View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by IntechOpen

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

3

Evolutionary Engineering of Artificial Proteins with Limited Sets of Primitive Amino Acids

Junko Tanaka, Hiroshi Yanagawa and Nobuhide Doi
Department of Biosciences and Informatics, Keio University
Japan

1. Introduction

Because present-day proteins are composed of 20 kinds of amino acids, the number of possible amino-acid sequences in a 100-residue protein is 20^{100} (approximately 10^{130}), which is larger than the total number of atoms in the universe (~10^{80}). The number of proteins that may have existed in nature throughout the history of life on the Earth has been estimated to be less than 10^{50} molecules (Mandecki, 1998) or 10^{43} molecules (Dryden et al., 2008). Thus, the vast sequence space available remains to be explored further, and the sequence space that remains unexplored provides an opportunity to create valuable proteins with novel structures and functions for biomedical and environmental applications. Evolutionary protein engineering or directed protein evolution has been used to create artificial proteins with novel functions (Bloom et al., 2005; Hoogenboom, 2005; Leemhuis et al., 2005; Romero & Arnold, 2009) by repeated mutation, selection and amplification, mimicking Darwinian evolution in the laboratory (Figure 1).

Using evolutionary engineering, several researchers have recently demonstrated in the laboratory how the steps of protein evolution might occur in nature. For example, Peisajovich et al. (2006) explored the plausibility of the permutation-by-duplication model of the evolution of the DNA-methyltransferase superfamily and indicated that new protein topologies can evolve gradually through multistep gene rearrangements while maintaining the function of the parent domain. Tokuriki and Tawfik (2009a) investigated the random mutational drift of several enzymes in the presence of overexpressed chaperonin and revealed that protein stability is a major constraint in protein evolution and is a buffering mechanism by which chaperonin can alleviate this constraint. Huang et al. (2008) indicated that new protein functions can be generated by combining unrelated domains and subsequently optimizing the domain interface. However, these studies have mainly focused on the relatively recent evolutionary pathways of modern proteins; none of the hypotheses regarding the early evolution of primitive proteins has yet been tested.

In this chapter, we focused on the hypothesis that proteins consisted of fewer amino acid types during the early stage of protein evolution. Although modern proteins consist of 20 amino acid types, it has been proposed that primordial proteins consisted of a smaller set of "primitive" amino acids that could have been abundantly formed on the prebiotic Earth. Additional, "new" amino acids were then gradually recruited into the genetic code (Section 2). To test this hypothesis, we used the powerful tool "mRNA display" (Section 3) and

examined the rate at which folding ability (Section 4) and function (Section 5) occurred in artificial proteins consisting of limited sets of amino acids. An improved understanding of protein evolutionary pathways can provide more efficient tools for the creation of artificial proteins with novel functions and structures.

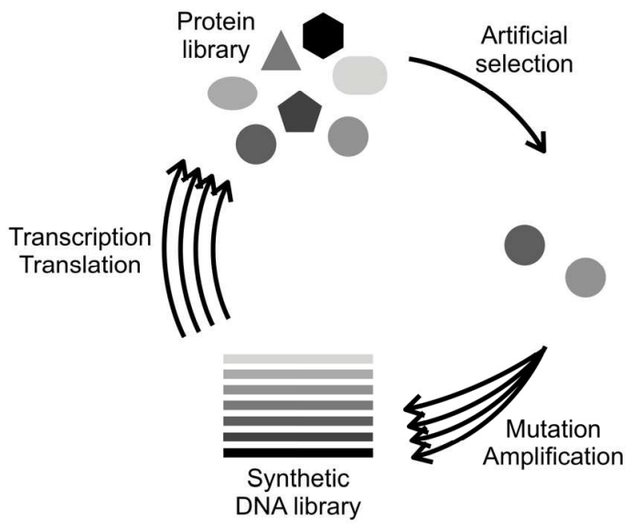


Fig. 1. The cycle used in the Darwinian evolution of proteins. A pool of genotype molecules (a synthetic DNA library) is converted to a pool of phenotype molecules (a protein library), from which proteins with a desired property are selected (artificial selection). The genotype molecules corresponding to the selected proteins are then amplified and mutated for further selection cycles.

2. A hypothesis regarding the origin and early evolution of proteins and the genetic code

According to Oparin's chemical evolutionary hypothesis regarding the Origin of Life (Oparin, 1961), a soup of nutrient organic compounds was available to the first organism on the primitive Earth. Once a self-replicating molecule formed from the primordial soup, this early replicator could have evolved to a primordial cell. Since the discovery of RNA enzymes (ribozymes) (Altman, 1981; Cech et al., 1981), RNA molecules have been postulated to be the first self-replicating molecule to play the following two roles: storing genetic information and catalyzing a chemical reaction. Later, RNA acquired various catalytic activities (in the RNA world), and protein synthesis could have been established. Finally, the location of genetic information moved from RNA to DNA because DNA is more stable than RNA. At present, proteins are synthesized based on the information contained in DNA and have particular structures and functions that provide various kinds of biological activities.

How were present-day proteins and the genetic code generated in the RNA world? It has been proposed that primordial proteins consisted of a small set of amino acids such as Ala, Gly, Asp and Val, which could have been abundantly formed early during chemical evolution (Miller, 1987). Interestingly, the codons for these amino acids all have guanosine (G) as the first nucleotide; for this reason, the codons GNC and GNN, where N denotes U, C, A or G, were proposed to have formed the early genetic code (Eigen & Schuster, 1978).

Thereafter, according to the coevolution theory (Wong, 1975, 1988, 2005), the genetic code coevolved with amino acid biosynthetic pathways, and additional amino acids were introduced after production through their synthetic pathways. Recently, Trifonov (2004) deduced a list of the consensus order in which amino acids were incorporated into the genetic code on the basis of 60 criteria (Figure 2A). This list revealed that the amino acids synthesized in Miller's spark discharge experiments (Gly, Ala, Asp, Val, Pro, Ser, Glu, Thr, Leu and Ile) appeared first, and that the amino acids associated with codon capture events (His, Cys, Phe, Tyr, Met and Trp) came last (Trifonov, 2004). Jordan et al. (2005) verified this list using comparative genome sequence analysis of orthologous proteins in the genomes of bacteria, archaea and eukaryotes. These authors clarified that the frequencies of Gly, Ala, Glu and Pro consistently decrease in proteins while the frequencies of Ser, His, Cys, Met and Phe increase during protein evolution (Figure 2B). They took into consideration the concept that the amino acids with decreasing frequencies are thought to have been the first amino acids incorporated into the genetic code; conversely, all amino acids with increasing frequencies, except Ser, are probably late recruitments (Jordan et al., 2005). The trend of transitioning amino acid composition (Figure 2B) corresponds well to Trifonov's list of the order of incorporation of amino acids into the genetic code (Figure 2A).

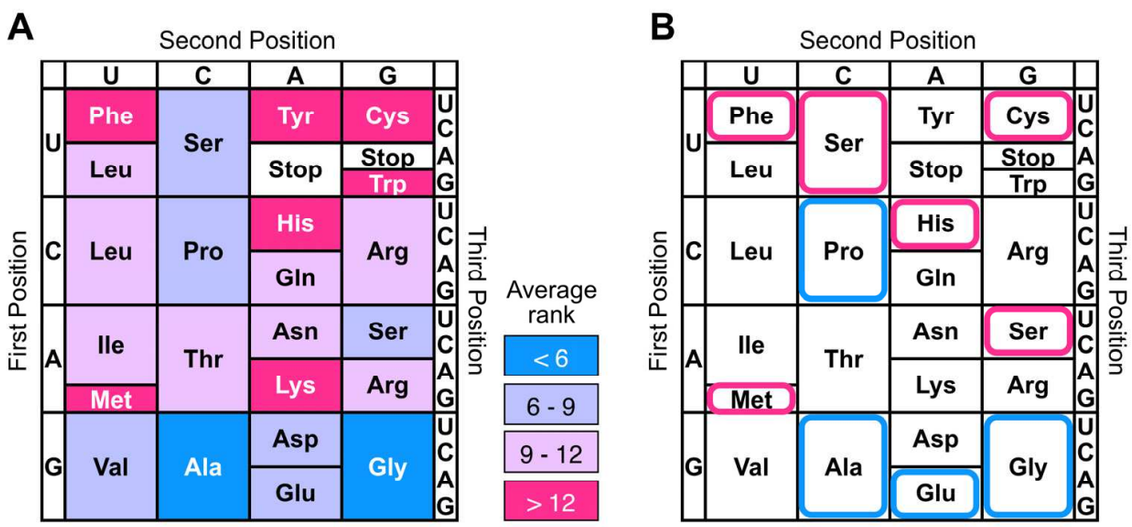


Fig. 2. The universal genetic code. (A) The average rank represents the chronological order of amino acid addition to the genetic code (Trifonov, 2004). The ranking values calculated based on 60 criteria were Gly, 3.5; Ala, 4.0; Asp, 6.0; Val 6.3; Pro, 7.3; Ser, 7.6; Glu, 8.1; Thr, 9.4; Leu, 9.9; Arg, 11.0; Asn, 11.3; Ile, 11.4; Gln, 11.4; His, 13.0; Lys, 13.3; Cys, 13.8; Phe, 14.2; Tyr, 15.2; Met, 15.4 and Trp, 16.5. (B) The trend of amino acid gain and loss during protein evolution (Jordan et al., 2005). The amino acids for which the frequencies consistently decrease (*i.e.*, primitive amino acids) are highlighted in blue, and the amino acids for which the frequencies consistently increase (*i.e.*, amino acids that were probably recruited late into the genetic code) are highlighted in red.

Can native protein structure and function be achieved with such reduced alphabets? Several researchers have demonstrated that the amino acid usage of various natural globular proteins and enzymes can be restricted to 5–9 members while retaining their structures and functions (Table 1). For example, Riddle et al. (1997) simplified SH3 domains in which 90%

of the sequence employed just five types of amino acids. Silverman et al. (2001) restricted 78% of the sequence of the prototypical $(\beta/\alpha)_8$ barrel enzyme, triosephosphate isomerase, to seven types of amino acids. Akanuma et al. (2002) generated variants of orotate phosphoribosyl transferase in which 88% of the sequence used just nine types of amino acids. Finally, Walter et al. (2005) created an active enzyme, chorismate mutase, which was constructed entirely from nine types of amino acids. Other researchers have attempted to produce *de novo* proteins from designed combinatorial libraries. Hecht's group created four helix bundle proteins based on binary patterning using five types of nonpolar amino acids (Val, Met, Ile, Phe and Leu) and six kinds of polar amino acids (Asp, Glu, Asn, Lys, Gln and His) (Go et al., 2008; Kamtekar et al., 1993; Patel et al., 2009). Jumawid et al. (2009) produced *de novo* proteins with an α3β3 structure using a simplified binary combination of hydrophobic amino acids (Val, Ile and Leu) and hydrophilic amino acids (Ala, Glu, Lys and Thr). These experiments support the hypothesis that the full amino-acid alphabet set is not essential for the structure and biological function of proteins. However, these experiments were not focused on whether the limited sets of amino acids used are primitive or not. Thus, the hypothesis that primordial proteins originally consisted of a small repertoire of primitive amino acids that gradually increased by coevolution with amino acid biosynthetic pathways has been insufficiently supported by experimental data thus far. In the following section, we summarize molecular display technologies that can be used to experimentally demonstrate the hypothesis regarding the existence of primordial proteins.

Protein	Amino Acids (variety of amino acids)	Reference
Four α-helix bundle *de novo* protein	Asp, Val, Glu, Leu, Asn, Ile, Gln, His, Lys, Phe, Met (11)	Kamtekar et al., 1993
SH3 domain	Gly, Ala, Glu, Ile, Lys (5)	Riddle et al., 1997
Triosephosphate isomerase	Ala, Val, Glu, Leu, Gln, Lys, Phe (7)	Silverman et al., 2001
Orotate phosphoribosyl transferase	Gly, Ala, Asp, Val, Pro, Thr, Leu, Arg, Tyr (9)	Akanuma et al., 2002
Chorismate mutase	Asp, Glu, Leu, Arg, Asn, Ile, Lys, Phe, Met (9)	Walter et al., 2005
α3β3 *de novo* protein	Ala, Val, Glu, Thr, Leu, Ile, Lys (7)	Jumawid et al., 2009

Table 1. Proteins constructed using reduced sets of amino acids with retention of their biological functions or structures.

3. mRNA display for *in vitro* selection of proteins

In the selection of targeted functional biomolecules by directed evolution, the most important consideration is the ability to link genotype and phenotype. "Phenotype" refers to biological functions, whereas "genotype" refers to the nucleic acids coding for replication. The nucleic acid portions of RNA aptamers and ribozymes have roles in both function and replication. Proteins, however, have only functional roles and cannot be replicated. Therefore, the development of a molecular display technique that physically links genotype with phenotype is essential for directed protein evolution. As shown in Figure 3, various

molecular display techniques have been developed (Doi & Yanagawa, 2001; Matsumura et al., 2006). In 1985, Smith discovered that exogenous peptides could be displayed on a filamentous phage by fusing peptides of interest to the coat protein of a filamentous phage (Smith, 1985). This technology has been developed into the best-known display technique, phage display (Figure 3A). Phage display is a cell-based method in which proteins are expressed in *Escherichia coli*. Another display technique using living cells is cell-surface display (Figure 3B) in which proteins are displayed on the surface of living cells, such as yeast (Georgiou et al., 1997; Murai et al., 1997) or mammalian cells (Wolkowicz et al., 2005). These cell-based display techniques have some weaknesses; the library size is limited by the number of cells and transformation efficiency (typically below 10^9), and some proteins that are toxic to the cell are excluded from the library. To overcome such weaknesses, completely *in vitro* techniques have been developed, such as ribosome display (Hanes & Plückthun, 1997; Figure 3C), mRNA display (Nemoto et al., 1997; Roberts & Szostak, 1997; Figure 3D) and DNA display (Doi & Yanagawa, 1999; Figure 3E). Each display technique has been improved and applied to functional selection for peptides and proteins (Matsumura et al., 2006).

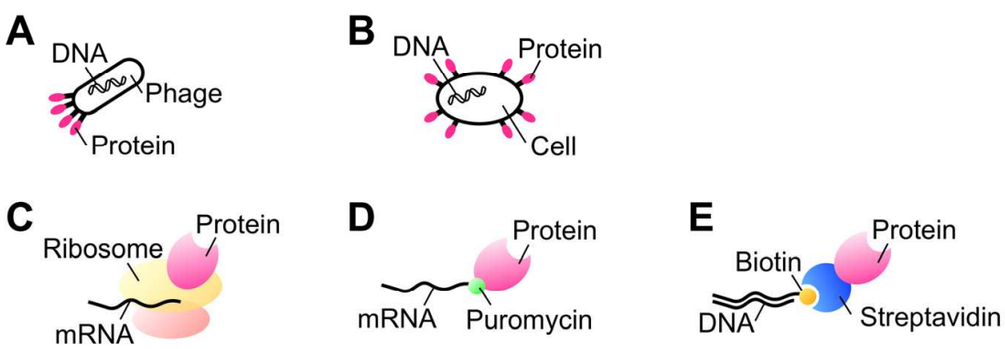


Fig. 3. Molecular display techniques. (A) Phage display (Smith, 1985). Proteins are displayed on a filamentous phage by fusing them to coat proteins of the phage. (B) Cell-surface display (Georgiou et al., 1997; Murai et al., 1997; Wolkowicz et al., 2005). Proteins are displayed on the surface of living cells. (C) Ribosome display (Hanes & Plückthun, 1997). Individual nascent proteins are coupled to their corresponding mRNA through ribosomes. (D) mRNA display (Nemoto et al., 1997; Roberts & Szostak, 1997). The protein is covalently linked with its corresponding mRNA *via* puromycin. (E) DNA display (Doi & Yanagawa, 1999). The protein is linked with its corresponding DNA by streptavidin-biotin interaction in water-in-oil emulsion.

Techniques for mRNA display have been developed in our laboratory and independently in that of Szostak (Nemoto et al., 1997; Roberts & Szostak, 1997). In this technique, each cell-free translated polypeptide (phenotype) in a library is covalently linked with its corresponding mRNA (genotype) *via* puromycin. This antibiotic is an analogue of the 3′ end of aminoacyl-tRNA (Figure 4A) and causes premature termination of translation by binding to the C-terminus of the nascent polypeptide chain. When its concentration is very low, puromycin is transferred to the C-terminus of the full-length protein (Miyamoto-Sato et al., 2000). Based on this property of puromycin, when mRNA lacking a stop codon is ligated with puromycin at the 3′ end and translated using a cell-free translation system, an mRNA (genotype) and full-length protein (phenotype) conjugate is produced (Figure 4B).

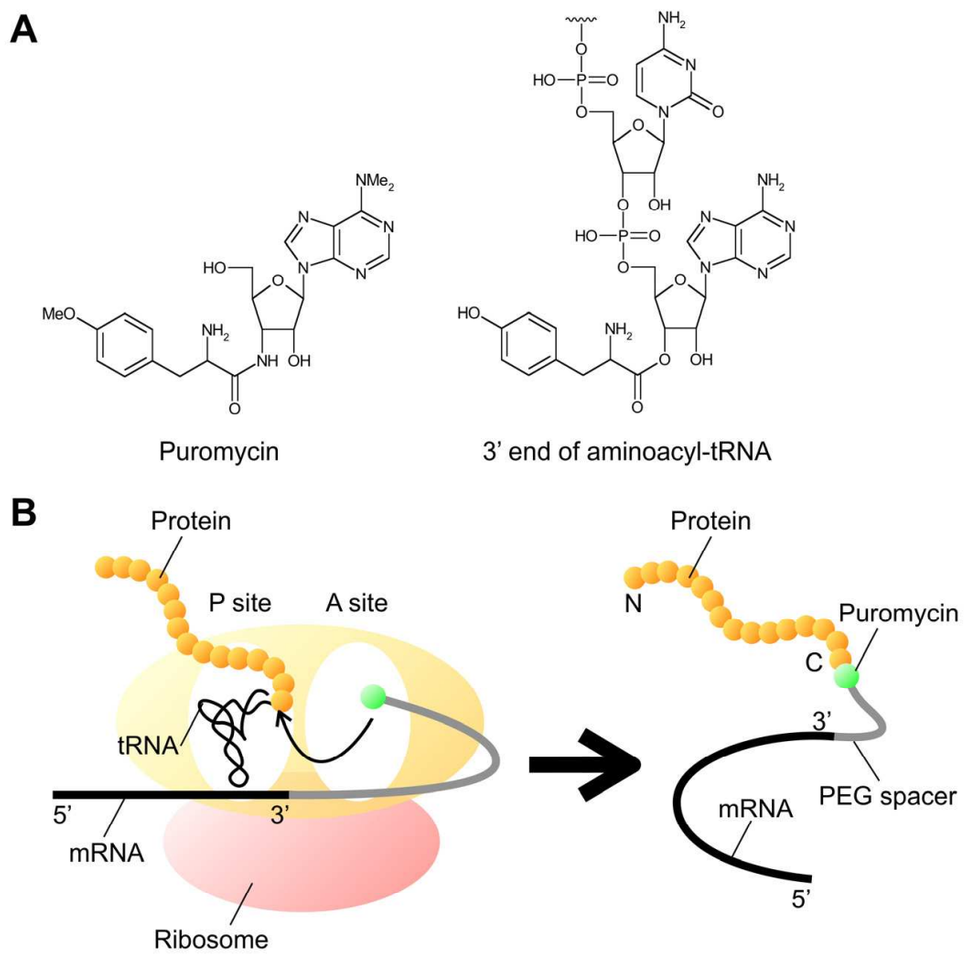


Fig. 4. The principle behind the formation of an mRNA-displayed protein. (A) The structure of puromycin and the 3′ end of aminoacyl-tRNA. (B) Puromycin ligated to the 3′-terminal end of an mRNA *via* a polyethyleneglycol (PEG) spacer can enter the ribosomal A site to bind covalently to the C-terminal end of the protein that it encodes (Nemoto et al., 1997).

In mRNA display, a larger number of molecules (approximately 10^{12-13}) can be handled than is possible using other cell-based display techniques such as phage display. This enables the enrichment of active sequences with low abundance from libraries with high diversity and complexity.

The typical scheme of *in vitro* selection using mRNA display is shown in Figure 5. Proteins are displayed on mRNA by cell-free translation of modified mRNA as described above. After affinity selection *via* the protein portion of an mRNA-displayed protein from the library, selected proteins can be easily identified by amplification and sequencing of the mRNA portion. Moreover, targeted proteins with low-copy numbers can be also detected by iterative selection. In the following sections, we describe the application of mRNA display to the construction of random-sequence protein libraries with a limited set of amino acids (Section 4) (Tanaka et al., 2010) and to the selection of functional proteins from partially randomized libraries with a limited set of amino acids (Section 5) (Tanaka et al., 2011).

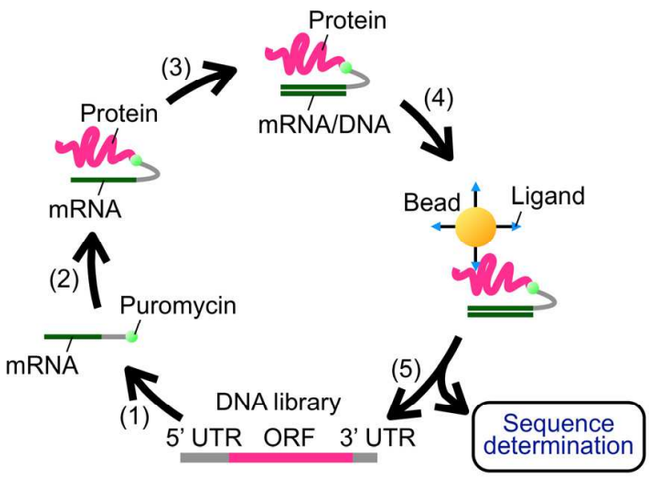


Fig. 5. Schematic representation of the mRNA-display selection of ligand-binding proteins. (1) A DNA library is transcribed and ligated with a polyethyleneglycol-puromycin spacer. (2) The modified mRNA library is translated using a cell-free translation system. (3) The resulting mRNA-protein conjugates are purified and reverse-transcribed. (4) The mRNA/DNA-protein conjugates are incubated with the ligand-immobilized beads, washed and competitively eluted with the free ligand. (5) The DNA portion of the eluted molecules is amplified using PCR to form a DNA library for the next round.

4. Random-sequence proteins with primitive amino acids

How frequently did functional or folded proteins occur in the RNA world? To answer this question, Keefe & Szostak (2001) selected novel ATP binding proteins from a random-sequence protein library based on the 20-amino acid alphabet using mRNA display. The authors roughly estimated that the frequency of occurrence of functional proteins is 1 in 10^{11}. Over the last decade, no functional protein has been obtained from random-sequence libraries. One of the difficulties in functional selection is that random-sequence proteins that use 20 types of amino acids tend to aggregate (Mandecki, 1990; Prijambada et al., 1996; Watters & Baker, 2004). Because primordial proteins presumably consisted of a smaller set of amino acids that could have been abundantly formed during early chemical evolution as mentioned above, random-sequence proteins that use 20 types of amino acids may have different physical properties from primordial proteins.

As shown in Table 2, random-sequence proteins that use a limited set of amino acids reportedly have different properties from random-sequence proteins that use 20 kinds of amino acids. Although random-sequence proteins based on three kinds of amino acids (QLR proteins, which consist of Gln, Leu and Arg) tend to strongly aggregate (Davidson & Sauer, 1994; Davidson et al., 1995), random-sequence proteins with the primitive amino acids Ala, Gly, Val, Asp and Glu, which are encoded by codons of the form GNN (N = T, C, A or G), demonstrated extremely high solubility (Doi et al., 2005). Using mRNA display, we constructed three classes of random-sequence libraries consisting of limited sets of amino acids (Tanaka et al., 2010); these libraries were encoded using the codons GNN, RNN (R = A or G, encoding a 12-amino acid alphabet) and NNN (encoding the full set of amino acids). When proteins that were arbitrarily chosen from these libraries were expressed in *Escherichia*

coli, all proteins from the GNN library were present in the soluble fraction, all of the proteins from the NNN library were present in the insoluble fraction, and the proteins from the RNN library were intermediate in character, *i.e.*, one out of 14 RNN proteins was expressed only in the soluble fraction, 11 RNN proteins were expressed only in the insoluble fraction, and two were expressed in both fractions (Tanaka et al., 2010).

Protein	Variety of Amino Acids	Solubility	Secondary Structure Content	Reference
QLR protein	3	Quite low	Strong	Davidson & Sauer, 1994; Davidson et al., 1995
GNN protein	5	High	Low	Doi et al., 2005; Tanaka et al., 2010
RNN protein	12	Medium	Low	Tanaka et al., 2010
NNN protein	20	Low	Low	Yamauchi et al., 1998; Tanaka et al., 2010

Table 2. Biophysical properties of random-sequence proteins constructed using reduced alphabets.

What causes such difference in solubilities? To investigate this question, we examined the relationship between the solubility of random-sequence proteins and several properties of the amino acid sequences (Tanaka et al., 2010). It has been suggested that protein solubility is strongly affected by net charge and the fraction of turn-forming residues (Gly, Asp, Pro, Ser and Asn) and is weakly affected by hydrophobicity and protein size (Wilkinson & Harrison, 1991). We found no relation between solubility and the fraction of turn-forming residues, hydrophobicity [calculated based on the index of Kyte and Doolittle (1982)], or protein size for GNN, RNN and NNN proteins (Tanaka et al., 2010). The high solubility of GNN proteins could be attributed to net charge because all GNN proteins lack positively charged amino acids. Soluble RNN proteins have higher net charge and lower hydrophobicity than insoluble RNN proteins. However, the low solubility of NNN proteins with high net charge and low hydrophobicity cannot be easily explained.

Random-sequence proteins with limited sets of amino acids have been structurally characterized. QLR proteins, which have three kinds of amino acids, exhibited strong α-helical content in aqueous solution but tended to aggregate, and the addition of a denaturing agent is necessary for solubilization (Davidson & Sauer, 1994; Davidson et al., 1995). Soluble RNN proteins largely adopted random coil conformations but formed α-helical structures in a hydrophobic environment (Tanaka et al., 2010). Hence, these results indicate that RNN proteins have the potential to form at least partial secondary structures, similar to random-sequence proteins based on 20 amino acid types (Yamauchi et al., 1998). Thus, there may be a trade-off between secondary structure formation and high solubility among random-sequence proteins (Table 2). Other experiments showed the presence of

hydrophobic clusters in GNN and RNN proteins (Doi et al., 2005; Tanaka et al., 2010). Furthermore, GNN and RNN proteins formed monomeric structures with more compact shapes than the random-coil structures adopted by denatured proteins of similar molecular weight but had more extended shapes than the globular structures of natural proteins. That is, they probably form molten globule-like structures (Figure 6). However, random-sequence proteins based on 3- and 20-amino acid alphabets have been reported to form oligomeric structures due to their tendency toward aggregation (Davidson & Sauer, 1994; Davidson et al., 1995; Yamauchi et al., 1998).

Recently, a large number of intrinsically unstructured domains that become structured only during binding to the target (*i.e.*, induced fit) have been identified in nature (Wright & Dyson, 1999). Moreover, artificial proteins that form well-folded structures after interaction with their target were produced (Walter et al., 2005; Vamvaca et al., 2004; Chaput & Szostak, 2004). Such partially structured polypeptides might have been the first evolutionary intermediates, and their functions and structures would have coevolved (Tokuriki & Tawfik, 2009b). Thus, random-sequence proteins based on the set of amino acids encoded by the codon RNN may include such evolutionary intermediates because these proteins contain partial secondary structures and hydrophobic clusters.

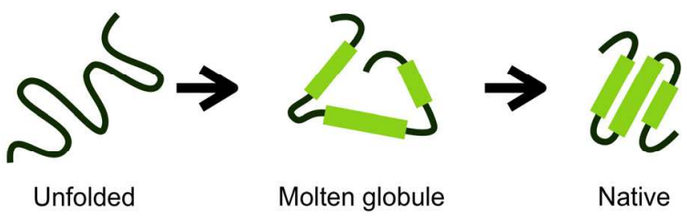


Fig. 6. Protein folding. In the unfolded state, the polypeptide chain adopts an entirely random conformation. In the folded state, the protein takes on a unique conformation. The protein folds into the compact native structure through an intermediate state, *i.e.*, a molten globule state (Ohgushi & Wada, 1983), in which much of the secondary structure (light green) is present.

5. Functional proteins consisting of primitive amino acids

As described in the previous section, random-sequence proteins constructed with subsets of the putative primitive amino acids (5 and 12-amino acid alphabets) have higher solubility than those constructed using the natural 20-member alphabet, although other biophysical properties remain very similar. Because the solubility of globular proteins is an important factor in the exertion of their function, it is of interest to test whether functional proteins occur more frequently in a library based on a limited set of primitive amino acids than in a library based on the 20-amino acid alphabet or other non-primitive alphabets.

To address this question, we attempted to compare the frequencies with which functional proteins occur in libraries based on various sets of amino acids (Tanaka et al., 2011). First, we designed randomized *src* SH3 gene libraries in which approximately half the residues of the SH3 gene were replaced by various kinds of randomized codons (Figure 7A). We utilized three limited sets of amino acids: (1) the set coded by the lower half of the genetic code (RNN) contains mainly putative primitive amino acids (*e.g.*, Gly and Ala); (2) the set

coded by the upper half of the genetic code (YNN, where Y = T or C) contains many putative new amino acids (*e.g.*, Cys, Phe, Tyr and Trp); and (3) the set coded using all bases (NNN) contains all 20 kinds of amino acids, used as a control. Subsequently, functional SH3 sequences that can bind to the SH3 ligand peptide were selected from each library using mRNA display as described in Section 3. After three rounds of *in vitro* selection, the contents of active SH3 domains in each round were analyzed using an enzyme-linked immunosorbent assay (ELISA) (Figure 7B). Functional SH3 sequences were enriched from the natural NNN library and the RNN library rich in "primitive" amino acids but not from the YNN library rich in "new" amino acids (Figure 7B).

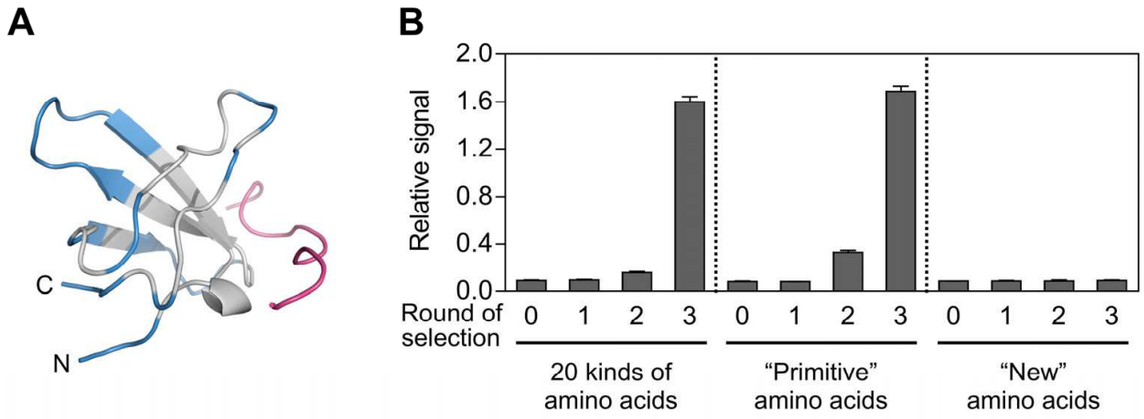


Fig. 7. *In vitro* selection of functional SH3 proteins using mRNA display. (A) The three-dimensional structure of the *src* SH3 domain (blue and gray) complexed with its peptide ligand VSL12 (red). The SH3 domain was partially randomized (blue). The structure was visualized using PyMol (PDBid, 1QWF; Feng et al., 1995). (B) The fraction of functional SH3 sequences at each round of mRNA-display selection (see Figure 5). The total amount of the three libraries [*i.e.*, those based on 20 kinds of amino acids (NNN), putative "primitive" amino acids (RNN) and putative "new" amino acids (YNN)] that bound to the peptide ligand before (0) and after 1–3 rounds of selection were quantified using ELISA. Error bars indicate the s.d. of four samples (Tanaka et al., 2011).

This result experimentally supports, for the first time, *in silico* simulations showing that modern proteins might be simplified more easily using a set of putative primitive amino acids than by a set of putative new amino acids (Babajide et al., 1997). We propose that this result cannot be explained based on differences in the typical biophysical properties (*e.g.*, charge and hydrophobicity) of individual amino acids coded by RNN versus those coded by YNN for the following reasons. First, we reconstructed a randomized SH3 domain in which highly conserved positions [*e.g.*, the ligand-binding region, the hydrophobic core and the polar surface region (Larson & Davidson, 2000)] were fixed. Second, the amino acid compositions of the randomized regions were designed to roughly equalize the biophysical properties (*i.e.*, the proportion of hydrophobic residues present and β-sheet propensity) among three kinds of random codons and resemble those of modern proteins. Thus, the reason behind the utility of primitive amino acids remains unknown but may reflect the evolutionary constraint that primordial proteins consisted of a small set of primitive amino acids and gradually acquired new amino acids during the course of neutral evolution.

Functional SH3 sequences were enriched during the second round in the library containing primitive amino acids but not during the second round in the library based on the 20-amino acid alphabet (Figure 7B). Because the biophysical properties (in particular, β-sheet propensity) were considered to be almost equal, it is not reasonable to suggest that particular amino acids that were not included in the library containing primitive amino acids, such as Pro, prevented the formation of secondary structure and peptide binding. Further study showed that the proteins selected from both libraries have similar biophysical properties, including ligand specificity, ligand affinity and thermostability (Tanaka et al., 2011). Therefore, the library rich in putative primitive amino acids included a slightly larger number of functional SH3 sequences than the randomized library based on the full set of amino acids.

Interestingly, proteins selected from the library based on the primitive amino acids were more likely to be expressed in the soluble fraction in *E. coli* than those selected from the library based on the 20-amino acid alphabet, in agreement with the results obtained using random-sequence proteins mentioned in Section 4. Thus, increasing the content of primitive amino acids in proteins may improve not only the frequency at which folded and functional proteins occur but also their solubility.

Recently, it has been reported that such limited sets of amino acids are effective for functional selection from randomized libraries (Reetz et al., 2008; Wu et al., 2010; Zheng & Reetz, 2010), although only a few amino acids were randomized in the active sites in these studies. Reetz et al. compared the quality of randomized libraries in which five amino acid residues around the active site of epoxide hydrolase were replaced by 20 kinds of amino acids encoded by NNK (K = T or G) or 12 kinds of amino acids (Gly, Asp, Val, Ser, Leu, Arg, Asn, Ile, His, Cys, Phe and Tyr) encoded by NDT (D = T, A or G). The NDT library produced many more variants with high activity than the NNK library (Reetz et al., 2008). Moreover, the authors succeeded in modifying other enzymes using a library based on the randomized codon NDT, for example, by inducing allosteric effects into Baeyer–Villiger monooxygenase (Wu et al., 2010) and manipulating the stereoselectivity of limonene epoxide hydrolase (Zheng & Reetz, 2010). Fellouse et al. (2004, 2005) demonstrated that the performance of a randomized antibody library was maintained when the number of amino acid types constituting part of a randomized complementarity-determining region (CDR) in the library was reduced to just four (Ala, Ser, Asn and Tyr) or even two (Ser and Tyr).

Although protein engineering using a limited set of primitive amino acids might improve protein folding ability and the frequency of occurrence of functional proteins, the need remains to determine the most appropriate subset of amino acids for functional selection because our study (Tanaka et al., 2011) and those of Reetz's group (Reetz et al., 2008; Wu et al., 2010; Zheng & Reetz, 2010) and Fellouse et al. (2004, 2005) simultaneously compared only a few subsets of amino acids. Some putative new amino acids may be essential for some structures and functions. For example, Cys, which might be a late recruit into the amino acid repertoire, improves structural stability by forming intra- and intermolecular disulfide bonds. A new amino acid, His, is also significant because of its role at the active center of enzymes where it binds to metal ions through the imidazole group. In the course of protein evolution, the recruitment of putative new amino acids may have generated new catalytic activities and more complicated and stable structures. This would be a reason for proteins to have employed an expanded set of amino acids rather than limiting themselves

to primitive amino acids. Thus, not only putative primitive amino acids but also certain new amino acids, such as Cys and His as described above, may have been needed in the design of artificial proteins, depending on the target function.

6. Conclusion

Soluble, functional proteins tend to occur more frequently in libraries based on limited sets of primitive amino acids than in libraries based on limited sets of new amino acids and library based on the full set of 20 amino acids. Thus, the evolutionary engineering of proteins using limited sets of primitive amino acids may be an effective tool for the creation of artificial proteins, such as industrial enzymes and monoclonal antibodies that are used in the pharmaceutical industry.

7. Acknowledgments

We thank members of our laboratory at Keio University for helpful comments and discussions. We also thank Drs Hideaki Takashima, Kenichi Horisawa, Seiji Tateyama and Etsuko Miyamoto-Sato in particular for their help with mRNA display and Dr Toru Tsuji for help with characterization of the SH3 domain variants. This work was supported in part by a Grant-in-Aid for Scientific Research (19657073) from the JSPS (Japan Society for the Promotion of Science) and a Grant-in-Aid from the Keio University Global Center of Excellence (G-COE) Program entitled 'Center of Human Metabolomic Systems Biology' from MEXT (the Ministry of Education, Culture, Sports, Science and Technology) of Japan.

8. References

Akanuma, S.; Kigawa, T. & Yokoyama, S. (2002). Combinatorial mutagenesis to restrict amino acid usage in an enzyme to a reduced set. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 99, No. 21, pp. 13549-13553.

Altman, S. (1981). Transfer RNA processing enzymes. *Cell*, Vol. 23, No. 1, pp. 3-4.

Babajide, A.; Hofacker, I.L.; Sippl, M.J. & Stadler, P.F. (1997). Neutral networks in protein space: a computational study based on knowledge-based potentials of mean force. *Folding & Design*, Vol. 2, No. 5, pp. 261-269.

Bloom, J.D.; Meyer, M.M.; Meinhold, P.; Otey, C.R.; MacMillan, D. & Arnold F.H. (2005). Evolving strategies for enzyme engineering. *Current Opinion in Structural Biology*, Vol. 15, No. 4, pp. 447-452.

Cech, T.R.; Zaug, A.J. & Grabowski, P.J. (1981). *In vitro* splicing of the ribosomal RNA precursor of Tetrahymena: involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell*, Vol. 27, No. 3, pp. 487-496.

Chaput, J.C. & Szostak, J.W. (2004). Evolutionary optimization of a nonbiological ATP binding protein for improved folding stability. *Chemistry & Biology*, Vol. 11, No. 6, pp. 865-874.

Davidson, A.R. & Sauer, R.T. (1994). Folded proteins occur frequently in libraries of random amino acid sequences. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 91, No. 6, pp. 2146-2150.

Davidson, A.R.; Lumb, K.J. & Sauer, R.T. (1995). Cooperatively folded proteins in random sequence libraries. *Nature Structural Biology*, Vol. 2, No. 10, pp. 856-864.

Doi, N. & Yanagawa, H. (1999). STABLE: protein-DNA fusion system for screening of combinatorial protein libraries in vitro. *FEBS Letters*, Vol. 457, No. 2, pp. 227-230.

Doi, N. & Yanagawa, H. (2001). Genotype-phenotype linkage for directed evolution and screening of combinatorial protein libraries. *Combinatorial Chemistry & High Throughput Screening*, Vol. 4, No. 6, pp. 497-509.

Doi, N.; Kakukawa, K.; Oishi, Y. & Yanagawa, H. (2005). High solubility of random-sequence proteins consisting of five kinds of primitive amino acids. *Protein Engineering, Design & Selection*, Vol. 18, No. 6, pp. 279-284.

Dryden, D.T.F.; Thomson, A.R. & White, J.H. (2008). How much of protein sequence space has been explored by life on Earth? *Journal of The Royal Society Interface*, Vol. 5, No. 25, pp. 953-956.

Eigen, M. & Schuster, P. (1978). The hypercycle. A principle of natural self-organization. Part C: The realistic hypercycle. *Naturwissenschaften*, Vol. 65, No. 7, pp. 341-369.

Fellouse, F.A.; Li, B.; Compaan, D.M.; Peden, A.A.; Hymowitz, S.G. & Sidhu, S.S. (2005). Molecular recognition by a binary code. *Journal of Molecular Biology*, Vol. 348, No. 5, pp. 1153-1162.

Fellouse, F.A.; Wiesmann, C. & Sidhu, S.S. (2004). Synthetic antibodies from a four-amino-acid code: a dominant role for tyrosine in antigen recognition. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 101, No. 34, pp. 12467-12472.

Feng, S.; Kasahara, C.; Rickles, R.J. & Schreiber, S.L. (1995). Specific interactions outside the proline-rich core of two classes of Src homology 3 ligands. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 92, No. 26, pp. 12408-12415.

Georgiou, G.; Stathopoulos, C.; Daugherty, P.S.; Nayak, A.R.; Iverson, B.L. & Curtiss, R. 3rd. (1997). Display of heterologous proteins on the surface of microorganisms: from the screening of combinatorial libraries to live recombinant vaccines. *Nature Biotechnology*, Vol. 15, No. 1, pp. 29-34.

Go, A.; Kim, S.; Baum, J. & Hecht, M.H. (2008). Structure and dynamics of de novo proteins from a designed superfamily of 4-helix bundles. *Protein Science*, Vol. 17, No. 5, pp. 821-832.

Hanes, J. & Plückthun, A. (1997). *In vitro* selection and evolution of functional proteins by using ribosome display. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 94, No. 10, pp. 4937-4942.

Hoogenboom, H.R. (2005). Selecting and screening recombinant antibody libraries. *Nature Biotechnology*, Vol. 23, No. 9, pp. 1105-1116.

Huang, J.; Koide, A.; Makabe, K. & Koide, S. (2008). Design of protein function leaps by directed domain interface evolution. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 105, No. 18, pp. 6578-6583.

Jordan, I.K.; Kondrashov, F.A.; Adzhubei, I.A.; Wolf, Y.I.; Koonin, E.V.; Kondrashov, A.S. & Sunyaev, S. (2005). A universal trend of amino acid gain and loss in protein evolution. *Nature*, Vol. 433, No. 7026, pp. 633-638.

Jumawid, M.T.; Takahashi, T.; Yamazaki, T.; Ashigai, H. & Mihara, H. (2009). Selection and structural analysis of *de novo* proteins from an α3β3 genetic library. *Protein Science*, Vol. 18, No. 2, pp. 384-398.

Kamtekar, S.; Schiffer, J.M.; Xiong, H.; Babik, J.M. & Hecht, M.H. (1993). Protein design by binary patterning of polar and nonpolar amino acids. *Science*, Vol. 262, No. 5140, pp. 1680-1685.

Keefe, A.D. & Szostak, J.W. (2001). Functional proteins from a random-sequence library. *Nature*, Vol. 410, No. 6829, pp. 715-718.

Kyte, J. & Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, Vol. 157, No. 1, pp.105-132.

Larson, S.M. & Davidson, A.R. (2000). The identification of conserved interactions within the SH3 domain by alignment of sequences and structures. *Protein Science*, Vol. 9, No. 11, pp. 2170-2180.

Leemhuis, H.; Stein, V.; Griffiths, A.D. & Hollfelder, F. (2005). New genotype-phenotype linkages for directed evolution of functional proteins. *Current Opinion in Structural Biology*, Vol. 15, No. 4, pp. 472-478.

Mandecki, W. (1990). A method for construction of long randomized open reading frames and polypeptides. *Protein Engineering*, Vol. 3, No. 3, pp. 221-226.

Mandecki, W. (1998). The game of chess and searches in protein sequence space. *Trends in Biotechnology*, Vol. 16, No. 5, pp. 200-202.

Matsumura, N.; Doi, N. & Yanagawa, H. (2006). Recent progress and future prospects in protein display technologies as tools for proteomics. *Current Proteomics*, Vol. 3, No. 3, pp. 199-215.

Miller, S.L. (1987). Which organic compounds could have occurred on the prebiotic earth? *Cold Spring Harbor Symposia on Quantitative Biology*, Vol. 52, pp. 17-27.

Miyamoto-Sato, E.; Nemoto, N.; Kobayashi, K. & Yanagawa, H. (2000). Specific bonding of puromycin to full-length protein at the C-terminus. *Nucleic Acids Research*, Vol. 28, No. 5, pp. 1176-1182.

Murai, T.; Ueda, M.; Atomi, H.; Shibasaki, Y.; Kamasawa, N.; Osumi, M.; Kawaguchi, T.; Arai, M. & Tanaka, A. Genetic immobilization of cellulase on the cell surface of Saccharomyces cerevisiae. (1997). *Applied Microbiology and Biotechnology*, Vol. 48, No. 4, pp. 499-503.

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 Letters*, Vol. 414, No. 2, pp. 405-408.

Ohgushi, M. & Wada, A. (1983). 'Molten-globule state': a compact form of globular proteins with mobile side-chains. *FEBS Letters*, Vol. 164, No. 1, pp. 21-24.

Oparin, A.I. (1961). *Life, its nature, origin and development*, Academic Press, New York.

Patel, S.C.; Bradley, L.H.; Jinadasa, S.P. & Hecht, M.H. (2009). Cofactor binding and enzymatic activity in an unevolved superfamily of *de novo* designed 4-helix bundle proteins. *Protein Science*, Vol. 18, No. 7, pp. 1388-1400.

Peisajovich, S.G.; Rockah, L. & Tawfik, D.S. (2006). Evolution of new protein topologies through multistep gene rearrangements. *Nature Genetics*, Vol. 38, No. 2, pp. 168-174.

Prijambada, I.D.; Yomo, T.; Tanaka, F.; Kawama, T.; Yamamoto, K.; Hasegawa, A.; Shima, Y.; Negoro, S. & Urabe, I. (1996). Solubility of artificial proteins with random sequences. *FEBS Letters*, Vol. 382, No. 1-2, pp. 21-25.

Reetz, M.T.; Kahakeaw, D. & Lohmer, R. (2008). Addressing the numbers problem in directed evolution. *ChemBioChem*, Vol. 9, No. 11, pp. 1797-1804.

Riddle, D.S.; Santiago, J.V.; Bray-Hall, S.T.; Doshi, N.; Grantcharova, V.P.; Yi, Q. & Baker, D. (1997). Functional rapidly folding proteins from simplified amino acid sequences. *Nature Structural Biology*, Vol. 4, No. 10, pp. 805-809.

Roberts, R.W. & Szostak, J.W. (1997). RNA-peptide fusions for the *in vitro* selection of peptides and proteins. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 94, No. 23, pp. 12297-12302.

Romero, P.A. & Arnold, F.H. (2009). Exploring protein fitness landscapes by directed evolution. *Nature Reviews Molecular Cell Biology*, Vol. 10, No. 12, pp. 866-876.

Silverman, J.A.; Balakrishnan, R. & Harbury, P.B. (2001). Reverse engineering the $(\beta/\alpha)_8$ barrel fold. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 98, No. 6, pp. 3092-3097.

Smith, G.P. (1985). Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science*, Vol. 228, No. 4705, pp. 1315-1317.

Tanaka, J.; Doi, N.; Takashima, H. & Yanagawa, H. (2010). Comparative characterization of random-sequence proteins consisting of 5, 12, and 20 kinds of amino acids. *Protein Science*, Vol. 19, No. 4, pp. 786-795.

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*, Vol. 6, No. 3, e18034.

Tokuriki, N. & Tawfik, D.S. (2009a). Chaperonin overexpression promotes genetic variation and enzyme evolution. *Nature*, Vol. 459, No. 7247, pp. 668-673.

Tokuriki, N. & Tawfik, D.S. (2009b). Protein dynamism and evolvability. *Science*, Vol. 324, No. 5924, pp. 203-207.

Trifonov, E.N. (2004). The triplet code from first principles. *Journal of Biomolecular Structure & Dynamics*, Vol. 22, No. 1, pp. 1-11.

Vamvaca, K.; Vögeli, B.; Kast, P.; Pervushin, K. & Hilvert, D. (2004). An enzymatic molten globule: efficient coupling of folding and catalysis. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 101, No. 35, pp. 12860-12864.

Walter, K.U.; Vamvaca, K. & Hilvert, D. (2005). An active enzyme constructed from a 9-amino acid alphabet. *The Journal of Biological Chemistry*, Vol. 280, No. 45, pp. 37742-37746.

Watters, A.L. & Baker, D. (2004). Searching for folded proteins *in vitro* and *in silico*. *European Journal of Biochemistry*, Vol. 271, No. 9, pp. 1615-1622.

Wilkinson, D.L. & Harrison, R.G. (1991). Predicting the solubility of recombinant proteins in *Escherichia coli*. *Biotechnology*, Vol. 9, No. 5, pp. 443-448.

Wolkowicz, R.; Jager, G.C. & Nolan, G.P. (2005). A random peptide library fused to CCR5 for selection of mimetopes expressed on the mammalian cell surface via retroviral vectors. *The Journal of Biological Chemistry*, Vol. 280, No. 15, pp. 15195-15201.

Wong, J.T. (1975). A co-evolution theory of the genetic code. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 72, No. 5, pp. 1909-1912.

Wong, J.T. (1988). Evolution of the genetic code. *Microbiological Sciences*, Vol. 5, No. 6, pp. 174-181.

Wong, J.T. (2005). Coevolution theory of the genetic code at age thirty, *BioEssays*, Vol. 27, No. 4, pp. 416-425.

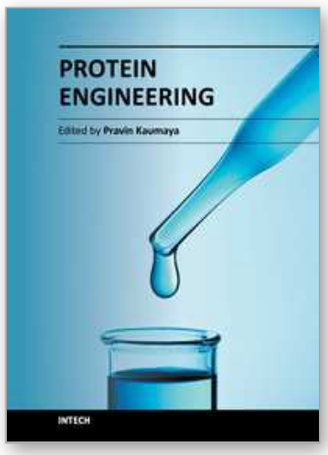
Wright, P.E. & Dyson, H.J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of Molecular Biology*, Vol. 293, No. 2, pp. 321-331.

Wu, S.; Acevedo, J.P. & Reetz, M.T. (2010). Induced allostery in the directed evolution of an enantioselective Baeyer-Villiger monooxygenase. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 107, No. 7, pp. 2775-2780.

Yamauchi, A.; Yomo, T.; Tanaka, F.; Prijambada, I.D.; Ohhashi, S.; Yamamoto, K.; Shima, Y.; Ogasahara, K.; Yutani, K.; Kataoka, M. & Urabe, I. (1998). Characterization of soluble artificial proteins with random sequences. *FEBS Letters*, Vol. 421, No. 2, pp. 147-151.

Zheng, H. & Reetz, M.T. (2010). Manipulating the stereoselectivity of limonene epoxide hydrolase by directed evolution based on iterative saturation mutagenesis. *Journal of the American Chemical Society*, Vol. 132, No. 44, pp. 15744-15751.

IntechOpen



Protein Engineering

Edited by Prof. Pravin Kaumaya

ISBN 978-953-51-0037-9
Hard cover, 344 pages
Publisher InTech
Published online 24, February, 2012
Published in print edition February, 2012

A broad range of topics are covered by providing a solid foundation in protein engineering and supplies readers with knowledge essential to the design and production of proteins. This volume presents in-depth discussions of various methods for protein engineering featuring contributions from leading experts from different counties. A broad series of articles covering significant aspects of methods and applications in the design of novel proteins with different functions are presented. These include the use of non-natural amino acids, bioinformatics, molecular evolution, protein folding and structure-functional insight to develop useful proteins with enhanced properties.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Junko Tanaka, Hiroshi Yanagawa and Nobuhide Doi (2012). Evolutionary Engineering of Artificial Proteins with Limited Sets of Primitive Amino Acids, Protein Engineering, Prof. Pravin Kaumaya (Ed.), ISBN: 978-953-51-0037-9, InTech, Available from: http://www.intechopen.com/books/protein-engineering/evolutionary-engineering-of-artificial-proteins-with-limited-sets-of-primitive-amino-acids

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

© 2012 The Author(s). Licensee IntechOpen. This is an open access article distributed under the terms of the Creative Commons Attribution 3.0 License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

IntechOpen

IntechOpen