BIOCHEMISTRY
including biophysical chemistry & molecular biology

Current Topic

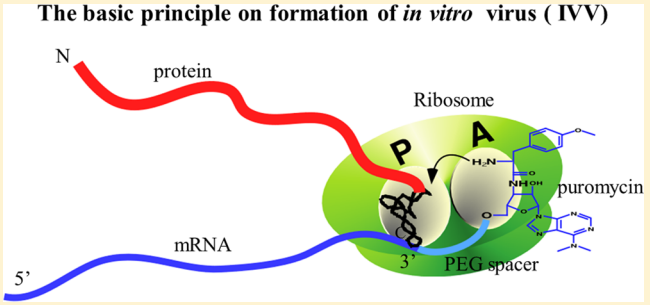
pubs.acs.org/biochemistry

Exploration of the Origin and Evolution of Globular Proteins by mRNA Display

Hiroshi Yanagawa*

Department of Biosciences and Informatics, Faculty of Sciences and Technology, Keio University, 3-14-1, Hiyoshi, Kohoku-ku, Yokohama 223-8522, Japan

ABSTRACT: The questions of how proteins first appeared on the primitive earth and how they evolved into functional proteins are fundamental. If we can understand the origins and evolution of proteins, we should be able to create novel functional proteins. Evolutionary protein engineering or directed protein evolution has been used to create artificial proteins with novel functions by repeated mutation, selection, and amplification, mimicking Darwinian evolution in the laboratory. For this purpose, display technology, such as mRNA display, to link genotype with phenotype is extremely important. Here I focus on three hypotheses regarding the origin and evolution of proteins. First, Eigen's GNC hypothesis proposes that the early genetic code began from the directionless codons GNC and GNN, where N denotes U, C, A, or G. Second, Ohno's gene duplication theory proposes that gene duplication produces two functionally redundant, paralogous genes, of which one retains the original function, leaving the second free to evolve adaptively. Third, Gilbert's exon shuffling theory proposes that new genes are formed through shuffling of small segments corresponding to exons. I then review various experimental approaches to evolutionary protein engineering using mRNA display, such as the creation of functional proteins from random sequences with limited sets of amino acids, randomly mutated folded proteins, and block-shuffled sequence proteins, and I discuss the results in relation to these three hypotheses.



HYPOTHESES REGARDING THE ORIGINS AND EARLY EVOLUTION OF PROTEINS

In 1953, Miller and Urey discovered that protein-forming amino acids such as glycine and alanine can easily be formed by electric discharges through a primitive atmosphere model consisting of a mixture of methane, ammonia, and water.[1] Since then, abiotic synthesis of biological building blocks, such as amino acids, nucleobases, and sugars that constitute protein, RNA, and DNA, has been examined in many chemical evolution experiments. According to the chemical evolution hypothesis,[2] soups of nutrient organic compounds formed in various environments of the primitive Earth, such as tidelands,[3] warm oceans,[4] and submarine hydrothermal vents.[5] Once a self-replicating molecule formed in the primordial soup, this early replicator could have evolved into a primordial cell.[4]

Since the discovery of RNA enzymes (ribozymes),[6,7] RNA molecules have been postulated to be the first self-replicating molecules capable of storing genetic information and catalyzing chemical reactions. Later, RNA acquired multiple catalytic activities in the "RNA world", and protein synthesis could have been established. Finally, the location of genetic information moved from RNA to DNA through the "RNA-protein (RNP) world", because DNA is more stable than RNA. Present living organisms mainly use not RNA but proteins as enzymes or constituent materials, because proteins are more flexible than nucleic acids for building biopolymers with various functions. When the RNA world shifted to the RNA-protein (RNP) world, what kind of function and structure did the first emerging proteins have? Moreover, can we evaluate the likelihood of such a shift? Here, I will describe an approach to these questions based on *in vitro* evolutionary protein engineering. First, after reviewing our basic knowledge of proteins, I will introduce some hypotheses about their origin and evolution.

With regard to protein biosynthesis and the origin of the genetic code, the following hypothesis has been proposed on the basis of circumstantial evidence. Namely, primordial proteins were composed of a small set of amino acids, such as glycine, alanine, aspartic acid, and valine, which could have been abundantly formed by electric discharges through the primitive atmosphere[1] in the primeval earth environment, and this set was extended to the current 20 amino acids because of the evolution of amino acid biosynthetic pathways. Interestingly, the codons for these amino acids all have guanosine (G) as the first nucleotide (Figure 1); for this reason, Eigen et al.[9] hypothesized that the early genetic code began from the directionless codons GNC and GNN, where N denotes U, C, A, or G.

The best-known hypotheses about protein-encoding gene evolution are the gene duplication theory[10] and the exon shuffling theory.[11] Ohno proposed the gene duplication theory in 1970, representing the emergence and evolution of new

Received: December 25, 2012
Revised: May 13, 2013
Published: May 16, 2013

dx.doi.org/10.1021/bi301704x | *Biochemistry* 2013, 52, 3841−3851

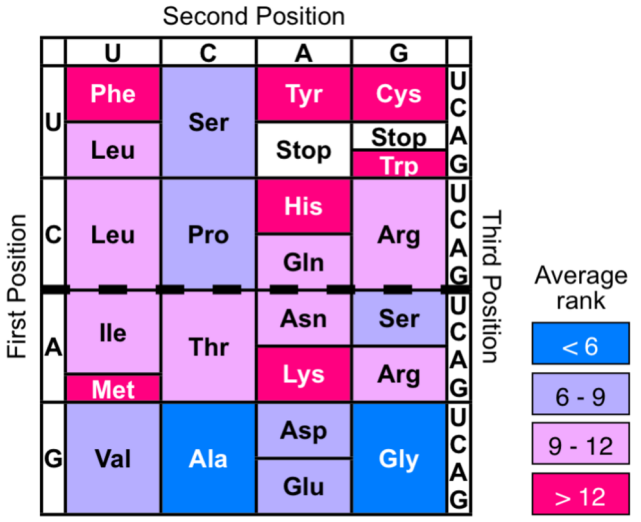


Figure 1. Universal genetic code. The average rank represents the chronological order of addition of an amino acid to the genetic code. The values calculated from 60 criteria were 3.5 for Gly, 4.0 for Ala, 6.0 for Asp, 6.3 for Val, 7.3 for Pro, 7.6 for Ser, 8.1 for Glu, 9.4 for Thr, 9.9 for Leu, 11.0 for Arg, 11.3 for Asn, 11.4 for Ile, 11.4 for Gln, 13.0 for His, 13.3 for Lys, 13.8 for Cys, 14.2 for Phe, 15.2 for Tyr, 15.4 for Met, and 16 for Trp.[8] The top half of the table of the genetic code corresponding to codon YNN (Y = T or C; N = T, C, A, or G) contains many newly added amino acids (e.g., Phe and Cys), while the bottom half corresponding to codon RNN (R = A or G) contains the most primitive amino acids (Ala and Gly).

genes as "one creation followed by hundreds of plagiarisms".[10] That is, he suggested that the creation of a new gene with a new function is a very rare occurrence and instead many genes arise by plagiarism, in which the original gene is copied (duplicated) and subsequently the duplicate is modified in part to change its function. Ohno argued that gene duplication has been the most important evolutionary driver since the emergence of the universal common ancestor. It can be defined as duplication of any region of DNA that contains a gene; it may occur as an error in homologous recombination, a retrotransposition event, or duplication of an entire chromosome. The second copy of the gene is often free from selection pressure; that is, mutations in this copy have no deleterious effect on the host organism. Thus, mutations accumulate in it faster than in a functional single-copy gene, over multiple generations of organisms. It was not until the late 1990s, by which time many genome sequences had been determined and analyzed, that the prevalence and importance of gene duplication had been clearly demonstrated.[12] Through genomic sequence analysis, population genetic modeling, and molecular experimentation, rapid progress has also been made in understanding the mechanisms by which duplicate genes diverge in function and contribute to evolution. However, Bershtein and Tawfik recently measured the frequency of potentially adaptive mutations under various mutational drifts, and their data supported a revision of Ohno's duplication model in the spirit of previously proposed alternatives, in which gene duplication promotes divergence by alleviating the selection pressure, rather than totally relieving it.[13]

On the other hand, immediately after the discovery of exon/intron structure in 1977, Gilbert proposed that gene evolution occurs by shuffling of exons.[11] According to the exon theory of genes advocated by Gilbert, the first genes were formed from small pieces of DNA corresponding to exons, and their products were short polypeptides 15−20 amino acids long that served as folding and functional elements. New genes were formed by exon shuffling after the emergence of a certain initial number of genes.

These hypotheses are both consistent with the idea that "hundreds of plagiarisms" occurred after a certain number of genes had initially emerged. To test and compare these hypotheses, we require a methodology for conducting *in vitro* evolution. Therefore, I next review our mRNA display technology, which represents as a powerful tool for selection of functional proteins from different protein libraries.

A STRATEGY FOR *IN VITRO* SELECTION OF FUNCTIONAL PROTEINS BY THE *IN VITRO* VIRUS (MRNA DISPLAY) METHOD

The objective of evolutionary protein engineering or directed protein engineering is to create a desired protein function through successive rounds of mutation, selection, and amplification starting from a parent protein that exhibits a related function, mimicking Darwinian evolution in the laboratory.[14] In the selection of desired functional proteins by evolutionary protein engineering, the most important consideration is the ability to link genotype and phenotype. "Phenotype" refers to biological functions, whereas "genotype" refers to the coding nucleic acids required 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 replicate. Therefore, the development of a molecular display technique that physically links genotype with phenotype is essential for directed protein evolution. To date, various cell-based display techniques, such as cell-surface display (including phage display) and the yeast two-hybrid method, have been developed.[15,16] However, these cell-based display techniques have some weaknesses; i.e., the library size is limited by the number of cells (typically $<10^9$) and the transformation efficiency, and some proteins that are toxic to the cell are excluded from the library. Totally *in vitro* display technologies, such as ribosome display,[17] mRNA display,[18−20] and DNA display,[21] can overcome these weaknesses because they do not require living cells.

In mRNA display, a conjugate in which mRNA (genotype) binds to protein (phenotype) through puromycin in a cell-free translation system is constructed, and thus, the mRNA moiety can be amplified by means of reverse transcription polymerase chain reaction (RT-PCR) after affinity selection via the protein moiety of the conjugate. Via the performance of iterative selection, very low copy number proteins can be routinely picked up from large-scale cDNA libraries, in the range of 10^{13} members. Puromycin, an analogue of the 3′ end of aminoacyl-tRNA, is transferred nonspecifically to growing polypeptide chains, causing premature termination of translation.[22−29] However, at very low concentrations, puromycin is transferred specifically to the carboxyl terminus of the full-length protein.[30] On the basis of 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. In mRNA display, a larger number of molecules (approximately 10^{12}−10^{13}) can be handled versus what is possible using other cell-based display techniques, such as phage display. This allows the enrichment of active sequences with low abundance from libraries with high diversity and

dx.doi.org/10.1021/bi301704x | *Biochemistry* 2013, 52, 3841−3851

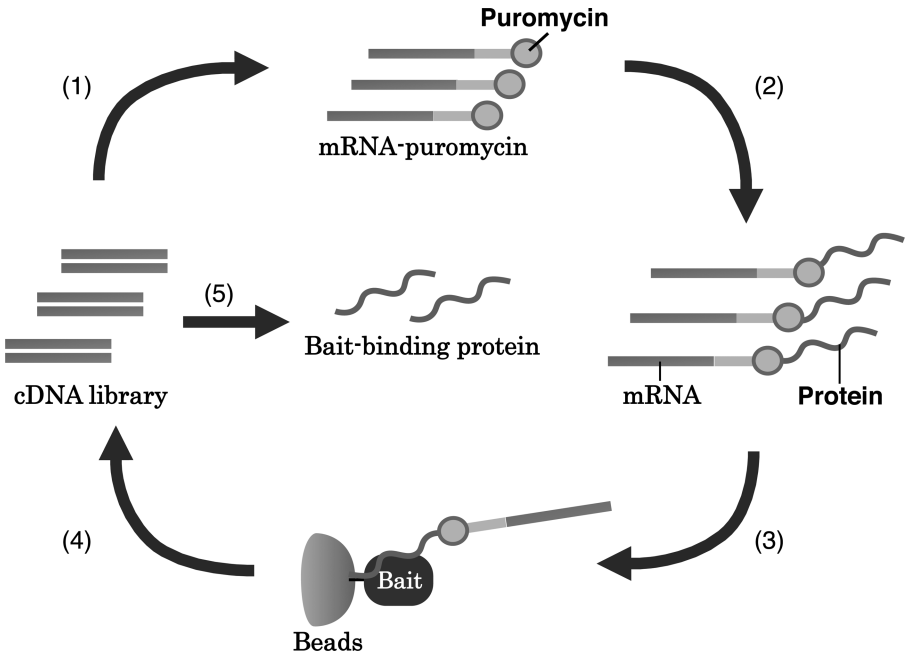


Figure 2. Schematic representation of *in vitro* selection of bait-binding proteins by means of the mRNA display (IVV) method. A cDNA library (random sequence, block-shuffled or derived from eukaryotic cells) is ligated with a PEG-Puro spacer (1) and translated *in vitro* (2) to form a protein−mRNA conjugate (IVV) library. This library is incubated with biotinylated bait-immobilized beads (3), and unbound molecules are washed away. The bound molecules are eluted with bait, and the mRNA portion of these selected molecules is amplified by RT-PCR (4). The resulting DNA is used for the next round of selection and analyzed by cloning and sequencing. Finally, the bait-binding proteins are identified (5). Any material (e.g., protein, peptide, DNA, drug, plastic, or ceramic) can be used as bait.

complexity. mRNA display was originally developed in our laboratory in 1997 and independently reported shortly thereafter by Szostak; the conjugate of protein with its encoding mRNA was named *in vitro* virus (IVV) by us[18,19] and RNA−peptide fusion by them.[20] mRNA display methods were originally developed for evolutionary protein engineering based on *in vitro* translation systems, and they were subsequently applied for a variety of functional analyses, including protein−protein,[31−34] DNA−protein,[35] RNA−protein,[36,37] drug−protein,[38−40] peptide−protein,[41−44] and antigen−antibody[45−47] interactions.

A typical scheme of *in vitro* selection using mRNA display is shown in Figure 2. Proteins are displayed on mRNA by cell-free translation of modified mRNA as described above. After affinity selection via the protein portions of mRNA-displayed proteins from the library, selected proteins can be easily identified by amplification and sequencing of the mRNA portions. Moreover, targeted proteins with low copy numbers can be detected by iterative selection.

In particular, I next focus attention on the application of mRNA display to the selection of functional proteins from fully randomized and partially randomized libraries with a limited set of amino acids,[48,49] and selection of functional proteins from block-shuffled libraries,[50−52] which are constructed on the basis of the hypotheses described above. Such application of mRNA display allows the validity of the hypotheses to be evaluated in terms of the numbers of functional proteins obtained from these libraries.

■ "ONE CREATION": SEARCH FOR THE ORIGIN OF GLOBULAR STRUCTURE IN PROTEINS

Globular proteins are thermodynamically folded into an intrinsically stable structure as a consequence of their amino acid sequence and can exhibit a function by recognizing a molecule with a complementary structure, like a lock and key. The number of possible amino acid sequences of a 100-residue protein is 20^{100} (approximately 10^{130}), which is larger than the total number of atoms in the universe ($\sim 10^{80}$). This number is enormous compared with the number of proteins that may have existed in nature throughout the history of life on the Earth, which has been estimated to be $<10^{50}$ molecules[53] or $<10^{43}$ molecules.[54] Thus, a vast sequence space still remains available to be explored, providing an opportunity to create useful proteins with novel structures and functions for biomedical and environmental applications.

The rate of occurrence of amino acid sequences with a function and structure in the sequence space is an extremely important issue, not only for the origin of protein structure and function but also for the creation of novel functional proteins. If functional proteins cannot be easily obtained because of the low appearance frequency of functional proteins in the sequence space, then the evolutionary search process may not have been random. If it was not random, an understanding of the mechanisms involved would be helpful for the creation of artificial proteins by evolutionary engineering.

In evolutionary RNA engineering, screening of functional RNAs from random sequences has proven to be a successful strategy.[55−59] However, in evolutionary protein engineering, it is technically difficult to conduct screening in the same way. Some searches for protein sequences with function and conformation from artificial protein libraries with random amino acid sequences have been performed. However, no protein with a function and conformation equivalent to those of natural globular proteins was obtained, probably because artificial proteins consisting of random amino acid sequences generally aggregate easily, and the sequence space covered by conventional screening technology is quite small.[60−62] In 2001,

dx.doi.org/10.1021/bi301704x | *Biochemistry* 2013, 52, 3841−3851

Keefe and others[63] obtained artificial ATP-binding proteins from a random-sequence library (based on the 20-amino acid alphabet), including 10^{12} sequences, and roughly estimated that the frequency of occurrence of functional proteins is 1 in 10^{11}. From the viewpoint of the vast sequence space available, this value is not so low, but it is very low from the viewpoint of actually finding new functional proteins. There are very few examples of the creation of random-sequence proteins with function and conformation comparable to those of natural proteins. So, how did the present-day variety of functional proteins arise in the early evolution of life?

Eigen's GNC hypothesis[9] that the genetic code began from the directionless codons GNC and GNN can be tested with both reductionistic and constitutive approaches. First, as a reductionistic approach, Jordan et al.[64] conducted comparative genome sequence analysis of orthologous proteins in the genomes of bacteria, archaea, and eukaryotes. They concluded that the frequencies of Gly, Ala, Glu, and Pro consistently decreased in proteins while the frequencies of Ser, His, Cys, Met, and Phe consistently increased during protein evolution. 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 later recruits.[64] On the other hand, all 20 amino acids are not necessarily indispensable for the maintenance of various conformations of today's naturally occurring proteins, and it has been experimentally shown by many researchers that the conformation and function of proteins can be maintained with only five to nine different amino acids. This is an example of a constitutive approach. Although random-sequence proteins based on three kinds of amino acids (QLR proteins, which consist of Gln, Leu, and Arg) tend to aggregate strongly,[65,66] 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), showed extremely high solubility.[69] Using mRNA display, we constructed three classes of random-sequence libraries consisting of limited sets of amino acids;[48] 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. For instance, one 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.[48] Thus, a random-sequence library consisting of "primitive" amino acids may include functional sequences at a higher rate than a library based on 20 amino acids, because water solubility is an important factor in protein function.

What causes such differences in solubility? To investigate this question, we examined the relationship between the solubility of random-sequence proteins and several properties of the amino acid sequences.[48] 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.[67] We found no relation between solubility and the fraction of turn-forming residues, hydrophobicity (calculated on the basis of the index of Kyte and Doolittle[68]), or protein size for GNN, RNN, and NNN proteins.[48] 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 charges and lower hydrophobicities than insoluble RNN proteins. However, the low solubility of NNN proteins with high net charges and low hydrophobicities cannot be easily explained. The constitutive approach thus provided circumstantial evidence that supports the hypothesis that early proteins started from a small number of primitive amino acids, as well as offering a construction method for libraries that might contain useful new proteins.

Although modern proteins consist of 20 different amino acids, it has been proposed that primordial proteins consisted of a small set of amino acids, and additional amino acids have gradually been recruited into the genetic code. This hypothesis has recently been supported by comparative genome sequence analysis, but there has been no direct experimental evidence. We utilized a novel experimental approach to test the hypothesis that nativelike globular proteins could be simplified by employing a set of putative primitive amino acids, with retention of both structure and function. We performed *in vitro* selection of a functional SH3 domain as a model from partially randomized libraries with different sets of amino acids using mRNA display. Indeed, a library rich in putative primitive amino acids included a larger number of functional SH3 sequences than a library rich in putative new amino acids. Further, the functional SH3 sequences were enriched from the primitive library slightly earlier than from a randomized library with the full set of amino acids, while the function and structure of the selected SH3 proteins with the primitive alphabet were comparable with those obtained from the 20-amino acid alphabet. Application of this approach to various combinations of codons in protein sequences may be useful not only for clarifying the precise order of amino acid expansion in the early stages of protein evolution but also for efficiently creating novel functional proteins in the laboratory.

As described above, 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 important for 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 nonprimitive 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.[49] First, we designed randomized src SH3 gene libraries in which approximately half the residues of the SH3 gene were replaced with various kinds of randomized codons. We utilized three limited sets of amino acids: (1) the set encoded by the lower half of the genetic code (RNN), which contains mainly putative primitive amino acids (e.g., Gly and Ala); (2) the set encoded by the upper half of the genetic code (YNN, where Y = T or C), shown in Figure 1, which contains many putative new amino acids (e.g., Cys, Phe, Tyr, and Trp); and (3) the set encoded using all bases (NNN), which contains all 20 kinds of amino acids, used as a control. Functional SH3 sequences that can bind to the SH3 ligand peptide were selected from each library using mRNA display. After three rounds of *in vitro* selection, the contents of active SH3 domains in each round were analyzed using an enzyme-linked immunosorbent assay

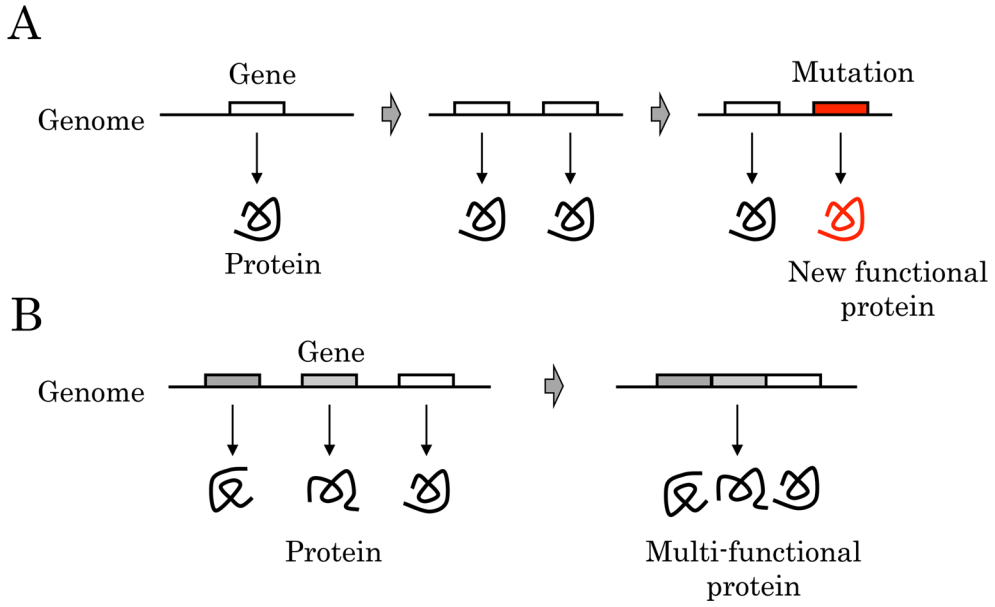


Figure 3. Evolution by gene copying and shuffling. Scheme A illustrates gene duplication.[10] Because the original function of the encoded protein may be lost in response to gene mutation, it is unlikely that a protein with a new function will be successfully created by mutation when only a single gene encoding a protein with an important function for survival exists in the genome. However, when two or more copies of the same gene are present, proteins with a new function derived from accumulated mutation of one copy can arise, while the original function is maintained by the remaining intact gene (right). Scheme B illustrates gene shuffling.[11] A gene encoding a protein with two or more functions may be created when plural genes or gene fragments with different functions are connected to each other by reorganization of the genome.

(ELISA). 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. These results appear to be the first experimental evidence supporting the idea that primitive proteins were indeed composed of those amino acids considered to be primitive.

Moreover, the sequence space containing many primitive amino acids includes more SH3 ligand binding activities than that containing all 20 kinds of amino acids, as indicated by the fact that the ligand-binding sequences were condensed slightly earlier from the former library containing many primitive amino acids than from the library containing all 20 amino acids. Furthermore, two of eight proteins selected from the library containing many primitive amino acids were expressed in the soluble fraction of *E. coli*, while seven of seven proteins selected from the library containing all 20 amino acids were expressed in the insoluble fraction in *E. coli*. These results are consistent with the experimental results obtained with the random-sequence proteins described above. Thus, it is possible that a set containing many primitive amino acids not only increases the appearance frequency of functional proteins but also increases the solubility of proteins. Because the proteins selected from the library containing many primitive amino acids showed much the same ligand affinity, specificity, and thermostability as those selected from the library containing all 20 amino acids, it seems that the former proteins are not necessarily inferior to the latter in functional and structural properties.[49] These experimental results are consistent with the hypothesis that proteins at an early stage of biological evolution consisted of primitive amino acids.

If this is the case, and functional and structural proteins were originally formed from only primitive amino acids, why did the number of amino acids increase to the present value of 20? The Cys residue, considered to have been a late addition to the repertoire, can improve the structural stability of proteins by forming intramolecular disulfide bonds. The His residue, also considered to be a late addition, has an imidazole group, which can combine with a metal ion and thereby serve as an active center of enzymes. Thus, it can be considered that introduction of such amino acids into the set of primitive amino acids would have assisted in the stabilization of protein structure and the acquisition of new functionality.

In future work, we intend to compare the frequencies with which functional proteins occur in proteins other than the SH3 domain from the three libraries, as well as in completely random-sequence proteins, using the various sets of amino acids described above.[49] Moreover, although only 12 kinds of amino acids encoded by the codon RNN were used as primitive amino acids in this research, it might be possible to identify the order in which amino acids were introduced into the genetic code by performing similar experiments with sets of various combinations of amino acids.

"HUNDREDS OF PLAGIARISMS": EVOLUTION OF PROTEINS BY COPYING AND SHUFFLING GENES

It is a scenario of Ohno's gene duplication theory[10] that a new functional gene can easily be created by copying a gene encoding the amino acid sequence of an existing protein with a stable structure and introducing mutations that alter the active site of the copied protein (Figure 3A). Recent sequencing of the whole genomes of various living organisms has led to the identification of many genes that are considered to have been created by gene duplication.[70] Moreover, in protein engineering, there have been many reports of examples in which the substrate specificity of enzymes was changed by the replacement of amino acids at the active site,[71] or a totally new active site was created on an unrelated protein.[72] Ohno called this "one creation followed by hundreds of plagiarisms"; in other words, after the high barrier of acquiring stable structure has been cleared, it is easy to create various functional

dx.doi.org/10.1021/bi301704x | *Biochemistry* 2013, 52, 3841−3851

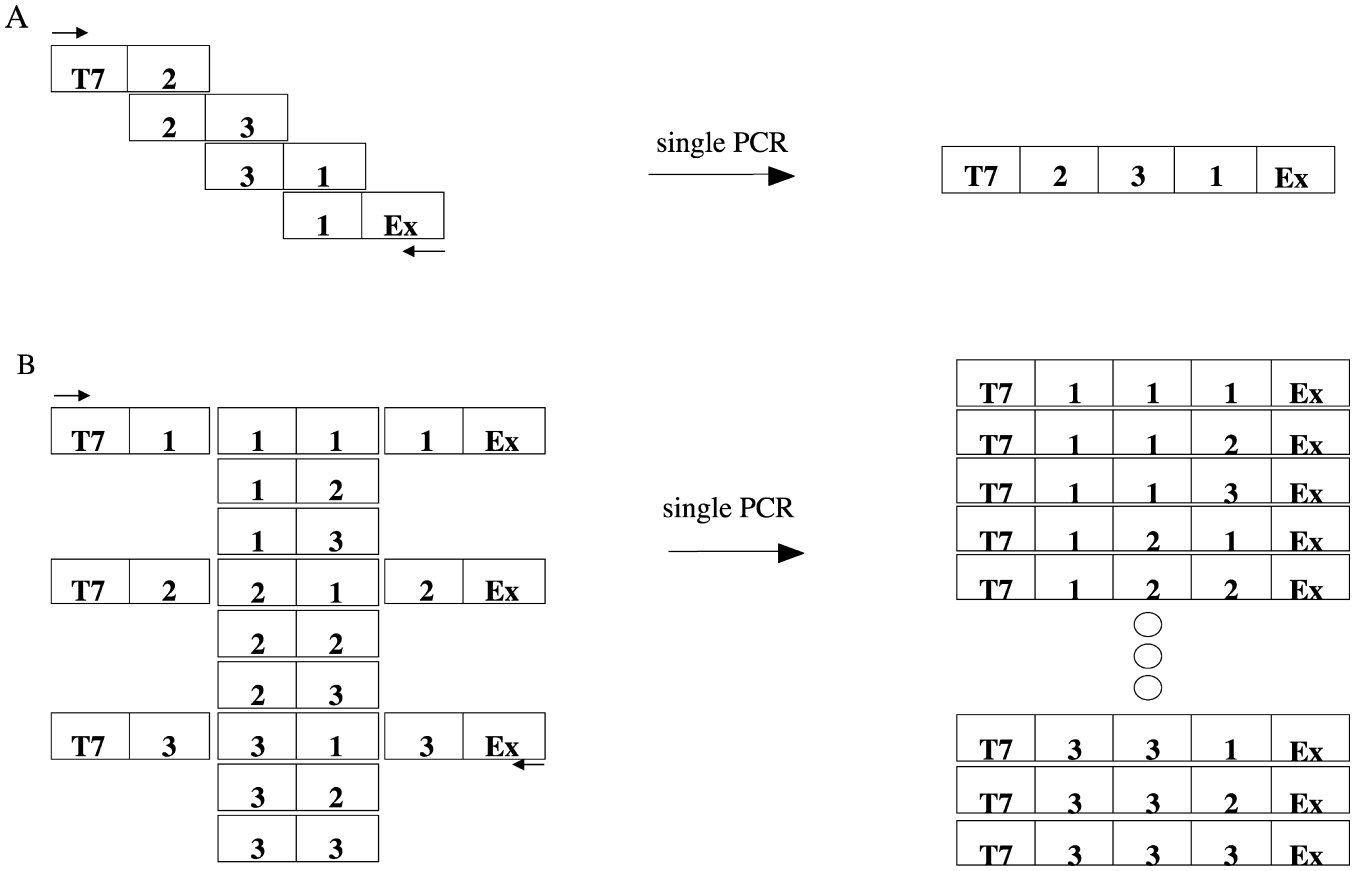


Figure 4. Schematic diagrams of multirecombinant PCR (A) and random multirecombinant PCR (RM-PCR) (B). In RM-PCR, different sequences consisting of several building blocks arranged in different orders may be synthesized in a single PCR. T7 and Ex are 5′ and 3′ consensus sequences where primers anneal to prime the extension.[48]

proteins by using the same framework. Moreover, new genes encoding multifunctional proteins with two or more functions can be dynamically created not only by introducing point mutations into the copy gene but also by combining the copy genes (Figure 3B).

Gilbert put forward the exon shuffling theory, that exon shuffling is promoted by introns acting as a "paste", leading to genes encoding new proteins.[11] It was shown that exon shuffling was effective in creating multidomain proteins with repeated linking of two or more domains, such as the cell adhesion molecules required for the evolution of multicell eukaryotes. Exon shuffling may have been effective not only for the creation of multidomain proteins by combination of domains but also for "one creation" of genes encoding domains, because the region encoded by an exon corresponds to a peptide sequence that is smaller than a unit of structure formation. However, introns are present universally in eukaryotes but are less often found in prokaryotes, and two theories have been proposed to explain this: one is the "introns early" theory that introns existed in the common ancestors and contributed to creation of the first genes but were lost later in prokaryotes,[73] and the other is the "introns late" theory that introns were inserted into the genome of eukaryotes afterward as parasitic genes and contributed only to domain shuffling.[74]

Comparison of the genomes of today's living organisms cannot easily distinguish the two possibilities. However, a constitutive approach can be used to examine whether it is possible to create a new protein by combining short peptide fragments. We have developed an RM (random multi-recombinant)-PCR method[50,51] capable of constructing easily connected genes by combining genes corresponding to peptide fragments with different structural motifs (Figure 4). We applied it to the construction of artificial proteins and analyzed their structure forming capability. When blocks including secondary structures such as α-helix or β-sheet were combined as "structural motifs", greater structure forming capability and enzymatic activity were found.[52,75,76]

Recently, several methods have been developed to combine multiple DNA fragments without homologous sequences. Structure-based combinatorial protein engineering (SCOPE) employs chimeric primers to combine structural elements from different proteins that have similar folds but have a low level of sequence identity.[77] Diverse linker sequences can be used to connect the structural elements and can be selected as required in this method. Hiraga and Arnold developed the sequence-independent site-directed chimeragenesis (SISDC) method using a type IIb restriction enzyme, which also allows for combining distantly related or unrelated proteins at multiple discrete sites, although one or two amino acids at each crossover position must be fixed.[78] Nonhomologous random recombination (NRR) is a method for creating combinatorial libraries by block shuffling based on random digestion and blunt-end ligation of parent genes in the presence of appropriate amounts of hairpin sequences.[79] This method appears to be simple, but blunt-end ligation yields DNA fragments containing misoriented structural elements. To minimize the impact of this problem, the constructed DNA fragments were fused to chloramphenicol acetyltransferase, and

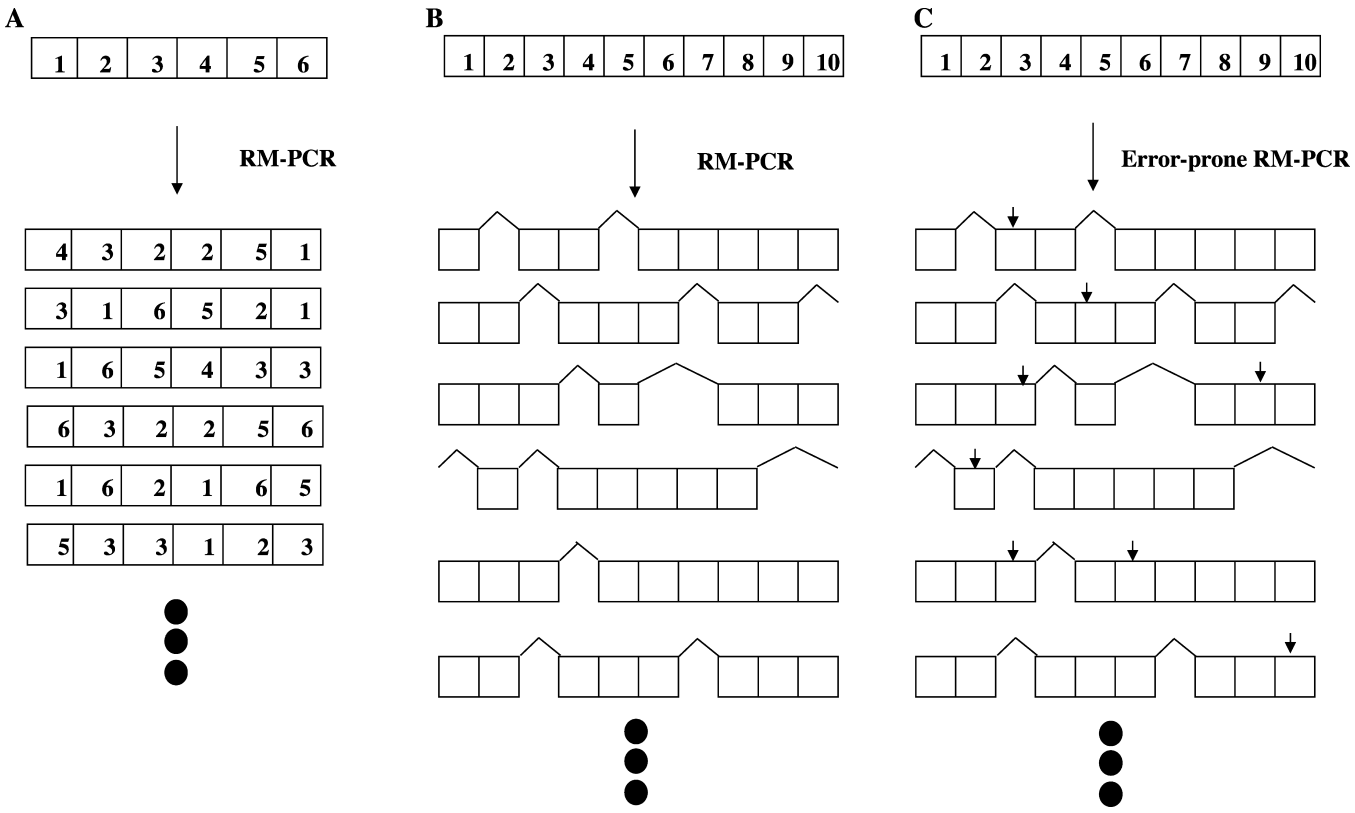


Figure 5. (A) Construction of a random shuffling library by RM-PCR. The six building blocks are shuffled and combined to yield many different structural genes. The random shuffling library should contain different block sequences whose every position has an equal probability of encoding any of the building blocks. In principle, this is attained by RM-PCR using equal amounts of dimer templates prepared by ligating two building blocks in all possible combinations. (B) Construction of an alternative splicing library by RM-PCR. Ten building blocks from a gene are spliced alternatively at the DNA level to yield different structural genes. RM-PCR with reaction mixtures containing dimer templates in appropriate ratios and amounts can also be used to create alternative splicing libraries. The ratios are determined theoretically by the frequencies of dimer templates in the desired library.[49] (C) Alternative splicing library created by RM-PCR under error-prone conditions where the created block sequences have different point mutations (indicated by arrows on the block sequences).

preselection was performed in the presence of chloramphenicol to obtain only structural genes with a correct open reading frame. Y-Ligation-based block shuffling (YLBS) can also be used to combine different DNA blocks.[80] Arnold et al. developed a computational algorithm to identify the fragments of proteins that can be recombined without disturbing the integrity of the three-dimensional structure.[81] Moreover, Shiba et al. created genes that repeated a "functional motif" and produced artificial proteins inducing apoptosis of cancer cells[82] or controlling growth of an inorganic crystal.[83]

There is another advantage to the exon−intron structure. Alternative splicing is present in ≥60% of human genes and is considered to be effective in producing various proteins from a small number of genes. Although the partial sequences that are inserted and deleted by alternative splicing are constructed with domain units having intrinsic structure forming ability, analysis of alternative splicing of many genes identified by genome sequencing showed that many more insertions and deletions of partial short sequences than would be expected are found in the domains.[84] Therefore, alternative splicing, which operates through simple point mutation of a splicing control domain without any large-scale conversion process (such as reorganization of the genome, e.g., exon shuffling), is an effective process for constructing new domain structures by insertion and deletion.[84]

The efficiency of splicing newly introduced by point mutation is not 100%, because splicing to produce the original gene also occurs simultaneously, so that loss of original gene function does not occur. Therefore, it is possible to try new sequences and still maintain the original gene function, even in the absence of gene duplication to generate a copy of the original gene. On the basis of this concept of alternative splicing, we divided the gene encoding human estrogen receptor α ligand binding domain (hERαLBD) into 10 blocks at the boundaries between secondary structures or exons and constructed an artificial alternative splicing DNA library by means of RM-PCR as described above (Figure 5). We then examined whether totally new functional proteins could be obtained by using our mRNA display method[18,19,31−33] and GTP affinity chromatography combined with quantitative real-time PCR. After three rounds of screening, we obtained three clones capable of binding specifically to GTP, whose structure is completely different from that of the original ligand, estradiol.[85] Block-shuffled mutants binding to a well-known immune suppressor, a macrolide compound FK506, whose structure is also completely different from that of the original ligand, estradiol, were also screened from an identical artificial alternative splicing library by mRNA display. These block-shuffled mutants binding to GTP and FK506 were stabilized as oligomers. Recently, it was confirmed by means of X-ray crystallographic analysis that a block-shuffled mutant of barnase

dx.doi.org/10.1021/bi301704x | *Biochemistry* 2013, 52, 3841−3851

forms a three-dimensional (3D) domain-swapped dimer (personal communication from Prof. Katayanagi). Thus, the block-shuffled mutants binding to GTP and FK506 may form 3D domain-swapped oligomers.[86] These results may imply that alternative splicing has contributed substantially to the diversification of protein function during evolution, and that domain swapping serves as a mechanism for evolution of some oligomeric proteins. Further, dividing natural proteins into structural subunits and recombining them may be a more efficient means of creating totally new functional proteins, as compared with searching completely random sequences.[63]

CHALLENGES AND PROSPECTS FOR THE FUTURE

As shown in Table I, by using a cDNA library of natural proteins, mRNA display can be applied also to functional analysis of natural proteins binding to a target molecule (bait) fixed on the beads. It has already been shown that protein complexes,[31−34] transcription factors,[35] drug-target proteins,[38−40] signal peptides,[41−44] and antibodies[45−47] can be screened from a cDNA library by using protein, DNA, drug, or antigen as bait.

Table I. Summary of Functional Screening of Various Proteins by mRNA Display

DNA library	interaction (bait/prey)	output example
artificial DNA library (random sequence and block shuffling)	peptide/protein	SH domains with limited sets of amino acids (GNN, RNN, and NNN)[49]
	protein/peptide	NLS,[a,41] Bcl-X_L, and MDM2-binding peptides[43,44]
	compound/protein	ATP and GTP-binding proteins[63,85]
cDNA library	protein/protein	Jun interactors,[31] Fos interactors,[32] transcription factor interactors,[33] PKM2[34]
	DNA/protein	transcription factors[35]
	antigen/antibody	anti-fluorescein scFv,[b,45] anti-p53 and MDM2 scFvs,[46] anti-fluorescein Fab[c,47]
	drug/protein	FKBP12,[d,38] NPM,[e,39] hCAP-G2[f,40]
	protein/RNA	*Dcx*[g,36]

[a]NLS, nuclear localization signal. [b]scFv, single-chain fragment variable. [c]Fab, fragment antigen binding. [d]FKBP12, FK506-binding protein 12 kDa. [e]NPM, nucleophosmin. [f]hCAP-G2, a subunit of condensing II. [g]*dcx*, doublecortin mRNA.

mRNA display with the IVV method can handle a larger number of molecules (approximately 10^{12}−10^{13}) than other cell-based display techniques, such as phage display. This allows the enrichment of active sequences with low abundance from libraries with high diversity and complexity. Furthermore, mRNA display can be used to screen for functional proteins with increased stability and solubility. As compared with other display methods, the mRNA−protein conjugates used in the IVV method are relatively resistant to denaturants, such as guanidium salt and urea, and heat. Thus, to evolve thermostable proteins (enzymes), IVV screening is conducted under conditions of increasing selection pressure, i.e., gradually increasing concentration of urea or guanidinium salt, or gradually increasing temperature. Soluble proteins can be obtained spontaneously, because insoluble and aggregated IVVs are removed in the process of IVV screening. We have applied off-rate selection as a selection pressure for affinity maturation of proteins with binding activity, such as the scFv antibody. For off-rate selection, mRNA-displayed scFvs that bound to antigen-immobilized beads were washed with a large excess of free antigen to prevent rebinding of antibodies to the beads. Because the off rate (k_{off}) depends on the half-life ($\tau_{1/2}$) of the antigen−antibody complex, higher-affinity binders with a low k_{off} can remain on the beads for a longer washing time. We set the washing time periods of off-rate selection to decrease the k_{off} about several-fold per round (3, 20, 96, and 336 h). Although mRNA is considered to be a labile molecule, it is stable for at least 2 weeks at low temperatures. After four rounds of off-rate selection, the total binding activity of *in vitro*-translated products of the library DNA at each round was analyzed by a competitive ELISA. The binding activity gradually increased in successive rounds of selection, whereas little or no activity was observed in the presence of a competitor, indicating enrichment of specific binders with high affinity in the library. Even materials such as glass, metal, ceramic, and plastic can be used as bait for mRNA display, so it is possible to create novel proteins that bind specifically and strongly to these materials by application of mRNA display. For screening of protein−RNA interaction, the mRNA−protein fusion library has to be screened after conversion of the RNA moieties into RNA−DNA hybrids, because an RNA molecule used as bait competes with the RNA moieties of the IVV library. On the other hand, mRNA display can be applied to screen regulatory RNA sequences binding to RNA-binding protein, because the IVV consists of RNA and protein.[36,37]

We further developed a large-scale, high-throughput IVV screening system by using a combination of a biorobot, microfluidic tip, and tiling array. Large-scale data sets (∼1000 interactors) of protein−protein interaction of 50 transcription factors were obtained using the combination of mRNA display and the biorobot. We also combined mRNA display with a microfluidic system for *in vitro* selection and evolution of antibodies and achieved an ultrahigh enrichment efficiency of 10^6-fold per round. Although mRNA display is a powerful screening tool for protein interaction analysis, the final cloning and sequencing processes represent a bottleneck, resulting in many false negatives. However, we have utilized a tiling array[87] and next-generation sequencing technology[88] to identify specifically binding proteins selected with mRNA display technology. This approach allowed us to detect minute amounts of many binding proteins at an ultrahigh sensitivity of $>10^6$-fold compared to that of an existing capillary sequencing method.

Directed protein evolution in the laboratory has the potential to test evolutionary theories and to reproduce evolutionary scenarios.[89] For example, duplication and circular permutation[90] and homologous recombination[91] have already been examined in the laboratory. In this review, I have shown that soluble, functional proteins tend to occur more frequently in libraries based on limited sets of primitive amino acids than in a library based on the full set of 20 amino acids, and that artificial alternative splicing (block shuffling) is likely to have contributed efficiently to the evolution of new protein domains. Thus, the evolutionary engineering of proteins using limited sets of primitive amino acids or artificial alternative splicing libraries may be an effective tool for the creation of artificial functional proteins that could find application in the pharmaceutical industry. It is also possible to combine the random-sequence library and block shuffling library approaches by using mRNA display; at first, hydrophobic or hydrophilic

fragments are screened and catalogued from a random-sequence library, and then soluble and foldable proteins are screened after shuffling. In the future, I think that a promising strategy would be to select functional proteins for an identical target from three types of libraries based on random sequences with limited sets of amino acids, randomly mutated folded proteins, and block-shuffled sequence proteins by using mRNA display and to compare their structure forming ability and the frequency of appearance of artificial functional proteins.

AUTHOR INFORMATION

Corresponding Author

*Telephone: 81 + 45-500-1731. E-mail: hyan@bio.keio.ac.jp.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

I thank Drs. Nobuhide Doi and Junko Tanaka for reading the manuscript prior to publication and for their valuable comments. I also thank the members of our laboratory, Dr. Toru Tsuji, Michiko Onimaru, Dr. Noriko Tabata, Yuko Sakuma-Yonemura, and Dr. Junko Tanaka, for their hard work and support.

REFERENCES

(1) Miller, S. L. (1953) A production of amino acids under possible primitive earth conditions. *Science 117*, 528−529.

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

(3) Yanagawa, H., Kojima, K., Ito, M., and Handa, N. (1990) Synthesis of polypeptides by microwave heating: I. Formation of polypeptides during repeated hydration-dehydration cycles and their characterization. *J. Mol. Evol. 31*, 180−186.

(4) Yanagawa, H., Ogawa, Y., Kojima, K., and Ito, M. (1988) Construction of protocellular structures under simulated primitive earth conditions. *Origins Life Evol. Biospheres 18*, 179−207.

(5) Yanagawa, H., and Kobayashi, K. (1992) An experimental approach to chemical evolution in submarine hydrothermal systems. *Origins Life Evol. Biospheres 22*, 147−159.

(6) Altman, S. (1981) Transfer RNA processing enzymes. *Cell 23*, 3−4.

(7) Cech, T. R., Zaug, A. J., and 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 27*, 487−496.

(8) Trifonov, E. N. (2004) The triplet code from first principles. *J. Biomol. Struct. Dyn. 22*, 1−11.

(9) Eigen, M., and Schuster, P. (1978) The hypercycle. A principle of natural self-organization. Part C: The realistic hypercycle. *Naturwissenschaften 65*, 341−369.

(10) Ohno, S. (1970) *Evolution of gene duplication*, Springer-Verlag, New York.

(11) Gilbert, W. (1987) The exon theory of genes. *Cold Spring Harbor Symp. Quant. Biol. 52*, 901−905.

(12) Zhang, J. (2003) Evolution by gene duplication: An update. *Trends Ecol. Evol. 18*, 292−298.

(13) Bershtein, S., and Tawfik, D. S. (2008) Ohno's model revisited: Measuring the frequency of potentially adaptive mutations under various mutation drifts. *Mol. Biol. Evol. 25*, 2311−2318.

(14) Bloom, J. D., Meyer, M. M., Meinhold, P., Otey, C. R., MacMillan, D., and Arnold, F. H. (2005) Evolving strategies for enzyme engineering. *Curr. Opin. Struct. Biol. 15*, 447−452.

(15) Doi, N., and Yanagawa, H. (2001) Genotype-phenotype linkage for directed evolution and screening of combinatorial protein libraries. *Comb. Chem. High Throughput Screening 4*, 497−509.

(16) Matsumura, N., Doi, N., and Yanagawa, H. (2006) Recent progress and future prospects in protein display technologies as tools for proteomics. *Curr. Proteomics 3*, 199−215.

(17) Hanes, J., and Plückthun, A. (1997) *In vitro* selection and evolution of functional proteins by using ribosome display. *Proc. Natl. Acad. Sci. U.S.A. 94*, 4937−4942.

(18) Nemoto, N., Miyamoto-Sato, E., Husimi, Y., and 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.

(19) Miyamoto-Sato, E., Takashima, H., Fuse, S., Sue, K., Ishizaka, M., Tateyama, S., Horisawa, K., Sawasaki, T., Endo, Y., and Yanagawa, H. (2003) Highly stable and efficient mRNA templates for mRNA-protein fusions and C-terminally labeled proteins. *Nucleic Acids Res. 31*, e78.

(20) Roberts, R. W., and Szostak, J. W. (1997) RNA-peptide fusions for the *in vitro* selection of peptides and proteins. *Proc. Natl. Acad. Sci. U.S.A. 94*, 12297−12302.

(21) Doi, N., and Yanagawa, H. (1999) STABLE: Protein-DNA fusion system for screening of combinatorial protein libraries *in vitro*. *FEBS Lett. 457*, 227−230.

(22) Yarmolinsky, M. B., and Haba, G. L. (1959) Inhibition by puromycin of amino acid incorporation into protein. *Proc. Natl. Acad. Sci. U.S.A. 45*, 1721−1729.

(23) Allen, D. W., and Zamecnik, P. C. (1962) The effect of puromycin on rabbit reticulocyte ribosomes. *Biochim. Biophys. Acta 55*, 865−874.

(24) Nathans, D. (1964) Puromycin inhibition of protein synthesis: Incorporation of puromycin into peptide chains. *Proc. Natl. Acad. Sci. U.S.A. 55*, 585−592.

(25) Traut, R. R., and Monro, R. E. (1964) The puromycin reaction and its relation to protein synthesis. *J. Mol. Biol. 10*, 63−72.

(26) Zamir, A., Leder, P., and Elson, D. (1966) A ribosome-catalyzed reaction between N-formylmethionyl-tRNA and puromycin. *Proc. Natl. Acad. Sci. U.S.A. 56*, 1794−1801.

(27) Nathans, D., and Neidle, A. (1963) Structural requirements for puromycin inhibition of protein synthesis. *Nature 197*, 1076−1077.

(28) Steiner, G., Kuechler, E., and Barta, A. (1988) Photo-affinity labelling at the peptidyl transferase centre reveals two different positions for the A- and P-sites in domain V of 23S rRNA. *EMBO J. 7*, 3949−3955.

(29) Kirillov, S., Porse, B. T., Vester, B., Woolley, P., and Garrett, R. A. (1997) Movement of the 3′-end of tRNA through the peptidyl transferase centre and its inhibition by antibiotics. *FEBS Lett. 406*, 223−233.

(30) Miyamoto-Sato, E., Nemoto, N., Kobayashi, K., and Yanagawa, H. (2000) Specific bonding of puromycin to full-length protein at the C-terminus. *Nucleic Acids Res. 28*, 1176−1182.

(31) Horisawa, K., Tateyama, S., Ishizaka, M., Matsumura, N., Takashima, H., Miyamoto-Sato, E., Doi, N., and Yanagawa, H. (2004) *In vitro* selection of Jun-associated proteins using mRNA display. *Nucleic Acids Res. 32*, e169.

(32) Miyamoto-Sato, E., Ishizaka, M., Horisawa, K., Tateyama, S., Takashima, H., Fuse, S., Sue, K., Hirai, N., Masuoka, K., and Yanagawa, H. (2005) Cell-free co-translation and selection using *in vitro* virus for high-throughput analysis of protein-protein interactions and complexes. *Genome Res. 15*, 710−717.

(33) Miyamoto-Sato, E., Ishizaka, M., Fujimori, S., Hirai, N., Masuoka, K., Saito, R., Ozawa, Y., Hino, K., Washio, T., Tomita, M., Yamashita, T., Oshikubo, T., Akasaka, H., Sugiyama, J., Matsumoto, Y., and Yanagawa, H. (2010) A comprehensive resource of interacting protein regions for refining human transcription factor networks: Domain-based interactome. *PLoS One 5*, e9289.

(34) Tamada, M., Nagano, O., Tateyama, S., Ohmura, M., Yae, T., Ishimoto, T., Sugihara, E., Onishi, N., Yamamoto, T., Yanagawa, H., Suematsu, M., and Saya, H. (2012) Modulation of glucose metabolism by CD44 contributes to antioxidant status and drug resistance in cancer cells. *Cancer Res. 72*, 1438−1448.

(35) Tateyama, S., Horisawa, K., Takashima, H., Miyamoto-Sato, E., Doi, N., and Yanagawa, H. (2006) Affinity selection of DNA-binding protein complexes using mRNA display. *Nucleic Acids Res. 34*, e27.
(36) Horisawa, K., Imai, T., Okano, H., and Yanagawa, H. (2009) 3′-Untranslated region of doublecortin mRNA is a binding target of the Musashi1 RNA-binding protein. *FEBS Lett. 583*, 2429−2434.
(37) Horisawa, K., Imai, T., Okano, H., and Yanagawa, H. (2010) The Musashi family RNA-binding proteins in stem cells. *Biomolecular Concepts 1*, 59−66.
(38) Doi, N., Takashima, H., Wada, A., Oishi, Y., Nagano, T., and Yanagawa, H. (2007) Photocleavable linkage between genotype and phenotype for rapid and efficient recovery of nucleic acids encoding affinity-selected proteins. *J. Biotechnol. 131*, 231−239.
(39) Shiheido, H., Terada, F., Tabata, N., Hayakawa, I., Matsumura, N., Takashima, H., Ogawa, Y., Du, W., Yamada, T., Shoji, M., Sugai, T., Doi, N., Iijima, S., Hattori, Y., and Yanagawa, H. (2012) A phthalimide derivative that inhibits centrosomal clustering is effective on multiple myeloma. *PLoS One 7*, e38878.
(40) Shiheido, H., Naito, Y., Kimura, H., Genma, H., Takashima, H., Tokunaga, M., Ono, T., Hirano, T., Du, W., Yamada, T., Doi, N., Iijima, S., Hattori, Y., and Yanagawa, H. (2012) An anilinoquinazoline derivative inhibits tumor growth through interaction with hCAP-G2, a subunit of condensin II. *PLoS One 7*, e44889.
(41) Kosugi, S., Hasebe, M., Matsumura, N., Takashima, H., Miyamoto-Sato, E., Tomita, M., and Yanagawa, H. (2009) Six classes of nuclear localization signals specific to different binding grooves of importin*α*. *J. Biol. Chem. 284*, 478−485.
(42) Kosugi, S., Hasebe, M., Tomita, M., and Yanagawa, H. (2009) Systematic identification of cell cycle-dependent nucleocytoplasmic shuttling proteins by prediction of composite motifs. *Proc. Natl. Acad. Sci. U.S.A. 106*, 10171−10176.
(43) Matsumura, N., Tsuji, T., Sumida, T., Kokubo, M., Onimaru, M., Doi, N., Takashima, H., Miyamoto-Sato, E., and Yanagawa, H. (2010) mRNA display selection of a high-affinity, Bcl-X_L-specific binding peptide. *FASEB J. 24*, 2201−2210.
(44) Shiheido, H., Takashima, H., Doi, N., and Yanagawa, H. (2011) mRNA display selection of an optimized MDM2-binding peptide that potently inhibits MDM2-p53 interaction. *PLoS One 6*, e17898.
(45) Fukuda, I., Kojoh, K., Tabata, N., Doi, N., Takashima, H., Miyamoto-Sato, E., and Yanagawa, H. (2006) *In vitro* evolution of single-chain antibodies using mRNA display. *Nucleic Acids Res. 34*, e127.
(46) Tabata, N., Sakuma, Y., Honda, Y., Doi, N., Takashima, H., Miyamoto-Sato, E., and Yanagawa, H. (2009) Rapid antibody selection by mRNA display on a microfluidic chip. *Nucleic Acids Res. 37*, e64.
(47) Sumida, T., Yanagawa, H., and Doi, N. (2012) *In vitro* selection of Fab fragments by mRNA display and gene-linking emulsion PCR. *J. Nucleic Acids 2012*, 371379.
(48) Tanaka, J., Doi, N., Takashima, H., and Yanagawa, H. (2010) Comparative characterization of random-sequence proteins consisting of 5, 12, and 20 kinds of amino acids. *Protein Sci. 19*, 786−795.
(49) Tanaka, J., Yanagawa, H., and Doi, N. (2011) Comparison of the frequency of functional SH3 domains with different limited sets of amino acids using mRNA display. *PLoS One 6*, e18034.
(50) Tsuji, T., Onimaru, M., and Yanagawa, H. (2001) Construction of combinatorial protein libraries by random multi-recombinant PCR. *Nucleic Acids Res. 29*, e97.
(51) Tsuji, T., Onimaru, M., Kitagawa, M., Kojoh, K., Tabata, N., and Yanagawa, H. (2004) Random multirecombinant polymerase chain reaction. *Methods Enzymol. 388*, 61−75.
(52) Tsuji, T., Onimaru, M., and Yanagawa, H. (2006) Towards the creation of novel proteins by block shuffling. *Comb. Chem. High Throughput Screening 9*, 259−269.
(53) Mandecki, W. (1998) The game of chess and searches in protein sequence space. *Trends Biotechnol. 16*, 200−202.
(54) Dryden, D. T. F., Thomson, A. R., and White, J. H. (2008) How much of protein sequence space has been explored by life on Earth? *J. R. Soc., Interface 5*, 953−956.
(55) Ellington, A. D., and Szostak, J. W. (1990) *In vitro* selection of RNA molecules that bind specific ligands. *Nature 346*, 818−822.
(56) Tuerk, C., and Gold, L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science 249*, 505−510.
(57) Fukuda, K., Vishnuvardhan, D., Sekiya, S., Hwang, J., Kakiuchi, N., Taira, K., Shimotohno, K., Kumar, P. K. R., and Nishikawa, S. (2000) Isolation and characterization of RNA aptamers specific for the hepatitis C virus nonstructural protein 3 protease. *Eur. J. Biochem. 267*, 3685−3694.
(58) Khati, M., Schuman, M., Ibrahim, J., Sattentau, Q., Gordon, S., and James, W. (2003) Neutralization of infectivity of diverse R5 clinical isolates of human immunodeficiency virus type 1 by gp120-binding 2′F-RNA aptamers. *J. Virol. 77*, 12692−12698.
(59) Rhie, A., Kirby, L., Sayer, N., Wellesley, R., Disteret, P., Sylvester, I., Gill, A., Hope, J., James, W., and Tahiri-Alaoui, A. (2003) Characterization of 2′-fluoro-RNA aptamers that bind preferentially to disease-associated conformations of prion protein and inhibit conversion. *J. Biol. Chem. 278*, 39697−39705.
(60) Mandecki, W. (1990) A method for construction of long randomized open reading frames and polypeptides. *Protein Eng. 3*, 221−226.
(61) Prijambada, I. D., Yomo, T., Tanaka, F., Kawama, T., Yamamoto, K., Hasegawa, A., Shima, Y., Negoro, S., and Urabe, I. (1996) Solubility of artificial proteins with random sequences. *FEBS Lett. 382*, 21−25.
(62) Watters, A. L., and Baker, D. (2004) Searching for folded proteins *in vitro* and *in silico*. *Eur. J. Biochem. 271*, 1615−1622.
(63) Keefe, A. D., and Szostak, J. W. (2001) Functional proteins from a random-sequence library. *Nature 410*, 715−718.
(64) Jordan, I. K., Kondrashov, F. A., Adzhubei, I. A., Wolf, Y. I., Koonin, E. V., Kondrashov, A. S., and Sunyaev, S. (2005) A universal trend of amino acid gain and loss in protein evolution. *Nature 433*, 633−638.
(65) Davidson, A. R., and Sauer, R. T. (1994) Folded proteins occur frequently in libraries of random amino acid sequences. *Proc. Natl. Acad. Sci. U.S.A. 91*, 2146−2150.
(66) Davidson, A. R., Lumb, K. J., and Sauer, R. T. (1995) Cooperatively folded proteins in random sequence libraries. *Nat. Struct. Biol. 2*, 856−864.
(67) Wilkinson, D. L., and Harrison, R. G. (1991) Predicting the solubility of recombinant proteins in *Escherichia coli*. *Biotechnology 9*, 443−448.
(68) Kyte, J., and Doolittle, R. F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol. 157*, 105−132.
(69) Doi, N., Kakukawa, K., Oishi, Y., and Yanagawa, H. (2005) High solubility of random-sequence proteins consisting of five kinds of primitive amino acids. *Protein Eng., Des. Sel. 18*, 279−284.
(70) Lynch, M., and Conery, J. S. (2000) The evolutionary fate and consequences of duplicate genes. *Science 290*, 1151−1155.
(71) Leatherbarrow, R. J., and Fersht, A. R. (1986) Protein engineering. *Protein Eng. 1*, 7−16.
(72) Park, H. S., Nam, S. H., Lee, J. K., Yoon, C. N., Mannervik, B., Benkovic, S. J., and Kim, H. S. (2006) Design and evolution of new catalytic activity with an existing protein scaffold. *Science 311*, 535−5381.
(73) Blake, C. C. F. (1978) Do genes-in pieces imply protein in pieces? *Nature 273*, 267.
(74) Cavallier-Smith, T. (1978) Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth of rate, and the solution of the DNA C-value paradox. *J. Cell Sci. 34*, 247−278.
(75) Tsuji, T., Yoshida, K., Satoh, A., Kohno, T., Kobayashi, K., and Yanagawa, H. (1999) Foldability of barnase mutants obtained by permutation of modules or secondary structure units. *J. Mol. Biol. 286*, 1581−1596.
(76) Tsuji, T., Nagata, T., and Yanagawa, H. (2008) N- and C-termninal fragments of a globular protein constructed by elongation of

modules as a units associated for functional complementation. *J. Biochem. 144*, 513−521.
(77) O'Maille, P. E., Bakhtina, M., and Tsai, M.-D. (2002) Structure-based combination protein engineering (SCOPE). *J. Mol. Biol. 321*, 677−691.
(78) Hiraga, K., and Arnold, F. H. (2003) General method for sequence-independent site-directed chimeragenesis. *J. Mol. Biol. 330*, 287−296.
(79) Bittker, J. A., Le, B. V., Liu, J. M., and Liu, D. R. (2004) Directed evolution of protein enzymes using nonhomologous random recombination. *Proc. Natl. Acad. Sci. U.S.A. 101*, 7011−7016.
(80) Kitamura, K., Kinoshita, Y., Narasaki, S., Nemoto, N., Husimi, Y., and Nishigaki, K. (2002) Construction of block-shuffled libraries of DNA for evolutionary protein engineering: Y-ligation-based block shuffling. *Protein Eng. 15*, 843−853.
(81) Voigt, C. A., Martinez, C., Wang, Z.-G., Mayo, S. L., and Arnold, F. H. (2002) Protein building blocks preserved by recombination. *Nat. Struct. Biol. 9*, 553−558.
(82) Saito, H., Minamisawa, T., Yamori, T., and Shiba, K. (2008) Motif-programmed artificial protein induces apoptosis in several cancer cells by disrupting mitochondria. *Cancer Sci. 99*, 398−406.
(83) Sano, K., Sasaki, H., and Shiba, K. (2005) Specificity and biomineralization activities of Ti-binding peptide-1 (TBP-1). *Langmuir 21*, 3090−3095.
(84) Wen, F., Li, F., Xia, H., Lu, X., Zhang, X., and Li, Y. (2004) The impact of very short alternative splicing on protein structures and functions in the human genome. *Trends Genet. 20*, 232−236.
(85) Tsuji, T., Onimaru, M., Doi, N., Miyamoto-Sato, E., Takashima, H., and Yanagawa, H. (2009) *In vitro* selection of GTP-binding proteins by block shuffling of estrogen-receptor fragments. *Biochem. Biophys. Res. Commun. 390*, 689−693.
(86) Bennett, M. J., Schlunegger, M. P., and Eisenberg, D. (1995) 3D domain swapping: A mechanism for oligomer assembly. *Protein Sci. 4*, 2455−2468.
(87) Horisawa, K., Doi, N., and Yanagawa, H. (2008) Use of cDNA tiling arrays for identifying protein-interactions selected by *in vitro* display technologies. *PLoS One 3*, e1646.
(88) Fujimori, S., Hirai, N., Ohashi, H., Masuoka, K., Nishikimi, A., Fukui, Y., Washio, T., Oshikubo, T., Yamashita, T., and Miyamoto-Sato, E. (2012) Next-generation sequencing coupled with a cell-free display technology for high-throughput production of reliable interactome data. *Sci. Rep. 2*, 691.
(89) Peisajovich, S. G., and Tawfik, D. S. (2007) Protein engineers turned evolutionists. *Nat. Methods 4*, 991−994.
(90) Peisajovich, S. G., Rockah, L., and Tawfik, D. S. (2006) Evolution of new protein topologies through multistep gene rearrangements. *Nat. Genet. 38*, 168−174.
(91) Carbone, M. N., and Arnold, F. H. (2007) Engineering by homologous recombination: Exploring sequence and function within a conserved fold. *Curr. Opin. Struct. Biol. 17*, 454−459.