JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

Communication

pubs.acs.org/JACS

GFP Variants with Alternative β-Strands and Their Application as Light-driven Protease Sensors: A Tale of Two Tails

Keunbong Do* and Steven G. Boxer*

Department of Chemistry, Stanford University, Stanford, California 94305-5012, United States

Supporting Information

ABSTRACT: Green fluorescent protein (GFP) variants that carry one extra strand 10 (s10) were created and characterized, and their possible applications were explored. These proteins can fold with either one or the other s10, and the ratio of the two folded forms, unambiguously distinguished by their resulting colors, can be systematically modulated by mutating the residues on s10 or by changing the lengths of the two inserted linker sequences that connect each s10 to the rest of the protein. We have discovered robust empirical rules that accurately predict the product ratios of any given construct in both bacterial and mammalian expressions. Exploiting earlier studies on photodissociation of cut s10 from GFP (Do and Boxer, 2011), ratiometric protease sensors were designed from the construct by engineering a specific protease cleavage site into one of the inserted loops, where the bound s10 is replaced by the other strand upon protease cleavage and irradiation with light to switch its color. Since the conversion involves a large spectral shift, these genetically encoded sensors display a very high dynamic range. Further engineering of this class of proteins guided by mechanistic understanding of the light-driven process will enable interesting and useful application of the protein.

Green fluorescent protein (GFP) and its color variants are widely used as genetically encoded fluorescent reporters for cellular imaging and sensing.[1−5] To further extend the versatility of these proteins, we describe the design of a series of unusual GFP variants that carry two strand 10 (s10) sequences,[6] one at the N- and the other at the C-terminal end of a circularly permuted GFP, such that the resulting folded protein can have either green or yellow fluorescence depending on which s10 it binds to (Figure 1).[7−10] One way of achieving the spectral distinction between the two folded forms is to place 203T on one strand, which corresponds to the sequence of the wild type GFP, and 203Y on the other, which is the key mutation that generates a class of yellow fluorescent proteins (YFPs).[11,12] Since there are two strands carrying different residues at position 203 that quite significantly affects the absorption and fluorescence of the protein (e.g., Figure 2A, 3B, and S1, Supporting Information), a two-letter notation will be used in this communication to indicate the two 203 residues in a construct from N- to C-terminus. For instance, the construct given as an example in Figure 1 will be denoted as "TY", which means that the N-terminal 203 residue is threonine and the C-terminal 203 residue is tyrosine; the reverse is denoted "YT".

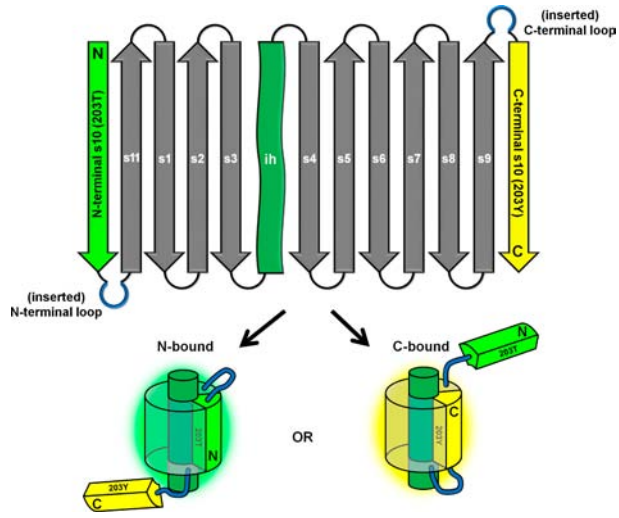

Figure 1. Schematic illustrations of the primary topology string of the protein (top) and the two possible forms of the folded protein (bottom). The inner helix containing the amino acids that become the chromophore is denoted as ih (top) and illustrated as a green cylinder surrounded by the 11 β-strands (bottom). The identity of the amino acid at position 203 determines whether the folded protein exhibits GFP or YFP fluorescence, and this is illustrated schematically by coloring s10 and the halo of the protein green or yellow, respectively. The N- and the C-terminal loops which are color coded in blue indicate the extra sequence inserted within the native loops.[14] In the absence of any bound s10, the fluorescence intensity is greatly reduced.[13] All sequences, expression conditions, and spectral properties of the proteins can be found in SI 1−3 and 5.

When the protein with two alternative s10s is expressed in *E. coli*, the product is found to be a mixture of the two bound forms, one binding to the N- and the other to the C-terminal s10. As shown in Figure 2A, the absorption spectrum of the mixture can be fit by a linear combination of the two basis spectra taken as described in the first section of the Supporting Information (SI 1), to accurately determine the composition of the mixture. Interestingly, the two bound forms are separable by anion-exchange chromatography, and each purified form has spectral properties indistinguishable from those of the corresponding GFP or YFP analogs (with just 11 β-strands) irrespective of the lengths of the added loops (see SI 4). Once

Received: April 15, 2013
Published: July 2, 2013

ACS Publications © 2013 American Chemical Society

dx.doi.org/10.1021/ja4037274 | J. Am. Chem. Soc. 2013, 135, 10226−10229

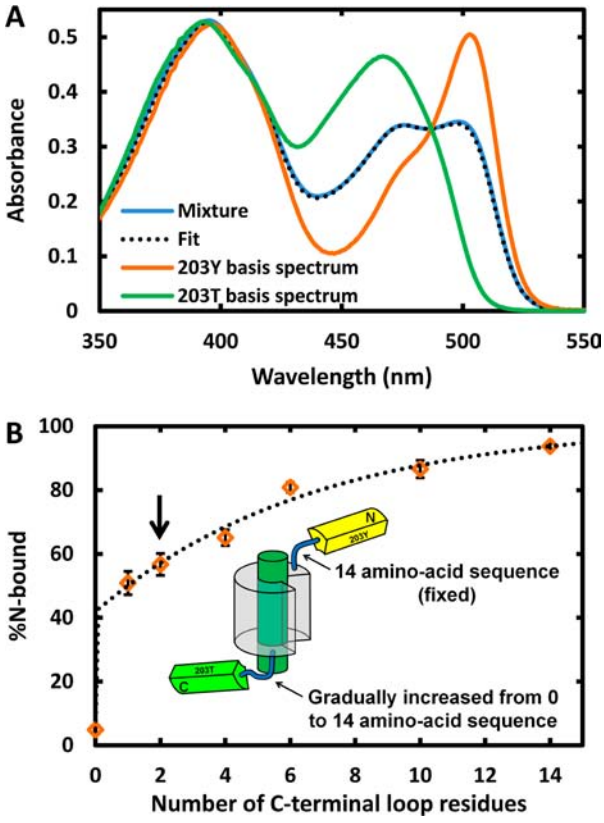

Figure 2. Determining and controlling the relative populations of the two bound forms. (A) The isolated mixtures of GFPs expressed in *E. coli* at 23 °C with alternative strands can be analyzed by fitting their absorption spectra by a linear combination of the YFP and the GFP basis spectra (pH 8.0).[18] This particular sample was a YT construct whose N-terminal loop is GSSGSGSSGSGSSG and C-terminal loop is GS, which corresponds to the plot in panel B for two C-terminal loop residues (pointed with a vertical arrow in panel B). The two basis spectra are normalized at the isosbestic point around 487 nm and shown in orange and green solid lines. The linear combination gives a 58:42 ratio of YFP to GFP in the expression mixture. These populations can be fully separated (see SI 4). (B) Relative N-bound population of YT constructs as a function of the number of C-terminal loop residues.[14] Data were collected from three independent expressions in *E. coli*, and the standard deviation is used for the error bars. The N-terminal loop is fixed as GSSGSGSSGSGSSG in all constructs, and the C-terminal loop is mostly a repeat of G and S (see SI 2 for full sequences). The dotted line is drawn as a guide to the eye.

separated, the absorption and fluorescence spectra do not change over many days at room temperature, indicating that internal s10 exchange is extremely slow (see SI 6).

The relative populations of the two folded forms as expressed are found to be systematically modulated by asymmetrically changing s10 residues. For instance, if the inserted loops have similar sequence and length, $s10^{203Y}$ binds to the remainder of the protein more favorably than $s10^{203T}$ in general. As an example, a YT construct expresses as approximately 94% N-bound, and a TY construct expresses as approximately 92% C-bound when the two inserted loops both consist of 14 amino-acid residues. Furthermore, a destabilizing mutation such as K209Q on the N-terminal s10 of the YT construct lowers the isolated N-bound population down to around 50%.

The relative populations of the two bound forms in the expression mixture can also be modulated by varying the number of residues on the two inserted loops. As schematically described by the inset cartoon in Figure 2B, YT constructs with various C-terminal loop lengths were prepared, while the N-terminal loop was fixed to a 14-amino-acid-long sequence. Interestingly, the N-bound population is around 5% (i.e., 95% C-bound) when there is no inserted loop on the C-terminal side, but insertion of even a single residue (glycine) increases the relative population to around 51% (i.e., 49% C-bound). If the composition of an expression mixture reflects that of the system at thermodynamic equilibrium during some stage of protein folding and chromophore maturation, the large change in the relative population caused by this single-residue insertion suggests that most of the favorable interaction that exists within the native loop structure is lost even by the slightest perturbation.[14] As the number of C-terminal loop residues is further increased, the relative N-bound population also increases. Compared to the first single insertion, the following insertions show a smaller but systematic impact per added residue on the relative population. Based on a simple analysis of the trend, the composition of a given construct can be accurately predicted (see SI 7).

When the proteins are denatured in guanidine hydrochloride solution and refolded, the newly set composition is very different from the composition of the expression mixture (see SI 8). This suggests that the free energy landscape of protein folding becomes very different when a fully mature chromophore is present compared to when the nascent protein folds immediately after synthesis from the ribosome prior to chromophore maturation.[15−17] In other words, the composition of the expression mixture might reflect that of the system at thermodynamic equilibrium during a certain stage of chromophore maturation, but it does not reflect that of the system at equilibrium with the fully mature chromophore. With this distinction in mind, the system can provide a well-defined unimolecular binding assay with convenient optical readout to analyze the differential thermodynamic contribution of each residue and the loop length, and the overall effect of the circular permutation with or without the mature chromophore; a more complete description and thermodynamic analysis of the system will be reported separately.

Irrespective of the detailed physical origin, it is notable that the composition of the expression mixture can be predicted very accurately when the lengths of the loops and the types of residues on the two terminal strands are specified (see SI 7). Furthermore, we have found that the rules are independent of cell type. Several of the constructs were expressed in mammalian cells (HEK293 and U205), where the ratios of the two bound forms were identical with those estimated for the expression mixture from *E. coli* (i.e., as in Figure 2B; see SI 9). This shows that folding of these GFP variants is largely independent of the expression machinery specific to each cell type and that the same rules apply for both prokaryotes and eukaryotes concerning the composition of the expression mixture. Such general predictability is especially useful to initialize the protein population (likely as all in one or the other color) in cell-based assays or to use it as a tool for modulating interactions in cells.

Since the competing strands are at opposite ends of the protein, we hypothesized that cotranslational folding could favor binding of the N-terminal strand while the C-terminal strand is still in the ribosome tunnel. To test this, a delay sequence of five leucines was introduced at the C-terminus of the protein. Two of these constructs were made, which were identical in every way except that one used a repeat of the most

dx.doi.org/10.1021/ja4037274 | J. Am. Chem. Soc. 2013, 135, 10226−10229

common leucine codon in the delay sequence while the other used the least common codon in it. However, when the proteins were expressed in *E. coli*, no noticeable difference was found in the expression compositions of the two expression mixtures (see SI 10). This suggests that the delay is short relative to the time it takes for the protein to become kinetically trapped in the N-bound form. Note that the construct, with its unambiguous bimodal folding and clear optical readout would serve as an ideal model system to investigate the effect of synonymous codons in protein folding and to study the principles of other cotranslational folding processes in general.[19−21]

One possible application of these alternative β-strand GFPs is to create a novel type of protease sensor exploiting the photodissociation phenomenon described in our earlier work, where the dissociation rate of cut s10 could be enhanced dramatically with light irradiation.[13] As illustrated schematically in Figure 3A, if the loop connected directly to the bound s10 is engineered to contain a proteolytic site, upon being cut by the protease, the cut strand can be irreversibly photodissociated and replaced by the other strand that causes the color to shift. This can be readily adapted as a selective (by choice of loop sequence) and genetically encoded protease sensor with a large dynamic range. Figure 3B shows the implementation of a prototypical thrombin sensor, where the thrombin cleavage site (LVPRGS sequence) is inserted into the loop directly connected to the C-terminal s10 in a TY construct. The protein expresses in *E. coli* with over 90% C-bound (i.e., spectrally 90% YFP). In the presence of active thrombin and light, the spectrum rapidly shifts from that of YFP to that of GFP as shown in Figure 3B. The half-life of the conversion process is 5 min from a single exponential fit as shown in Figure 3C (see the caption of Figure 3 and SI 11 for specifications). The sensor can be adapted to any protease simply by inserting the appropriate recognition sequence into the loop; for example, we obtained similar results with trypsin and caspase (see SI 11). Most importantly, since the combined presence of protease activity and light results in a conversion of YFP to GFP or vice versa, a very large ratiometric dynamic range is achievable. For instance, in the constructs used in this communication, GFP emits 40 times stronger than YFP at 490 nm when excited with 440 nm light, while YFP emits 60 times stronger than GFP at 530 nm when excited with 515 nm light. This simple consideration gives a ratiometric dynamic range of over 2,000 fold. Even if the detector emission wavelength is fixed, for example to 530 nm, and excitations at 440 and 515 nm are compared, the contrast is over 100 fold, which is a much higher dynamic range than conventional genetically encoded FRET sensors with dynamic range around 2−4 fold.[5] Lastly, because light is required for the conversion, protease activity detection and the release of the cut peptide can be spatially and temporally controlled.

The very flat control baseline in Figure 3C implies that the light-driven conversion from one bound form to the other for this construct is a process with very low quantum yield when the covalent bonds are intact. The contrast shown between the cut and the uncut protein stems in part from the molecularity of the corresponding reactions, where the former involves two fragments that are driven apart whereas the latter undergoes a process that is strictly intramolecular. Such stability of the uncut species against light irradiation can be advantageous for certain applications, for instance a protease sensor with low background. On the other hand, it is observed that for certain

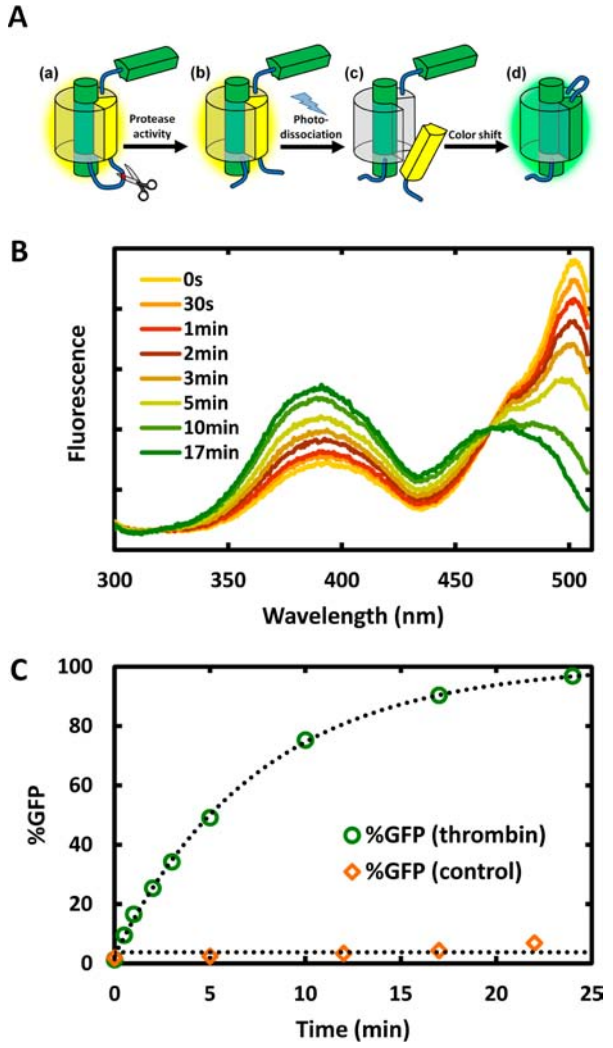

Figure 3. (A) Schematic illustration of the ratiometric protease sensor. When the protease cuts the cleavage site inserted in the loop connected to the bound s10, the cut s10 remains associated with the protein until it is photodissociated.[13] Photodissociation is followed by irreversible intramolecular replacement by the alternative s10, which results in the color shift by virtue of the residue at position 203. (B) After the protease sensor was incubated with 20 units mL^{-1} thrombin for 10 min, excitation spectra (emission collected at 520 nm) were taken at various times for a 3-mL sample of 2 μM thrombin sensor at 35 °C irradiated with 13 mW of 405 nm cw diode laser light with a 3 cm path length (see SI 2 and SI 11 for full sequence of the protein and more detailed procedure). (C) The percentage of GFP at each time point was spectrally estimated (as in Figure 2A) and fit by a single exponential function. The control and the thrombin samples were prepared in exactly the same way except that thrombin (Plasminogen-Free, Bovine, EMD Millipore) was added to a concentration of 20 units mL^{-1} 10 min prior to light irradiation. The control sample was exposed to the same light and temperature conditions, and the spectrum barely changed over 20 min. The spectrum does not change within the measurement time if the cut protein is left in the dark.

constructs the light-driven intramolecular strand swap is more feasible (see SI 12). It should be possible to engineer variants with higher quantum yields for the light-driven intramolecular strand swap, so that the protein can reversibly photoswitch between the two bound forms (as illustrated in Figure 1), changing its color and the binding partner (and anything attached to the binding partner) upon light irradiation. To

dx.doi.org/10.1021/ja4037274 | J. Am. Chem. Soc. 2013, 135, 10226−10229

achieve this, it might be necessary to introduce destabilizing mutations along s10, and deeper mechanistic understanding of how the chromophore excitation couples to the association and dissociation of the strand would provide useful guidance.

In summary, we designed and expressed GFP variants with one extra s10. The relative populations of the two bound forms were determined by the residues on s10 as well as by the lengths of the loop sequences connecting the two strands to the rest of the protein, and the composition of the expression mixture could be accurately predicted for a given construct. The composition of the bound forms was independent of the observed cell types where *E. coli*, HEK293, and U205 cells were used for expressions. With the unambiguous optical readout to estimate the result of its bimodal folding, the construct can potentially serve as an ideal model system to study alternate frame folding and cotranslational folding, as well as individual amino acid and loop contributions to protein folding energetics in a general and quantitative way. Finally, a prototype of a genetically encoded ratiometric protease sensor was designed from the construct and demonstrated to have very large dynamic range.

Many concepts and applications beyond the protease sensor are suggested by the results reported here. For instance, it should be straightforward to incorporate a peptide, a binding domain, or a target sequence (e.g., a phosphorylation site) into one of the two strands such that the strand no longer binds to the rest of the GFP when the attached sequence binds to a target molecule or when the incorporated target sequence is modified. In this way the presence of the corresponding binding partner or the activity of the modifier (e.g., a kinase) could be monitored by the color shift when the other strand binds to the rest of the GFP. Furthermore, the construct can be developed into genetically encoded and light-addressable modulator of access to the active site of an enzyme or of protein−protein interactions to control enzymatic activity or localization of molecules with light.[22−24] Better understanding of the underlying physical mechanisms will be essential to guide further development.

■ ASSOCIATED CONTENT

Supporting Information

Protein sequences, expression and purification methods, instrumentation, and other procedures are detailed. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

kbd0810@gmail.com; sboxer@stanford.edu

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Luke Oltrogge and Professor Michael Lin for many helpful discussions and comments, and Dr. Jiwon Hwang in the Kopito lab at Stanford for help with mammalian cell culture. This research was supported in part by a grant from the NIH (GM27738). K.D. is supported by a John Stauffer Stanford Graduate Fellowship and the Korea Foundation for Advanced Studies.

■ REFERENCES

(1) Tsien, R. Y. *Annu. Rev. Biochem.* **1998**, *67*, 509−44.
(2) Chudakov, D. M.; Lukyanov, S.; Lukyanov, K. A. *Trends Biotechnol.* **2005**, *23*, 605−613.
(3) Wiedenmann, J.; Oswald, F.; Nienhaus, G. U. *Life* **2009**, *61*, 1029−1042.
(4) Miyawaki, A. *Curr. Opin. Neurobiol.* **2003**, *13*, 591−596.
(5) Nagai, T.; Yamada, S.; Tominaga, T.; Ichikawa, M.; Miyawaki, A. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 10554−10559.
(6) In this communication, the numbering of strands and residues follows that of the original GFP crystal structure entries (PDB ID: 1EMA, 1YFP), see refs 11 and 12.
(7) This is reminiscent of a more general idea of alternate frame folding; see refs 8−10.
(8) Stratton, M. M.; Mitrea, D. M.; Loh, S. N. *Chem. Biol.* **2008**, *3*, 723−732.
(9) Mitrea, D. M.; Parsons, L. S.; Loh, S. N. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 2824−2829.
(10) Stratton, M. M.; Loh, S. N. *Protein Sci.* **2011**, *20*, 19−29.
(11) Ormo, M.; Cubitt, A. B.; Kallio, K.; Gross, L. A.; Tsien, R. Y.; Remington, S. J. *Science* **1996**, *273*, 1392−1395.
(12) Wachter, R. M.; Elsliger, M.; Kallio, K.; Hanson, G. T.; Remington, S. J. *Structure* **1998**, *6*, 1267−1277.
(13) Do, K.; Boxer, S. G. *J. Am. Chem. Soc.* **2011**, *133*, 18078−18081.
(14) In this communication, the term "loop" will be used generally to indicate the inserted linkers. This is not to be confused with the loop structure already present in the native GFP structure unless explicitly mentioned. See SI 1, 2 for all sequences.
(15) Reid, B. G.; Flynn, G. C. *Biochemistry* **1997**, *36*, 6786−6791.
(16) Andrews, B. T.; Schoenfish, A. R.; Roy, M.; Waldo, G.; Jennings, P. A. *J. Mol. Biol.* **2007**, *373*, 476−490.
(17) Andrews, B. T.; Gosavi, S.; Finke, J. M.; Onuchic, J. N.; Jennings, P. A. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 12283−12288.
(18) Composition can be estimated using fluorescence excitation spectra in a similar manner.
(19) Ugrinov, K. G.; Clark, P. L. *Biophys. J.* **2010**, *98*, 1312−1320.
(20) Cortazzo, P.; Cerveñansky, C.; Marín, M.; Reiss, C.; Ehrlich, R.; Deana, A. *Biochem. Biophys. Res. Commun.* **2002**, *293*, 537−541.
(21) Komar, A. A. *Trends Biochem. Sci.* **2009**, *34*, 16−24.
(22) Wu, Y. I.; Frey, D.; Lungu, O. I.; Jaehrig, A.; Schlichting, I.; Kuhlman, B.; Hahn, K. M. *Nature* **2009**, *461*, 104−108.
(23) Levskaya, A.; Weiner, O. D.; Lim, W. A.; Voigt, C. A. *Nature* **2009**, *461*, 997−1001.
(24) Zhou, X. X.; Chung, H. K.; Lam, A. J.; Lin, M. Z. *Science* **2012**, *338*, 810−814.