⁸ GFP residue numbering refers to that of sfGFP (PDB: 2B3P). The corresponding AT-GFP residue numbering differs by two from sfGFP after position 89.

ACKNOWLEDGMENTS

We thank Joachim Jaeger and the Wadsworth Center's Macromolecular Crystallography Core for use of the facilities. Thanks to Donna E. Crone for help with cloning, and to Steven Macari for growing crystals. Funding was provided by the NIH grants 5R21GM088838 and 5R01 GM099827 to C.B.

REFERENCES

1. Tsien RY (1998) The green fluorescent protein. *Ann Rev Biochem* 67:509–544.
2. Zimmer M (2002) Green fluorescent protein (GFP): Applications, structure, and related photophysical behavior. *Chem Rev* 102:759–781.
3. Crone DE, Huang YM, Pitman DJ, Schenkelberg C, Fraser K, Macari S, Bystroff C. GFP-Based biosensors. In: Rinken T, Ed. (2013) *State of the art in biosensors—General aspects*. ISBN: 978-953-51-1004-0. InTech.
4. Zimmer M (2009) GFP: From jellyfish to the Nobel prize and beyond. *Chem Soc Rev* 38:2823.
5. Jackson SE, Craggs TD, Huang J (2006) Understanding the folding of GFP using biophysical techniques. *Expert Rev Proteom* 3:545–559.
6. Fukuda H, Arai M, Kuwajima K (2000) Folding of green fluorescent protein and the cycle3 mutant. *Biochemistry* 39:12025–12032.
7. Yang TT, Cheng L, Kain SR (1996) Optimized codon usage and chromophore mutations provide enhanced sensitivity with the green fluorescent protein. *Nucleic Acids Res* 24:4592–4593.
8. Cramer A, Whitehorn EA, Tate E, Stemmer WPC (1996) Improved green fluorescent protein by molecular evolution using DNA shuffling. *Nat Biotechnol* 14:315–319.
9. Pédelacq J-D, Cabantous S, Tran T, Terwilliger TC, Waldo GS (2006) Engineering and characterization of a superfolder green fluorescent protein. *Nat Biotechnol* 24:79–88.
10. Cabantous S, Terwilliger TC, Waldo GS (2005) Protein tagging and detection with engineered self-assembling fragments of green fluorescent protein. *Nat Biotechnol* 23:102–107.
11. Andrews BT, Schoenfish AR, Roy M, Waldo G, Jennings PA (2007) The rough energy landscape of superfolder GFP is linked to the chromophore. *J Mol Biol* 373:476–490.
12. Andrews BT, Roy M, Jennings PA (2009) Chromophore packing leads to hysteresis in GFP. *J Mol Biol* 392:218–227.
13. Brandts JF, Halvorson HR, Brennan M (1975) Consideration of the possibility that the slow step in protein denaturation reactions is due to *cis-trans* isomerism of proline residues. *Biochemistry* 14:4953–4963.
14. Wedemeyer WJ, Welker E, Scheraga HA (2002) Proline *cis-trans* isomerization and protein folding. *Biochemistry* 41:14637–14644.
15. Kiefhaber T, Grunert HP, Hahn U, Schmid FX (1990) Replacement of a *cis* proline simplifies the mechanism of ribonuclease T1 folding. *Biochemistry* 29:6475–6480.
16. Schultz Da, Schmid FX, Baldwin RL (1992) *Cis* proline mutants of ribonuclease A. II. Elimination of the slow-folding forms by mutation. *Protein Sci* 1:917–924.
17. Odeley C, Mayr L, Schmid F (1995) Non-prolyl *cis-trans* peptide bond isomerization as a rate-determining step in protein unfolding and refolding. *J Mol Biol* 245:69–78.
18. Eyles SJ, Gierasch LM (2000) Multiple roles of prolyl residues in structure and folding. *J Mol Biol* 301:737–747.
19. Enoki S, Saeki K, Maki K, Kuwajima K (2004) Acid denaturation and refolding of green fluorescent protein. *Biochemistry* 43:14238–14248.
20. Steiner T, Hess P, Bae JH, Wiltzsch B, Moroder L, Budisa N (2008) Synthetic biology of proteins: Tuning GFPs folding and stability with fluoroproline. *PloS One* 3:e1680.
21. Andrews BT, Gosavi S, Finke JM, Onuchic JN, Jennings PA (2008) The dual-basin landscape in GFP folding. *Proc Natl Acad Sci USA* 105:12283–12288.
22. Labute P (2008) The generalized Born/volume integral implicit solvent model: Estimation of the free energy of hydration using London dispersion instead of atomic surface area. *J Comput Chem* 29:1693–1698.
23. Wang J, Cieplak P, Kollman PA (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J Comp Chem* 21:1049–1074.
24. Huang YM, Bystroff C (2012) Exploring objective functions and cross-terms in the optimization of an energy function for protein design. New York: ACM Press, pp 155–162.
25. Samarkina ON, Popova AG, Gvozdkik EY, Chkalina AV, Zvyagin IV, Rylova YV, Rudenko NV, Lusta Ka, Kelmanson IV, Gorokhovatsky AY, Vinokurov LM (2009) Universal and rapid method for purification of GFP-like proteins by the ethanol extraction. *Protein Exp Purif* 65:108–113.
26. Luft JR, Collins RJ, Fehrman NA, Lauricella AM, Veatch CK, DeTitta GT (2003) A deliberate approach to screening for initial crystallization conditions of biological macromolecules. *J Struct Biol* 142:170–179.
27. Battye TG, Kontogiannis L, Johnson O, Powell HR, Leslie AG (2011) iMOSFLM: A new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallogr D* 67:271–281.
28. 1994) The CCP4 suite: Programs for protein crystallography. *Acta Crystallogr D* 50:760–763.
29. Vagin A, Teplyakov A (2010) Molecular replacement with MOLREP. *Acta Crystallogr D* 66:22–25.
30. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL (1998) Crystallography and NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D* 54:905–921.
31. Brunger AT (2007) Version 1.2 of the crystallography and NMR system. *Nat Protocols* 2:2728–2733.
32. Emsley P, Cowtan K (2004) Coot: Model-building tools for molecular graphics. *Acta Crystallogr D* 60:2126–2132.
33. Chen VB, Arendall WB, 3rd, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC (2010) MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallogr D* 66:12–21.
34. Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A (1995) NMRPipe: A multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* 6:277–293.

35. Wigner E (1938) The transition state method. *Trans Faraday Soc* 34:29–41.
36. Griep S, Hobohm U (2010) PDBselect 1992–2009 and PDBfilter-select. *Nucleic Acids Res* 38:D318–D319.
37. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25: 3389–3402.
38. Yuan X, Bystroff C (2005) Non-sequential structure-based alignments reveal topology-independent core packing arrangements in proteins. *Bioinformatics* 21: 1010–1019.
39. Huang YM, Bystroff C (2009) Complementation and reconstitution of fluorescence from circularly permuted and truncated green fluorescent protein. *Biochemistry* 48:929–940.
40. Voityuk AA, Michel-Beyerle ME, Rosch N (1998) Quantum chemical modeling of structure and absorption spectra of the chromophore in green fluorescent proteins. *Chem Phys* 231:13–25.
41. Chattoraj M, King BA, Bublitz GU, Boxer SG (1996) Ultra-fast excited state dynamics in green fluorescent protein: Multiple states and proton transfer. *Proc Natl Acad Sci USA* 93:8362–8367.
42. Bystroff C, Baker D (1998) Prediction of local structure in proteins using a library of sequence–structure motifs. *J Mol Biol* 281:565–577.
43. Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations. *J Mol Biol* 7:95–99.
44. Kaya H, Chan HS (2003) Origins of chevron rollovers in non-two-state protein folding kinetics. *Phys Rev Lett* 90:258104.