Comparative characterization of random-sequence proteins consisting of 5, 12, and 20 kinds of amino acids

Junko Tanaka, Nobuhide Doi,* Hideaki Takashima, and Hiroshi Yanagawa*

Department of Biosciences and Informatics, Keio University, Yokohama 223-8522, Japan

Received 29 October 2009; Revised 21 January 2010; Accepted 24 January 2010

DOI: 10.1002/pro.358

Published online 16 February 2010 proteinscience.org

Abstract: Screening of functional proteins from a random-sequence library has been used to evolve novel proteins in the field of evolutionary protein engineering. However, random-sequence proteins consisting of the 20 natural amino acids tend to aggregate, and the occurrence rate of functional proteins in a random-sequence library is low. From the viewpoint of the origin of life, it has been proposed that primordial proteins consisted of a limited set of amino acids that could have been abundantly formed early during chemical evolution. We have previously found that members of a random-sequence protein library constructed with five primitive amino acids show high solubility (Doi et al., *Protein Eng Des Sel* 2005;18:279–284). Although such a library is expected to be appropriate for finding functional proteins, the functionality may be limited, because they have no positively charged amino acid. Here, we constructed three libraries of 120-amino acid, random-sequence proteins using alphabets of 5, 12, and 20 amino acids by preselection using mRNA display (to eliminate sequences containing stop codons and frameshifts) and characterized and compared the structural properties of random-sequence proteins arbitrarily chosen from these libraries. We found that random-sequence proteins constructed with the 12-member alphabet (including five primitive amino acids and positively charged amino acids) have higher solubility than those constructed with the 20-member alphabet, though other biophysical properties are very similar in the two libraries. Thus, a library of moderate complexity constructed from 12 amino acids may be a more appropriate resource for functional screening than one constructed from 20 amino acids.

Keywords: random DNA library; primitive amino acids; reduced alphabet; protein solubility; mRNA display; *in vitro* selection

Introduction

How were present-day proteins selected from the huge sequence space during protein evolution? The size of the protein sequence space is typically estimated to be 20^{100} (1.3×10^{130}) for a 100-residue polypeptide composed of the 20 kinds of natural amino acids. This number is enormous, compared with the number of

Abbreviations: aa, amino acid; bis-ANS, 4,4'-dianilino-1,1'-binaphthyl-5,5'-disulfonic acid; BSA, bovine serum albumin; CD, circular dichroism; *E. coli*, *Escherichia coli*; LB, Luria-Bertani; mRNA, messenger RNA; Ni-NTA, nickel-nitrilotriacetic acid; PCR, polymerase chain reaction; PEG, polyethylene glycol; RNase, ribonuclease; RT, reverse transcription; SDS-PAGE, sodium dodecyl sulfate-polyacrylamide gel electrophoresis; TBS, tris-buffered saline; TFE, trifluoroethanol; TIM, triosephosphate isomerase.

Grant sponsor: Japan Society for the Promotion of Science (JSPS; Grant-in-Aid for Scientific Research); Grant number: 19657073; Grant sponsor: Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT; Grant-in-Aid for the Keio University Global Center of Excellence (G-COE) Program Entitled "Center of Human Metabolomic Systems Biology").

*Correspondence to: Hiroshi Yanagawa, Department of Biosciences and Informatics, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama 223-8522, Japan. E-mail: hyana@bio.keio.ac.jp or Nobuhide Doi, Department of Biosciences and Informatics, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama 223-8522, Japan. E-mail: doi@bio.keio.ac.jp

proteins that may have existed in nature throughout the history of life on the Earth, which has been estimated to be less than 10^{50} molecules¹ or 10^{21} – 10^{43} molecules.² Thus, the nature has tested only a tiny fraction of the possible protein sequence space since the origin of life. However, the size of the protein sequence space at the early stage of protein evolution may have been overestimated. It is commonly believed that the 20 amino acids that comprise the alphabet of present-day proteins did not appear simultaneously on the evolutionary scene.³ It has been proposed that primordial proteins were constructed from a limited set of amino acids, which could have been formed in abundance at an early stage of chemical evolution, and that primitive cells have gradually introduced new amino acids into the repertoire for protein synthesis.⁴

Several researchers have demonstrated that the amino acid usage of various natural globular proteins and enzymes can be restricted to 5–13 members without substantial alteration of their structures and biological functions.^{5–8} Riddle *et al.* simplified the sequence of a small β -sheet protein, the SH3 domain, by using phage display selection, and produced two SH3 variants in which 90% of the sequence, excluding the binding region, utilized only five amino acids (Ala, Gly, Glu, Ile, and Lys).⁵ Silverman *et al.* generated variants of the prototypical $(\beta/\alpha)_8$ barrel enzyme, triosephosphate isomerase (TIM), in which the amino acids at 142 of 182 structural positions were simplified to seven kinds (Ala, Glu, Val, Lys, Phe, Leu, and Gln) by means of *in vivo* selection for TIM activity.⁶ Akanuma *et al.* fabricated 88% of the 213-residue orotate phosphoribosyltransferase with a reduced set of nine amino acids (Ala, Gly, Asp, Val, Arg, Thr, Leu, Pro, and Tyr) by means of growth-related phenotype selection, though the total number of amino acid types was 13.⁷ Walter *et al.* focused on an α -helical protein, chorismate mutase, and created an active enzyme constructed entirely from a set of nine amino acids (Asp, Glu, Arg, Met, Ile, Asn, Lys, Phe, and Leu) by *in vivo* selection.⁸

Other researchers have also produced folded *de novo* proteins from pattern sequence libraries constructed from limited numbers of amino acids. Hecht's group designed four helix bundle proteins by arranging five kinds of nonpolar amino acids (Val, Met, Ile, Phe, and Leu) and six kinds of polar amino acids (Asp, Glu, Asn, Lys, Gln, and His) in an appropriate order.^{9–11} Jumawid *et al.* have constructed $\alpha 3 \beta 3$ *de novo* proteins through binary combination of simplified hydrophobic (Val, Ile, and Leu) and hydrophilic (Ala, Glu, Lys, and Thr) amino acid sets.¹² These findings all support the view that the full amino acid alphabet set is not essential for protein folding.

As an alternative approach, several researchers have shown that random-sequence proteins con-

structed from limited alphabets have different characters from random-sequence proteins with the 20-amino acid alphabet. Although random-sequence proteins with the 20-amino acid alphabet showed no remarkable secondary structure,^{13–16} random-sequence proteins formed with only three kinds of amino acids (Gln, Leu, and Arg) had helical structure but show a strong tendency to aggregate.^{17,18} We have reported that random-sequence proteins consisting of five primitive amino acids (Ala, Gly, Val, Asp, and Glu), encoded by GNN codons (N = T, C, A, or G), show high solubility.¹⁹ Thus, random-sequence libraries with limited alphabets, especially the GNN codon-based library, may be appropriate for functional protein searching. However, the previously constructed GNN codon-based library included many sequences with frameshift and stop codons, and the lengths of the sequences varied. Further, the functionality of GNN codon-based proteins may be limited, because the alphabet contains no positively charged amino acid.

In this study, we constructed random-sequence gene libraries with limited amino acid alphabets by using mRNA display to eliminate incomplete sequences with frameshift and stop codons from the libraries.^{20,21} In the mRNA display technique, each mRNA in a library is covalently bound to its corresponding protein through puromycin,^{22,23} but mRNA sequences containing a stop codon or frameshift cannot form mRNA–protein conjugates or cannot be properly translated to generate the C-terminal tag, respectively. Thus, they can be washed away in affinity selection based on a property of the peptide portion. We constructed a new random-sequence gene library containing RNN (R = A or G) codons encoding a 12-amino acid alphabet, that is, ANN codons for three primitive amino acids (Ser, Thr, and Arg) and four advanced amino acids (Met, Ile, Asn, and Lys) and GNN codons for five primitive amino acids (Ala, Gly, Val, Asp, and Glu). Therefore, the RNN codon-based library contains not only two kinds of negatively charged amino acids, Asp and Glu, but also two kinds of positively charged amino acids, Arg and Lys. We also constructed a GNN library and an NNN library so that we could compare the physical properties of random-sequence proteins with alphabets consisting of 5, 12, and 20 amino acids.

Results and Discussion

Construction of random-sequence protein libraries with limited alphabets

We constructed three kinds of random-sequence DNA libraries with more than 100 contiguous random codons GNN (for Ala, Gly, Asp, Glu, and Val), RNN (GNN plus Ile, Met, Thr, Asn, Lys, Ser, and

Figure 1. Schematic representation of preselection for removing sequences containing stop codons and frameshifts from a library by mRNA display. (1) Random DNA library is transcribed and ligated with a PEG spacer bearing puromycin. (2) The RNA library is translated *in vitro*. An RNA template without a stop codon and frameshift displays a full-length protein with the C-terminal FLAG tag (red). An RNA template with a stop codon displays no protein. An RNA template with a frameshift displays the corresponding frame-shifted polypeptide (blue) lacking the FLAG tag. (3) The resulting mRNA–protein conjugates are purified with anti-FLAG antibody-immobilized beads. (4) The RNA portion of the bound molecules is amplified by RT-PCR to form a random DNA library without stop codons and frameshifts.

Arg), or NNN (all 20 kinds of amino acids). We designed DNA cassettes consisting of 15 consecutive random codons between constant sequences to be assembled for construction of DNA libraries with 120 random codons, as previously described.¹⁹

When the synthesized DNA cassettes of NNN, RNN, and GNN were sequenced, not only the NNN cassette but also the GNN and RNN cassettes, included stop codons (NNN, 66%; GNN, 20%; RNN, 12%), probably due to chemical synthetic errors. Thus, we planned to eliminate sequences containing stop codons and frameshifts from the synthesized random-sequence DNA library by “preselection” using mRNA display, as previously described.²¹ When a library of mRNA-displayed proteins with a C-terminal FLAG tag is purified by using anti-FLAG antibody-immobilized beads, sequences containing stop codons or frameshifts can be eliminated from the library, because mRNA sequences containing a stop codon cannot associate with any protein on the ribosome through puromycin, and sequences containing frameshifts cannot be translated to polypeptides with the C-terminal FLAG tag (Fig. 1). After preselection and RT-PCR, the fraction of DNA sequence containing stop codons and frameshifts was effectively decreased, and no such sequences were observed in 30 samples sequenced from each library.

The preselected DNA cassettes were then assembled through digestion and ligation, as previously described.¹⁹ The random region of each library

was flanked by fixed sequences encoding affinity tags, the N-terminal T7-tag, and the C-terminal FLAG tag. The lengths of sequences arbitrarily chosen from the constructed random-sequence libraries were almost equal (Fig. 2), whereas the lengths varied in the previous study without preselection.¹⁹

Expression of arbitrarily chosen random-sequence proteins with limited amino acid alphabets

As solubility in aqueous solution is one of the most important properties of globular proteins, we examined the solubility of random-sequence proteins arbitrarily chosen from each library. The random-sequence proteins were overexpressed in *E. coli*, and proteins in the soluble fraction and the insoluble fraction were detected by Western blotting with anti-T7-tag antibody. As shown in Figure 3 and Table I, all of the GNN proteins (G1–7 except G3, which was not expressed) from the GNN library were present in the soluble fraction [Fig. 3(A)], in agreement with the previous study,¹⁹ whereas all of the NNN proteins (N1–10 except N1, N3, N6, and N9, which were not expressed) from the NNN library were present in the insoluble fraction [Fig. 3(C)]. The RNN proteins (R1–16 except R6 and R11, which were not expressed) from the RNN library were intermediate in character; that is, one RNN protein (R16) was expressed only in the soluble fraction, 11 RNN proteins were expressed only in the insoluble fraction, and two (R4 and R7) were expressed in both fractions [Fig. 3(B)].

Some sequences, such as G3, R6, R11, N1, N3, N6, and N9, were not expressed in *E. coli*. Interestingly, the content of unexpressed proteins in the NNN library (40%) was higher than that in the RNN library (13%) or the GNN library (14%). It is well known that codon bias, tRNA abundance, and gene expression are correlated.²⁴ As the NNN includes many low-usage codons when compared with the RNN, the amount of protein production might be very less because of the presence of low-usage codon clusters. Recently, Kudla *et al.* found that strong secondary structure at the 5' end of an mRNA (up to around 30-nt downstream from the start codon) blocks ribosome binding and obstructs translation initiation.²⁵ In this study, however, a common T7-tag and linker sequence (51-nt) were introduced after the start codon for all three libraries (Fig. 2), and thus, this is not the reason for the bias. Another possibility is that translation products may be degraded in a sequence-specific manner in *E. coli*.

We examined the relationship between the solubility of random-sequence proteins and several properties of the amino acid sequences. Wilkinson and Harrison suggested that protein solubility is strongly affected by net charge and the fraction of turn-forming residues (Asp, Asn, Pro, Gly, and Ser)

Figure 3. Expression of random-sequence proteins from the GNN library (A), RNN library (B), and NNN library (C). The soluble (lanes S) and insoluble (lanes I) fractions of overexpressed proteins were analyzed by 16.5% Tricine SDS-PAGE. The proteins were detected by Coomassie brilliant blue staining (top) or Western blotting with anti-T7-tag antibody (bottom). The arrowheads indicate the positions of recombinant proteins.

proteins using a 20-amino acid alphabet¹⁴ have been reported to take helical structure only in the presence of TFE. Thus, random-sequence proteins using 12- and 20-amino acid alphabets showed similar pro-

pensity for secondary-structure formation. On the other hand, QLR proteins¹⁸ have helical structure in the absence of TFE, but most QLR proteins tend to aggregate, and addition of GuHCl is necessary for

Table I. Numbers of Random-Sequence Proteins in Various Solubility Ranges

Library	Solubility (%)			0
	100	≥50	<50	
GNN	2	4	0	0
RNN	1	1	1	11
NNN	0	0	0	6

solubilization. Thus, there may be a trade-off between secondary-structure formation and high solubility in the case of random-sequence proteins.

A bis-ANS binding experiment indicated the presence of hydrophobic clusters in the RNN proteins (Fig. 6). Bis-ANS is a fluorescent probe that detects molten globule states of a protein through binding to an accessible hydrophobic core, resulting in enhanced fluorescence emission.²⁹ As the fluorescence change in the presence of R16 was larger than that in the presence of R7 (Fig. 6), R16 may have more extensive hydrophobic clusters than R7.

Table II. Parameters of Random-Sequence Proteins

Protein	Solubility ^a	Kyte-Doolittle ^b	Net Charge ^c	Turn ^d
G1	100	+0.38	-29	47
G2	100	+0.27	-29	42
G3	—	+0.03	-39	49
G4	≥50	+0.41	-30	50
G5	≥50	-0.20	-39	56
G6	≥50	+0.31	-29	43
G7	≥50	+0.20	-31	50
R1	0	-0.19	+4	42
R2	0	-0.44	-8	52
R3	0	-0.36	-13	41
R4	<50	-0.05	-4	49
R5	0	-0.03	+5	42
R6	—	-0.27	-7	48
R7	≥50	-0.70	-13	45
R8	0	-0.03	-3	39
R9	0	-0.56	+4	44
R10	0	-0.06	-11	37
R11	—	-0.27	-7	48
R12	0	+0.17	-6	40
R13	0	+0.05	-2	43
R14	0	-0.44	-7	47
R15	0	-0.29	-5	40
R16	100	-1.05	-12	52
N1	—	-0.49	+4	43
N2	0	-0.74	+16	44
N3	—	-0.49	+7	41
N4	0	-1.04	+8	46
N5	0	-0.17	+7	45
N6	—	-0.64	+9	39
N7	0	-0.82	+13	52
N8	0	-0.22	+5	59
N9	—	-0.80	+2	42
N10	0	-0.74	+1	63

^a —, Not expressed.

^b Calculated based on the index.²⁷

^c The number of Arg and Lys minus the number of Asp and Glu.

^d The number of Asp, Asn, Pro, Gly, and Ser.

Figure 4. Purification of two soluble proteins from the RNN library. The random-sequence proteins, R7 and R16 with His₆ tag, were overexpressed in *E. coli*, and the soluble fractions of the crude lysate were purified on Ni-NTA resins. The samples before (N, nonpurified) and after purification (P, purified) were resolved by 15% SDS-PAGE and stained with Coomassie brilliant blue. M, marker.

Size exclusion chromatography showed that two RNN proteins form monomeric structures with more compact shapes than the random-coil structures of denatured proteins with similar molecular weight, but more extended shapes than the globular structures of natural proteins (Fig. 7). The Stokes radius of R16 is smaller than that of R7, whereas the molecular weight of R16 is larger than that of R7. This result indicates that R16 forms a more compact shape than R7, and this is consistent with the results of the bis-ANS binding experiment. GNN proteins¹⁹ formed monomeric structures with extended shapes, as did RNN proteins, but random-sequence proteins using a 20-amino acid alphabet¹⁴ have been reported to form oligomeric structures, owing to their tendency to aggregate.

Thus, the two RNN proteins have partial secondary structure and a hydrophobic core, and their estimated sizes indicate that they are more compact than would be expected for expanded random-coil structures. These properties are similar to those reported for other soluble random-sequence proteins with a 5 or 20 alphabet.^{14,19} Although the proteins do not exhibit extensive, well-folded structure, a large number of intrinsically unstructured domains, which become structured only during binding to the target (i.e., induced fit), have already been identified in nature³¹ and designed by protein engineers.^{8,30,32} From the viewpoint of molecular evolution, such partially structured polypeptides might have been the first evolutionary intermediates, and their function and structure would have coevolved.³³

Figure 5. Circular dichroism spectra of random-sequence proteins R7 (A) and R16 (B). CD spectra of purified proteins were measured at different concentrations of TFE (0, 20, and 40%).

Toward the creation of novel functional proteins with reduced amino acid alphabets

Although the frequency of occurrence of functional proteins in a randomized library has been estimated to be about 1 in 1×10^{11} ,^{15,34} the frequency of functional proteins in the sequence space will depend on the choice of structural motif and the level of catalytic activity required for selection. For instance, Taylor *et al.* calculated the required library size for selection of chorismate mutase activity as 5×10^{23} ; this size far exceeds what is currently accessible in laboratory experiments.³⁵ Therefore, the possibility that the choice of an appropriate restricted amino acid alphabet could be a powerful tool in designing randomized libraries for screening is of great interest. Reetz *et al.* compared the quality of randomized libraries, in which five different sites of epoxide hydrolase were replaced by codon NNK (20 kinds of amino acids) or codon NDT (12 kinds of amino acids; Val, Asp, Gly, Ile, Asn, Ser, Arg, Leu, His, Phe, Tyr, and Cys), and they found that the NDT library showed a much higher frequency of positive variants and yielded a greater improvement of catalytic activity in comparison with the NNK library.³⁶ The amino acid composition of the NDT library is different from that of the RNN library, which has no aromatic residues; however, it is believed that aromatic amino acids did not exist in the early stage of evolution, because they would have been quickly destroyed by ultraviolet irradiation.³⁷ As the RNN library was completely randomized, whereas the NDT library was randomized only for three residues from the active site, the results obtained with the two libraries cannot be directly compared. Moreover, Reetz *et al.* and we did not investigate other libraries with a 12 alphabet. Thus, the best set of amino

acids for functional screening still remains to be identified, and further investigation is required.

In this study, we have shown that random-sequence proteins exhibit unique biophysical characters that depend on the amino acid alphabet employed. Random-sequence proteins from the RNN library tended to have higher solubility than those from the NNN library. Because other biophysical properties seemed to be almost the same among the RNN proteins and the NNN proteins, it will be of interest to test whether the frequency of occurrence of functional proteins in a random-sequence library based on the codon RNN, which mainly encodes primitive amino acids, is higher than that in a

Figure 6. Fluorescence spectra of 10 μ M bis-ANS (excitation at 393 nm) in the absence or presence of 2 μ M R7 protein or 2 μ M R16 protein.

Figure 7. Stokes radius plot for R7 and R16 proteins. The vertical arrows indicate the elution volumes determined by gel filtration of purified R7 and R16 proteins. The circles indicate the elution volumes of four control proteins (BSA, 67.0 kDa; ovalbumin, 43.0 kDa; chymotrypsinogen A, 25.0 kDa; and RNase, 13.7 kDa) plotted against their known Stokes radii, and the line represents an empirical equation relating Stokes radius to elution position. The Stokes radius of unfolded protein and folded globular protein (horizontal arrows) was calculated from the literature equation.³⁰

library based on a 20-amino acid alphabet. Experiments along this line are in progress in our laboratory by using mRNA display for *in vitro* selection of functional proteins from randomized RNN and NNN libraries. It should be noted that the higher solubility of the RNN proteins in *E. coli* may be particularly favorable for screening methods based on *E. coli* expression systems, including phage display, when compared with those based on *in vitro* translation systems.

Materials and Methods

Construction of random-sequence DNA library

DNAs consisting of 15 consecutive random codons between constant sequences [5'-GGTAGATCTGGACCTGCAGGATGGGNN (GNN, RNN, or NNN)₁₄ TGGGCGAGACCGCTCGAGGTTC-3'] were synthesized by Fasmac, Japan. The random-sequence cassettes were amplified by PCR using primers 5'-ATGGCTAGCATGACTGGTGGACAGCAAATGGGTAGATCTGGACCTGCAGGATG-3' and 5'-TTTTTTTTTCTTGTCGTCATCGTCCTTGTAGTCAAGAACCTCGAGCGGTCTCGC-3' with KOD-plus DNA polymerase (Toyobo). The PCR products were purified with a QIAquick PCR purification kit (Qiagen). The purified DNA was reamplified by PCR using primers 5'-ATTTAGGTGACACTATAGAACAACAACAACAACAACAACAACAATAATGGCTAGCATGACTGGTGGAC-3'

and 5'-TTTTTTTTTCTTGTCGTCATCGTCCTTGTAG-3'. The DNA was purified with a QIAquick PCR purification kit.

mRNA display selection was performed as previously described.²⁰ Briefly, the purified DNA was transcribed with a RiboMax large-scale RNA production system-SP6 (Promega). The resulting RNA was purified with an RNeasy mini kit (Qiagen) and ligated with polyethylene glycol (PEG)-Puro spacer [p(dCp)₂-T(Fluor)p-PEGp-(dCp)₂-puromycin] using T4 RNA ligase (Takara). The ligated RNA was purified with the RNeasy mini kit. A 40-μL aliquot of a wheat germ extract reaction mixture (Promega) containing 10 pmol of the ligated random-sequence RNA cassette, 80 μM amino acid mixture, 76 mM potassium acetate, and 40 U of RNase inhibitor (Invitrogen) was incubated for 1 h at 26°C. The reaction mixture containing mRNA-displayed proteins (mRNA-protein conjugates) was added to 60 μL of anti-FLAG M2 antibody-immobilized agarose beads (Sigma) preequilibrated with 40 μL of TBST and mixed on a rotator at 4°C for 1 h. The beads were washed with 300 μL of TBST three times. The mRNA-displayed proteins were eluted with TBST containing 1 mg/mL FLAG M2 peptide (Sigma) at 4°C for 1 h. The mRNA portion of eluted mRNA-displayed proteins was amplified by RT-PCR with a One-Step RT-PCR kit (Qiagen) using primers 5'-GGTAGATCTGGACCTGCAGGATG-3' and 5'-GAACCTCGAGCGGTCTCGC-3'. The RT-PCR products were purified with a QIAquick PCR purification kit.

The purified products were separated into two equal aliquots that were restricted with either BsaI or BfuAI. The resulting fragments were purified by 8% PAGE, ligated together using T4 DNA ligase (Promega), and amplified by PCR using primers 5'-GGTAGATCTGGACCTGCAGGATG-3' and 5'-GAACCTCGAGCGGTCTCGC-3'. Repeating this procedure three times yielded a final library with eight contiguous random regions.

Cloning, expression, and purification of the random-sequence proteins

Randomly selected clones from the DNA libraries were sequenced with an ABI PRISM 3100 (Applied Biosystems). The random-sequence regions of the clones were digested with BglII and XhoI and subcloned into the pET20 vector (Novagen) containing the N-terminal T7-tag sequence and the C-terminal His₆ tag sequence. The individual plasmids were transformed into *Escherichia coli* BL21(DE3)-Codon-Plus cells (Stratagene). The bacteria were grown in LB broth containing 100 μg/mL ampicillin and 40 μg/mL chloramphenicol at 37°C, and protein expression was induced by adding 0.1 mM isopropylthio-β-D-galactoside. After an additional 3 h of growth, the bacteria were harvested by centrifugation and lysed in a BugBuster (Novagen) containing a protease inhibitor cocktail (Sigma). The centrifuged

supernatants were used as soluble fractions. The pellets were resuspended in a buffer containing 8M urea, and the supernatants after centrifugation were used as insoluble fractions. The proteins in these fractions were analyzed by 16.5% Tricine SDS-PAGE and detected by Coomassie brilliant blue staining or Western blotting with anti-T7-tag antibody. The proteins in soluble fractions were purified by affinity chromatography using Ni-NTA Superflow resin (Qia-gen), from which they were eluted with a pH gradient. The protein molar concentration was determined by a BCA protein assay kit (Pierce).

CD spectroscopy

CD spectra of purified proteins were measured using a J-820 spectropolarimeter (Jasco) at 4°C in the presence of different concentrations of 2,2,2-trifluoroethanol (TFE, Wako). The light pathlength used was 2 mm. The results were expressed as mean residue molar ellipticity [θ].

Fluorescence spectroscopy

The fluorescence spectra of 10 μM 4,4'-dianilino-1,1'-binaphthyl-5,5'-disulfonic acid (bis-ANS, Molecular Probes) in the absence and presence of 1 μM protein were measured at 4°C on a FP-777 spectrofluorometer (Jasco) with excitation at 393 nm.

Size exclusion chromatography

Gel filtration experiments on purified proteins were performed using Superdex-75 (Amersham). Molecular weights were determined by linear regression analysis using a Gel Filtration LMW Calibration Kit (Amersham). The Stokes radii of the random-sequence proteins and the control proteins (bovine serum albumin, ovalbumin, chymotrypsinogen, and ribonuclease) were calculated from their elution volumes as previously described.³⁸

References

- Mandecki W (1998) The game of chess and searches in protein sequence space. *Trends Biotechnol* 16:200–202.
- Dryden DTF, Thomson AR, White JH (2008) How much of protein sequence space has been explored by life on earth? *J R Soc Interface* 5:953–956.
- Wong JT (2005) Coevolution theory of the genetic code at age thirty. *BioEssays* 27:416–425.
- Trifonov EN (2004) The triplet code from first principles. *J Biomol Struct Dyn* 22:1–11.
- Riddle DS, Santiago JV, Bray-Hall ST, Doshi N, Grantcharova VP, Yi Q, Baker D (1997) Functional rapidly folding proteins from simplified amino acid sequences. *Nat Struct Biol* 4:805–809.
- Silverman JA, Balakrishnan R, Harbury PB (2001) Reverse engineering the (αβ)₈ barrel fold. *Proc Natl Acad Sci USA* 98:3092–3097.
- Akanuma S, Kigawa T, Yokoyama S (2002) Combinatorial mutagenesis to restrict amino acid usage in an enzyme to a reduced set. *Proc Natl Acad Sci USA* 99:13549–13553.
- Walter KU, Vamvaca K, Hilvert D (2005) An active enzyme constructed from a 9-amino acid alphabet. *J Biol Chem* 280:37742–37746.
- Kamtekar S, Schiffer JM, Xiong H, Babik JM, Hecht MH (1993) Protein design by binary patterning of polar and nonpolar amino acids. *Science* 262:1680–1685.
- Go A, Kim S, Baum J, Hecht MH (2008) Structure and dynamics of de novo proteins from a designed superfamily of 4-helix bundles. *Protein Sci* 17:821–832.
- Patel SC, Bradley LH, Jinadasa SP, Hecht MH (2009) Cofactor binding and enzymatic activity in an unevolved superfamily of de novo designed 4-helix bundle proteins. *Protein Sci* 18:1388–1400.
- Jumawid MT, Takahashi T, Yamazaki T, Ashigai H, Mihara H (2009) Selection and structural analysis of de novo proteins from an α3β3 genetic library. *Protein Sci* 18:384–398.
- Mandecki W (1990) A method for construction of long randomized open reading frames and polypeptides. *Protein Eng* 3:221–226.
- Yamauchi A, Yomo T, Tanaka F, Prijambada ID, Ohhashi S, Yamamoto K, Shima Y, Ogasahara K, Yutani K, Kataoka M, Urabe I (1998) Characterization of soluble artificial proteins with random sequences. *FEBS Lett* 421:147–151.
- Keefe AD, Szostak JW (2001) Functional proteins from a random-sequence library. *Nature* 410:715–718.
- Watters AL, Baker D (2004) Searching for folded proteins in vitro and in silico. *Eur J Biochem* 271:1615–1622.
- Davidson AR, Sauer RT (1994) Folded proteins occur frequently in libraries of random amino acid sequences. *Proc Natl Acad Sci USA* 91:2146–2150.
- Davidson AR, Lumb KJ, Sauer RT (1995) Cooperatively folded proteins in random sequence libraries. *Nat Struct Biol* 2:856–864.
- Doi N, Kakukawa K, Oishi Y, Yanagawa H (2005) High solubility of random-sequence proteins consisting of five kinds of primitive amino acids. *Protein Eng Des Sel* 18:279–284.
- Miyamoto-Sato E, Takashima H, Fuse S, Ishizaka M, Tateyama S, Horisawa K, Sawasaki T, Endo Y, Yanagawa H (2003) Highly stable and efficient mRNA templates for mRNA-protein fusions and C-terminally labeled proteins. *Nucleic Acids Res* 31:e78.
- Cho G, Keefe AD, Liu R, Wilson DS, Szostak JW (2000) Constructing high complexity synthetic libraries of long ORFs using in vitro selection. *J Mol Biol* 297:309–319.
- Nemoto N, Miyamoto-Sato E, Husimi Y, Yanagawa H (1997) In vitro virus: bonding of mRNA bearing puromycin at the 3'-terminal end to the C-terminal end of its encoded protein on the ribosome in vitro. *FEBS Lett* 414:405–408.
- Roberts RW, Szostak JW (1997) RNA-peptide fusions for the in vitro selection of peptides and proteins. *Proc Natl Acad Sci USA* 94:12297–12302.
- Ikemura T (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J Mol Biol* 151:389–409.
- Kudla G, Murray AW, Tollervey D, Plotkin JB (2009) Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* 324:255–258.
- Wilkinson DL, Harrison RG (1991) Predicting the solubility of recombinant proteins in *Escherichia coli*. *Biotechnology* 9:443–448.
- Kyte J, Doolittle RF (1982) A simple method for displaying the hydrophobic character of a protein. *J Mol Biol* 157:105–132.

28. Shiraki K, Nishikawa K, Goto Y (1995) Trifluoroethanol-induced stabilization of the alpha-helical structure of beta-lactoglobulin: implication for non-hierarchical protein folding. *J Mol Biol* 245:180–194.
29. Stryer L (1965) The interaction of a naphthalene dye with apomyoglobin and apohemoglobin. A fluorescent probe of non-polar binding sites. *J Mol Biol* 13:482–495.
30. Vamvaca K, Vögeli B, Kast P, Pervushin K, Hilvert D (2004) An enzymatic molten globule: efficient coupling of folding and catalysis. *Proc Natl Acad Sci USA* 101:12860–12864.
31. Wright PE, Dyson HJ (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* 293:321–331.
32. Chaput JC, Szostak JW (2004) Evolutionary optimization of a nonbiological ATP binding protein for improved folding stability. *Chem Biol* 11:865–874.
33. Tokuriki N, Tawfik DS (2009) Protein dynamism and evolvability. *Science* 324:203–207.
34. Seelig B, Szostak JW (2007) Selection and evolution of enzymes from a partially randomized non-catalytic scaffold. *Nature* 448:828–831.
35. Taylor SV, Walter KU, Kast P, Hilvert D (2001) Searching sequence space for protein catalysts. *Proc Natl Acad Sci USA* 98:10596–10601.
36. Reetz MT, Kahakeaw D, Lohmer R (2008) Addressing the numbers problem in directed evolution. *ChemBiochem* 9:1797–1804.
37. Wong JT, Evolution and mutation of the amino acid code. In: Ricard J, Cornish-Bowden A, Eds. (1984) *Dynamics of biochemical systems*. New York: Plenum Press, pp 247–257.
38. Uversky VN (1993) Use of fast protein size-exclusion liquid chromatography to study the unfolding of proteins which denature through the molten globule. *Biochemistry* 32:13288–13298.