**Protein Engineering, Design & Selection** vol. 20 no. 2 pp. 69–79, 2007
Published online February 2, 2007 doi:10.1093/protein/gzl056

# The creation of a novel fluorescent protein by guided consensus engineering

**Mingha Dai*, Hugh E.Fisher*, Jamshid Temirov*,
Csaba Kiss, Mary E.Phipps, Peter Pavlik,
James H.Werner and Andrew R.M.Bradbury[1]**

Bioscience Division, Los Alamos National Laboratory, Los Alamos,
NM 87545, USA

[1]To whom correspondence should be addressed.
E-mail: amb@lanl.gov
*These authors contributed equally to this work.

**Consensus engineering has been used to increase the stability of a number of different proteins, either by creating consensus proteins from scratch or by modifying existing proteins so that their sequences more closely match a consensus sequence. In this paper we describe the first application of consensus engineering to the *ab initio* creation of a novel fluorescent protein. This was based on the alignment of 31 fluorescent proteins with >62% homology to monomeric Azami green (mAG) protein, and used the sequence of mAG to guide amino acid selection at positions of ambiguity. This consensus green protein is extremely well expressed, monomeric and fluorescent with red shifted absorption and emission characteristics compared to mAG. Although slightly less stable than mAG, it is better expressed and brighter under the excitation conditions typically used in single molecule fluorescence spectroscopy or confocal microscopy. This study illustrates the power of consensus engineering to create stable proteins using the subtle information embedded in the alignment of similar proteins and shows that the benefits of this approach may extend beyond stability.**

*Keywords*: fluorescent proteins/consensus engineering/Azami green/protein stability

## Introduction

Consensus engineering (Steipe *et al.*, 1994; Steipe, 2004) is an approach to increase the stability of proteins by modifying a protein sequence so that it more closely resembles a consensus derived from the alignment of numerous proteins of a particular family. This was initially applied to immunoglobulin variable domains (Steipe *et al.*, 1994; Ohage and Steipe, 1999; Wirtz and Steipe, 1999), and subsequently to scFvs (Knappik *et al.*, 2000; Visintin *et al.*, 2002; Arndt *et al.*, 2003), scFv fusions (McDonagh *et al.*, 2003), whole antibodies (Whitcomb, 2003) and isolated CH3 domains (Demarest *et al.*, 2004). This has proved particularly helpful in the use of scFvs as intracellular inhibitors of protein function. As antibodies are secreted proteins, with a large part of their stability conferred by disulfide bonds, the reducing environment of the cytoplasm prevents the formation of such disulfide bonds, so diminishing cytoplasmic stability and functionality. scFvs which have been found to be naturally stable in the cytoplasm tend to have sequences resembling consensus V-region sequences (Visintin *et al.*, 2002), and the mutation of scFvs to sequences closer to the consensus significantly increases their stability, allowing them to function under reducing conditions (Ohage and Steipe, 1999; Wirtz and Steipe, 1999). The use of this concept has not been limited to antibody domains, but has also been applied to the stabilization of GroEL minichaperones (Nikolova *et al.*, 1998; Wang *et al.*, 1999, 2000), WW (Jiang, 2001) and SH3 domains (Maxwell, 1998). Recently, a consensus version of gp120 was designed and expressed for use as a vaccine (Gao *et al.*, 2005), and although stability data were not reported, it was extremely well expressed compared to gp120 derived from single parental isolates.

A strikingly successful example of consensus engineering was the application to phytases (Lehmann *et al.*, 2000a, 2000b, 2001, 2002), enzymes used in animal feed technology. In initial experiments, a consensus sequence based on 13 sequences was produced by gene assembly. This showed an increase in stability of 22°C over the best parental sequence (Lehmann *et al.*, 2000b), which could be increased an additional 30°C (to 90.4°C) by eliminating a number of destabilizing mutations and including more sequences in the consensus. The final consensus sequences differed from the closest parental sequences by 18% (80 residues over 444). Unlike other applications of consensus engineering, which tend to be gradualist, with the introduction of mutations in small numbers, the first consensus sequence (Lehmann *et al.*, 2000b) in this study was produced using an *ab initio* approach with the gene synthesized directly. Although a similar approach was used to generate the V genes used in a phage antibody library containing modular consensus frameworks (Knappik *et al.*, 2000), the consensus sequences used were relatively similar to the closest individual V genes, and changes in stability were not studied.

The stabilization of proteins is important for improved half-life and function under the adverse conditions encountered in biotechnological and pharmacological applications. It has also been regarded as important in the use of proteins as scaffolds to generate libraries of specific binders. It has been reasoned that if the starting scaffold is more stable, it will be more tolerant to the destabilizing effects of mutations, or insertions, used to mediate binding. This was the rationale used in the creation of the phage antibody library described above. More recently, it has been applied to ankyrins. These are widely distributed, soluble repeat proteins generally involved in protein–protein interactions, with each polypeptide repeat composed of 31 amino acids. A consensus sequence created by the alignment of over 2000 ankyrins was used as a scaffold for diversity by mutation of specific exposed residues within three repeats (Kohl *et al.*, 2003). The high stability of the consensus ankyrin (denaturation requires boiling in urea) explains why artificial ankyrins selected from these libraries, with Tms of 65–85°C (Binz *et al.*, 2003), are more stable than most other scaffolds.

Although consensus engineering has not been applied to fluorescent proteins, a recent study used gene synthesis to recreate putative ancestral fluorescent proteins (Ugalde *et al.*, 2004; Chang *et al.*, 2005). Three different evolutionary algorithms (Yang, 1997) were used, with amino acids, codons or nucleotides as sequence information input. These algorithms identify sequences which often resemble putative consensus sequences, but which are usually different, with the degree of difference depending crucially upon the evolutionary model chosen. As the goal of these published studies was to determine whether the appearance of red fluorescence was the result of a single evolutionary event, or the result of convergent evolution, the effects of such engineering on other protein properties was, unfortunately, not examined. One reason why consensus engineering may not have been applied to fluorescent proteins is because of the relative ease with which mutations can cause a loss of fluorescence. This is due to the sensitivity of the fluorophore to its local environment (Tsien, 1998), which under normal circumstances is surrounded by the protective beta can. Consensus engineering could result in distortion of this co-operative structure if mismatched beta sheet amino acid pairs which normally co-vary are introduced. In the work described here, we have applied consensus engineering to create a functional fluorescent protein *ab initio*, using proteins bearing homology to the Azami green fluorescent protein, to create the consensus and monomeric Azami green (mAG) itself as a 'guide' protein to resolve ambiguities, and hence maintain appropriate co-variance. In doing so, we have created a highly functional novel fluorescent protein which differs from any of the component sequences by at least 10%.

## Materials and methods

### Gene synthesis

Genes for mAG and CGP were synthesized by Blue Heron Biotechnology (Seattle), using *Escherichia coli* preferred codons. Dimeric (dAG) and tetrameric (tAG) versions of mAG were created from mAG by mutation using the Quick Change kit (Stratagene) and the following oligonucleotides: primer pDAG-A188Y/K190F (AAA AAA GAT GTT CGT CTT CCA GAT TAC CAC TTC GTG GAC CAC CGC ATT GAA ATC) was used to create the dimeric mutant and pTAG-T123V(5′-TGA TAT TCG CTT TGA TGG AGT GAA CTT CCC CCC GAA CG-3′) was used to create the tetrameric (native) version from the dimeric mutant. Mutations were confirmed by DNA sequencing.

### Expression and purification

All expression plasmids (based on pET28b) were transformed into *E. coli* BL-21 (DE3) Gold, plated on 2XTY/Kan/3% glucose and grown overnight at 37°C. Individual colonies were picked and grown overnight in liquid 2XTY/Kan/glucose at 37°C. One milliliter of confluent culture was used to inoculate 50 ml 2XTY/Kan/IPTG in 250 ml shaker flasks for expression overnight.

Proteins were purified by low-salt immobilized metal affinity chromatography (IMAC). Cultures were harvested by centrifugation, sonicated and resuspended in 10 mM Tris pH 8.0, and recentrifuged at 3000 *g* for 30 min at 4°C. The supernatant was applied to IMAC columns pre-equilibrated with Tris buffer (10 mM pH8) for initial binding. The flow-through was reapplied three further times and washed with 20 bed volumes Tris buffer (10 mM pH8). An additional wash of 20 bed volumes of 10 mM Tris buffer/300 mM NaCl/10% glycerol was performed preceding a final Tris buffer (10 mM pH8) wash before elution in 600 mM Imidazole. The buffers were exchanged (into 10 mM Tris pH8) from the eluted proteins three times using 10 000 MWCO Amicon Ultra filtration devices at 4°C. The desalted proteins were stored at 4°C or room tempaerature preceding further evaluation. Protein samples for SDS–PAGE comparison were diluted for equivalent fluorescence utilizing a Tecan Spectrafluor Plus plate fluorometer equipped with 485 nm excitation and 535 nm emission filters prior to standard denaturation and gel loading. The OD280 absorption was calculated from protein sequences, and this was used to calculate the concentrations of purified proteins. For each protein a calibration curve of fluorescence versus protein concentration was created, and this was used to calculate the expression levels of bacterial cultures. For photophysical characterization, the proteins were centrifuged in an Eppendorf 5415 at maximum RPM for 10 min to remove minor protein aggregates that had formed during storage.

### Gel filtration

Size-exclusion chromatography experiments were performed with a Superdex 75 column (bed volume, *V*t of 24 ml; Amersham Biosciences, Uppsala, Sweden) run over an FPLC apparatus. The column was equilibrated and eluted with 0.05 M Sodium phosphate buffer, 0.15 M NaCl, pH 7.2. The flow rate was 0.5 ml/min and 0.2 ml protein sample, at 2 mg/ml, was injected. The column was calibrated under the conditions outlined above, using the following molecular mass markers (Amersham Biosciences): carbonic anhydrase*c* (29 000 Da), ovalbumin (43 000 Da), bovine serum albumin (67 000 Da) and blue dextran (2000 kDa).

### Thermal stability

The thermal stability of the proteins was examined using a Stratagene Mx3005P real time PCR machine. Prior to analysis, samples were diluted for equivalent fluorescence as described above. The temperature of the samples was increased at 0.1°C/s. Fluorescence readings (with excitation and emission at 492 nm and 516 nm, respectively) were taken each 0.5°C from 55 to 95°C. Measurements were exported to Excel and graphs produced. The graphs were normalized by first subtracting the background level (defined as the measured fluorescence of the denatured sample) from each series, and then normalized so that the fluorescence at 55°C was identical for each of the samples.

### Spectral characterization

*Absorption measurements*

Absorption spectra were collected on a Spectronic Genesys2 (Thermo Fisher Scientific, Waltham, MA) and exported to Microsoft Excel and normalized to the point of maximum absorption for each sample. Excitation and emission spectra were generated by a QuantumMaster 6SE (Photon Technologies Incorporated; Edison, NJ) spectrofluorometer utilizing 1 cm-square cuvettes. Excitation scans were evaluated with 509 nm emission wavelength except in the case of CGP, which was evaluated at 514 nm. Emission scans were performed identically, but with the wavelengths indicated. All emission scans were normalized to the maximum value obtained at the main emission peak for each sample. pH titration was carried out according to (Levine and Ward, 1982). Samples were incubated for 5 min, at the indicated pHs, and fluorescence at 535 was measured after excitation at 485. Buffers for the pH 2–11 range contained 50 mM each of sodium phosphate, citric acid and glycine. The pH 11.0–11.9 range buffers contained 25 mM sodium phosphate, the pH 12–13 range buffers contained 50 mM potassium chloride and sodium hydroxide and the pH 13–14 range contained appropriate dilutions of sodium hydroxide.

### *Fluorescence correlation spectroscopy*

FCS (Elson and Magde, 1974; Magde *et al.*, 1974) was performed on a home-built confocal microscope. Samples consisted of a 4 μl sample droplet of the protein at ~10 nM concentration. This drop was placed on a No. 1 glass coverslip and suspended below a 60 × 1.2 NA water immersion microscope objective (Nikon, Japan). A small fraction (typically 20 μW) of the 488 nm light emitted from an argon ion laser was reflected by a dichroic mirror (505 DRLP, Omega Optical) into the back of the microscope objective and focused to a near diffraction limited spot in the sample droplet ~25 μm away from the solvent-glass interface. The fluorescence emitted from the probe volume was collected by the objective, passed through the excitation dichroic and was spatial (50 micron pinhole) and spectrally (530DF30, Omega Optical) filtered before being focused onto the active area of a single photon counting avalanche photodiode (SPCM 200 PQ, Perkin Elmer Optoelectronics, Quebec, Canada).

The output pulse from the detector was suitably conditioned prior to being fed into the input of a digital correlator card (ALV-5000E, ALV-Laser Vertriebsgeselleschaft, m-b.H., Langen FRG). Normalized autocorrelations $G(\tau)$ of the detector counts were performed using the supplied ALV-5000 software. In FCS, deviations from the average light intensity are autocorrelated and normalized. The normalized autocorrelation function being given by:

$$G(\tau) = \frac{< \delta I(t)\delta I(t+\tau) >}{< I(t) >^2} \qquad (1)$$

In the above equation, angle bracketed quantities denote time averages and $I(t)$ is the time dependent light intensity of the detection channel. Deviations in the fluorescence light intensity from the average ($\delta I = I(t) - < I(t) >$) can result from diffusion of fluorescent molecules into and out of the probe volume, or chemical reactions or photochemical reactions, such as the formation of triplet or dark states, that change the fluorescence intensity of the species in the volume. The expected correlation function due to the diffusion of molecules into and out of a three-dimensional ellipsoidal Gaussian probe volume was treated by Aragón and Pecora (1976). We have fit the measured correlation functions to this model and have accounted for triplet or dark-state transitions by the method of Widengren *et al.* (1995). The following functional form was used to fit the measured data:

$$G(\tau) = \frac{1}{N} \cdot \frac{1}{1+\tau/\tau_c} \cdot \frac{1}{\sqrt{1+(\omega_o/\omega_z)^2 \frac{\tau}{\tau_c}}} \cdot \left[1 - f + f \cdot \exp\left(\frac{-\tau}{\tau_d}\right)\right] + d.c. \qquad (2)$$

where the five parameters of the fit are: $N$, the average number of molecules in the probe volume; $\tau_c$, the correlation time; $(\omega_o/\omega_z)$, the ratio of $1/e^2$ radius of the probe volume to the axial $1/e^2$ radius; $f$, the fraction of irradiated molecules in the dark state; $\tau_d$, the dark-state lifetime and d.c., a constant offset. The ratio of the radial to axial dimensions of the probe volume, $(\omega_o/\omega_z)$ was determined based upon the cross-correlation functions generated by 10 nM solutions of Rhodamine 110 and held fixed in all subsequent fitting procedures of the proteins' correlation curves.

For this work, the main parameter of interest from the fit to the data is the correlation time, $\tau_c$, which relates inversely to the diffusion constant, $D$, as $\tau_c = \omega_o^2/4D$. From the measured correlation time of Rhodamine 110 in water (40 μs) and its known diffusion coefficient, 280 μm$^2$/s (Rigler *et al.*, 1993), $\omega_o$ can be directly determined to be 0.21 μm for our FCS apparatus. With the radius of the probe volume determined, the measured correlation time can be used to extract a diffusion coefficient and hence a hydrodynamic radius via the Stokes–Einstein relation:

$$D = \frac{kT}{6\pi\eta R} \qquad (3)$$

where in the above equation $k$ is Boltzmann's constant, T is the solution temperature in Kelvin, $\eta$ is the viscosity of the solution and $R$ is the hydrodynamic radius of the diffusing species.

### *Quantum yield determinations*

Five solutions of the standard (fluorescein in 0.1 M NaOH) and the unknowns (our proteins in 1X PBS buffer) with increasing concentration were prepared. Corrected absorption spectra of all solutions were recorded with JASCO V-530 UV-visible spectrophotometer and absorbance values at 488 nm were noted. Then corrected fluorescence emission spectra of all solutions were measured in PTI QM-6 spectrofluorometer with 488 nm excitation. A graph of a total number of emitted photons (area under the fluorescence emission curves) versus absorbance (at 488 nm) was generated. By comparing the linear plot gradients of the standard ($s$) and unknowns ($u$) the quantum yields were calculated according to the equation (Velapoldi and Tonnesen, 2004):

$$QY_u = QY_s \frac{Grad_u \lambda_{exs} (\eta_u)^2}{Grad_s \lambda_{exu} (\eta_s)^2} \qquad (4)$$

since the same excitation wavelength ($\lambda_{ex}$) was used for all the samples and refractive indices ($\eta$) of the solvents are essentially the same, they cancel each other and the quantum yields are calculated based only on the ratio of the gradients of the unknowns and the standard. The quantum yield of fluorescein was taken to be 0.91 (Karasawa *et al.*, 2003).

### *Fluorescence lifetime measurements*

As part of spectral characterization, the fluorescence lifetimes of the proteins were also studied using time-correlated single photon counting. A pulsed semiconductor diode laser (PDL-800, Picoquant) operating at 437 nm with a 1 MHz repetition rate and nominally 50 ps pulse width was used for fluorescence excitation. Collection of the excitation was performed orthogonal to the excitation beam using magic angle conditions. Fluorescence was wavelength resolved using an Acton spectrometer before being detected by a single photon counting PMT module (Hamamatsu). Fluorescence decays were measured by a Becker-Hickl SPC-330 correlated photon counting card at the emission maximum of each protein.

## Results

### Design strategy

It was impossible to apply the consensus strategy to GFP, as there are very few genes of sufficient similarity, which are not merely GFP mutated derivatives, to create a valid consensus. On the other hand, at the time this work was initiated there were 39 other translated protein sequences in Genbank with a homology >62% to mAG. This is a monomeric derivative of a fluorescent protein derived from the stony coral Galaxeidae (Karasawa *et al.*, 2003), with homology to GFP of <6% at the amino acid level. After removal of sequences differing by few amino acids, 31 sequences remained (Fig. 1A). These were used to calculate a consensus sequence as described in the figure legend. Briefly, this involved using the consensus sequence calculated by Vector NTI (Invitrogen) for each position, unless there was no consensus, or the mAG amino acid comprised >40% of all sequences, in which case the mAG amino acid was used instead. Amino acids previously identified as being important for monomerization (Karasawa *et al.*, 2003; Verkhusha and Lukyanov, 2004) were also retained. We term this slight modification of consensus engineering, 'guided consensus engineering', since rather than taking the commonest amino acid at each individual position, the guide sequence is used to arbitrate amino acid ambiguity. We reasoned that using a guide sequence in this way would eliminate problems related to amino acid co-variance. This final protein, termed consensus green protein (CGP) differed from mAG by 23 amino acids (10.2%), wild-type AG by 26 amino acids (11.6%) and by 76 amino acids (34%) from the most distant protein used to create the consensus (GFP2 from *Agaricia fragilis* – AAU06857). The differences between CGP and the AG proteins are predominantly located in the N-terminal 120 residues, where 16.7% of amino acids differ (compared to 2.9% for the C-terminal remainder), and the differences are mainly found in the beta strands (19/23) rather than the loops (3/23) or the central alpha helix (1/23). Evaluation of the amino acids changed in the consensus sequence using MacPymol (DeLano Scientific) showed that most (20/23) were on the surface. The evolutionary relationship of CGP to the proteins used to derive the consensus is shown in Fig. 1B.

### Protein expression and properties

The genes for mAG and CGP were synthesized (Blue Heron Biotechnology, Seattle) using codons biased towards *E. coli*, and cloned into a pET28b vector derivative. In the original paper describing mAG (Karasawa *et al.*, 2003), the dimeric version was created from the tetrameric version by mutation of a single amino acid (V123T), and the mAG version by mutation of an additional two amino acids (Y188A, F190K). We created dimeric (dAG) and tetrameric (tAG, wild type) versions from mAG using oligonucleotides to insert the same mutations in two stages, as described in the *Materials and methods*.

All four proteins were expressed in *E. coli* BL21 by induction with IPTG. It was striking that bacteria expressing the CGP gene were more fluorescent than those carrying the mAG gene at all temperatures tested (Fig. 2A). On the assumption that the specific fluorescence within a bacterial cell is similar to that of the purified protein, we were able to estimate the intracellular expression levels for all four proteins by reference to a calibration curve constructed using purified proteins. At 37°C, CGP was the protein expressed at the highest levels, reaching over 600 mg/l of which over 300 mg/l could be purified using the His6 tag. At 30°C, its expression was second only to tAG. In Fig. 2B, polyacrylamide gel analysis (PAGE) of the purified proteins is shown, with the fluorescence level of each protein normalized prior to loading. The similar intensity of the Coomassie blue bands, indicates that, at a first approximation, the intrinsic fluorescence levels of the four proteins are similar.

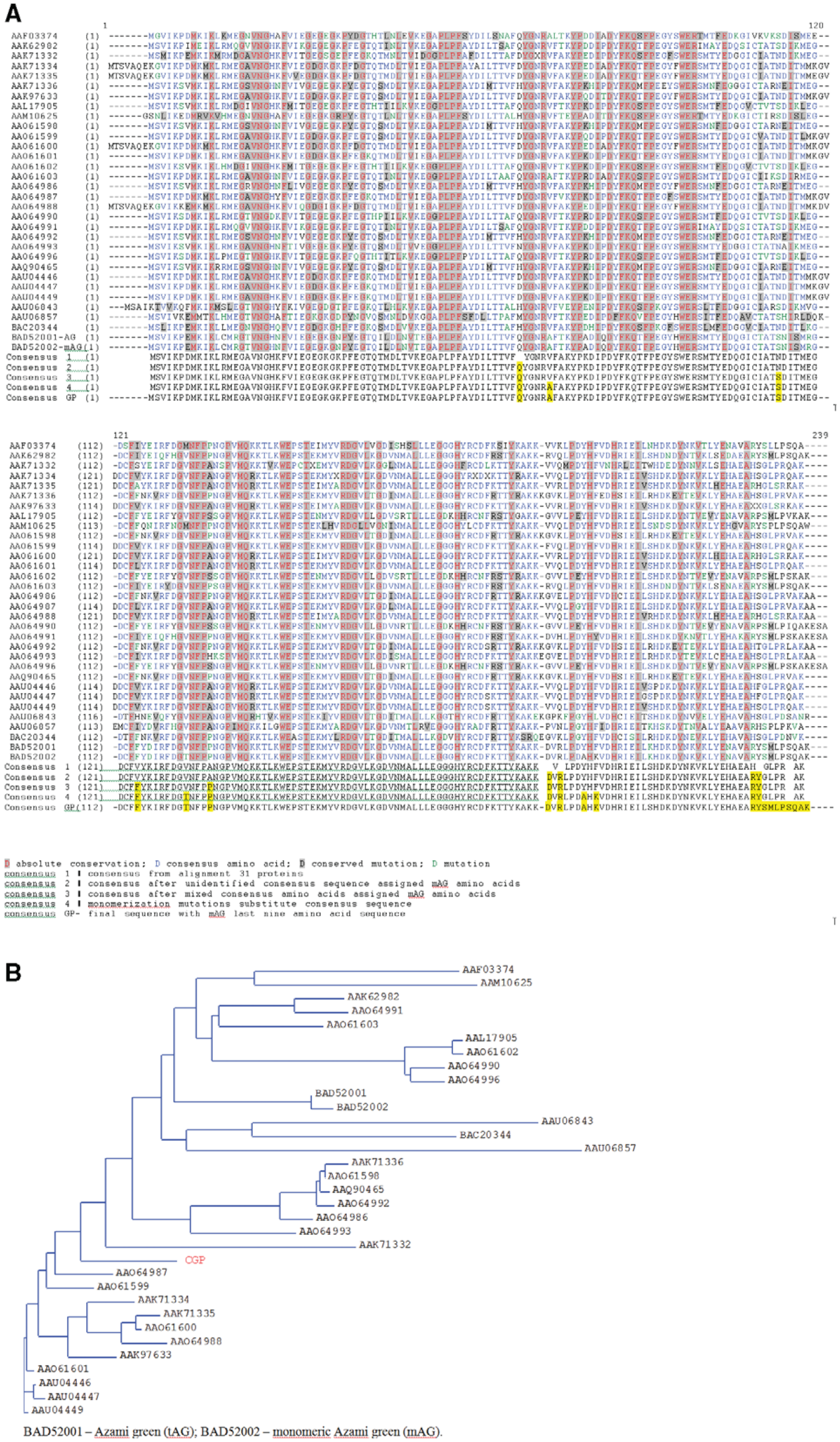
### *Thermal stability*

Thermal stability was assessed using a real-time PCR machine (Stratagene MX3005), with heat denaturation monitored by observing the fluorescence emission at 516 nm after excitation at 492 nm (Fig. 3). CGP, mAG and tAG show biphasic melting curves with cooperative unfolding in the second phase. For each of these proteins, the cooperative melting phase was relatively tight, occurring over 8–10°C. tAG is the most stable protein, with the midpoint of the cooperative transition at ~90°C. This is followed by mAG (84.5°C) and CGP (79°C). dAG on the other hand, did not show cooperative unfolding, with fluorescence decreasing almost linearly with increasing temperature. This, increased stability of the tetrameric form, is similar to that previously observed in comparing dsRED to EGFP (Verkhusha *et al.*, 2003).

### *Oligomerization status*

As described above, amino acids thought to be important in the monomerization of mAG and other fluorescent proteins were retained in CGP. The oligomerization state of the four proteins was first examined using gel filtration. As shown in Fig. 4A, consistent with the previously reported results (Karasawa *et al.*, 2003) for the Azami green proteins, we were able to confirm that tAG, the wild-type protein is tetrameric, whereas dAG is dimeric and mAG is monomeric. CGP was also monomeric, confirming that the preservation of those amino acids previously identified as being monomerizing (Karasawa *et al.*, 2003; Verkhusha and Lukyanov, 2004), within CGP were sufficient to maintain a monomeric state. The oligomerization state of the different proteins was also examined by fluorescence correlation spectroscopy (FCS), which can be used to directly 'size' molecules and is an independent measure that complements gel filtration in determining a protein's oligomerization state. It was one of the first experimental methods to verify that DsRed is a tetramer (Heikal *et al.*, 2000). Figure 4B shows auto-correlation curves for the four fluorescent proteins measured at 20 μW. From the fits to the data, the correlation (diffusion) times (in microseconds) are:

$$\text{mAG} = 90 \pm 1; \quad \text{dAG} = 135 \pm 1; \quad \text{tAG} = 188 \pm 3;$$
$$\text{CGP} = 90 \pm 1.$$



```
AAF03374       (1)  -------MGVIKPDMKIKLKMEGNVNGHAFVIEGEGEGKPYDGTHTLNLEVKEGAPLPFSYDILSNAFQYGNRALTKYPDDIADYFKQSFPEGYSWERTMTFEDKGIVKVKSDISMEE--
AAK62982       (1)  -------MSVIKPIMEIKLRMQGVVNGHKFVIKGEGEGKPFEGTQTINLTVKEGAPLPFAYDILTSAFQYGNRVFTKYPDDIPDYFKQTFPEGYSWERIMAYEDQSICTATSDIKMEG--
AAK71332       (1)  -------MSMIKPEMKIKMRMDGAVNGHKFVITGEGSGEPFEGKQTMNLTVIDGGPLPFAFDILTTAFDYGXRVFAKYPEDIPDYFKQSFPEGFSWERSMTYEDGGICIATNDIKMEG--
AAK71334       (1)  MTSVAQEKGVIKPDMKMKLRMEGAVNGHKFVIEGDGKGKPFDGTQTMDLTVIEGAPLPFAYAILTTVFDYGNRVFAKYPEDIADYFKQTFPEGYFWERSMTYEDQGICIATNDITMMKGV
AAK71335       (1)  MTSVAQEKGVIKPDMKMKLRMEGAVNGHKFVVEGDGKGKPFDGTQTMDLTVIEGAPLPFAYDILTTVFDYGNRVFAKYPEDIADYFKQTFPEGYFWERSMTYEDQGICIATNDITMMEGV
AAK71336       (1)  -------MSVIKSVMKIKLRMEGSVNGHNFVIVGEGEGKPYEGTQSMDLTVKEGAPLPFAYDIMTTVFHYGNRVFAKYPKHIPDYFKQMFPEEYSWERSMNFEGGGICTARNEITMEG--
AAK97633       (1)  -------MSVIKPDMKIKLRMEGAVNGHKFVIEGDGKGKPFEGTQSMDLTVIEGAPLPFAYDILTTVFDYGNRVFAKYPQDITDYFKQTFPEGYFWERSMTYEDQGICIATNDITMMKGV
AAL17905       (1)  -------MSVIKSVMKIKLRMDGIVNGHKFMITGEGEGKPFEGTHTIILKVKEGGPLPFAYDILTTAFQYGNRVFTKYPKDIPDYFKQSFPEGYSWERSMTFEDQGVCTVTSDIKLEG--
AAM10625       (1)  ------GSNLIKEDMRVKVHMEGNVNGHAFVIEGEGKGRPYEGTQTLNLTVKEGAPLPFSYDILTTALHYGNRVFTEYPADITDYFKQSFPEGYSWERTMTYEDKGICTIRSDISLEG--
AAO61598       (1)  -------MSVIKSVMKIKLRMEGSVNGHNFVIVGEGEGKPYEGTQSMDLTVKEGAPLPFAYDIMTTVFHYGNRVFAKYPKHIPDYFKQMFPEGYSWERSMNFEDGGICTARNEITMEG--
AAO61599       (1)  -------MSVIKPDMKIKLRMEGAVNGHKFVIEGDGKGKPFEGTQSMDLTVKEGAPLPFAYDILTTVFDYGNRVFAKYPQDIPDYFKQTFPEGYSWERSMTYEDQGICVATNDITLMKGV
AAO61600       (1)  MTSVAQEKGVIKPDMKMKLRMEGAVNGHKFVIEGDGKGKPFDGTQTMDLTVIEGAPLPFAYDILTTVFDYGNRVFAKYPEDIADYFKQTFPEGYFWERSMTYEDQGICIATNDITMMKGV
AAO61601       (1)  -------MSVIKPDMKIKLRMEGAVNGHKFVIEGDGKGKPFEGKQTMDLTVIEGAPLPFAYDILTTVFDYGNRVFAKYPKDIPDYFKQTFPEGYSWERSMTYEDQGICIATNDITMMKGV
AAO61602       (1)  -------MSVIKSVMKIKLHMDGIVNGHKFMITGEGEGKPFEGTHTIILKVKEGGPLPFAYDILTTAFQYGNRVFTKYPKDIPDYFKQSFPEGYSWERSMTFEDQGVCTVTSDIKLEG--
AAO61603       (1)  -------MSVIKPDMKIKLRMEGAVNGHNFVIEGEGKGKPFEGTQTINLTVKEGGPLPFAYDILTAAFQYGNRAFTKYPRDIADYFKQSFPEGYSWERSMTYEDQGICIIKSDIRMEG--
AAO64986       (1)  -------MSVIKSVMKIKLRMEGRVNGHNFLIVGEGEGKPYEGTQSMDLTVKEGAPLPFAYDIMTTVFHYGNRVFAKYPKHIPDYFKQMFPEGYFWERSMNFEDGGICTARNEITMEG--
AAO64987       (1)  -------MSVIKPDMKMKLRMEGAVNGHYFVIEGEGKGKPFEGKQTMDLTVIEGAPLPFAYDILTTVFDYGNRVFVKYPKDIPDYFKQTFPEGFSWERSMTYEDQGICIATNDITMMKGV
AAO64988       (1)  MTSVAQEKGVIKKEMKMKLRMEGAVNGHKFVIQGDGIGKPFDGTQTMDLTVIEGAPLPFAYDILTTVFDYGNRVFAKYPEDIPDYFKQTFPEGYFWERSMTYEDQGICIATNDITMMKDV
AAO64990       (1)  -------MSVIKSDMKIKLRMEGTVNGDKFVITGEGEGKPFEGTHPIILKVKEGGPLPFAYDILTTAFQYGNRVFTKYPRDIPDYFKQSFPEGYSWERSMTFEDQGICTVTSDIKLEG--
AAO64991       (1)  -------MSVIKPDMKIKLRMQGVVNGHKFVIKGEGEGKPFEGTQTINLTVKEGAPLPFAYDILTSAFQYGNRVFTKYPDDIPDYFKQTFPEGYSWERIMAYEDQSICTATSDIKMEG--
AAO64992       (1)  -------MSVIKSVMKIKLRMEGSVNGHNFVIVGEGEGKPYEGTQSMDLTVKEGAPLPFAYDIMTTVFHYGNRVFAKYPKHIPDYFKQMFPEGYSWERSMNFEDGGICTARNEITMEG--
AAO64993       (1)  -------MSVIKSVMKIKLRMEGAVNGHNFTIVGEGEGKPYEGTQSMDLTVVEGGPLPFAYDILTTVFNYGNRVFAKYPKDIPDYFKQTFPEGYSWERSMTYEDGGICTATNAIKMEG--
AAO64996       (1)  -------MSVIKSDMKIKLPMEGTVNGHKFVITGEGEGKPFQGTHTITLKVKEGGPLPFPYDILTTAFQYGNRVFTKYPRDIPDYFKQSFPEGYSWERSMTFEDQGICTVTSDIKLEG--
AAQ90465       (1)  -------MSVIKSVMKIKRRMEGSVNGHNFVIVGEGEGKPYEGTQSMDLTVKEGAPLPFAYDIMTTVFHYGNRVFAKYPKHIPDYFKQMFPEGYSWERSMNFEDGGICIARNEITMEG--
AAU04446       (1)  -------MSVIKPDMKIKLRMEGAVNGHKFVIEGDGKGKPFEGKQTMDLTVIEGAPLPFAYDILTTVFDYGNRVFAKYPKDIPDYFKQTFPEGYSWERSMTYEDQGICIATNDITMMKGV
AAU04447       (1)  -------MSVIKPDMKIKLRMEGAVNGHKFVIEGDGKGKPFEGKQTMDLTVIEGAPLPFAYDILTTVFDYGNRVFAKYPKDIPDYFKQTFPEGYSWERSMTYEDQGICIATNDITMMKGV
AAU04449       (1)  -------MSVIKPDMKIKLRMEGAVNGHKFVIEGDGKGKPFEGKQTMDLTVIEGAPLPFAYDILTTVFDYGNRVFAKYPKDIPDYFKQTFPEGYSWERSMTYEDQGICIATNDITMMKGV
AAU06843       (1)  ---MSAIKTVVKQFMKIKMSLEGTVNGHYFKIVGEGDGTPFEGKQTLHLKVKEGAPLPFAYDILTTALHYGNRVFVEYPENIPDYFKQSFPKGYSWERSLTFEDGGICIARSDIKMVG--
AAU06857       (1)  -------MSVIVKEMMTKLHMEGTVNGHAFTIEGKGKGDPYNGVQSMNLDVKGGAPLPFSFDILTPAFMYGNRVFTKYPEDIPDFFKQVFPEGYHWERSITFEDQAVCTATSHIRLDQK-
BAC20344       (1)  -------MSLIKPEMKIKLLMEGNVNGHQFVIEGDGKGHPFEGKQSMDLVVKEGAPLPFAYDILTTAFHYGNRVFAKYPDHIPDYFKQSFPKGFSWERSLMFEDGGVCIATNDITLKG--
BAD52001-AG    (1)  -------MSVIKPEMKIKLCMRGTVNGHNFVIEGEGKGNPYEGTQILDLNVTEGAPLPFAYDILTTVFQYGNRAFTKYPADIQDYFKQTFPEGYHWERSMTYEDQGICTATSNISMRG--
BAD52002-mAG(1)  -------MSVIKPEMKIKLCMRGTVNGHNFVIEGEGKGNPYEGTQILDLNVTEGAPLPFAYDILTTVFQYGNRAFTKYPADIQDYFKQTFPEGYHWERSMTYEDQGICTATSNISMRG--
Consensus 1 (1)         MSVIKPDMKIKLRMEGAVNGHKFVIEGEGKGKPFEGTQTMDLTVKEGAPLPFAYDILTTVF YGNRVFAKYPKDIPDYFKQTFPEGYSWERSMTYEDQGICIATNDITMEG
Consensus 2 (1)         MSVIKPDMKIKLRMEGAVNGHKFVIEGEGKGKPFEGTQTMDLTVKEGAPLPFAYDILTTVFQYGNRVFAKYPKDIPDYFKQTFPEGYSWERSMTYEDQGICIATNDITMEG
Consensus 3 (1)         MSVIKPDMKIKLRMEGAVNGHKFVIEGEGKGKPFEGTQTMDLTVKEGAPLPFAYDILTTVFQYGNRVFAKYPKDIPDYFKQTFPEGYSWERSMTYEDQGICIATSDITMEG
Consensus 4 (1)         MSVIKPDMKIKLRMEGAVNGHKFVIEGEGKGKPFEGTQTMDLTVKEGAPLPFAYDILTTVFQYGNRAFAKYPKDIPDYFKQTFPEGYSWERSMTYEDQGICIATSDITMEG
Consensus GP (1)  -------MSVIKPDMKIKLRMEGAVNGHKFVIEGEGKGKPFEGTQTMDLTVKEGAPLPFAYDILTTVFQYGNRAFAKYPKDIPDYFKQTFPEGYSWERSMTYEDQGICIATSDITMEG--

AAF03374     (112)  -DSFIYEIRFDGMNFPPNGPVMQKKTLKWEPSTEIMYVRDGVLVGDISHSLLLEGGGHYRCDFKSIYKAKK-VVKLPDYHFVDHRIEILNHDKDYNKVTLYENAVARYSLLPSQA----
AAK62982     (112)  -DCFIYEIQFHGVNFPPNGPVMQKKTLKWEPSTEKMYVRDGVLKGDVNMALLLEGGGHYRCDFRSTYKAKK-RVQLPDYHFVDHRIEILSHDNDYNTVKLSEDAEARYSMLPSQAK---
AAK71332     (112)  -DCFSYEIRFDGVNFPANSPVMQKKTVKWEPCTXEMYVRDGVLKGGLNMALLLEGGGHFRCDLKTTYKAKK-VVQMPDYHFVNHRLEITWHDEDYNNVKLSEHAEAHSGLPRQAK----
AAK71334     (121)  DDCFVYKIRFDGVNFPANGPVMQRKTLKWEPSTEKMYXRDGVLKGDVNMALLLEGGGHYRXDXKTTYRAKK-VVQLPDYHFVDHRIEIVSHDKDYNKVKLYEHAEAHSGLPRQAK----
AAK71335     (121)  DDCFAYKIRFDGVNFPANGPVMQRKTLKWEPSTEIMYARDGVLKGDVNMALLLEGGGHYRCDFKTTYKAKK-VVRLPDYHFVDHRIEIVSHDKDYNKVKLHEHAEARHGLSRKAK----
AAK71336     (112)  -DCFFNKVRFDGVNFPPNGPVMQKKTLKWEPSTEKMYVRDGVLTGDINMALLLEGGGHYRCDFRTTYRAKKKGVKLPDYHFEDHSIEILRHDKEYTEVKLYEHAEAHSGLPRVAK----
AAK97633     (114)  DDCFVYKIRFDGVNFPANGPVMQRKTLKWEPSTEIMYARDGVLKGDVNMALLLEGGGHYRCDFKTTYKAKK-VVRLPDYHFVDHRIEIVSHDKDYNKVKLYEHAEAXXGLSRKAK----
AAL17905     (112)  -DCFFYEIRFYGVNFPSSGPVMQKKTLKWEPSTENMYVRDGVLLGDVSRTLLLEGDKHHRCNFRSTYGAKK-GVVLPEYHFVDHRIEILSHDKDYNTVEVYENAVARPSMLPVKAK---
AAM10625     (113)  -DCFFQNIRFNGMNFPPNGPVMQKKTLKWEPSTEKLHVRDGLLVGNINMALLLEGGGHYLCDFKTTYKAKK-VVQLPDYHFVDHRIEILSNDSDYNKVKLYEHGVARYSPLPSQAW---
AAO61598     (112)  -DCFFNKVRFDGVNFPPNGPVMQKKTLKWEPSTEKMYVRDGVLTGDINMALLLEGGGHYRCDFRTTYRAKKKGVKLPDYHFVDHSIEILRHDKEYTEVKLYEHAEAHSGLPRVAK----
AAO61599     (114)  DDCFVYKIRFDGVNFPANGPVMQKKTLKWEPSTEKMYVRDGVLKGDVNMALLLEGGGHYRCDFKTTYKAKK-FVQLPDYHFVDHRIEILSHDKDYNKVKLYEHAEAHSGLPRQAK----
AAO61600     (121)  DDCFVYKIRFDGVNFPANGPVMQRKTLKWEPSTEIMYVRDGVLKGDVNMALLLEGGGHYRCDFKTTYKAKK-VVRLPDYHFVDHRIEIVSHDKDYNKVKLHEHAEARHGLSRKAK----
AAO61601     (114)  DDCFLYKIRFDGVNFPANGPVMQRKTLKWEPSTEKMYVRDGVLKGDVNMALLLEGGGHYRCDFKTTYKAKK-VVQLPDYHFVDHRIEIVSHDKDYNKVKLYEHAEAHSGLPRQAK----
AAO61602     (112)  -DCFFYEIRFYGVNFPSSGPVMQKKTLKWEPSTENMYVRDGVLLGDVSRTLLLEGDKHHRCNFRSTYRAKK-GVVLPEYHFVDHRIEILSHDKDYNTVEVYENAVARPSMLPSKAK---
AAO61603     (112)  -DCFIYEIRYDGVNFPPSGPVMQKKTLKWEPSTEKMYVRDGVLKGDVNMALLLEGGGHYRCDFRSTYKAKK-RVQLPDYHFVDHRIEILSHDNDYNTVKLSENAEARYSMLPSQAK---
AAO64986     (112)  -DCFFNKVRFDGVNFPPNGPVMQKKTLKWEPSTEKMYVRDGVLTGDINMALLLEGGGHYRCDFRTTYRAKKKGVELPDYHFVDHCIEILSHDKDYNKVKLYEHAEAHSGLPRVAKAA--
AAO64987     (114)  DDCFLYKIRFDGVNFPANGPVMQKKTLKWEPSTEKMYVRDGVLKGDLNMALLLEGGGHYRCDFKTTYKAKK-VVQLPGYHFVDHRIEILSHDKDYNKVKLYEHAEAHSGLPRVAKAA--
AAO64988     (121)  DDCFVYKIRFDGVNFPANGPVMQRKTLKWEPSTEIMYARDGVLKGDVNMALLLEGGGHYRCDFKTTYKAKK-VVQLPDYHFVDHRIEIVRHDKDYNKVKLHEHAEARHGLSRKAKAA--
AAO64990     (112)  -DCFFYEIRFYGVNFPSSGPVMQKKTLKWEPSTENMYVRDGVLLGDVNRTLLLEGDKHHRCNFRSTYRAKK-GVVLPEYHFVDHRIEILSHDKDYNTVEVYENAVARPSMLPIQAKESA
AAO64991     (112)  -DCFIYEIQFHGVNFPPNGPVMQKKTLKWEPSTEKMYVRDGVLKGDVNMALLLEGGGHYRCDFRSTYKAKK-DVHLPDYHYVDHRIEILSHDKDYKNVTLYEHAKARYSMLPSKAKESA
AAO64992     (112)  -DCFFNKVRFDGVNFPPNGPVMQKKTLKWEPSTEKMYVRDGVLTGDINMALLLEGGGHYRCDSRTTYRAKKKGVKLPDYHFVDHSIEILRHDKEYTEVKLYEHAEAHTGLPRLAKAA--
AAO64993     (112)  -DCFIYKIRFDGVNFPHKSPVMQKKTLKWEPSTEKMYVRDGVLKGDISMALLLEGGGHYRCDFKTTYKAKKEGVELPDYHFVDHRIEILRHDKDYTEVKLCEHAEAHSGLPRLAKAA--
AAO64996     (112)  -DCFFYEIRFYGVNFPSNGPVMQKKTLKWEPSTENMYVRDGVLLGDVNRTLLLEGDKHHRCNFRSTYRAKK-GVVLPEYHFVDHRIEILSHDKDYNTVEVYENAVARPSMLPSKAKESA
AAQ90465     (112)  -DCFFNKVRFDGVNFPPNGPVMQKKTLKWEPSTEKMYVRDGVLTGDINMALLLEGGGHYRCDFRTTYRAKKKGVKLPDYHFVDHSIEILRHDKEYTEVKLYEHAEAHSGLPRVAK----
AAU04446     (114)  DDCFVYKIRFDGVNFPANGPVMQRKTLKWEPSTEKMYVRDGVLKGDVNMALLLEGGGHYRCDFKTTYKAKK-VVQLPDYHFVDHRIEIVSPDKDYNKVKLYEHAEAHFGLPRQAK----
AAU04447     (114)  DDCFVYKIRFDGVNFPANGPVMQRKTLKWEPSTEKMYVRDGVLKGDVNMALLLEGGGHYRCDSKTTYKAKK-VVQLPDYHFVDHRIEIGSPDKDYNKVKLYEHAEAHFGLPRQAK----
AAU04449     (114)  DDCFVYKIRFDGVNFPANGPVMQRKTLKWEPSTEKMYVRDGVLKGDVNMALLLEGGGHYRCDSKTTYKAKK-VVQLPDYHFVDHRIEIVSHDKDYNKVKLYEHAEAHSGLPRQAK----
AAU06843     (116)  -DTFHNEVQFYGVNFPANGPVMQRHTVKWEPSTEKIYVRDGVLTGDITMALLLKGGTHYRCDFRTTYKAKEKGPKFPGYHLVDHCIEITSHDKDYNVVELYEHAVAHSGLPDSANR---
AAU06857     (113)  EMCFIYDVRFHGVNFPANGPIMQKKILGWEPSTEKMYARDGVLKGDVNMTLRVEGGGHYRADFRTTYKAKK-PVNLPGYHFIDHRIEITKHSKDYTNVALYEAAVARHSPLPKVA----
BAC20344     (112)  -DTFFNKVRFDGVNFPPNGPVMQKKTLKWEASTEKMYLRDGVLTGDITMALLLKGDVHYRCDFRTTYKSRQEGVKLPGYHFVDHCISILRHDKDYNEVKLYEHAVAHSGLPDNVK----
BAD52001     (112)  -DCFFYDIRFDGVNFPPNGPVMQKKTLKWEPSTEKMYVRDGVLKGDVNMALLLEGGGHYRCDFKTTYKAKK-DVRLPDYHFVDHRIEILKHDKDYNKVKLYENAVARYSMLPSQAK---
BAD52002     (112)  -DCFFYDIRFDGTNFPPNGPVMQKKTLKWEPSTEKMYVRDGVLKGDVNMALLLEGGGHYRCDFKTTYKAKK-DVRLPDAHKVDHRIEILKHDKDYNKVKLYENAVARYSMLPSQAK---
Consensus 1 (121)   DCFVYKIRFDGVNFPANGPVMQKKTLKWEPSTEKMYVRDGVLKGDVNMALLLEGGGHYRCDFKTTYKAKK  V LPDYHFVDHRIEILSHDKDYNKVKLYEHAEAH GLPR AK
Consensus 2 (121)   DCFVYKIRFDGVNFPANGPVMQKKTLKWEPSTEKMYVRDGVLKGDVNMALLLEGGGHYRCDFKTTYKAKK DVRLPDYHFVDHRIEILSHDKDYNKVKLYEHAEARYGLPR AK
Consensus 3 (121)   DCFFYKIRFDGVNFPPNGPVMQKKTLKWEPSTEKMYVRDGVLKGDVNMALLLEGGGHYRCDFKTTYKAKK DVRLPDYHFVDHRIEILSHDKDYNKVKLYEHAEARYGLPR AK
Consensus 4 (121)   DCFFYKIRFDGTNFPPNGPVMQKKTLKWEPSTEKMYVRDGVLKGDVNMALLLEGGGHYRCDFKTTYKAKK DVRLPDAHKVDHRIEILSHDKDYNKVKLYEHAEARYGLPR AK
Consensus GP(112)  -DCFFYKIRFDGTNFPPNGPVMQKKTLKWEPSTEKMYVRDGVLKGDVNMALLLEGGGHYRCDFKTTYKAKK-DVRLPDAHKVDHRIEILSHDKDYNKVKLYEHAEARYSMLPSQAK----

D absolute conservation; D consensus amino acid; D conserved mutation; D mutation
consensus 1 | consensus from alignment 31 proteins
consensus 2 | consensus after unidentified consensus sequence assigned mAG amino acids
consensus 3 | consensus after mixed consensus amino acids assigned mAG amino acids
consensus 4 | monomerization mutations substitute consensus sequence
consensus GP- final sequence with mAG last nine amino acid sequence
```

BAD52001 – Azami green (tAG); BAD52002 – monomeric Azami green (mAG).

**Fig. 1.** Comparison of proteins closely related to mAG. **(A)** alignment of 31 proteins with homology to mAG of 62% or more. Consensus 1 was calculated using the Vector NTI AlignX program (Invitrogen). A consensus amino acid was identified if it, or a chemically similar amino acid (e.g. aspartate and glutamate), was present in 50% or more of the sequences at each individual position. Consensus 1 was further modified as follows. (i) Where no consensus amino acid could be defined, because no amino acid comprised >50% of the sequences, the mAG sequence was used instead (consensus 2). (ii) Where the number of genes carrying the mAG sequence comprised of 40% or more of the sequences, the mAG sequence was used rather than the consensus sequence (consensus 3). (iii) Amino acids previously identified as being important for the monomerization of mAG (Karasawa *et al.*, 2003) and other fluorescent proteins (Verkhusha, and Lukyanov, 2004) were used rather than the consensus sequence (consensus 4). (iv) The last eight amino acids were assigned the mAG sequence (consensus green protein, CGP). For each consensus, differences to the previous consensus are indicated by highlighting in yellow. Wild-type Azami green (AG) is shown for reference. **(B)** The phylogenetic tree for the sequences used to create CGP, as well as CGP and wild-typeAG, are indicated.

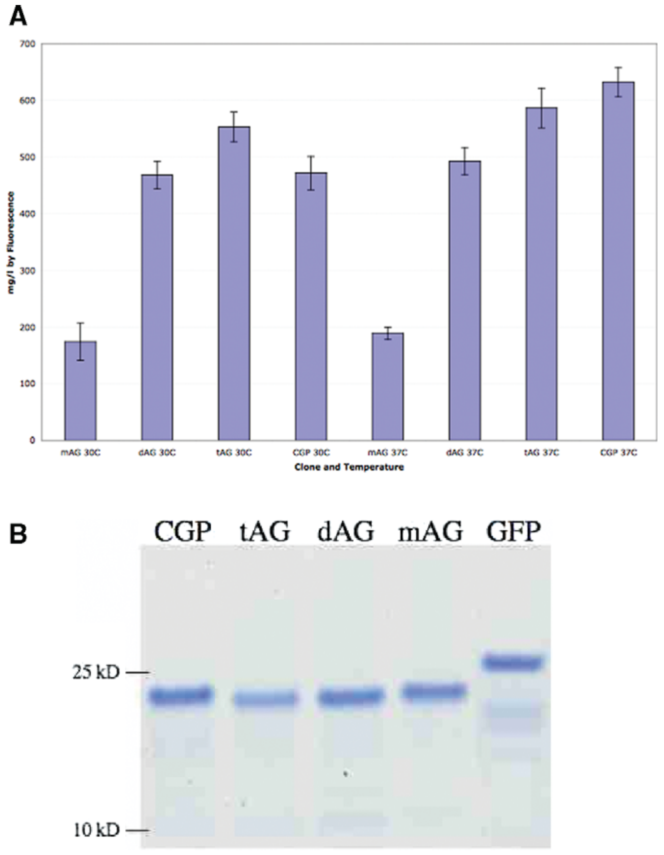


**Fig. 2.** (**A**) Expression levels of the different proteins expressed in bacteria at different temperatures, calculated by reference to a titration curve of fluorescence versus protein concentration carried out for each purified protein. (**B**) Purified CGP and AG proteins analyzed by PAGE. The fluorescence of each protein preparation was determined, and the amounts loaded normalized for fluorescence.

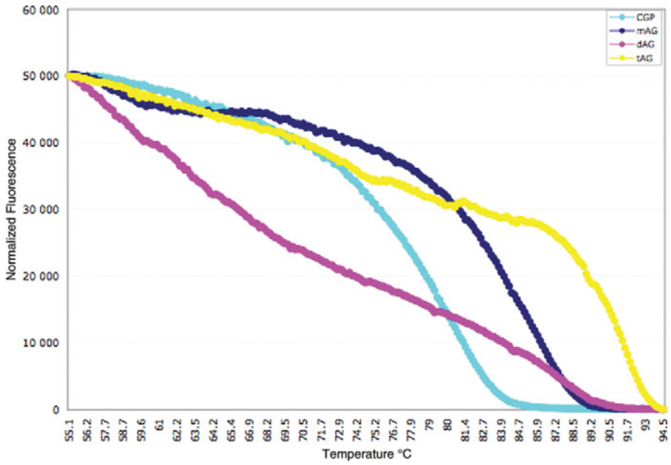


**Fig. 3.** Thermal stability of CGP and AG proteins. The thermal stability was determined by observing the fluorescence emission at 516 nm after excitation at 492 nm using a real-time PCR machine (Stratagene). Emission was measured approximately each 0.5°C, and a three point running average is plotted.

Corresponding (to the correlation times) hydrodynamic radii were calculated using equation 3 and are given (in Angstroms) below:

$$\mathrm{mAG} = 20 \pm 0.1; \quad \mathrm{dAG} = 30 \pm 0.2;$$
$$\mathrm{tAG} = 42 \pm 0.6; \quad \mathrm{CGP} = 20 \pm 0.1.$$

The ratios of the hydrodynamic radii for the proteins are as follows:

$$\frac{\mathrm{tAG}}{\mathrm{dAG}} = 1.4; \quad \frac{\mathrm{dAG}}{\mathrm{mAG}} = \frac{\mathrm{dAG}}{\mathrm{CGP}} = 1.5;$$

indicating that CGP is indeed a monomer, and also confirming the dimeric and tetrameric structures of dAG and tAG.

*Spectral properties*

An examination of the absorption spectra of CGP, shown in Fig. 5A, shows similarity to the Azami green proteins, with a red shift of 11 nm (to a peak excitation of 503 nm) compared to the Azami green proteins (peak at 492 nm). The emission properties of CGP (Fig. 5B) are also slightly different to the other green fluorescent proteins, showing a red shift peak emission to 514 nm, beyond that of both GFP (509 nm) and the Azami green proteins (504 nm), accounting for its slightly yellowish tinge under visible light. As was found in the case of the original Azami clones (Kurosawa *et al.*, 2003), CGP proved difficult to excite with UV irradiation between 350 and 400 nm, suggesting that the chromophore is maintained exclusively in the anionic form.

The quantum yields of the different proteins were measured using the relative method comparing the absorbances and fluorescence intensities of the proteins to fluorescein measured in the same instrument under identical conditions (Velapoldi and Tonnesen, 2004). The quantum yields of the Azami green proteins (Table I) were similar to one another and somewhat higher than those previously reported (Karasawa *et al.*, 2003). The quantum yield of CGP was 0.55.

The responses of the four proteins to changes in pH were tested by incubation for 5 min at the pH levels shown in Fig. 5C. Between pH4 and 5 CGP was more stable than either mAG or dAG, while it rapidly lost fluorescence between pH11 and 12. mAG and dAG, on the other hand, were more stable at the higher pHs, whereas tAG was stable at both high and low pHs.

## Photophysical single molecule properties

For some time, fluorescent proteins have been used as labels for confocal fluorescence microscopy and are seeing increasing use for single molecule studies of protein trafficking or gene expression. As such, we investigated the photophysical properties of these proteins under the excitation conditions typically used for single molecule fluorescence imaging. In particular, we investigated the effect of excitation laser power on the dark-state lifetime, the fraction of molecules in the dark state and number of photons emitted by the proteins. These parameters were extracted from the fits of the correlations curves. Figure 6A plots the number of photons per molecule per second (‘brightness’) versus the excitation laser power. The brightness per fluorophore reported in Fig. 6A is the product of the average count rate measured during the correlation run for

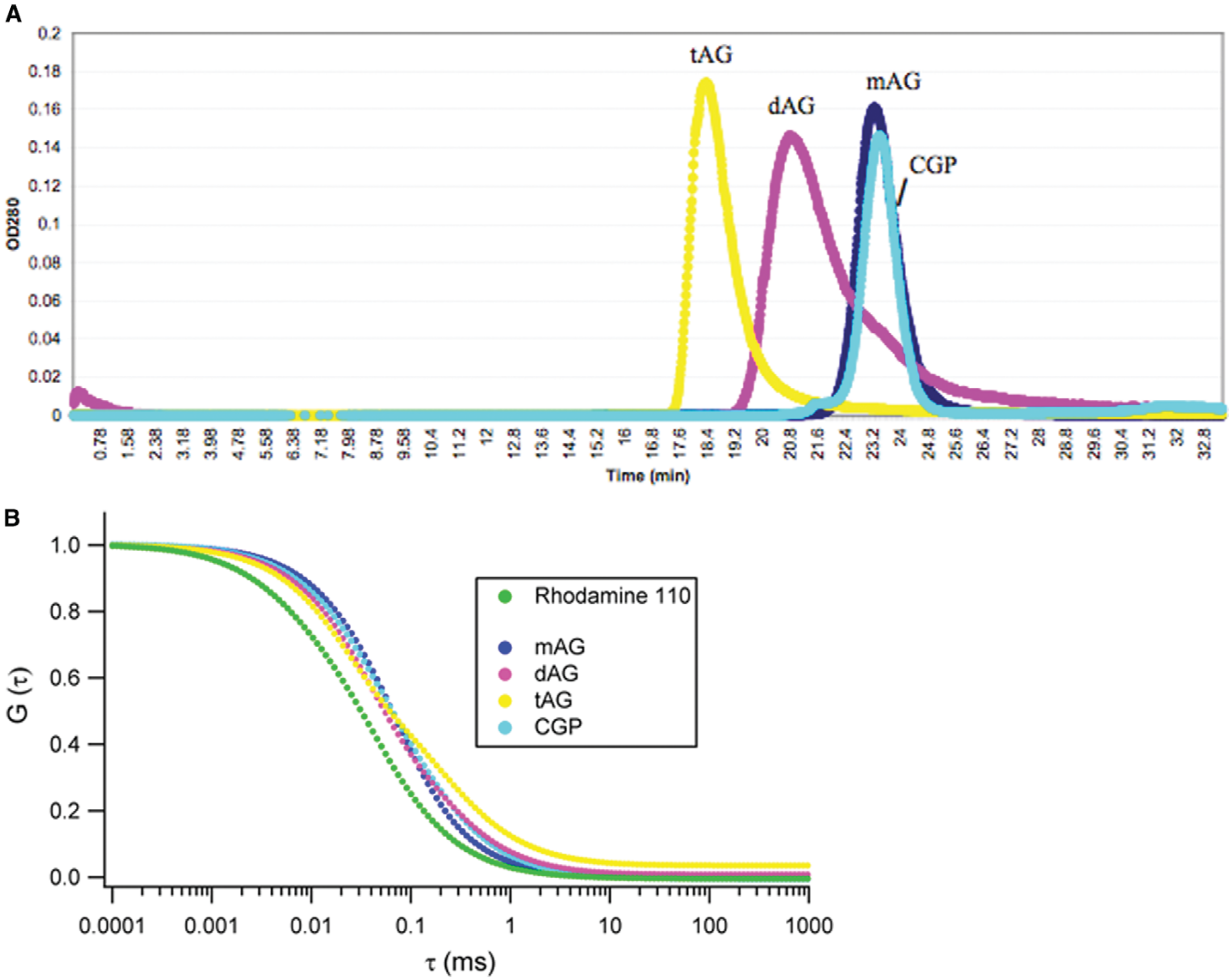


**Fig. 4.** Analysis of oligomerization status of CGP and AG proteins. (**A**) The purified proteins were analyzed by gel filtration on a Superdex 75 column, after calibration, using absorption at 260 nm. (**B**) 'Sizing' of the different proteins by FCS.

that fluorophore with the amplitude of the correlation function (which is inversely related to the number occupancy of the optical probe volume). Solid lines show an exponential fit to the data points to aid in visualization. As shown, dimeric and tetrameric AG molecules reach a plateau (optically saturate) in the excitation intensity range studied, whereas mAG and CGP emission rates continue to climb, indicating higher saturation intensities. However, when the monomers are compared, the number of emitted photons per CGP molecule increases more rapidly with increasing laser power than mAG, eventually emitting 50% more photons per molecule per second at the highest laser power (60 μW). In Fig. 6B and C, the reciprocal of the dark-state lifetime (flicker rate) and the fraction of molecules in the dark-state are plotted against the excitation laser power (lines in Fig. 6C represent the mean value of the fraction). dAG and tAG showed similar non-linear behavior, while for CGP and mAG the flicker rates increased almost linearly with laser power. Among the four proteins, CGP exhibited the strongest dependence of the dark-state lifetime on the excitation power (Fig. 6B) and the smallest fraction of molecules in the dark state (Fig. 6C). Again, comparing the monomers we observed that the fraction of molecules in the dark state and the lifetime of this dark state for mAG molecules were significantly large compared to with CGP molecules. The fluorescence decay curves and calculated lifetime values for the different proteins are presented in Fig. 7. The fluorescence lifetime of CGP (3.03 ns) is less than that of mAG (3.50 ns). While shortening of the lifetimes for the oligomeric proteins (dAG 3.06 ns, and tAG 3.13 ns) are likely due to interaction and self-quenching of the monomeric units within the multimeric state, the reduction in lifetime for CGP is likely to be an indication of a slightly different micro-environment.

## Discussion

This work describes only the second time that consensus engineering has been applied to the creation of a novel protein in which the protein was created *de novo* without examining the effects of intermediate mutations. Of the 31 proteins used to create the consensus, CGP is closest to two *Montastraea cavernosa* GFPs (AA061601 and AA061599) and mAG (BAD52002). Differences between these four

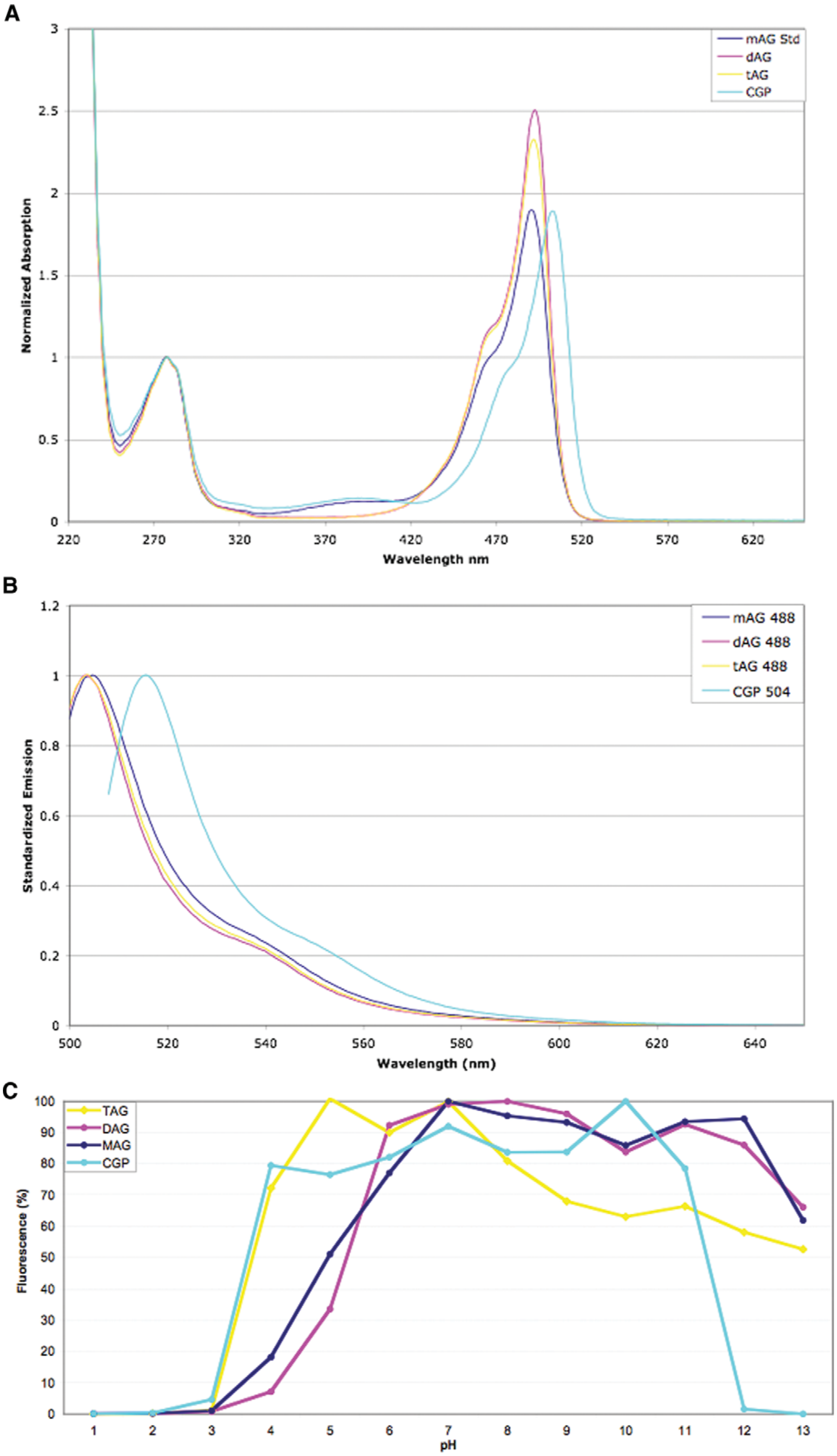


**Fig. 5.** Spectral properties of CGP and AG proteins. (**A**) Relative excitation spectra of different proteins normalized at 280 nm. (**B**) The normalized emission spectra of the fluorescent proteins upon 488 nm (mAG, dAG, tAG) or 504 nm (CGP) excitation. (**C**) Effect of pH on the fluorescence of each of the proteins, expressed as a percentage of maximal fluorescence.

**Table I.** Quantum yields of CGP and Azami proteins

| | Quantum yield (this study) (%) | Quantum yield (Karasawa *et al.*, 2003) |
|---|---|---|
| CGP | 0.55 ± 2.2 | Not appropriate |
| mAG | 0.88 ± 2.4 | 0.81 |
| dAG | 0.87 ± 2.1 | Not determined |
| tAG | 0.90 ± 2.0 | 0.67 |

The quantum yields determined here, and those provided in the first description of the AG proteins (Karasawa *et al.*, 2003) are provided.

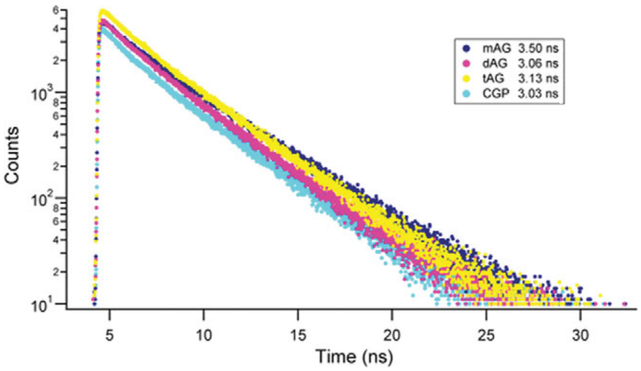


**Fig. 7.** Fluorescence decay spectra of CGP and AG proteins. Fluorescence lifetime values (shown) are extracted by fitting the curves with single-exponential equation.

proteins can be found at 42 positions and CGP differs from each of these three proteins by ~10%. Notwithstanding this sequence similarity, an examination of the evolutionary relationship between these four different proteins (Fig. 1B) shows them to be spread over the phylogenetic tree, indicating that the derived consensus sequence is not a putative ancestral sequence.

The success of consensus engineering depends crucially upon the sequences used to derive the consensus, and illustrates one of the differences between evolutionary and consensus sequences. The inclusion of a new sequence which branches near the basal node of an evolutionary tree is likely to have a strong influence on a putative ancestral sequence, but as a single sequence will have little bearing on a consensus sequence. In contrast, a consensus is heavily influenced by the manner in which sequences are sampled, and can be easily biased towards a particular sequence if too many highly similar sequences are included. It is for this reason that synthetic mutational derivatives of natural sequences and sequences differing by few amino acids were excluded. In this work, we first chose sequences on the basis of homology >62% to mAG. Almost 75% of the inter-protein comparisons showed homology of 80% or less, and only 11.6% showed homology >90%. Although the data shown here, as well as that described for the phytase consensus (Lehmann *et al.*, 2000a, 2000b, 2001, 2002), show that consensus engineering can be applied in an *ab initio* approach, it is not clear which level of percentage homology should be used to create such functional proteins.

It is possible that the reason CGP is so functional is because we only used consensus amino acids, where they exceeded 60% of the available sequences, and at positions in which no consensus amino acid could be identified, used a guide protein (mAG) to direct sequence selection, with the goal of maintaining cooperative structure.

A recent paper (Socolich *et al.*, 2005) examined more carefully the concept of consensus engineering within the context of the information required to specify a specific protein fold, and revealed that for WW domains, pure consensus sequences, in which the most abundant amino acid at each position is chosen, do not lead to native folded structures. Instead, the co-variance of amino acids at key positions must also be considered. The approach we, and many others applying consensus engineering for biotechnological purposes, have taken did not take co-variance into account explicitly. This suggests either that co-variance is especially important in small proteins (the WW domain is only 36 amino acids), but not larger proteins, or that information

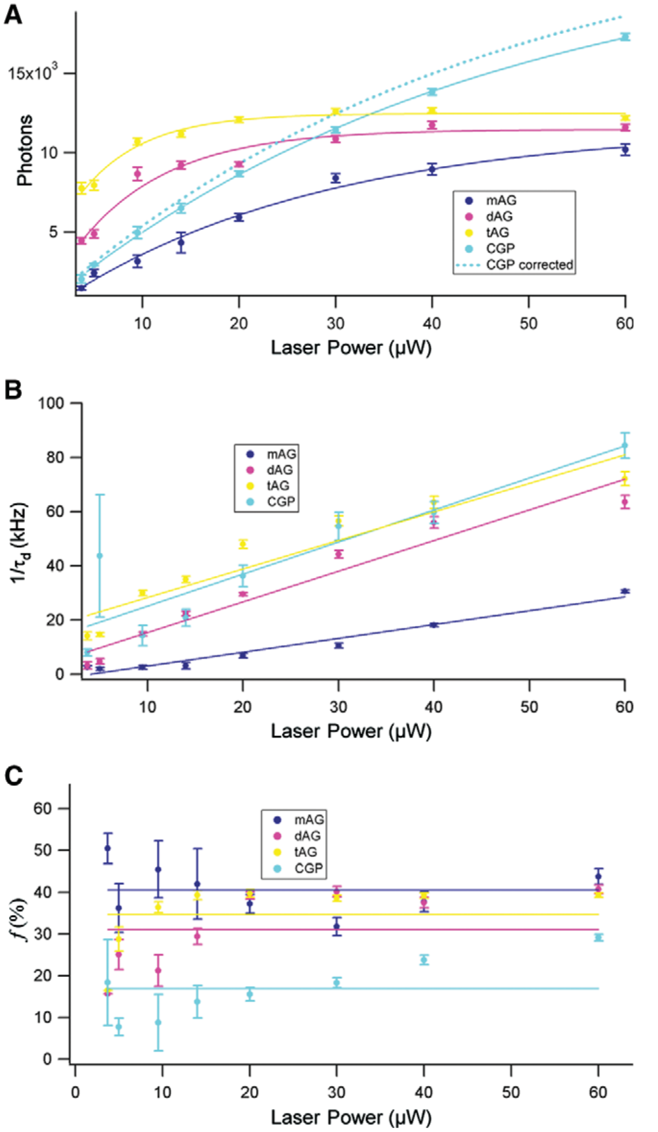


**Fig. 6.** Photophysical properties of CGP and AG proteins. (**A**) Number of photons per molecule per second versus laser power. Dashed line shows the ‘brightness’ of CGP corrected for the transmission of the filters (dichroic and bandpass) and ratios of the extinctions coefficients relative to mAG. (**B**) Flickering rate ($1/\tau_d$) versus laser power. (**C**) Fraction of molecules in the dark state versus laser power. Lines represent the mean value of all points.

on covariance in larger proteins is intrinsically encoded by the consensus sequence, especially when a guide sequence (mAG) is used to resolve weak consensus positions.

It is interesting to note that over 85% (20/23) of the amino acids changed from mAG to CGP are predicted to be found on the surface. This is similar to the proportion changed in the creation of the phytase consensus (Lehmann *et al.*, 2000b), and is likely to reflect the observation that surface amino acids evolve more quickly than buried amino acids (Goldman *et al.*, 1998; Tseng and Liang, 2006), unless they are in active sites (Tseng and Liang, 2006), or involved in protein–protein interactions (Mintseris and Weng, 2005). As a result, surface residues in general are likely to constitute a higher proportion of those amino acids changed in consensus engineering.

Although we were disappointed that CGP had a lower quantum yield than mAG (measured in bulk), it was interesting to observe the different behavior of the two proteins at the laser powers typically used in confocal fluorescence microscopy or single molecule fluorescence applications (Fig. 6). In particular, CGP appears to be brighter than mAG, with this trend increasing with increasing laser powers (Fig. 6A). Also shown in Fig. 6A is corrected' brightness per fluorophore of CGP (dashed line). This correction takes into account the different extinction coefficients of mAG and CGP at 488 nm and the different transmission efficiencies of the emitted fluorescence spectra through the excitation dichroic and band-pass filters used in the FCS measurements. Note that at all laser powers used (with and without corrections), CGP is brighter than mAG in the FCS measurements on a per molecule basis. The apparent contradiction between the bulk measured quantum yield and the increased brightness of CGP at the single molecule level is likely to be due to different excited state dynamics of the two proteins at the large excitation rates necessary for single fluorophore detection. At the laser fluences required for single molecule FCS, CGP spends a lower fraction of time spent in a dark state (Fig. 6C), and has a shorter dark-state lifetime (Fig. 6B) than mAG and thus appears brighter. We note the bulk quantum yield measurements were performed with excitation fluences of 10 $\mu W/cm^2$, whereas the *lowest* excitation fluence used in Fig. 6 was $\sim 3.0\ kW/cm^2$: over 100 millionfold higher than that used for the bulk measurements. We expect that the brightness per fluorophore for the two proteins must cross between 10 $\mu W/cm^2$ and 3 $kW/cm^2$, although we were unable to confirm this experimentally, as FCS cannot be practically carried out much below 1 $\mu W$. While the exact nature of the dark state in CGP and mAG is at present unknown, the existence of such a photo-induced dark state is consistent with other light-induced photophysical changes in other fluorescent proteins, such as color shifts (Lukyanov *et al.*, 2005). Under the intense illumination used in the FCS experiments, mAG molecules may be pumped to this non-fluorescent state more readily than CGP. Additional investigations are required to reveal the photophysical and photochemical properties of these proteins at the single molecule level and such investigations are currently under way. Depending upon the laser powers used, and the application, CGP may be a more suitable fluorophore than mAG, with the caveat that at the high laser powers used in FCS the photobleaching rate of CGP is faster than mAG.

Since this work was initiated, a large number of additional fluorescent protein sequences have been deposited in the database, and 82 unique proteins can now be identified as having homology to mAG of >62%. A preliminary examination of the consensus sequence generated by these 82 proteins reveals a number of differences with respect to CGP. Some of these would revert some amino acids back to mAG, whereas others would make changes at additional sites. In the creation of the consensus sequence described here, no account was taken of the color of the fluorescent protein. With the availability of more fluorescent protein sequences, it may now be possible to restrict the component sequences used to create the consensus to proteins of particular colors. However, whether this would lead to correspondingly colored consensus proteins awaits to be seen.

The work described here shows that consensus engineering does not always lead to more stable proteins, especially if applied in an *ab initio* approach. Lehmann *et al.* (2000a, 2000b, 2001, 2002) found that when they revisited their improved phytase, by examining the roles of individual amino acids, they were able to increase stability still further, by removing some of the consensus changes, indicating that it would be worth examining the roles of individual amino acids changed in CGP compared to the proteins which are closest. Although CGP is slightly less stable than mAG (Tm 79 versus 84.5°C), and in that sense failed to fulfill one of the goals of this work, it is remarkable that the application of consensus engineering to a class of proteins whose function is so sensitive to mutation yields such a functional protein. In addition to being brighter under excitation conditions used for single molecule studies or confocal flourescent microscopy, CGP is better expressed than the guide protein. We believe our success is, at least in part, related to the use of a guide protein to resolve ambiguities in the consensus. By using the guide protein sequence for those positions at which a consensus cannot be derived, or the consensus comprises <60% of the sequences, covariance of structurally and functionally important amino acids should be maintained.

This second application of *ab initio* consensus engineering confirms the validity of this approach to create novel proteins which do not exist in nature. We expect this to have significant biotechnological implications.

## Acknowledgements

We are grateful to Mark Prescott for helpful discussions, and DOE (GTL) for financial support to ARMB.

## References

Aragon,S. and Pecora,R. (1976) *J. Chem. Phys.*, **64**(4), 1791–1803.

Arndt,M.A., Krauss,J., Schwarzenbacher,R., Vu,B.K., Greene,S. and Rybak,S.M. (2003) *Int. J. Cancer*, **107**(5), 822–829.

Binz,H.K., Stumpp,M.T., Forrer,P., Amstutz,P. and Pluckthun,A. (2003) *J. Mol. Biol.*, **332**(2), 489–503.

Chang,B.S., Ugalde,J.A. and Matz,M.V. (2005) *Methods Enzymol.*, **395**, 652–670.

Demarest,S.J., Rogers,J. and Hansen,G. (2004) *J. Mol. Biol.*, **335**(1), 41–48.

Elson,E.L. and Magde,D. (1974) *Biopolymers*, **13**(1), 1–27.

Gao,F., Weaver,E.A., Lu,Z., Li,Y., Liao,H.X., Ma,B., Alam,S.M., Scearce,R.M., Sutherland,L.L. and Yu,J.S., *et al.* (2005) *J. Virol.*, **79**(2), 1154–1163.

Goldman,N., Thorne,J.L. and Jones,D.T. (1998) *Genetics*, **149**(1), 445–458.

Heikal,A.A., Hess,S.T., Baird,G.S., Tsien,R.Y. and Webb,W.W. (2000) *Proc. Natl. Acad. Sci. USA.*, **97**(22), 11996–12001.

Jiang,X., Kowalski,J. and Kelly,J.W. (2001) *Protein Sci.*, **10**(7), 1454–1465.

Karasawa,S., Araki,T., Yamamoto-Hino,M. and Miyawaki,A. (2003) *J. Biol. Chem.*, **278**(36), 34167–34171.

Knappik,A., Ge,L., Honegger,A., Pack,P., Fischer,M., Wellnhofer,G., Hoess,A., Wolle,J., Pluckthun,A. and Virnekas,B. (2000) *J. Mol. Biol.*, **296**(1), 57–86.

Kohl,A., Binz,H.K., Forrer,P., Stumpp,M.T., Pluckthun,A. and Grutter,M.G. (2003) *Proc. Natl. Acad. Sci. USA.*, **100**(4), 1700–1705.

Lehmann,M., Pasamontes,L., Lassen,S.F. and Wyss,M. (2000a) *Biochim Biophys Acta*, **1543**(2), 408–415.

Lehmann,M., Kostrewa,D., Wyss,M., Brugger,R., D'Arcy,A., Pasamontes,L. and van Loon,A.P. (2000b) *Protein Eng.*, **13**(1), 49–57.

Lehmann,M. and Wyss,M. (2001) *Curr. Opin. Biotechnol.*, **12**(4), 371–375.

Lehmann,M., Loch,C., Middendorf,A., Studer,D., Lassen,S.F., Pasamontes,L., van Loon,A.P. and Wyss,M. (2002) *Protein Eng.*, **15**(5), 403–411.

Levine,L.D. and Ward,W.W. (1982) *Comp. Biochem. Physiol.*, **77**, 77–85.

Lukyanov,K.A., Chudakov,D.M., Lukyanov,S. and Verkhusha,V.V. (2005) *Nat. Rev. Mol. Cell Biol.*, **6**(11), 885–891.

Magde,D., Elson,E.L. and Webb,W.W. (1974) *Biopolymers*, **13**(1), 29–61.

Maxwell,K.L. and Davidson,A.R. (1998) *Biochemistry*, **37**(46), 16172–16182.

McDonagh,C.F., Beam,K.S., Wu,G.J., Chen,J.H., Chace,D.F., Senter,P.D. and Francisco,J.A. (2003) *Bioconjug. Chem.*, **14**(5), 860–869.

Mintseris,J. and Weng,Z. (2005) *Proc. Natl. Acad. Sci. USA.*, **102**(31), 10930–10935.

Nikolova,P.V., Henckel,J., Lane,D.P. and Fersht,A.R. (1998) *Proc. Natl. Acad. Sci. USA.*, **95**(25), 14675–14680.

Ohage,E. and Steipe,B. (1999) *J. Mol. Biol.*, **291**(5), 1119–1128.

Rigler,R., Mets,U., Widengren,J. and Kask,P. (1993) *Eur. Biophys. J.*, **22**(3), 169–175.

Socolich,M., Lockless,S.W., Russ,W.P., Lee,H., Gardner,K.H. and Ranganathan,R. (2005) *Nature*, **437**(7058), 512–518.

Steipe,B. (2004) *Methods Enzymol.*, **388**, 176–186.

Steipe,B., Schiller,B., Pluckthun,A. and Steinbacher,B. (1994) *J. Mol. Biol.*, **240**(3), 188–192.

Tseng,Y.Y. and Liang,J. (2006) *Mol. Biol. Evol.*, **23**(2), 421–436.

Tsien,R.Y. (1998) *Annu. Rev. Biochem.*, **67**, 509–544.

Ugalde,J.A., Chang,B.S. and Matz,M.V. (2004) *Science*, **305**(5689), 1433.

Velapoldi,R.A. and Tonnesen,H.H. (2004) *J. Fluoresc.*, **14**(4), 465–472.

Verkhusha,V.V. and Lukyanov,K.A. (2004) *Nat. Biotechnol.*, **22**(3), 289–296.

Verkhusha,V.V., Kuznetsova,I.M., Stepanenko,O.V., Zaraisky,A.G., Shavlovsky,M.M., Turoverov,K.K. and Uversky,V.N. (2003) *Biochemistry*, **42**(26), 7879–7884.

Visintin,M., Settanni,G., Maritan,A., Graziosi,S., Marks,J.D. and Cattaneo,A. (2002) *J. Mol. Biol.*, **317**(1), 73–83.

Wang,Q., Buckle,A.M. and Fersht,A.R. (2000) *J. Mol. Biol.*, **298**(5), 917–926.

Wang,Q., Buckle,A.M., Foster,N.W., Johnson,C.M. and Fersht,A.R. (1999) *Protein Sci.*, **8**(10), 2186–2193.

Whitcomb,E.A., Martin,T.M. and Rittenberg,M.B. (2003) *J. Immunol.*, **170**(4), 1903–1909.

Widengren,J., Mets,U. and Rigler,R. (1995) *J. Phys. Chem.*, **99**(36), 13368–13379.

Wirtz,P. and Steipe,B. (1999) *Protein Sci.*, **8**(11), 2245–2250.

Yang,Z. (1997) *Comput. Appl. Biosci.*, **13**(5), 555–556.

**Received November 29, 2006; accepted December 6, 2006**

**Edited by Jim Marks**