Orig Life Evol Biosph (2014) 44:75–86
DOI 10.1007/s11084-014-9371-8

Experimental Evolution of a Green Fluorescent Protein Composed of 19 Unique Amino Acids without Tryptophan

Akio Kawahara-Kobayashi · Mitsuhiro Hitotsuyanagi · Kazuaki Amikura · Daisuke Kiga

Received: 10 October 2013 / Accepted: 25 September 2014 /
Published online: 16 November 2014

Abstract At some stage of evolution, genes of organisms may have encoded proteins that were synthesized using fewer than 20 unique amino acids. Similar to evolution of the natural 19-amino-acid proteins GroEL/ES, proteins composed of 19 unique amino acids would have been able to evolve by accumulating beneficial mutations within the 19-amino-acid repertoire encoded in an ancestral genetic code. Because Trp is thought to be the last amino acid included in the canonical 20-amino-acid repertoire, this late stage of protein evolution could be mimicked by experimental evolution of 19-amino-acid proteins without tryptophan (Trp). To further understand the evolution of proteins, we tried to mimic the evolution of a 19-amino-acid protein involving the accumulation of beneficial mutations using directed evolution by random mutagenesis on the whole targeted gene sequence. We created active 19-amino-acid green fluorescent proteins (GFPs) without Trp from a poorly fluorescent 19-amino-acid mutant, S1-W57F, by using directed evolution with two rounds of mutagenesis and selection. The N105I and S205T mutations showed beneficial effects on the S1-W57F mutant. When these two mutations were combined on S1-W57F, we observed an additive effect on the fluorescence intensity. In contrast, these mutations showed no clear improvement individually or in combination on GFPS1, which is the parental GFP mutant composed of 20 amino acids. Our results provide an additional example for the experimental evolution of 19-amino-acid proteins without Trp, and would help understand the mechanisms underlying the evolution of 19-amino-acid proteins. (236 words)

Keywords Directed evolution · Green fluorescent protein · Tryptophan · 19 amino acids · Genetic code

Electronic supplementary material The online version of this article (doi:10.1007/s11084-014-9371-8) contains supplementary material, which is available to authorized users.

A. Kawahara-Kobayashi · M. Hitotsuyanagi · K. Amikura · D. Kiga (✉)
Department of Computational Intelligence and Systems Science, Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, Kanagawa 226-8503, Japan
e-mail: kiga@dis.titech.ac.jp

K. Amikura · D. Kiga
Earth-Life Science Institute, Tokyo Institute of Technology, Meguro, Tokyo 152-8551, Japan

Introduction

Proteins present in existing organisms are composed of 20 different amino acids, encoded by the universal genetic code. This code is considered to have evolved from simpler forms (Crick 1968; Wong 1975; Fournier and Gogarten 2007), such as one that encoded 19 amino acids. Similarly, genes encoding proteins composed of 20 unique amino acids are considered to have evolved from those encoding proteins composed of 7–13 unique amino acids (Wong 1976; Osawa et al. 1992; Baumann and Oro 1993) available in the primitive earth conditions (Miller 1953). Along with the establishment of novel cellular biosynthetic pathways, new amino acids were gradually introduced into the repertoire for synthesizing diverse proteins (Wong 1976). This theory is supported by the phylogenies of the aminoacyl-tRNA synthetases (aaRSs) (Woese et al. 2000).

Tryptophan (Trp) is thought to be the last amino acid included within the universal genetic code. Biosynthetic pathway of Trp pointed to the possibility that this amino acid was introduced during the late stages of the evolution of genetic code (Wong 1975). Indeed, Trp is hardly generated during chemical evolution experiments imitating primordial conditions (Miller 1953). Further, sequence analysis of aaRSs showed that tryptophanyl-tRNA synthetase (TrpRS), as well as tyrosyl-tRNA synthetase (TyrRS), are the youngest among 20 aaRSs (Ribas de Pouplana et al. 1996). At present, Trp is believed to be the last amino acid introduced during evolution (Trifonov 2004).

Proteins composed of 19 amino acids would have been able to evolve by accumulating beneficial mutations without using the 20th canonical amino acid. A previous report has described the experimental evolution of two naturally occurring proteins, GroEL and GroES, composed of 19 unique amino acids without Trp (Wang et al. 2002). Directed evolution improved the ability of *E. coli*-derived GroEL and GroES to assist folding of GFP. In spite of random mutagenesis to the DNAs, mutants with improved properties did not harbor Trp. Directed evolution has been successfully used for improving the properties of 19-amino-acid proteins without Cys. A single-chain-dimer streptavidin had been evolved to improve binding activity to biotin, without introducing Cys (Aslan et al. 2005).

To further understand the evolution of proteins composed of 19 distinct amino acids, we created active 19-amino-acid green fluorescent proteins (GFPs) without Trp using directed evolution with repeated rounds of random mutagenesis and selection. Here, we describe a less active mutant composed of 19 amino acids without Trp that generated by replacing Trp^{57}. This mutant gained activity by accumulating beneficial mutations through two rounds of experimental evolution. We found that the combined effect of the two beneficial mutations on the initial mutant was additive, like in the case of GroEL, where the stability of the protein improved because of the combined effect of mutations (Wang et al. 1999). Recording the improvement in the activity of the 19-amino-acid protein without Trp using experimental evolution, where it is easy to observe the accumulation of beneficial mutations, would provide additional insights into the late stages of protein evolution.

Materials and Methods

DNA Construction

The sequences of the primers used in DNA construction are shown in Table S1. Site-directed mutagenesis was performed using the Gene Tailor site-directed mutagenesis system (Life Technologies).

For construction of the plasmid DNA encoding the GFP mutants, a coding region of GFPS1 on pGFPS1 (Seki et al. 2008) was amplified using primers named Placq-*Bgl*_II-RBS-F and Placq-*Spe*_I-R (Table S1). The amplified fragment was cloned into Placq-GFP (Ayukawa et al. 2010) between the *Bgl* II and *Spe* I sites to replace the GFP coding sequence with GFPS1. We called this construct Placq-GFPS1. Site-directed mutagenesis on Placq-GFPS1 was performed to create Placq-S1-W57F and Placq-S1-W57Y by using forward primers named W57F-F and W57Y-F, respectively, and a common reverse primer named W57FY-R. Saturation mutagenesis using degenerate oligonucleotides (NNK codon with *N*=G/A/T/C and *K*=G/T) was performed on S1-W57F to introduce mutations to amino acids adjacent to Phe^{57}. For the mutation at Val^{29}, we used a pair of primers named V29NNK-F and V29NNK-R. Similarly, for other single-point mutations at Phe^{46}, Leu^{53}, Leu^{60}, and Leu^{64}, we used pairs of primers named F46NNK-F and F46NNK-R, L53NNK-F and L53NNK-R, L60NNK-F and L60NNK-R, L64NNK-F and L64NNK-R, respectively. For the mutation at the three consecutive positions, Asp^{216}, His^{217}, and Met^{218}, we used a pair of primers named D216H217M218NNK-F and D216H217M218NNK-R.

For the construction of the first and second round plasmid libraries, error-prone PCR was performed as previously described (Cadwell and Joyce 1994) on the S1-W57F coding region. Typically, a 50-μL reaction mixture contained 10 mM Tris–HCl (pH 8. 3), 50 mM KCl, 3 mM $MgCl_2$, 0.5 mM $MnCl_2$, 0.2 mM of each dATP and dGTP, 1.0 mM of each dTTP and dCTP, 0.2 μM of each primer set, 1 ng of template DNA, and 1.25 U of rTaq DNA polymerase (TOYOBO). The primers and procedure used for cloning were the same as those for Placq-GFPS1 construction. We found that mutation frequency was 6.7 mutations per 1,000 base pairs. The mixture of the random mutant plasmids was introduced into DH5α by electroporation. Mutant libraries were screened by FACS method. Typical mutant libraries for each FACS screening consisted of approximately 10^6 independent clones.

For the purpose of measuring the fluorescence intensity of the purified mutant proteins, we introduced a C-terminal His tag sequence YRYEFQLSAASKLAAALEHHHHHH into the sequences. In order to introduce a His tag sequence into the mutants derived from Placq-GFPS1, we applied the following three steps.

In step 1, we prepared the coding sequence of GFPS1-*Cla*_I-His that featured GFPS1 sequence followed by a *Cla* I recognition sequence and a His tag sequence. For this purpose, pGFPS1-*Cla*_I-His was constructed by introducing the *Cla* I site and His tag sequence simultaneously to pGFPS1 by site-directed mutagenesis with primers named pGFP-*Cla*_I-His-F and pGFP-*Cla*_I-His-R. The forward primer contained the *Cla* I site located between the GFPS1 coding region and the His tag to enable cloning as desired in the final step. By insertion of an additional nucleotide between the GFPS1 coding region and an out-of-frame His tag sequence, the primers generate the His tag sequence in frame.

In step 2, the GFPS1 coding sequence on Placq-GFPS1 was replaced with GFPS1-*Cla*_I-His. The GFPS1-*Cla*_I-His PCR fragment was amplified from pGFPS1-*Cla*_I-His by using primers named pGFPS1-F and GFP-His *Spe*_I-R. The fragment contained the *Nde*I site at the start codon of the GFPS1-*Cla*_I-His and *Spe* I site next to the stop codon of GFPS1-*Cla*_I-His. The PCR fragment was cloned into Placq-GFPS1 between the *Nde*I and *Spe*I sites to replace the GFPS1 coding sequence on Placq-GFPS1 with the GFPS1-*Cla*_I-His. We called this construct Placq-GFPS1-*Cla*_I-His.

In the third and final step, the coding sequences of the EGFP and GFPS1 mutants obtained from the selection in this work were amplified and cloned into Placq-GFPS1-*Cla*_I-His between the *Nde*I and *Cla*I sites to replace the GFPS1 coding region with the mutant sequences. For this purpose, we used Placq-*Bgl*_II-RBS-F, the same forward primer for the construction of Placq-GFPS1, and GFP-cloning-*Cla*_I-R.

Flow Cytometric Analysis and Cell Sorting

All flow cytometric analyses and cell sorting were carried out using a Becton-Dickinson FACSCalibur instrument with a 488-nm laser and a 515- to 545-nm emission filter. The throughput rate of cells was adjusted to 3,000 events per second. For the library constructed by saturation mutagenesis, 10^6 cells were analyzed. For the library constructed by random mutagenesis, 2×10^7 cells were sorted in exclusion mode. A gate in the fluorescence channel was set to recover highly fluorescent cells. Additional gates were set in the forward- and side-scatter channels to exclude events arising from large particles. The collected bacterial cells were centrifuged and rescued in LB medium. The cells were plated on LB dishes supplemented with 30 μg/mL kanamycin and incubated at 37 °C overnight. Colonies were picked, cultured, and analyzed using CellQuest (Becton Dickinson) and WinMDI 2.9 (http://en.bio-soft.net/other/WinMDI.html).

Protein Expression and Purification

Plasmids containing the GFP mutants bearing the His-tag-coding sequence were transformed into DH5α. Single colonies were picked and cultured at 37 °C overnight in 3 mL LB medium supplemented with 30 μg/mL kanamycin. Overnight culture (1 mL) was transferred into 100 mL of fresh LB medium containing 30 μg/mL kanamycin, and incubation was continued at 37 °C. When the optical density of cultures reached 0.6 at 590 nm, the temperature was lowered to 26 °C, and protein expression was continued for an additional 4 h. Following this, cells were collected by centrifugation for 5 min at 7,000×*g* and 4 °C. The harvested cells were suspended in 4 mL of buffer A (40 mM Tris–HCl [pH 7.6], 300 mM NaCl, 10 mM imidazole, 1 mM DTT) and disrupted by sonication. The disrupted cells were centrifuged for 10 min at 12,000×*g* and 4 °C. Aliquots (4 mL) of the supernatant were mixed with 6 mL buffer A, and filtered through Millex-GV 0.45-μm filters (Millipore). The filtrate was incubated with 500 μL bed volume of TALON resin (Clontech) at 4 °C for 1 h. The resin was then retrieved and washed with 30 mL buffer A. Bound proteins were eluted with elution buffer (buffer A containing 300 mM imidazole).

Fluorescence Intensity Measurement on Purified Mutant Proteins

For fluorescence measurements, 15.0 μg of purified protein was diluted in 2,000 μL PBS and analyzed at room temperature with an FP-6500 spectrofluorometer (JASCO). The excitation wavelength was set at 488 nm, and the emission wavelength between 500 and 730 nm (bandwidth =1 nm for both excitation and emission) was recorded.

Results

The loss of fluorescence due to mutation of a unique Trp in the parent GFP was not restored by a mutation of the Trp position itself or adjacent amino acids

As a starting point for mutagenesis in this work, we used a simplified GFP mutant named S1-W57F, in which Phe replaced the unique Trp^{57} (Fig. 1) of a brightness- and maturation speed-improved mutant GFPS1 (Seki et al. 2008). Enhanced GFP has been reported to not to allow Trp^{57} replacement by any of the other 19 amino acids (Steiner et al. 2008) or by non-canonical Trp-like amino acids (Budisa et al. 2004) without abolishing fluorescence. Accordingly, we confirmed that

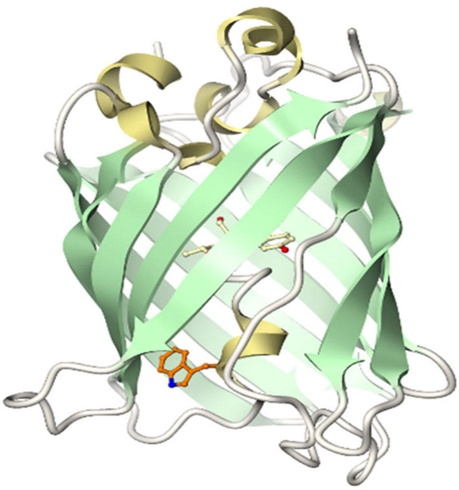

Fig. 1 Ribbon structure of EGFP (PDB entry 2y0g). The Trp residue and the central chromophore are highlighted as stick models. The Trp residue is shown in orange and the chromophore is shown in yellow. Molecular graphics were illustrated with CueMol 2 (http://www.cuemol.org/)

the replacement of Trp^{57} by the structurally similar residue Phe or Tyr abolished the fluorescence of GFPS1 (Fig. S1). We then attempted saturation mutagenesis at each position spatially close to Phe^{57} of S1-W57F to rescue the protein from structural perturbation introduced by the W57F mutation. Therefore, we prepared mutants in which codons for one of the amino acids Val^{29}, Phe^{46}, Leu^{53}, Leu^{60}, or Leu^{64} (Fig. S2) was replaced by NNK. Fluorescence intensities of *E. coli* libraries, consisting of 10^{5} colonies expressing each mutant were then measured by FACS method after 488-nm excitation. We found that none of the mutations made for correcting Trp^{57} mutation improved the fluorescence of S1-W57F. Further, we introduced saturation mutagenesis simultaneously at three sequential positions, Asp^{216}, His^{217}, and Met^{218}, which are also close to Phe^{57} and attempted to select active mutants from 10^{6} cells for 10^{5} clones, an adequate library size for the number of residues randomized. The strategy, however, did not improve the fluorescence of S1-W57F.

Repeated Rounds of Directed Evolution on 19-Amino-Acid GFP Without Trp

Because saturation mutagenesis at positions spatially close to Phe^{57} failed to yield any mutations that improved the fluorescence of S1-W57F, we used an error-prone PCR (epPCR) on the whole S1-W57F coding sequence. The epPCR products of the S1-W57F coding sequence were then cloned into Placq-GFP (Ayukawa et al. 2010). From an *E. coli* mutant library consisting of 10^{6} mutants carrying plasmids that harbor the epPCR product, we selected clones that showed higher fluorescence than S1-W57F mutant cells. Plasmids were isolated from 10 colonies, and their sequences were analyzed (Fig. S3 and Table 1). We found that these sequences contained 13 amino acid changes, distributed throughout the S1-W57F gene. Because some clones had the same amino-acid sequence, in all, 6 sequences were found to be varied in the analyzed clones. All of the 6 mutant S1-W57F sequences lacked Trp in their whole coding region including at position 57. The selected clones lacked deletions, insertions, or additional stop codons. The most frequent mutation was S205T, which is adjacent to the chromophore, neither adjacent nor close to Phe^{57}. This mutation was reported as one of the mutations that enhanced folding (Cubitt et al. 1999). A comparison of the fluorescence intensity of the mutant clones corresponding to the 6 distinct sequences revealed that mutants containing S205T mutation produced a protein with higher fluorescence intensity than the rest of the mutants (Table 1).

Table 1 Amino acid mutations, fluorescence intensities, and the number of clones for the mutants from the first round of the selection

Mutant	Amino acid mutations from S1-W57F	Fluorescence intensity*	Number of clones
S1-W57F	–	20	-
R1-1	S205T	98**	4
R1-2	K1I, S205T	100	1
R1-3	K166E, S205T	96	1
R1-4	Y145F, E235K	54	1
R1-5	Q184L, L236P	52	1
R1-6	K1R, Y145H, C200Y	89***	2

* The median cell fluorescence measured by a cell sorter equipped with an argon ion laser emitting at 488 nm

** Average value of 4 clones

*** Average value of 2 clones

Additional active 19-amino-acid mutants without Trp were generated from a second round library prepared from the R1-1 mutant harboring a single S205T mutation. After mutagenesis and selection as described for the first round, we sequenced 3 mutants and found that each contained 1 additional amino acid mutation (Table 2). One of the substitutions, N105I, was found at the same residue as the N105T on the superfolder mutant (Pedelacq et al. 2006) and N105Y on the superfast mutant (Fisher and DeLisa 2008). Further, these mutations were not close to the chromophore, Phe^{57}, or Thr^{205}. All three mutants had higher fluorescence intensity than that of the predecessor R1-1.

Fluorescence Intensities of Purified 19-Amino-Acid Mutants Without Trp

To purify mutant proteins and evaluate the fluorescence intensity, we introduced His-tag to 7 representative clones: previously reported GFPs (EGFP and GFPS1), the initial simplified mutant (S1-W57F), and selected mutants (R1-1, R2-1, R2-2, and R2-3). Upon excitation at 488 nm, all of the simplified mutants showed similar fluorescence spectra to that of EGFP and GFPS1 (Fig. S5). The order of the fluorescence intensity among the selected mutants was the same regardless of whether quantification was done using the *E. coli* cells expressing the GFP mutants or purified proteins. We also found that the fluorescence intensity of R1-1 was higher

Table 2 Amino acid mutations, fluorescence intensities, and the number of clones for the mutants from the second round of the selection

Mutant	Amino acid mutations from R1-1	Fluorescence intensity*,**,***	Number of clones
S1-W57F	–	4.8	–
R1-1	–	33	–
R2-1	G232D	56	1
R2-2	E172V	71	1
R2-3	N105I	84	1

* The median cell fluorescence measured by a cell sorter equipped with an argon ion laser emitting at 488 nm

** Detector voltage conditions, which determine the sensitivity, differ between Tables 1 and 2

*** Fluorescence histograms for cells expressing each selected clone from the second round are presented in Fig. S4 than that of S1-W57F and that of R2-3 was 4.3 times higher than that of R1-1 (Fig. S5). The most active 19-amino-acid mutant without Trp showed 32 and 50 % fluorescence intensity of GFPS1 and EGFP, respectively, which are proteins that consist of the 20 canonical amino acids.

To further evaluate the effect of the beneficial mutations on 19-amino-acid mutants without Trp, we constructed additional mutants N105I and/or S205T of the 19-amino-acid protein S1-W57F and the 20-amino-acid parental protein GFPS1. As expected, each of N105I and S205T improved fluorescence of S1-W57F (Fig. 2). Further, we found that the combined effect of N105I and S205T on S1-W57F was nearly additive. However, the beneficial effects of these mutations, individually and in combination, were not as apparent on GFPS1 as they were on S1-W57F.

Discussion

In this report, we provide an additional experimental example of the accumulation of beneficial mutations on a 19-amino-acid protein without Trp, which is thought to be the last amino acid

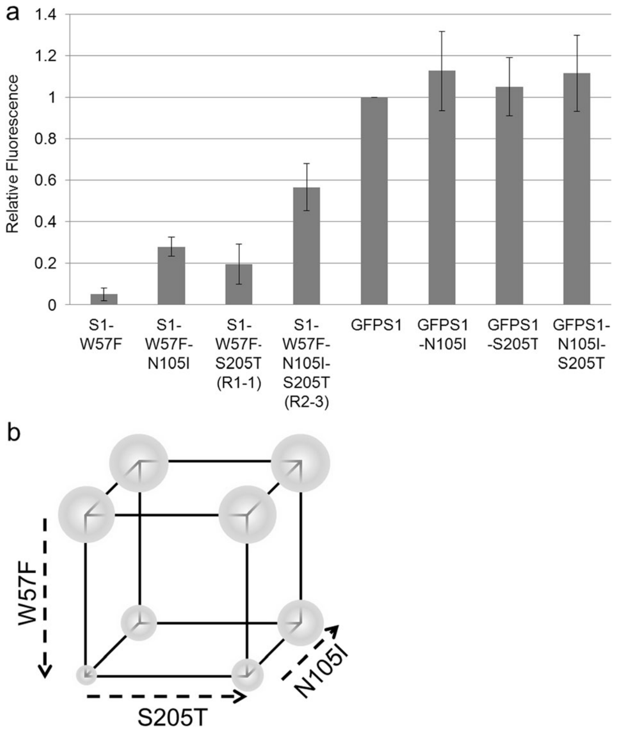

Fig. 2 Evaluation of the effect of mutations that improved the activity of 19-amino-acid mutants without Trp. **a**) Relative fluorescent intensity of the GFP mutants. N105I and S205T, the two mutations appeared in R2-3, were introduced both individually and in combination on the 19-amino-acid mutant S1-W57F and the 20-amino-acids parental mutant GFPS1. The purified proteins were excited at 488 nm, and emission intensity at 510 nm was recorded. Fluorescence intensity of GFPS1 was set at 1. The protein concentration for all measurements was set to 7.5 μg/ml in PBS. Results are presented as means±SD (n=3). **b**) Schematic diagram illustrating that beneficial mutation is detected more easily on a poorly active mutant than on a highly active mutant. The corners of the cube represent the genotypes of GFP mutants. The sizes of the spheres represent fluorescence activities. The bottom face indicates 19 amino-acid mutants. The top face indicates 20 amino-acid mutants

to be included within the universal genetic code. First, W57F mutation in GFPS1 abolished fluorescence (Fig. S1). Following the W57F replacement, using directed evolution with two rounds of random mutagenesis, we showed that S1-W57F, which is the initial 19-amino-acid mutant without Trp, could evolve and gain activity. Trp^{57}, a unique Trp residue in *Aequorea*-derived GFPs, is one of the residues responsible for chromophore formation, which forms the origin of green fluorescence. The process of chromophore formation requires a hydrophobic environment to allow the autocatalytic cyclization of the residues at positions 65–67. Trp^{57} is a part of the proline-rich "PVPWP" pentapeptide motif of GFP and contributes to maintaining the hydrophobic environment required to protect the chromophore from collisional quenching by oxygen or other diffusible ligands (Steiner et al. 2008). Fluorescence loss caused by the W57F mutation on GFPS1 may be attributed to a perturbation of the hydrophobic environment, provided that the mutations introduced for constructing GFPS1 (Seki et al. 2008) did not change the hydrophobic environment drastically. The reduction of chromophore formation observed when the W57F mutation was introduced in wild-type GFP (Bell et al. 2003) can also be attributed to the same reason.

Moreover, we observed an additive effect of the beneficial mutations on the 19-amino-acid protein without Trp using experimental evolution involving random mutagenesis to the whole gene sequence. Both the single mutations, N105I and S205T, improved the activity of S1-W57F (Fig. 2). When a double mutation (N105I and S205T) was introduced to S1-W57F, the fluorescence intensity was enhanced in a nearly additive manner. A previous report describes S205T as a mutation that enhances folding (Cubitt et al. 1999). N105I was found to occur at the same site as N105T and N105Y mutations, both of which showed increased tolerance against denaturants during a refolding experiment on the superfolder (Pedelacq et al. 2006) and superfast mutants (Fisher and DeLisa 2008), respectively. Therefore, S205T, and probably N105I also, might contribute to improve the folding of S1-W57F that had been disrupted by the W57F mutation. Although the combination of two individual beneficial mutations does not always show additive effects (Weinreich et al. 2005), we observed that S1-W57F accumulated two individual mutations and evolved to R2-3 with a simple uphill walk, which is one of the key features of directed evolution (Tracewell and Arnold 2009). Even on 19-amino-acid proteins, additivity of beneficial mutations in directed evolution was also suggested by the accumulation of beneficial amino acid replacements although this study selected the replacements by comparison of protein family sequences (Wang et al. 1999). Further, our present study in fact demonstrated the additivity of beneficial mutations that had accumulated by directed evolution involving repeated rounds of random mutagenesis to the whole gene sequence.

The active 19-amino-acid GFP mutants without Trp created in this study could be used as scaffolds that allow the introduction of Trp at any position. Such experiments would provide insights into the folding process of GFP. Because of its unique fluorescence properties, Trp has been utilized as a folding reporter to trace refolding processes. The fluorescence peak of Trp shifts to a longer wavelength when the protein of interest is denatured, because of the exposure of a Trp residue from the protein's interior to solvent water. The refolding process of GFP variants has been investigated using Trp^{57} and a chromophore (Enoki et al. 2004). Studies that involve the construction of various single Trp mutants by introducing a unique Trp at various positions throughout the structure into a Trp-free background might provide useful information about the protein folding process, as was shown on another protein (Maki et al. 2007). Re-introduction of Trp on a 19-amino-acid GFP without Trp would help understand the process of protein folding in detail. A 19-amino-acid GFP mutant without Cys has been created by combining two mutations, C48S and C70M, which were found by site-directed mutagenesis and saturation mutagenesis on Cys residues, respectively (Suzuki et al. 2012). This mutant improved the diffusional property of GFP in an oxidative environment and enhanced the applicability of the GFP tag in the analysis of the secretory pathway of proteins.

Because beneficial mutations on poorly active proteins can be readily detected when compared to that on highly active proteins containing 20 unique amino acids (Kotsuka et al. 1996; Zhou et al. 1996), we first created a 19 amino-acids mutant protein by site-directed mutagenesis. We then selected active mutants from a library prepared by random mutagenesis of the initial mutant. Indeed, N105I and S205T mutations found in R2-3 improved fluorescence of S1-W57F even when introduced individually (Fig. 2). In contrast, these mutations showed no clear improvement on the parental GFPS1. This asymmetric effect between different genetic backgrounds, known as epistasis (Lehner 2011; de Visser et al. 2011), is the basis of the "suppressor mutation method" in protein engineering (Kotsuka et al. 1996).

Trp was introduced into the genetic code likely because a bulky hydrophobic amino acid was needed for proteins to efficiently form a hydrophobic core. Although the universal genetic code has utilized Trp as one of the key building blocks, proteins can artificially evolve without Trp. For example, previous reports showed direct evolution of 19-amino-acid proteins, by introducing random mutagenesis to whole gene sequences (Wang et al. 2002; Aslan et al. 2005). Similarly, another report suggested that even an 18-amino-acid protein could evolve, while only part of the gene sequence was mutagenized in the report (Moroz et al. 2013). Proteins composed of 19 amino acids without Trp have also emerged through the natural evolution. For example, the percentage of *E. coli* proteins composed of 19 amino acids other than Trp is 11 %, which is the second largest number among *E. coli* protein composed of 19 amino acids, following 15 % for the percentage of proteins composed of 19 amino acids other than Cys (Han and Lee 2006). Furthermore, Trp is not essential for the catalytic moiety of *E. coli* enzymes (Pezo et al. 2013). If Trp does not provide catalytic activity, the question arises as to what is the significance of Trp addition to the genetic code. Introduction of Trp, which is the most bulky amino acids among the canonical ones, can strengthen hydrophobic interaction, as seen in another GroEL study (Muller et al. 2013). Fournier and Gogarten also reported that one of the most significant trends identified during the expansion of genetic code is an increase in the number and kind of aromatic amino acids, which can contribute to the packing of the protein hydrophobic core and helped in folding (Fournier and Gogarten 2007). If genes evolved to be longer to provide a new function, there will be a greater need for more amino acids that form the core of the protein than that form the surface to ensure proper folding (Supplementary Text). Therefore, the bulky hydrophobic amino acids would have contributed to occupy large volume of protein core (Alvarez-Carreno et al. 2013). With the length of the protein held constant, the occupation of the core by bulky hydrophobic amino acids allows allocation of more hydrophilic amino acid residues to the protein surface that is responsible for protein-protein interactions (Fig. S6). In this regard, addition of Trp, which has bulkiest side chain among the 20 canonical amino acids, was more significant than the addition of the other canonical amino acids.

Reappearance of excluded amino-acids often complicates the process of selecting an active simplified protein from the library that is constructed by random mutagenesis (Yamamoto et al. 2003). In our study, the reappearance of the UGG codon through random mutagenesis was relatively rare, since only one codon specifies Trp in the genetic code. However, if one attempts to exclude various types of amino acids through directed evolution, the reappearance of the codons for the specific amino acids through random mutagenesis would be inevitable. Reappearance of codons for the specific amino acids decreases the effective library size, thereby crippling the process of accumulating beneficial mutations. Thus, previous studies did not use directed evolution for the construction of simplified proteins. Reappearance of the codons for the specific amino acids during the directed evolution involving repeated rounds of random mutagenesis can be minimized using a simplified genetic code which assigns fewer than 20 amino acids to the sense codons (Kawahara-Kobayashi et al. 2012). Because the simplified genetic code completely excludes the specific amino acid from the genetic code, it is not incorporated into proteins translated from any mRNA. Although this code is currently an in vitro technique only, combination of this code with in vitro directed evolution method involving large library size, such as ribosome display, would lead efficient creation of simplified proteins. Therefore, the simplified code allows for the application of a conventional directed evolution strategy for the creation of simplified proteins to assess the functions of primordial proteins as well as to improve the utility of pharmaceuticals.

Creation of proteins composed of 19 unique amino acids could contribute to future creation of an organism that utilizes only 19 kinds of amino acids in its genetic code. To address the question whether the number of unique amino acids used in the proteome of an organism can be lowered, Pezo tried to eliminate Trp from *E. coli* proteome using a missense suppressor $\mathrm{tRNA}^{\mathrm{His}}$ bearing the anticodon for Trp (Pezo et al. 2013). If their trial would have succeeded and if one additionally eliminates TrpRS and/or $\mathrm{tRNA}^{\mathrm{Trp}}$ from the organism, the UGG codon becomes completely reassigned. Strategies involving global elimination of a specific amino acid in vivo could have lethal effects due to the severe damage to the host proteome. To create an organism that utilizes only 19 kinds of amino acids in its genetic code in future, we should create functional protein sequences without the specific amino acid and combine them into one genome by using genome engineering technology (Isaacs et al. 2011). In studies that use a combination of the simplified genetic code and directed evolution to facilitate selection, an active 19-amino-acid GFPs could be used as a folding reporter to distinguish folded or aggregated mutants of a target protein, when expressed as a C-terminal fusion construct of the target protein (Waldo et al. 1999). This would facilitate selection, especially if the initial 19-amino-acid inactive mutant or the mutants in the randomized library are aggregated. The creation of an organism that utilizes only 19 kinds of amino acids in its genetic code would give further implications about the significance of the appearance of the amino acids in the genetic code.

In conclusion, we improved a 19-amino-acid GFP without Trp through two rounds of random mutagenesis and selection. The combined effect of beneficial mutations in the poorly fluorescent 19-amino-acid GFP without Trp were additive. In contrast, these mutations showed no clear improvement individually or in combination on the parental GFP composed of 20 unique amino acids. The experimental evolution of proteins composed of fewer than 20 amino acids, in this and previous studies (Wang et al. 2002), suggested that proteins translated by genetic codes just before the completion of the universal genetic code would have evolved in the same manner as proteins composed of 20 amino acids. On the other hand, a study that severely limited the kinds of amino acids at non-conserved regions of a protein showed that some amino-acid sets were less effective than others in generating functional proteins (Tanaka et al. 2011). What amino-acid set is enough to retain evolvability equivalent to that of the 20-amino-acid set is an interesting question that warrants future research.

Acknowledgments The authors appreciate S. Yokoyama and T. Kigawa for providing the plasmid coding GFPS1. The authors thank S. Murata and K. Komiya for their technical advice. This research was supported by KAKENHI programs [19680016, 21680026, and 23119005 to D.K.]; Japan Society for the Promotion of Science (JSPS); Ministry of Education, Culture, Sports, Science and Technology (MEXT); Industrial Technology Research Grant Program in 2005 from the New Energy and Industrial Technology Development Organization (NEDO) to D.K; Research Fellowships of the Japan Society for the Promotion of Science for Young Scientists to A.K-K.

References

Alvarez-Carreno C, Becerra A, Lazcano A (2013) Norvaline and Norleucine may have been more abundant protein components during early stages of cell evolution. Orig Life Evol Biosph

Aslan FM, Yu Y, Mohr SC, Cantor CR (2005) Engineered single-chain dimeric streptavidins with an unexpected strong preference for biotin-4-fluorescein. Proc Natl Acad Sci U S A 102(24):8507–8512

Ayukawa S, Kobayashi A, Nakashima Y, Takagi H, Hamada S, Uchiyama M, Yugi K, Murata S, Sakakibara Y, Hagiya M, Yamamura M, Kiga D (2010) Construction of a genetic AND gate under a new standard for assembly of genetic parts. BMC Genomics 11:S16

Baumann U, Oro J (1993) Three stages in the evolution of the genetic code. Biosystems 29(2–3):133–141

Bell AF, Stoner-Ma D, Wachter RM, Tonge PJ (2003) Light-driven decarboxylation of wild-type green fluorescent protein. J Am Chem Soc 125(23):6919–6926

Budisa N, Pal PP, Alefelder S, Birle P, Krywcun T, Rubini M, Wenger W, Bae JH, Steiner T (2004) Probing the role of tryptophans in Aequorea victoria green fluorescent proteins with an expanded genetic code. Biol Chem 385(2):191–202

Cadwell RC, Joyce GF (1994) Mutagenic PCR. PCR Methods Appl 3(6):S136–S140

Crick FH (1968) The origin of the genetic code. J Mol Biol 38(3):367–379

Cubitt AB, Woollenweber LA, Heim R (1999) Understanding structure-function relationships in the Aequorea victoria green fluorescent protein. Methods Cell Biol 58:19–30

de Visser JA, Cooper TF, Elena SF (2011) The causes of epistasis. Proc Biol Sci 278(1725):3617–3624

Enoki S, Saeki K, Maki K, Kuwajima K (2004) Acid denaturation and refolding of green fluorescent protein. Biochemistry 43(44):14238–14248

Fisher AC, DeLisa MP (2008) Laboratory evolution of fast-folding green fluorescent protein using secretory pathway quality control. PLoS ONE 3(6):e2351

Fournier GP, Gogarten JP (2007) Signature of a primitive genetic code in ancient protein lineages. J Mol Evol 65(4):425–436

Han MJ, Lee SY (2006) The Escherichia coli proteome: past, present, and future prospects. Microbiol Mol Biol Rev 70(2):362–439

Isaacs FJ, Carr PA, Wang HH, Lajoie MJ, Sterling B, Kraal L, Tolonen AC, Gianoulis TA, Goodman DB, Reppas NB, Emig CJ, Bang D, Hwang SJ, Jewett MC, Jacobson JM, Church GM (2011) Precise manipulation of chromosomes in vivo enables genome-wide codon replacement. Science 333(6040):348–353

Kawahara-Kobayashi A, Masuda A, Araiso Y, Sakai Y, Kohda A, Uchiyama M, Asami S, Matsuda T, Ishitani R, Dohmae N, Yokoyama S, Kigawa T, Nureki O, Kiga D (2012) Simplification of the genetic code: restricted diversity of genetically encoded amino acids. Nucleic Acids Res 40(20):10576–10584

Kotsuka T, Akanuma S, Tomuro M, Yamagishi A, Oshima T (1996) Further stabilization of 3-isopropylmalate dehydrogenase of an extreme thermophile, Thermus thermophilus, by a suppressor mutation method. J Bacteriol 178(3):723–727

Lehner B (2011) Molecular mechanisms of epistasis within and between genes. Trends Genet 27(8):323–331

Maki K, Cheng H, Dolgikh DA, Roder H (2007) Folding kinetics of staphylococcal nuclease studied by tryptophan engineering and rapid mixing methods. J Mol Biol 368(1):244–255

Miller SL (1953) A production of amino acids under possible primitive earth conditions. Science 117(3046):528–529

Moroz OV, Moroz YS, Wu Y, Olsen AB, Cheng H, Mack KL, McLaughlin JM, Raymond EA, Zhezherya K, Roder H, Korendovych IV (2013) A single mutation in a regulatory protein produces evolvable allosterically regulated catalyst of nonnatural reaction. Angew Chem Int Ed Engl 52(24):6246–6249

Muller MM, Allison JR, Hongdilokkul N, Gaillon L, Kast P, van Gunsteren WF, Marliere P, Hilvert D (2013) Directed evolution of a model primordial enzyme provides insights into the development of the genetic code. PLoS Genet 9(1):e1003187

Osawa S, Jukes TH, Watanabe K, Muto A (1992) Recent evidence for evolution of the genetic code. Microbiol Rev 56(1):229–264

Pedelacq JD, Cabantous S, Tran T, Terwilliger TC, Waldo GS (2006) Engineering and characterization of a superfolder green fluorescent protein. Nat Biotechnol 24(1):79–88

Pezo V, Louis D, Guerineau V, Le Caer JP, Gaillon L, Mutzel R, Marliere P (2013) A metabolic prototype for eliminating tryptophan from the genetic code. Sci Rep 3:1359

Ribas de Pouplana L, Frugier M, Quinn CL, Schimmel P (1996) Evidence that two present-day components needed for the genetic code appeared after nucleated cells separated from eubacteria. Proc Natl Acad Sci U S A 93(1):166–170

Seki E, Matsuda N, Yokoyama S, Kigawa T (2008) Cell-free protein synthesis system from Escherichia coli cells cultured at decreased temperatures improves productivity by decreasing DNA template degradation. Anal Biochem 377(2):156–161

Steiner T, Hess P, Bae JH, Wiltschi B, Moroder L, Budisa N (2008) Synthetic biology of proteins: tuning GFPs folding and stability with fluoroproline. PLoS ONE 3(2):e1680

Suzuki T, Arai S, Takeuchi M, Sakurai C, Ebana H, Higashi T, Hashimoto H, Hatsuzawa K, Wada I (2012) Development of cysteine-free fluorescent proteins for the oxidative environment. PLoS ONE 7(5):e37551

Tanaka J, Yanagawa H, Doi N (2011) Comparison of the frequency of functional SH3 domains with different limited sets of amino acids using mRNA display. PLoS ONE 6(3):e18034

Tracewell CA, Arnold FH (2009) Directed enzyme evolution: climbing fitness peaks one amino acid at a time. Curr Opin Chem Biol 13(1):3–9

Trifonov EN (2004) The triplet code from first principles. J Biomol Struct Dyn 22(1):1–11

Waldo GS, Standish BM, Berendzen J, Terwilliger TC (1999) Rapid protein-folding assay using green fluorescent protein. Nat Biotechnol 17(7):691–695

Wang Q, Buckle AM, Foster NW, Johnson CM, Fersht AR (1999) Design of highly stable functional GroEL minichaperones. Protein Sci 8(10):2186–2193

Wang JD, Herman C, Tipton KA, Gross CA, Weissman JS (2002) Directed evolution of substrate-optimized GroEL/S chaperonins. Cell 111(7):1027–1039

Weinreich DM, Watson RA, Chao L (2005) Perspective: sign epistasis and genetic constraint on evolutionary trajectories. Evolution 59(6):1165–1174

Woese CR, Olsen GJ, Ibba M, Söll D (2000) Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. Microbiol Mol Biol Rev 64(1):202–236

Wong JT (1975) A co-evolution theory of the genetic code. Proc Natl Acad Sci U S A 72(5):1909–1912

Wong JT (1976) The evolution of a universal genetic code. Proc Natl Acad Sci U S A 73(7):2336–2340

Yamamoto Y, Tsutsumi Y, Yoshioka Y, Nishibata T, Kobayashi K, Okamoto T, Mukai Y, Shimizu T, Nakagawa S, Nagata S, Mayumi T (2003) Site-specific PEGylation of a lysine-deficient TNF-alpha with full bioactivity. Nat Biotechnol 21(5):546–552

Zhou HX, Hoess RH, DeGrado WF (1996) In vitro evolution of thermodynamically stable turns. Nat Struct Biol 3(5):446–451