J|A|C|S 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".

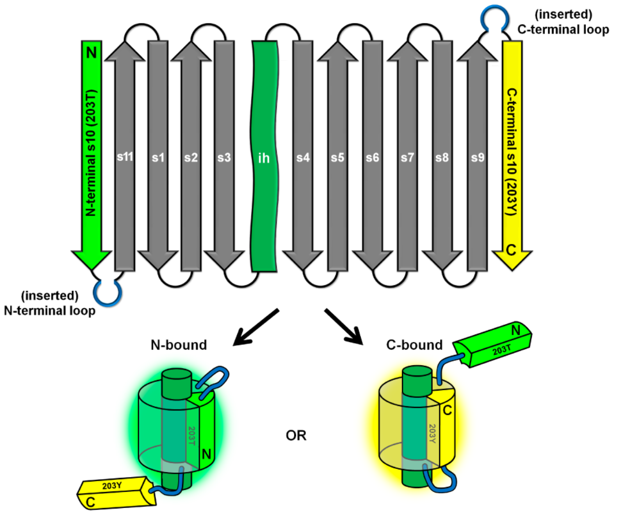

Figure 1. Schematic illustrations of the primary topology string of the protein (top) and the two possible forms of the folded protein (bottom). The inner helix containing the amino acids that become the chromophore is denoted as ih (top) and illustrated as a green cylinder surrounded by the 11 β-strands (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

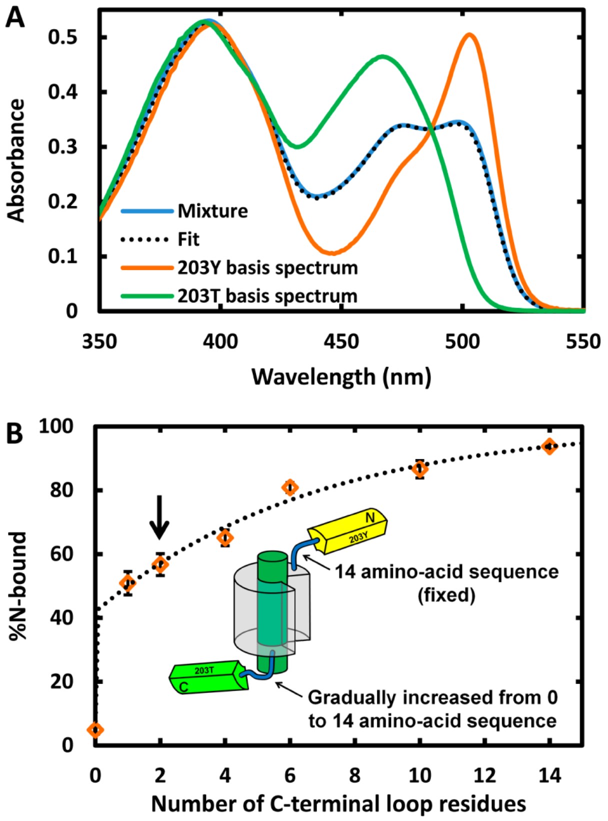

Figure 2. Determining and controlling the relative populations of the two bound forms. (A) The isolated mixtures of GFPs expressed in *E. coli* at 23 °C with alternative strands can be analyzed by fitting their absorption spectra by a linear combination of the YFP and the GFP basis spectra (pH 8.0).[18] 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 U2OS), 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 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.

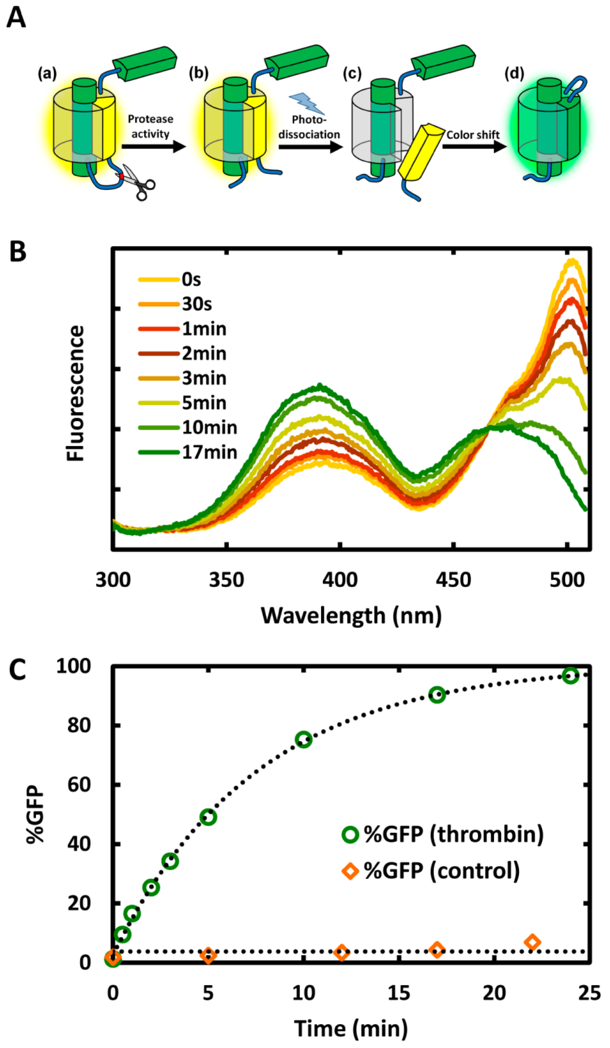

Figure 3. (A) Schematic illustration of the ratiometric protease sensor. When the protease cuts the cleavage site inserted in the loop connected to the bound s10, the cut s10 remains associated with the protein until it is photodissociated.[13] Photodissociation is followed by irreversible intramolecular 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.

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 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 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.

Supporting information for:
GFP Variants with Alternative β-strands and Application as a Light-driven Protease Sensor: A Tale of Two Tails

Keunbong Do* and Steven G. Boxer*
Department of Chemistry, Stanford University, Stanford, California 94305-5012, United States

Contents:

1. Protein sequences: for basis spectra

In this communication, the proteins were designed based on the sequence of GFP1-10OPT and GFP11M3,[1] which were derived from superfolder GFP.[2] Note also that the chromophore is formed from SYG sequence, which results in higher pKa than the chromophore formed from TYG sequence.

To obtain basis spectra and to find their isosbestic points, truncated GFP (~~s10:loop~~:GFP as described in our previous work[3]) was prepared, aliquoted into three, and excess s10 peptides with 203T, 203Y, or 203H were added respectively. The primary sequence of the three complexes are given below (from N- to C-terminus) where the added peptides are underlined, and their absorbance spectra is shown in Figure S1. Since the three complexes are prepared to have exactly the same concentrations, points where two spectra cross in Figure S1 (487 nm between 203T and 203Y, 421 nm between 203T and 203H, and 425 nm between 203Y and 203H) can be considered as isosbestic points.

Protein for the 203T basis spectra:

LPDNHYLS**T**QTVLSKDPNE·

HSGSGSKRDHMVLHEYVNAAGITHGMDELYGGTGGSASQGEELFTGVVPILVELDGDVNGHKFSVRGEGE
GDATIGKLTLKFISTTGKLPVPWPTLVTTLSYGVQAFSRYPDHMKRHDFFKSAMPEGYVQERTISFKDDGKY
KTRAVVKFEGDTLVNRIELKGTDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFTVRHNVEDGSVQ
LADHYQQNTPIGDGPVL

Protein for the 203Y basis spectra:

LPDNHYLS**Y**QTVLSKDPNE·

HSGSGSKRDHMVLHEYVNAAGITHGMDELYGGTGGSASQGEELFTGVVPILVELDGDVNGHKFSVRGEGE
GDATIGKLTLKFISTTGKLPVPWPTLVTTLSYGVQAFSRYPDHMKRHDFFKSAMPEGYVQERTISFKDDGKY
KTRAVVKFEGDTLVNRIELKGTDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFTVRHNVEDGSVQ
LADHYQQNTPIGDGPVL

Protein for the 203H basis spectra:

LPDNHYLS**H**QTVLSKDPNE·

HSGSGSKRDHMVLHEYVNAAGITHGMDELYGGTGGSASQGEELFTGVVPILVELDGDVNGHKFSVRGEGE
GDATIGKLTLKFISTTGKLPVPWPTLVTTLSYGVQAFSRYPDHMKRHDFFKSAMPEGYVQERTISFKDDGKY
KTRAVVKFEGDTLVNRIELKGTDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFTVRHNVEDGSVQ
LADHYQQNTPIGDGPVL

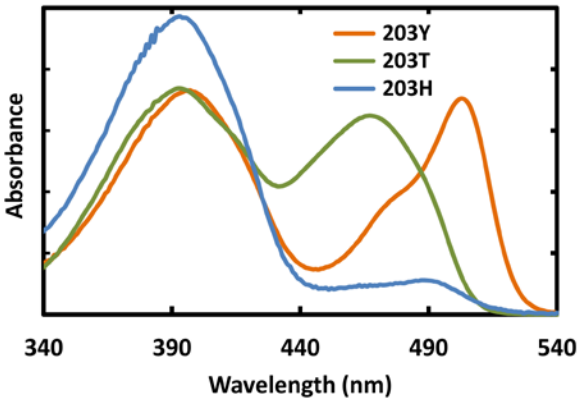

Figure S1. Absorbance spectra of GFP complexes with 203Y, 203T, and 203H at pH 8.0.

2. Protein sequences: variants with alternative strands

Primary sequence from N- to C-terminus.

Proteins for loop-dependent population control (Figure 2, S2, S3, S4, and S5):
MGHHHHHHSSGGLPDNHYLS**203Y**QTVLSKDPNE**GSSGSGSSGSGSSG**KRDHMVLHEYVNAAGITHGMDELYGGTGGSASQGEELFTGVVPILVELDGDVNGHKFSVRGEGEGDATIGKLTLKFISTTGKLPVPWPTLVTTLSYGVQAFSRYPDHMKRHDFFKSAMPEGYVQERTISFKDDGKYKTRAVVKFEGDTLVNRIELKGTDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFTVRHNVEDGSVQLADHYQQNTPIGDGPVL(**C-terminal loop**)LPDNHYLS**203T**QTVLSKDPNE

Amino acid sequences that correspond to s10 are underlined, and amino acids that correspond to the 203 residue are bolded with 203 written in front (note this is a YT construct). The N-terminal loop sequence (GSSGSGSSGSGSSG) is bolded, and the sequences for the C-terminal loop noted as **(C-terminal loop)** are given in Table S1 below for various constructs.

Number of inserted residues	Sequence (C-terminal loop)
0	(none)
1	G
2	GS
4	GSGS
6	GSGSGS
10	GSGSGSGSGS
14	GSSGSRGSDGSDGS

Table S1. Sequences of C-terminal loops. All YT constructs discussed in this communication share the common sequence as given above and the C-terminal loop is varied as given in this table.

Protease sensors (Figure 3 and S6):
MGHHHHHHSSGGLPDNHYLS**203T**QTVLSKDPNE**GSSGSGSSGSGSSG**KRDHMVLHEYVNAAGITHGMDELYGGTGGSASQGEELFTGVVPILVELDGDVNGHKFSVRGEGEGDATIGKLTLKFISTTGKLPVPWPTLVTTLSYGVQAFSRYPDHMKRHDFFKSAMPEGYVQERTISFKDDGKYKTRAVVKFEGDTLVNRIELKGTDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFTVRHNVEDGSVQLADHYQQNTPIGDGPVL**(C-terminal loop)** LPDNHYLS**203Y**QTVLSKDPNE

Amino acid sequences that correspond to s10 are underlined, and amino acids that correspond to the 203 residue are bolded with 203 written in front (note this is a TY construct). The N-terminal loop sequence (GSSGSGSSGSGSSG) is bolded, and the sequences for the C-terminal loop noted as **(C-terminal loop)** are given in Table S2 below for various types of sensors. The sequence known to be recognized by the protease is bolded, and the actual cut site is marked with a wedge (▼).

Sensor target	Sequence (C-terminal loop)
Trypsin activity	GSSGS**R▼**GSDGSDGS
Thrombin activity	GSG**LVPR▼GS**DG
Caspase-3 activity	GSGSG**DEVD▼**GSGSG

Table S2. Sequences of C-terminal loops. All TY constructs, including the protease sensors, discussed in this communication shares the common sequence as given above and the C-terminal loop is varied as given in this table.

3. Protein expression and purification in *E. coli*

One Shot® BL21(DE3) chemically competent *E. coli* (Invitrogen) were used to express proteins from the pET-15b vector. 5 mL of LB media (Broth Miller, EMD) with 100 mg·L^{-1} ampicillin was innoculated with a single colony and grown at 37^{o}C overnight with shaking at 160 rpm. This start-up culture was then poured directly into 1 L of LB solution with 100 mg·L^{-1} ampicillin and 0.25 g·L^{-1} IPTG, and incubated at 23^{o}C for approximately 24 hours with shaking at 180 rpm. The cells were spun down, and the resulting pellet was resuspended in lysis buffer (50 mM HEPES, 300 mM NaCl, and 10v% glycerol at pH 8.0) and lysed with a homogenizer. The cell lysate was spun down, and the supernatant was poured onto a Ni-NTA column equilibrated with the lysis buffer. Two column volumes of lysis buffer containing 20 mM imidazole was used for washing, and the same buffer with 200 mM imidazole was used to elute the protein. The eluate was further purified with anion-exchange chromatography (HiTrap^{TM} 5 mL Q HP, GE) as described below in SI 4.

4. Separation through anion-exchange chromatography

All expressed GFP variants discussed in this communication can be separated into two populations through anion-exchange chromatography. Based on spectral characterization of the two populations, the early eluting peak was found to be in the C-bound form, and the later eluting peak to be in the N-bound form of the folded protein in all observed cases. As an example, Figure S2A shows the chromatogram acquired by running a YT construct with a GS sequence inserted as the C-terminal loop (see above for full sequence of the protein) through an anion exchange column. From the chromatogram, it can be seen that there are two distinct populations eluting at different volumes; one peaked around 60 mL and the other around 73 mL in this specific case, where 216 mM and 253 mM NaCl concentration was reached respectively while maintaining the pH at 8.0 throughout. As shown in Figure S2B, the absorbance at 400 nm could be fit by summation of two Gaussian functions, and the ratio of the populations could be estimated as roughly 40:60 from the fit. This is similar to the ratio 42:58 which was estimated by fitting the absorbance spectrum of the collected mixture with the linear combination of the basis spectra as shown in Figure 2A in the main text. The ratio estimated from the linear fit of the spectra was used for all quantitative analyses in this communication to avoid the artifacts that could be introduced when highly overlapping peaks are fit by summation of multiple Gaussian functions. Finally, when fractions from the anion-exchange chromatography were taken as bracketed by two pairs of dotted lines in Figure S2A (fraction #1; 51-56 mL, fraction #2; 81-86 mL), the absorbance spectrum of each fraction was nearly indistinguishable from that of GFP or YFP, respectively, as shown in Figure S2C. This demonstrates that the first peak in the anion-exchange chromatogram corresponds to the population of the YT construct bound to its C-terminal s10, and the second peak to its N-terminal s10. We can speculate on one possible factor that enables the separation of the two bound forms, which is the differential exposure of the hexa-histidine tag at the N-terminal end. When the protein is in the C-bound form, the hexa-histidine tag, which would be positively charged, can interact with the positively-charged anion-exchange resins more freely because the whole s10 to which the tag is bound is also free. This will result in faster elution, whereas the protein in the N-bound form will have the hexa-histidine tag pinned down closer to the protein, resulting in less interaction between the tag and the resins and in turn in slower elution.

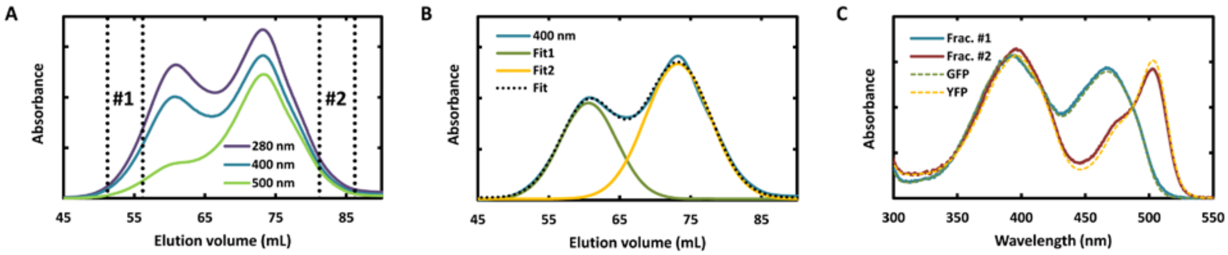

Figure S2. (A) Anion-exchange chromatogram of a YT construct with a GSSGSGSSGSGSSG sequence inserted as the N-terminal loop and GS as the C-terminal loop. The column was equilibrated with pH 8.0 buffer to have NaCl concentration of 108 mM and 5 mL·min^{-1} flow was applied while increasing the NaCl concentration by 1.96 mM per 1 mL flow and maintaining the pH at 8.0. (B) Absorbance at 400 nm was fit by addition of two Gaussian functions. (C) Absorbance spectra of the two fractions taken from anion-exchange chromatography. The corresponding fractions are bracketed by dotted lines in the above chromatogram in Figure S2A, where the first fraction was taken from 51 to 56 mL and the second from 81 to 86 mL. All spectra including the basis spectra of pure GFP and YFP are normalized at 487 nm, which is the known isosbestic point. This shows that the pKa of the chromophore is not affected by the loop lengths or by the presence of one extra s10. Also note that the native amino acid (serine) is present at position 65 so that appreciable populations of both the A and B state[5] are present (in contrast to the variants derived from the S65T mutation).

5. Spectral characterization

As described above in SI 4, the two bound forms of a YT construct with the N-terminal loop sequence of GSSGSGSSGSGSSG and the C-terminal loop sequence of GS can be separated through anion-exchange chromatography, and spectral properties of each bound form at pH 8.0 are summarized in Table S3 (see Figure S2C for absorbance spectra). For extinction coefficient estimation, absorption at 447 nm was compared after denaturing the proteins in 0.1 M NaOH,[4] and for fluorescence quantum yield measurement, fluorescein from the Reference Dye Sampler Kit (R-14782, Molecular Probes®) was used. The C-terminally bound fraction (indicated as "GFP fraction" in Table S3) was 96% C-bound and the N-terminally bound fraction (indicated as "YFP fraction" in Table S3) was 94% N-bound according to the basis spectra fit.

	$\lambda_{abs,A}$ (ε)	$\lambda_{em,A}$ (Φ)	$\lambda_{abs,B}$ (ε)	$\lambda_{em,B}$ (Φ)
YT C-loop: GS (GFP fraction)	394 (21,400)	504 (0.76)	467 (18,900)	502 (0.65)
YT C-loop: GS (YFP fraction)	396 (22,400)	512 (0.35)	503 (19,200)	520 (0.75)

Table S3. Spectral characterization of a YT construct. Peak wavelengths of the two absorption bands are given under $\lambda_{abs,A}$ and $\lambda_{abs,B}$. , where A and B correspond to the protonated and deprotonated form of the chromophore.[5] Extinction coefficients (ε) (in units of M^{-1}·cm^{-1}) and quantum yields (Φ) are given next to the absorption and emission wavelengths.

6. Kinetic stability of the folded proteins

When the relative composition of the N- or the C-bound form in a certain construct is enriched by anion exchange chromatography, neither purified species converts to the other even after several days at room temperature as evidenced by the unchanging absorbance spectrum. For instance, Figure S3 shows the absorbance spectra of a YT construct taken over time. The construct, which contains 14-residue-long loops on both sides (full sequence given above), initially expresses as 94% N-bound, and the composition does not change over several days at room temperature. When the relative N-bound population is dropped down to 54% by anion-exchange chromatography as estimated by the linear fit in Figure S2, the composition still does not change over several days at room temperature. The observation that either composition, 94% and 54% N-bound, stays unchanged for many days suggests that the intra-molecular s10 exchange is extremely slow, with a half-life well over a week.

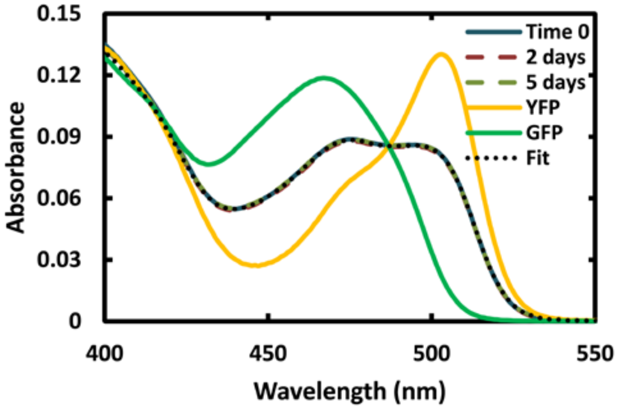

Figure S3. Absorbance spectra of YT construct with 14-residue-long loops on both sides (full sequence given in SI 2).

7. Predicting compositions

If the compositions given in Figure 2B are treated as equilibrium compositions, ΔG° = -RT·ln([C-bound]/[N-bound]), as plotted in Figure S3A for non-zero C-terminal loop lengths, can be interpreted as the free energy difference between the two bound forms. Then, the slope of the linear fit, 0.12 kcal·mol^{-1}, can be thought of as the energetic penalty imposed upon adding a single residue in the loop. When the length of the C-terminal loop is fixed and the N-terminal loop is varied, the same analysis gives a similar slope, 0.12 kcal·mol^{-1} (data not shown), suggesting that the difference in the number of loop residues on two sides determines the composition. Given that the HT construct with 14 amino acid residues on both loops expresses as approximately 68% N-bound, and assuming that this reflects an equilibrium composition, the energetic difference between the two bound forms of this construct can be estimated as 0.44 kcal·mol^{-1}. Then, if the C-terminal loop of this construct contains 2 or 10 residues instead of 14, binding of the C-terminal strand can be expected to gain energetic advantage of 0.12 times the difference in the number of loop residues, 0.12 × (14 - 2) = 1.44 or 0.12 × (14 - 10) = 0.48 kcal·mol^{-1}, leaving the C-terminally bound form 1.44 - 0.44 = 1.00 kcal·mol^{-1} or 0.48 - 0.44 = 0.04 kcal·mol^{-1} more stable than the N-terminally bound form, respectively. Since being 1.00 kcal·mol^{-1} or 0.04 kcal·mol^{-1} less stable translates into composing 15% or 48% of the population, one can make a *prediction* that HT construct with 14 N-terminal and 2 or 10 C-terminal loop residues will have 15% or 48% in the N-bound form, respectively. When these proteins were actually expressed in *E. coli* and when their compositions were estimated by fitting the absorbance spectra to the linear combination of basis spectra, as shown in Figure 3B, 14% and 47% were in the N-bound form, which are very close to the predicted values. Note that the predictability of compositions is strictly empirical at this point, and that although the system is treated here as if it is in thermodynamic equilibrium, it is difficult to conclude anything about the actual energetics underlying the folding of these proteins. More rigorous thermodynamic analysis of the system will be reported separately.

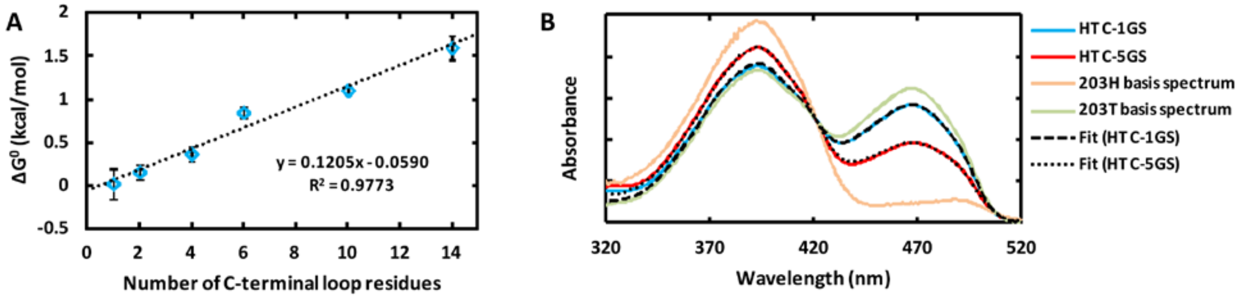

Figure S4. (A) -RT·ln([C-bound]/[N-bound]) as a function of the number of C-terminal loop residues in a YT construct with N-terminal loop that consists of 14 amino acid residues. (B) Absorbance spectra of HT constructs with 14 N-terminal and 2 (denoted as HT C-1GS) or 10 (denoted as HT C-5GS) C-terminal loop residues. The two basis spectra and the fits are shown together, and all spectra are normalized to have the same value at 421 nm, which is the isosbestic point of the two basis spectra.

8. *In vitro* refolding

Proteins were concentrated and mixed into 8 M guanidine hydrochloride buffer (8 M guanidine hydrochloride, 50 mM HEPES, and 50 mM NaCl at pH 8.0) so that the final concentration of guanidine hydrochloride is over 6 M. The mixture was kept at room temperature for more than 30 minutes for complete denaturation (the completion of denaturation was checked by making sure the absorbance spectrum fully converges to that of the denatured protein). The protein was refolded by exchanging back to lysis buffer so that the final concentration of guanidine hydrochloride was less than 1 mM. The composition of the refolded protein was estimated by absorbance or fluorescence as described in Figure 2A. Interestingly, the refolded proteins tend to favor the N-bound form compared to the initial mixture composition from expression. For example, the caspase sensor prototype (see Table S2) which expresses as ~90% C-bound turns into a mixture of ~30% C-bound (i.e. ~70% N-bound) once it is refolded *in vitro*.

9. Expression in mammalian cells

Two constructs, 1) a YT construct with 14-amino-acid insertion at the N-terminal side and no loop insertion at the C-terminal side (notated as C-0GS in Figure S5) and 2) a YT construct with 14-amino-acid insertion at the N-terminal side and 10-amino-acid insertion at the C-terminal side (denoted C-5GS in Figure S5), which are part of the set shown in Figure 2B, were expressed in mammalian cells as follows. HEK 293 and U205 cells growing on 6 cm plates were transfected using the calcium phosphate method with the vector pcDNA3.1. After 24 hours of incubation at 37^{o}C with 5% CO_2, the cells were washed with and resuspended in the lysis buffer (50 mM HEPES, 300 mM NaCl, and 10 v% glycerol at pH 8.0) to a final volume of 2mL. Concentrated (20%) SDS solution was added to the suspension for a final SDS concentration of 0.1%, and the mixture was passed through a 25 gauge needle three times to homogenize the cells. The lysed samples were spun down, and the supernatant from each sample was collected for fluorescence measurement. Non-transfected cells went through the same procedure for background subtraction. As shown in Figure S5, compositions could be estimated from fluorescence excitation spectra, where C-0GS expressed as 94% C-terminally bound and C-5GS expressed as 83% N-terminally bound, which are within the error of the values estimated from expressions in *E. coli*.

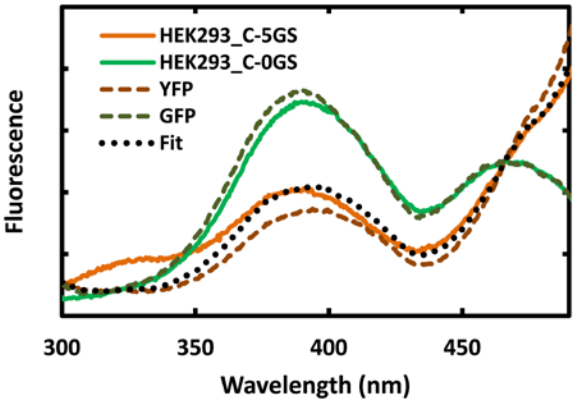

Figure S5. Excitation spectra of the supernatants were taken by scanning the excitation wavelength from 300 nm to 500 nm while collecting emission at 520 nm. The spectra of the two constructs expressed in HEK293 (solid lines), along with the basis spectra of GFP and YFP (dashed lines), are normalized to the isosbestic point at 466 nm. The spectrum of C-0GS is nearly indistinguishable from that of GFP, and the spectrum of C-5GS could be fit by a linear combination (dotted line) of the two basis spectra with the weighting factor of 0.17 for GFP and 0.83 for YFP.

10. Effect of rare codons

Two versions of plasmids encoding the protein with the following primary sequence were prepared:

MGHHHHHHSSGGLPDNHYLS**203T**QTVLSKDPNE**GSSGSGSSGSGSSG**KRDHMVLHEYVNAAGITHGMDELYGGTGGSASQGEELFTGVVPILVELDGDVNGHKFSVRGEGEGDATIGKLTLKFISTTGKLPVPWPTLVTTLSYGVQAFSRYPDHMKRHDFFKSAMPEGYVQERTISFKDDGKYKTRAVVKFEGDTLVNRIELKGTDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFTVRHNVEDGSVQLADHYQQNTPIGDGPVL**GSSGSRGSDGSDGS**LPDNHYLS**203H**QTVLSKDPNEGGSS**L**S**L**S**L**S**LL**.

The only difference between the two plasmids is the codon for the last five leucines, which are part of the GGSS**L**S**L**S**L**S**LL** sequence that follows the C-terminal s10. In one case, a CTA sequence, which corresponds to the rarest codon among those encoding leucine, and in the other case, a CTG sequence, which corresponds to the most abundant codon among those encoding leucine, was used for all five leucines. When the protein was expressed from each plasmid in identical conditions, however, the ratio of the two bound forms, as estimated from the two-Gaussian fit of the ion-exchange chromatography, was not changed by the synonymous mutation. Also, the overall expression yields were comparable. Note that there are 40 amino acids after s11 and before the first of the five C-terminal leucine (i.e. **SSGSRGSDGSDGS** LPDNHYLS**203H**QTVLSKDPNEGGSS). This would let the protein, up to s11, be outside the ribosome tunnel while the five leucines are being translated (the ribosomal tunnel occludes around 30 amino acids[6]).

11. Protease sensor experiments

Protease specifications
Thrombin: Plasminogen-free, Bovine , EMD Millipore
Caspase: 3 human, Sigma,
Trypsin: from bovine pancreas, Sigma

General Procedure

Into a 4 mL cuvette (4 cm tall), 3 mL of ~2 μM protein solution was added. The temperature of the sample was kept at 35 ^{o}C by using a temperature-controlled cuvette holder, and the temperature was checked with a thermocouple. While a magnetic stir bar was constantly stirring the sample, 405 nm laser light (10 ~ 25 mW) was directed vertically into the cuvette from above. At each time point, a 30 μL aliquot was taken from the irradiated sample and the photo-dissociation process was halted by transferring the aliquot directly into a fluorescence cuvette containing 2 mL of lysis buffer at room temperature. The fluorescence cuvettes containing the transferred aliquots were kept in the dark until fluorescence was measured, and the fluorescence was measured multiple times to make sure the measurement is not altering the composition.

Trypsin Sensor

K209Q mutation was introduced in the N-terminal strand to avoid ambiguous proteolysis after the lysine residue.

Caspase Sensor

As shown below in Figure S6, the general features are the same as those shown with the thrombin sensor in Figure 3. The altered baseline (~10% GFP) in Figure S6B indicates that the initial expression mixture of the caspase sensor contains more N-bound forms than the thrombin sensor, and the incomplete convergence to ~85% instead of 100% GFP indicates that some of the C-bound forms were not cut.

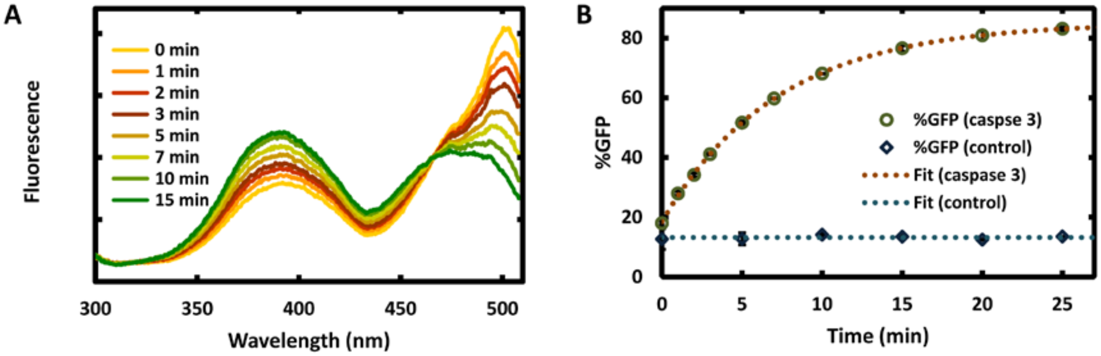

Figure S6. (A) Excitation spectra (emission collected at 520 nm) of the caspase sensor after being digested with caspase. (B) Composition change as a function of time.

Emission spectra of GFP and YFP

Since peak positions for GFP and YFP absorption and fluorescence emission are different, a large contrast can be seen in the ratio of two emissions. For instance, as shown in Figure S7, GFP emits 40 times stronger than YFP at 490 nm upon 440 nm excitation (Figure S7A), whereas YFP emits 60 times stronger than GFP at 530 nm upon 515 nm excitation (Figure S7B). Comparing the two gives over 2,000 fold dynamic range when GFP is converted to YFP or vice versa.

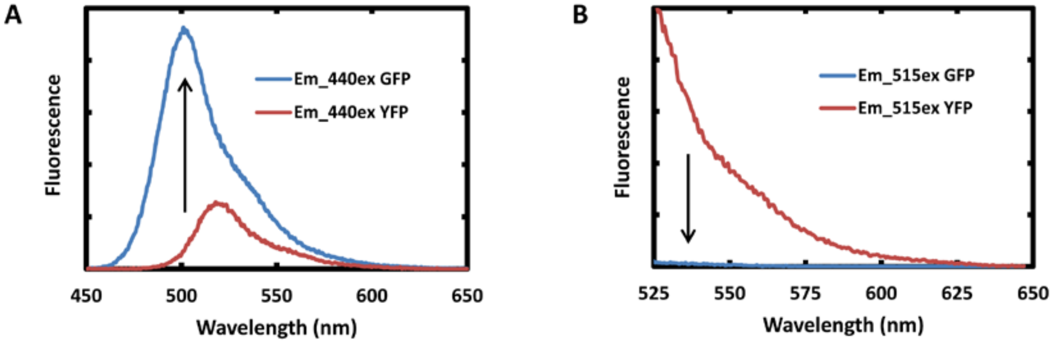

Figure S7. Emission spectra of GFP and YFP. Emission was collected while exciting at (A) 440 nm and (B) 515 nm.

12. Light-driven intra-molecular strand swap

When the N-bound fraction of the YT construct with 14 N-terminal and 2 C-terminal loop residues (GS) was separated by anion-exchange chromatography and irradiated with 405 nm light for 45 minutes at 35 ^{o}C, part of the population was converted into the C-bound form. Figure S8 shows the anion-exchange run with and without the irradiation, where the C-bound fraction (the first of the two peaks in Figure S8B) can be seen to grow in with irradiation. The two Gaussian fit of the 280 nm chromatogram in Figure S8B, using the method described in detail in SI 4, shows that 35% of the population is C-bound and 65% is N-bound. Note that this light-driven strand swap depends on the particular choice of the loop lengths, so any construct can be optimized to enhance this process or suppress it, as was the case for the protease sensor (Figure 3C).

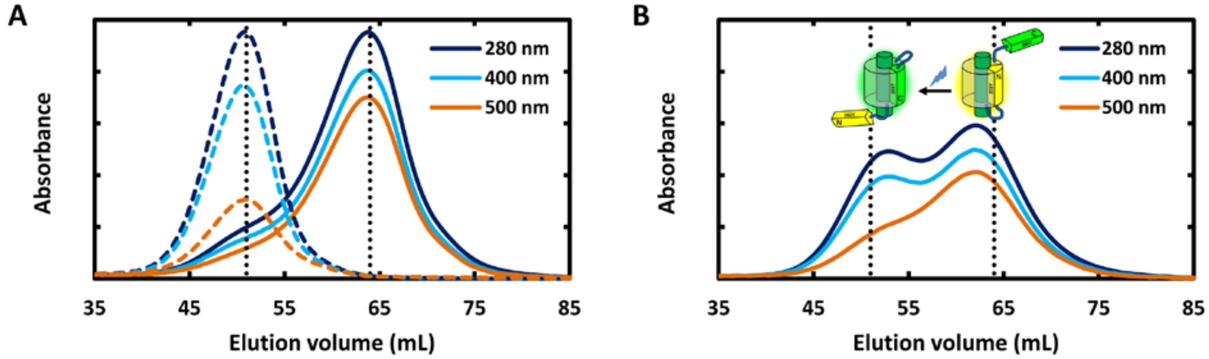

Figure S8. (A) Anion-exchange chromatogram of the C-bound (dotted lines) and the N-bound (solid lines) fraction. After separating the two bound forms as shown in Figure S2A, re-running each fraction through the anion-exchange column results in a peak at different elution volumes, where the C-bound peak is given in dotted lines and the N-bound peak in solid lines. (B) Anion-exchange chromatogram of the N-bound fraction after irradiation with 405 nm light. When the N-bound fraction is exposed to 405 nm light before re-running through the anion-exchange column, the first peak, which coincides with the C-bound peak as vertical dotted lines indicate, grows in while the second peak (the N-bound fraction) becomes smaller.

13. Instrumentation and other methods[3]

Fluorescence spectrometer
A LS55 from Perkin Elmer was used for all fluorescence measurements. Slit widths of 3~5 nm were typically used.

UV-vis spectrometer
A Lambda 25 from Perkin Elmer was used for all absorbance measurements.

Concentration determination
The relative absorbance at 447 nm before and after denaturing the protein in 0.1 M NaOH was used to estimate the extinction coefficient of the protein through the known 447 nm extinction coefficient of the GFP chromophore (44,100 $M^{-1}\cdot cm^{-1}$) in 0.1 M NaOH.[4] The estimated extinction coefficient was then used to determine concentrations.

References

(1) Cabantous, S.; Terwilliger, T. C.; Waldo, G. S. *Nature Biotechnol*. **2005**, *23*, 102-107.
(2) Pedelacq, J.; Cabantous, S.; Tran, T.; Terwilliger, T. C.; Waldo, G. S. *Nature Biotechnol*. **2006**, *24*, 79-88.
(3) Do, K.; Boxer, S. G. *J. Am. Chem. Soc*. **2011**, *133*, 18078-18081.
(4) Ward, W. W. *Bioluminescence and Chemiluminescence*, **1981**, (De Luca, M. and McElroy, D. W., eds.), Academic Press, New York, 235–242.
(5) Chattoraj, M.; King, B. A.; Bublitz, G. U.; Boxer, S. G. *Proc. Natl. Acad. Sci. U.S.A*. **1996**, *93*, 8362–83667
(6) Komar, A. A. *Trends Biochem. Sci*. **2009**, *34*, 16-24.