Towards the Creation of Novel Proteins by Block Shuffling

Toru Tsuji[1,2], Michiko Onimaru[1] and Hiroshi Yanagawa[1,*]

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

[2]*Department of Protein Engineering, Cancer Institute, Japanese Foundation for Cancer Research, Koutou, Tokyo 135-8550, Japan*

Abstract: We have been investigating the creation of novel proteins by means of block shuffling, where the term block refers to an amino acid sequence that corresponds to particular features of proteins, such as secondary structures, modules, functional motifs, and so on. Block shuffling makes it possible to explore the global sequence space, which is not feasible with conventional methods, such as DNA shuffling or family shuffling. To investigate what properties are required for the building blocks, we have analyzed the foldability and enzymatic activity of barnase mutants obtained by permutation of modules or secondary structure units. This reconstructive approach indicated that secondary structure units with mutual long-range interactions are more suitable than modules as building blocks, at least in the case of barnase. The results also suggested that proteins in evolutionarily intermediate states are created by block shuffling, and such proteins have the potential to be evolved into mature globular proteins. For the construction of combinatorial protein libraries, we have developed random multi-recombinant PCR (RM-PCR), which can combine different DNA fragments without homologous sequences. The libraries can be utilized for *in vitro* selection using *in vitro* virus (mRNA display) or stable (DNA display), which have also been developed in our laboratory. In this review article, we summarize our strategy to create novel proteins by block shuffling and review key literature in the field. Possible applications of the block shuffling strategy are also discussed.

1. INTRODUCTION

The questions of how proteins appeared on the primitive earth, and how they evolved into mature functional proteins are fundamental. If we can understand the processes of origin and evolution of proteins, we should be able to create novel functional proteins. We have been investigating the creation of novel proteins by block shuffling, where the term block means an amino acid sequence corresponding to a particular feature of proteins, such as secondary structures, modules, functional motifs, and so on. This strategy is based on the exon theory of genes proposed by Gilbert, who suggested that proteins acquired their functional diversity by combining the building blocks encoded by ancient exons in the early stages of protein evolution [1, 2]. The novel proteins initially produced by block shuffling would have only weak activity, but their functions could be improved by means of directed evolution [3-5].

DNA shuffling or family shuffling using homologous recombination among highly related sequences allows us to search local sequence space extensively [6-8] (Fig. **1A**), and represents a powerful approach to improve the functions of natural proteins and possibly also artificial proteins with moderate activity. On the other hand, block shuffling allows us to search the global sequence space, which can not be achieved by DNA (family) shuffling alone (Fig. **1B**). Natural proteins are only those proteins that have originated from historical events on the earth, and it is therefore very likely that new functional proteins that do not currently exist on earth are present in the global sequence space. Indeed, a completely novel ATP-binding protein was obtained from a random amino acid library [9]. However, the probability that such a functional protein is present in a random library has been estimated as only one per 10^{11} sequences. This probability is too low to make conventional selection experiments feasible. The combination of block shuffling and an optimization process is a possible strategy to explore the global sequence space efficiently. In other words, novel functional proteins with moderate activity obtained by block shuffling can then be subjected to directed evolution using DNA (family) shuffling to create novel functional proteins. It is therefore important to know what properties are required for the building blocks.

2. GO'S MODULES: CANDIDATES FOR THE BUILDING BLOCKS

The discovery of the exon/intron structure of DNA led the hypothesis that combining pre-existing smaller units encoded by ancient exons had created novel proteins [1, 2]. If the peptide segments encoded by the exons corresponded to independent structural and functional units, proteins with foldability and activity would have efficiently emerged by exon shuffling [10]. Go found a correlation of DNA exonic regions with protein structural units in several globular proteins [11, 12]. The unit consisting of a contiguous peptide chain forming a compact region in a globular protein was termed a module [13]. She proposed the module shuffling theory by combining her finding with the exon theory of genes proposed by Gilbert [14].

[*]Address correspondence to this author at the Department of Biosciences and Informatics, Faculty of Science and Technology, Keio University, 3-14-1, Hiyoshi, Kohoku-ku, Yokohama 223-8522, Japan; Tel: (81) (45) 566-1775; Fax: (81) (45) 566-1440; E-mail: hyana@bio.keio.ac.jp

1386-2073/06 $50.00+.00

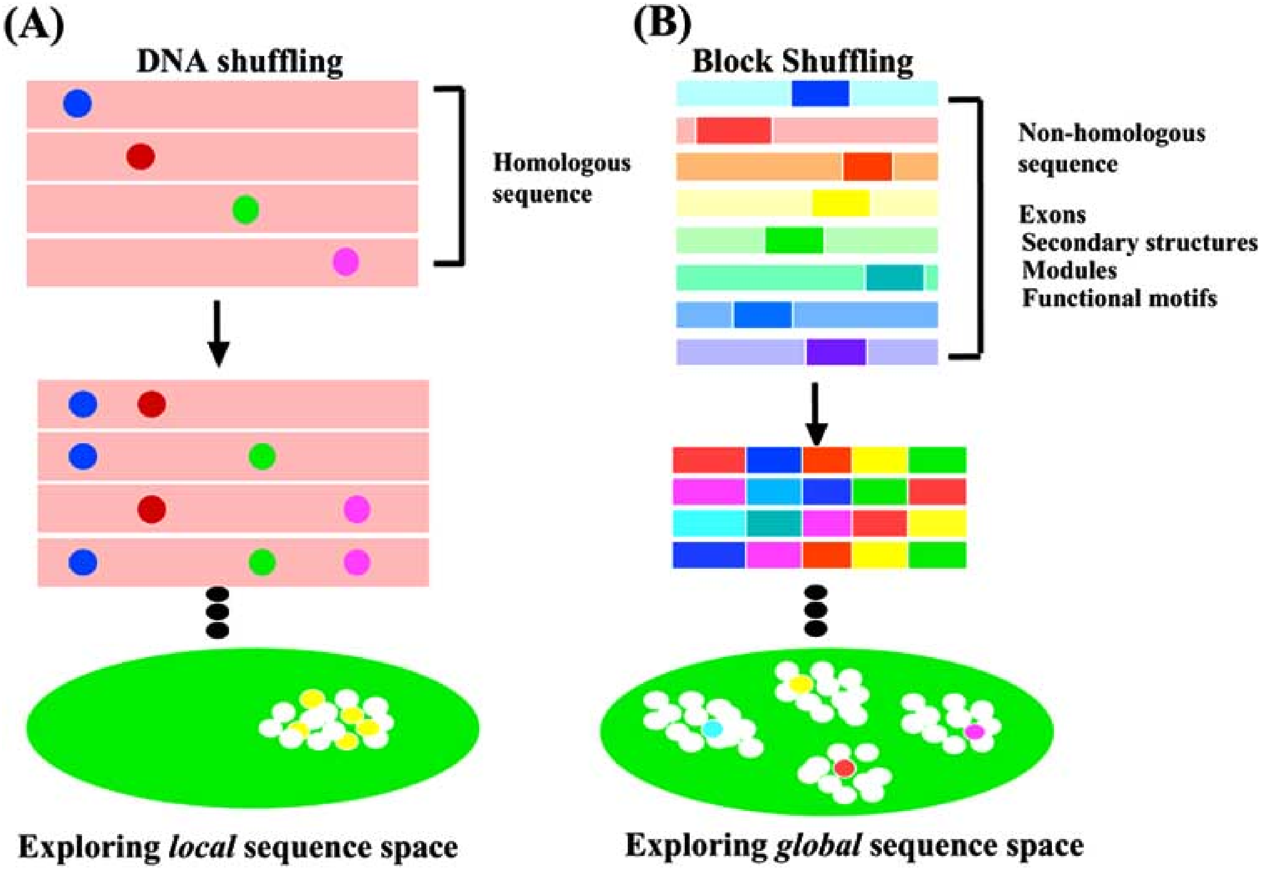

Fig. (1). Schematic diagrams of DNA (family) shuffling and block shuffling. (A) DNA shuffling combines several point mutations in parent sequences based on *in vitro* recombination at the homologous regions. DNA shuffling allows us to search local sequence space around parent sequences extensively [6, 7]. The sequence space explored by family shuffling is slightly larger than that explored by DNA shuffling [8]. (B) Block shuffling combines different building blocks such as secondary structures, modules, functional motifs, and so on. These building blocks would not have homologous sequences, and therefore novel technology that can combine different DNA fragments without homologous sequences is required. Block shuffling combined with DNA shuffling will allow us to search global sequence space thoroughly.

To elucidate biophysical properties of modules, barnase, a ribonuclease whose structural and enzymatic properties are well characterized [15-18], was divided into six modules (M1-M6) based on the centripetal profile [19] (Fig. **2A**), and the biophysical properties of these modules were investigated extensively. Three (M2, M3, and M6) out of the six modules were found to have RNase activity [20]. Local secondary structures of the dissected modules M1, M2, and M3 were observed at similar positions in the modules to those in the intact barnase [21, 22]. Takahashi *et al.* performed a molecular dynamics simulation study, and showed that five (M1, M2, M3, M4, and M5) of the six modules tended to retain native-like conformations [23]. Further, mini-barnase, constructed by deleting M2 from intact barnase, has some conformational properties that are similar to those of wild-type barnase [24, 25]. The six modules appeared to be stabilized mainly by intramodular hydrogen bonds rather than inter-modular interactions, suggesting that modules are independently folding units [19]. Indeed, certain modules in hemoglobin or β-xylanases have been successfully exchanged to create functional chimeric proteins [26-28]. These results support the idea that module shuffling would be useful way to create novel artificial proteins.

3. FOLDABILITY AND ENZYMATIC ACTIVITY OF BARNASE MUTANTS OBTAINED BY PERMUTATION OF MODULES OR SECONDARY STRUCTURE UNITS

3.1. Foldability

Barnase was divided into six extended structures that control the mutual relationships of the modules [29] (Fig. **2B**). The five minima in the centripetal profile were chosen as the boundaries of the six modules, while the five maxima were chosen as the boundaries of the extended structures. Because the extended structure often contains an intact secondary structure, it was termed a "secondary structure unit". On the other hand, module boundaries were often located in the center of β-strands (Fig. **2A**). We constructed 23 module mutants and 23 secondary structure unit mutants by means of permutation of the internal four modules (M2-M5) or secondary structure units (S2-S5), (Fig. **2C**). The structural and functional properties of these mutants were analyzed to gain new insight from the reconstructive approach, as opposed to the anatomical approach described above.

Although most of the 46 mutants did not fold into a stable conformation, two secondary structure unit mutants

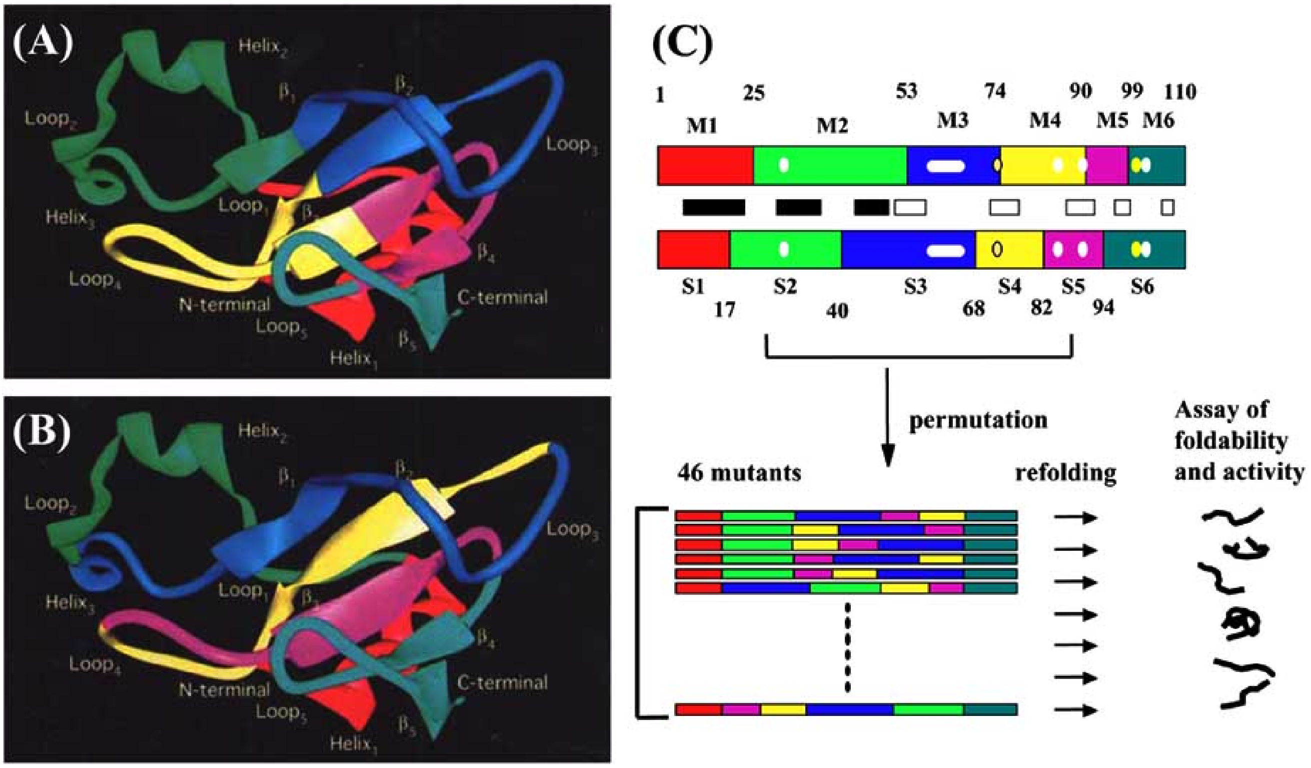

Fig. (2). Construction of barnase mutants obtained by permutation of modules or secondary structure units. (A) Six modules of wild-type barnase are shown in different colors. M1 is red; M2, green, M3, blue; M4, yellow; M5, pink; and M6, sky blue. (B) Six secondary structure units of wild-type barnase are shown in different colors. S1 is red; S2, green, S3, blue; S4, yellow; S5, pink; and S6, sky blue [29]. (C) The internal four blocks were permuted to construct 46 barnase mutants. The general acid and base residues, His-102 and Glu-73, are shown by yellow circles. Amino acids organizing the active site are shown by white circles. Black and white boxes indicate α-helices and β-sheets, respectively [31]. The 46 barnase mutants were expressed as inclusion bodies in *E. coli*, then refolded and characterized conformationally and enzymatically.

(S2543 and S2354H102A) were found to have secondary and tertiary structures [29, 30]. These structures unfolded cooperatively during urea- or thermal-induced unfolding experiments, suggesting that at least a part of their conformation is stabilized by long-range interactions. On the other hand, a locally structured region around a tryptophan residue was found in one of the module mutants (M3245), but the ^{1}H-NMR spectrum of the mutant showed a much smaller dispersion in the amide and methyl regions than those observed in S2543. Thus, none of the module mutants had distinct secondary or tertiary structure. Our reconstructive approach therefore indicates that secondary structure units are more suitable than modules as building blocks to create foldable proteins, at least in the case of barnase.

3.2. Enzymatic Activity

The RNase activity of the 46 barnase mutants was also investigated, and we found that eight secondary structure unit mutants and two module mutants had weak but distinct RNase activity [30, 31]. RNase activity of several of the mutants was further investigated at different pHs, and we found that they each showed a bell-shaped curve with an optimum pH. They also had a unique optimum temperature. Thus, they appear to have basal enzymatic properties, although their optimum temperature was quite low relative to that of natural enzymes.

3.3. Evolutionarily Intermediate States on the Fitness Landscape

Many of the mutants having RNase activity did not show ordered conformations, but their thermal denaturation profiles suggested that these mutants have RNase activities that depend on the stability of the conformation around the putative active sites. The proper conformations around active sites would be stabilized only when the substrate binds to them, namely an induced-folding mechanism. Indeed, a secondary structure unit mutant S4523, which is the most active mutant without ordered secondary and tertiary structures, showed clear changes of its far- and near-UV CD spectra when GMP, which is an inhibitor of the mutant, was added to the protein solution. The binding of GMP to wild-type barnase induces a change of local conformation around the active site, but the overall conformation is not affected [32]. This is due to the stable structure of wild-type barnase. On the other hand, because S4523 does not have such a stable conformation, the overall structure of the mutant would be easily affected by conformational change of the active site. Amino acids that are not involved directly in the activites would have been selected as scaffolds to generate active sites with thermal stability in the early stages of globular protein evolution [31, 33, 34]. If this is the case, the barnase mutants that have an active site, but not a stable scaffold can be considered as globular proteins in evolutionarily intermediate states.

The traditional approach to protein engineering is to improve the catalytic activities of natural proteins already having stable folds by using site-directed mutagenesis, saturation mutagenesis [35], error-prone PCR [36], and/or DNA (family) shuffling [6-8] (Fig. **3A**). On the other hand, protein engineering based on the new concept aims to evolve proteins having a nascent active site without conformational stability, which may be obtained by block shuffling, into fully active and stable proteins by directed evolution (Fig. **3B**).

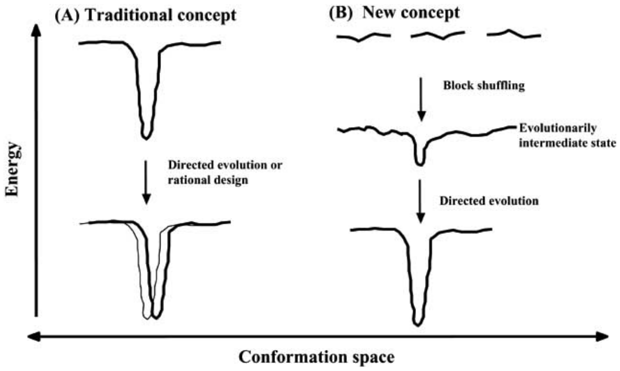

Fig. (3). Comparison of traditional and new concepts of protein engineering. (A) A traditional approach improves or alters the function of proteins having already stable conformation by site-directed mutagenesis, error-protein PCR, saturation mutagenesis, and/or DNA (family) shuffling. This approach changes the local configuration around the active site (It is also possible to construct more stable mutants). (B) A new concept of protein engineering aims to evolve proteins in evolutionarily intermediate states with a fragile structure, which are constructed by block shuffling, into mature functional and stable proteins.

3.4. Towards Describing a Fitness Landscape

The secondary structure unit mutant S2543 had the most ordered conformation of all the mutants, but its activity was the lowest among the active mutants. Active mutants with disordered conformations have potential to evolve further, because they can acquire appropriate scaffolds through directed evolution. However, S2543 may not readily evolve, because this mutant already has a stable scaffold to some extent. Active mutants with disordered conformation may be located at the foot of a high mountain, while weakly active mutants with ordered conformation may be located near the top of a low mountain in the fitness landscape (Fig. **4**). This is similar to the situation in which kinetic intermediates trapped at local energy minima have to break inappropriate bonds to attain a global minimum on the energy landscape in protein folding reactions [37]. Block shuffling resulting in evolutionarily intermediate states of many proteins, combined with further directed evolution, could be regarded as generating products with evolutional trajectories climbing from near the foot of the mountain to the top in the fitness landscape.

3.5. Nature of Building Blocks Required for the Construction of Globular Proteins

Our reconstructive approach showed that foldable and active mutants were obtained more readily from a library of secondary structure unit mutants, rather than a library of module mutants. Modules have been considered as independently folding units and candidate building blocks, but permutation of modules can disrupt intra- and/or inter-modular interactions, resulting in many unfoldable proteins. On the other hand, permutation of secondary structure units, which are extended regions contacting each other *via* long-range interactions, resulted in the emergence of two mutants having folded regions. We speculate that peptide fragments capable of contacting each other with long-range interactions represent possible building blocks for the construction of foldable proteins. A possible alternative strategy to create foldable proteins would be to use peptide fragments capable of taking different stable conformations by adapting themselves to the surrounding environment. It is known that some amino acid sequences that form secondary structures have sufficient plasticity to form alternative stable conformations, such as α-helix and β-sheet, depending on their environment [38-40]. Recently, a completely foldable and functional single chain variant of the Arc repressor homodimer was successfully constructed by permutation of secondary structure elements, which were connected with appropriately designed linker sequences [41]. This result strongly supports the idea that secondary structure units have sufficient plasticity to be rearranged into different orders.

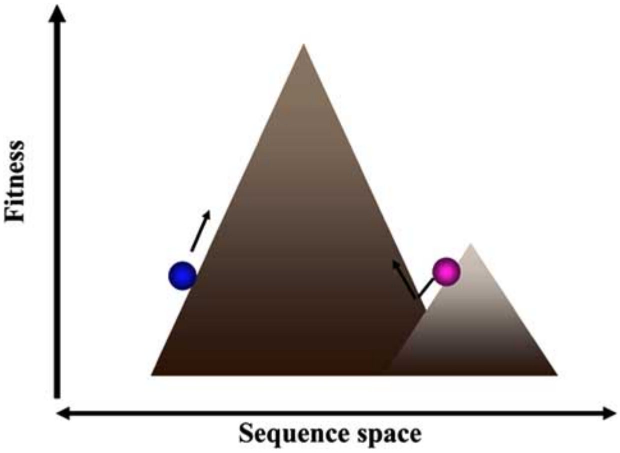

Fig. (4). Proteins in evolutionarily intermediate states located near the foot of a high mountain (blue circle) and near the top of a low mountain (pink circle) on the fitness landscape. Amino acids not directly involved in catalytic activity would presumably have been selected as scaffolds to stabilize the conformation of the pre-existing active site. If this is so, active mutants without stable conformation should have the ability to acquire mutations that stabilize the conformation of the active site, and thus to evolve into mature functional and stable proteins (blue circle). These mutants are present near the foot of the high mountain. On the other hand, foldable mutants with weak activity would not readily evolve, because they already have stable conformations. Such mutants appeared to be near the top of a low mountain (pink circle). Proteins trapped at local maxima have to return downhill first to attain the global maximum on the fitness landscape.

Iwakura *et al.* created all possible circular permutants of *E. coli* dihydrofolate reductase (DHFR), and tested their activity in an *in vivo* screening system to probe essential folding elements. They found that peptide segments involved in early folding events were conserved preferentially in the functional circular permutants [42]. Thus, peptide fragments that can trigger a folding reaction would be suitable building blocks.

Initiation sites of the folding reaction would be regions stabilized by local interactions on the primary sequence [43]. For example, a β-hairpin in protein G, termed G-peptide, which folds with a two-state transition as an isolated peptide fragment, can be considered as an independently folding unit [44] that would be a potential initiation site of a folding reaction. The structure of the β–hairpin is similar to those of modules in which two fragments of secondary structures are connected by a turn or loop. Therefore some modules may be independently folding units. However, modules were identified based only on the static three-dimensional structures of globular proteins, and therefore they might include different types of regions, including independently folding units like G-peptide, or regions containing fluctuating loops even in a folded state. We think it is desirable to classify modules into at least the above two categories, based on their role in the folding reaction.

Hiraga *et al*. created libraries containing all possible combinations of N- and C-terminal fragments of tryptophan synthase α-subunit, and identified combinations of the N- and C-terminal fragments that can express the activity by productive complementation [45]. Analysis of the fragments revealed that there were two distinct cleavable points in the subunit, and both were present on loops. This approach would be useful to identify essential units in the globular domain, and to find useful building blocks experimentally.

3.6. Interacting Ability of Barnase Mutants Obtained by Permutation of Secondary Structure Units

Among 46 barnase mutants, only S2354 was not expressed in *E. coli* under conditions where all other mutants were expressed. This suggests that S2354 has a strong RNase activity, and the activity prevented the growth of cells expressing the mutant gene. Indeed, wild-type barnase was expressed in *E. coli* when its inhibitor, barstar, was coexpressed [46]. To assess the foldability of S2354, we constructed S2354 in which His102 was substituted with alanine, which is situated in the active center of wild-type barnase. The mutant, S2354H102A, was successfully expressed in *E. coli*, and was found to have a partially folded conformation and a weak RNase activity. Surprisingly, S2354H102A interacted with other barnase mutants to generate enhanced RNase activity. Our results suggested that these interactions are based on complementation between secondary structure units of different protein molecules, and we concluded that S2354H102A uses His102 of other barnase mutants to enhance the activity [30].

Recently, 3D domain swapping has been widely accepted as a general mechanism for partially folded proteins to form homo-multimers through the interfaces between ordered regions of different protein molecules [47]. In the case of the barnase mutants, secondary structure units permuted in the mutants would maintain their intrinsic interacting ability, and the interacting ability would be used to form hetero-multimers (Fig. **5A**).

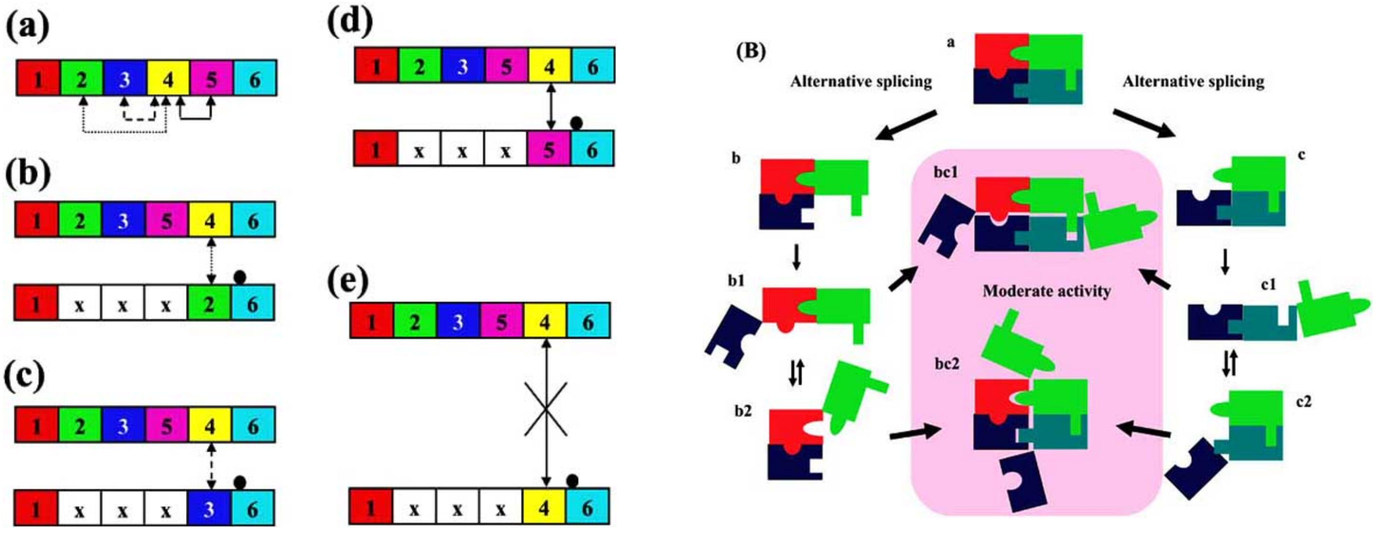

Fig. (5). (A) Interactions between secondary structure units. (a) Each secondary structure unit contacts each other unit in the globular conformation of wild-type barnase. Interactions between S4 and S2 (dotted line), S3 (broken line), and S5 (solid line) are indicated. (b), (c), and (d) Interactions between S2354H102A and barnase mutants having S2, 3, and 5 at the fifth block are indicated. These intermolecular interactions would allow His102 of other barnase mutants to contact a part of the putative active site of S2354H102A and generate RNase activity. (e) S4-S4 interaction would not occur, so S2354H102 can not use His 102 of barnase mutants having S4 at the fifth block. The black circle indicates His 102 of barnase mutants, and "x" indicates secondary structure units other than S2 (b), S3 (c), S5 (d), and S4 (e). (B) A proposed model for productive complementation between isoforms generated by alternative splicing. Alternative splicing generates two isoforms (b and c) from a foldable protein (a). Deletion of peptide fragments destabilizes the conformation of the isoforms, resulting in partially folded proteins (b1, b2, c1, and c2). If two different isoforms are present at the same time and place *in vivo*, interactions occur at the interfaces between ordered regions of the two kinds of isoforms to generate heterodimers (bc1 or bc2). Because complexes thus formed by productive complementation would show moderate activity, alternative splicing provides a mechanism to regulate the activity of proteins.

In living cells, alternative splicing might generate partially folded proteins, because insertions or deletions of peptide fragments encoded by exons would alter the conformational properties of proteins. Based on our experimental results, we speculate that isoforms generated by alternative splicing can also interact with each other to form hetero-multimers, and the complex formation might either enhance or weaken the activity (Fig. **5B**).

4. RANDOM MULTI-RECOMBINANT PCR (RM-PCR) FOR THE CONSTRUCTION OF COMBINATORIAL PROTEIN LIBRARIES BY BLOCK SHUFFLING

Block shuffling is a possible strategy to explore the global sequence space, and to create novel functional proteins. To realize this concept it was necessary to develop a novel strategy capable of combining different DNA fragments without homologous sequences. We developed random multi-recombinant PCR (RM-PCR) [48, 49] based upon overlap extension PCR [50]. Fig. **6A** shows a schematic diagram of multi-recombinant PCR, in which three building blocks are combined [29]. T7 and Ex are 5' and 3' consensus sequences, respectively, where forward and reverse primers anneal to prime the extension. The dimer templates of T7-2, 2-3, 3-1, and 1-Ex, having overlapping segments, are combined by a single PCR. Therefore, if many more dimer templates with overlapping segments are present in a tube, different multi-recombinant PCRs can proceed simultaneously, and many structural genes consisting of several building blocks arranged in different orders are obtained in a single PCR experiment (Fig. **6B**).

As shown in Fig. **6** the principle of RM-PCR is quite simple, and therefore the frequency of a certain building block in the library can be easily controlled. Fig. **7** shows a schematic overview of the application of RM-PCR to construct a random shuffling library where six building blocks, each encoding 25 amino acids, are combined. Equal amounts of the 36 dimer templates (each containing two building blocks) and 12 dimer templates containing 5' or 3' consensus sequences were mixed in the reaction mixture to obtain a library in which every position of the block sequences generated has an equal probability of encoding each of the six building blocks. Our previous study showed that many block sequences in which six building blocks were arranged into different orders were successfully constructed using this procedure [48]. Further, recently we confirmed that RM-PCR successfully combined 13 building blocks with different lengths, taken from different proteins.

Artificial alternative splicing libraries can also be constructed by mixing appropriate amounts of the dimer templates, according to the frequency of dimer templates in the desired library. Fig. **8** shows how one can determine the frequencies of the dimer templates. Fig. **8A** shows an example in which block sequences consisting of three building blocks are created from a structural gene that is divided into five building blocks. All dimer templates required are shown in Fig. **8B**, and the dimer templates are grouped into four classes. Each diagonal line covers each class of dimer templates. These classes are represented as block (N)-block (N+1), block (N)-block (N+2), block (N)-block (N+3), and block (N)-block (N+4). All possible block sequences consisting of three building blocks are shown in Fig. **8C**, and the figure also shows the frequency of each class of dimer templates in each block sequence. The sum of the frequencies for each class of dimer templates provides the molar ratio of the dimer template to be used in the

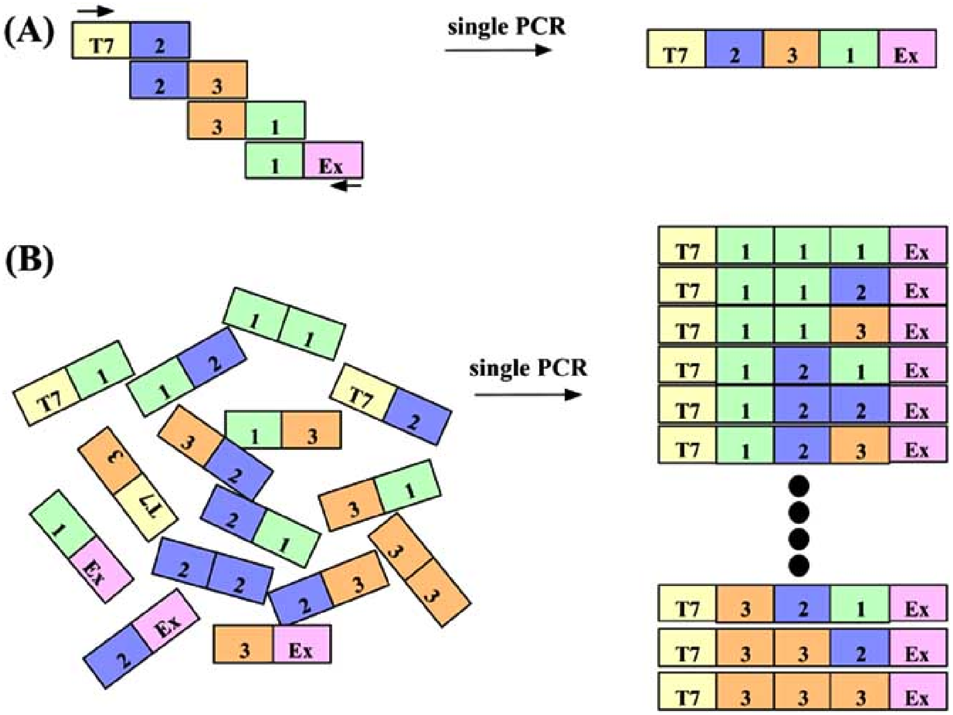

Fig. (6). A schematic diagram of multi-recombinant PCR (A) and random multi-recombinant PCR (RM-PCR) (B). In the multi-recombinant PCR dimer templates with overlapping segments are combined, resulting in one structural gene in a single PCR. In the RM-PCR different multi-recombinant PCRs proceed simultaneously, and many structural genes are constructed in a single PCR.

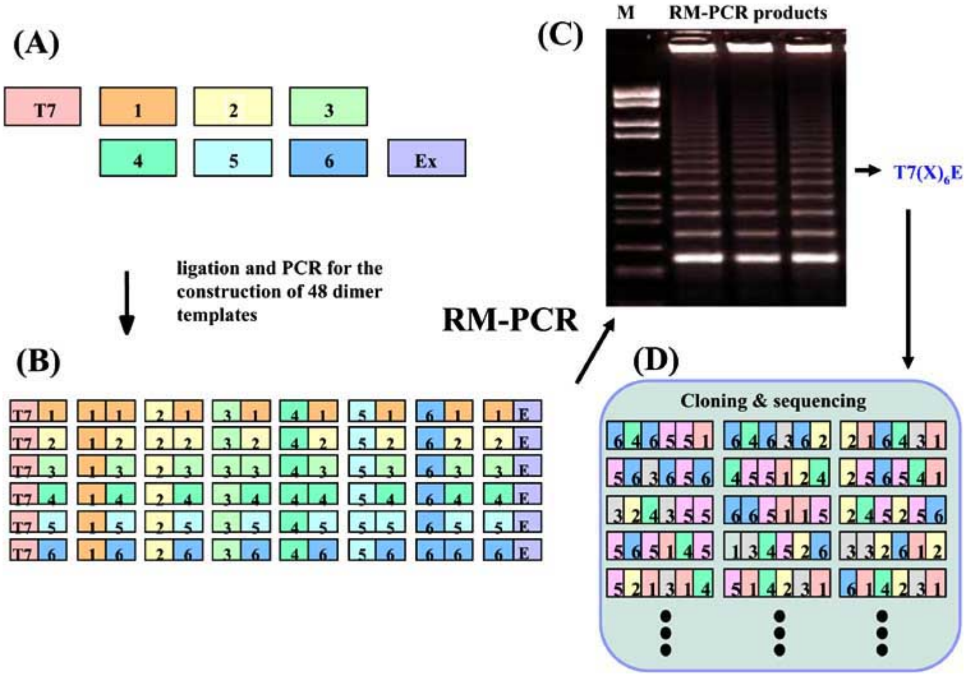
(A)
T7
1
2
3
4
5
6
Ex
ligation and PCR for the construction of 48 dimer templates
(B)
T7 1 | 1 1 | 2 1 | 3 1 | 4 1 | 5 1 | 6 1 | 1 E
T7 2 | 1 2 | 2 2 | 3 2 | 4 2 | 5 2 | 6 2 | 2 E
T7 3 | 1 3 | 2 3 | 3 3 | 4 3 | 5 3 | 6 3 | 3 E
T7 4 | 1 4 | 2 4 | 3 4 | 4 4 | 5 4 | 6 4 | 4 E
T7 5 | 1 5 | 2 5 | 3 5 | 4 5 | 5 5 | 6 5 | 5 E
T7 6 | 1 6 | 2 6 | 3 6 | 4 6 | 5 6 | 6 6 | 6 E
RM-PCR
(C)
M
RM-PCR products
T7(X)6E
(D)
Cloning & sequencing
6 4 6 5 5 1 | 6 4 6 3 6 2 | 2 1 6 4 3 1
5 6 3 6 5 6 | 4 5 5 1 2 4 | 2 5 6 5 4 1
3 2 4 3 5 5 | 6 6 5 1 1 5 | 2 4 5 2 5 6
5 6 5 1 4 5 | 1 3 4 5 2 6 | 3 3 2 6 1 2
5 2 1 3 1 4 | 5 1 4 2 3 1 | 6 1 4 2 3 1

Fig. (7). A schematic diagram for the construction of a random shuffling library by RM-PCR. Eight double-stranded DNA fragments encoding peptide sequences of interest were prepared. T7 and Ex are 5' and 3' consensus sequences, respectively (A). Thirty-six dimer templates of the six building blocks and 12 dimer templates of the consensus sequences were prepared and mixed in a tube, then PCR was performed (B). After electrophoresis, DNA fragments of the desired length were purified from the gel (C, DNA fragments containing six building blocks were purified in this case). The purified fragments were cloned, and sequenced (D), revealing the construction of many block sequences whose every position has an equal probability to encode each of the six building blocks.

reaction mixture. For example, the molar ratio of block (N)-block (N+1), block (N)-block (N+2), block (N)-block (N+3), and block (N)-block (N+4) is 6:3:1:0. These values can be represented systematically using binomial coefficients as shown in Fig. **8D**. We have successfully constructed four alternative splicing libraries, in which block sequences consisting mainly of five to eight building blocks were obtained from a parent protein that was divided into ten building blocks [49] (Fig. **9A-C**). In addition, it was shown that point mutations can be introduced at desired frequencies during block shuffling by performing RM-PCR with different concentrations of Mn^{2+} [49] (Fig. **9D**).

5. STRATEGIES FOR *IN VITRO* SELECTION OF FUNCTIONAL PROTEINS AND PEPTIDES: *IN VITRO* VIRUS (mRNA DISPLAY) AND STABLE (DNA DISPLAY)

Our laboratory has developed the *in vitro* virus (IVV) technique, in which mRNA and protein are linked in a cell-free translation system [51]. Robert and Szostak also independently developed the same method and reported it several months later [52]. In this mRNA display method, a protein (phenotype) is linked to its encoding mRNA (genotype) through puromycin, forming an mRNA-protein fusion molecule (IVV). The library of fusion molecules can be subjected to selection experiments; for example, by using affinity chromatography, proteins having binding ability to a target ligand are selected. The mRNA portions of the selected fusion molecules are reverse-transcribed, amplified by PCR, cloned, and finally sequenced. Proteins with the desired binding ability are thus identified easily from a library of different sequences. To improve the efficiency of formation of mRNA-protein fusion molecules, several refinements have been reported [53-58]. Miyamoto-Sato *et al*. reported that more than 70 % of input mRNA was successfully converted to mRNA-protein fusion molecules under the optimum conditions [54].

mRNA display has been utilized for different purposes by different groups. Robert and colleagues obtained several functional peptides containing non-natural amino acids [59-62]. Amino acid residues in three exposed loops of the tenth fibronectin type III domain were randomized, and the variants binding to TNF-α were selected [63]. A cDNA library was also subjected to mRNA display, and some sequences interacting with Bcl-x_L or FK506 were selected [64, 65]. Keefe and Szostak obtained novel ATP-binding proteins from a random amino acid library [9]. The structure of a selected protein was recently determined, revealing that it indeed has a novel fold [66]. This protein consisted of α-helices and β-strands like natural proteins, but the topology of the secondary structures was quite novel. The protein contains CXXC motifs binding to Zn^{2+}, which are often seen in natural proteins, but nucleotide binding proteins having these motifs have not yet been found in nature. Thus, elements of the protein are not new, but the particular combination of these elements gives the protein novelty. These results indicate that novel functional proteins are indeed present in the global sequence space, and have apparently not been accessed by natural evolutionary processes. In addition, it appears that novel proteins can be created by combining pre-existing structural and functional elements, such as secondary structures or functional motifs.

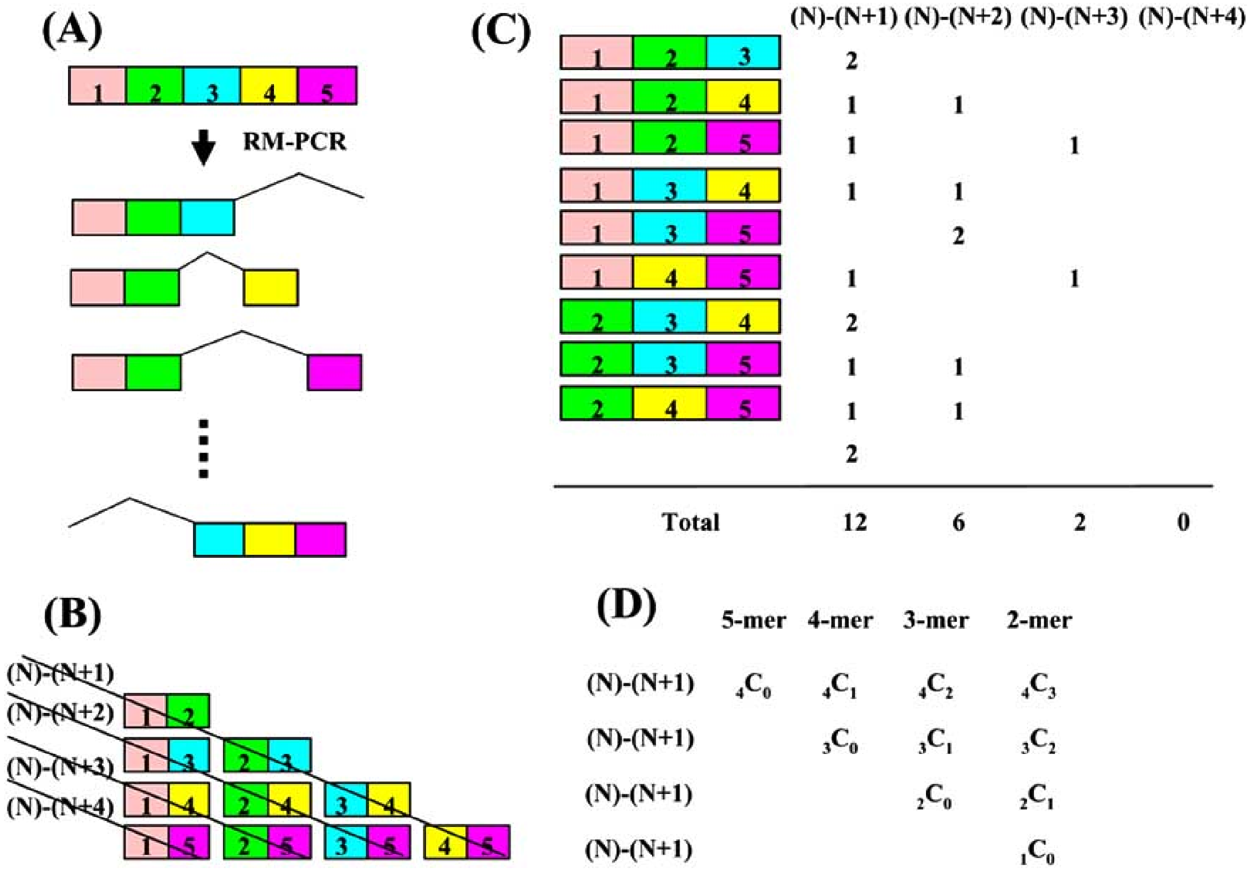

Fig. (8). Dimer templates for the construction of an alternative splicing library by RM-PCR. (A) Different block sequences consisting of three building blocks are created from a structural gene, which consists of five building blocks. (B) All dimer templates required for the construction of the library are shown. These dimer templates can be grouped into four classes. Diagonal lines cover each class of dimer template. (C) The frequency of each class of dimer template appearing in all possible block sequences consisting of three building blocks is indicated. The sum of the frequencies for each class of dimer template gives the required molar ratio of dimer templates in the reaction mixture for RM-PCR. (D) Relative frequencies of dimer templates in all possible block sequences with a certain number of building blocks can be indicated systemically by using binomial coefficients.

In our laboratory, mRNA display has been applied for mapping protein-protein interaction networks. Horisawa *et al.* identified 16 novel Jun-associated protein candidates from a mouse brain cDNA library [67], and further monitored successive enrichment or elimination of clones during iterative selection rounds using quantitative real-time PCR [68]. Miyamoto-Sato *et al.* reported more than 10 Fos-interacting proteins, including proteins that interacted with Fos indirectly [69].

During the development of the *in vitro* virus technique, puromycin was found to bind to the C-terminal end of proteins specifically on the ribosome under certain conditions [70, 71]. This phenomenon gave us the idea that proteins could be labeled specifically at the C-terminal end by using puromycin derivatives modified with fluorescent dyes. We have established a high-throughput assay technology for detecting protein-protein interactions *in vitro* by combining this labeling method, fluorescence cross-correlation spectroscopy (FCCS) and protein microarrays [72, 73].

Doi and Yanagawa developed a DNA-protein fusion system termed STABLE (STreptavidin-Biotin Linkage in Emulsions) for *in vitro* selection of proteins [74]. This method relies on an *in vitro* transcription/translation reaction compartmentalized in water-in-oil emulsions. In each emulsion compartment, streptavidin-fused polypeptides are synthesized and attached to the encoding DNA *via* their biotin label. The resulting DNA-protein fusion molecules can be subjected to *in vitro* selection experiments similarly to mRNA display libraries, but a process of reverse-transcription is not required. STABLE was successfully used to find more than 20 FLAG-like peptides, which can be recognized by anti-FLAG antibody, from a random amino acids library [75]. The selected peptides were encoded by DNA sequences that were not similar to each other, indicating that convergent evolution had been successfully simulated *in vitro*.

As described above, we have established experimental conditions for subjecting a block-shuffling library to *in vitro* selection. Our goal is to obtain structurally and functionally immature proteins from a combinatorial library created by block shuffling, and then to evolve these proteins to structurally and functionally mature proteins, which would allow us to define evolutional trajectories on the fitness landscape. The combination of random multi-recombinant PCR and DNA shuffling is a possible strategy to attain this

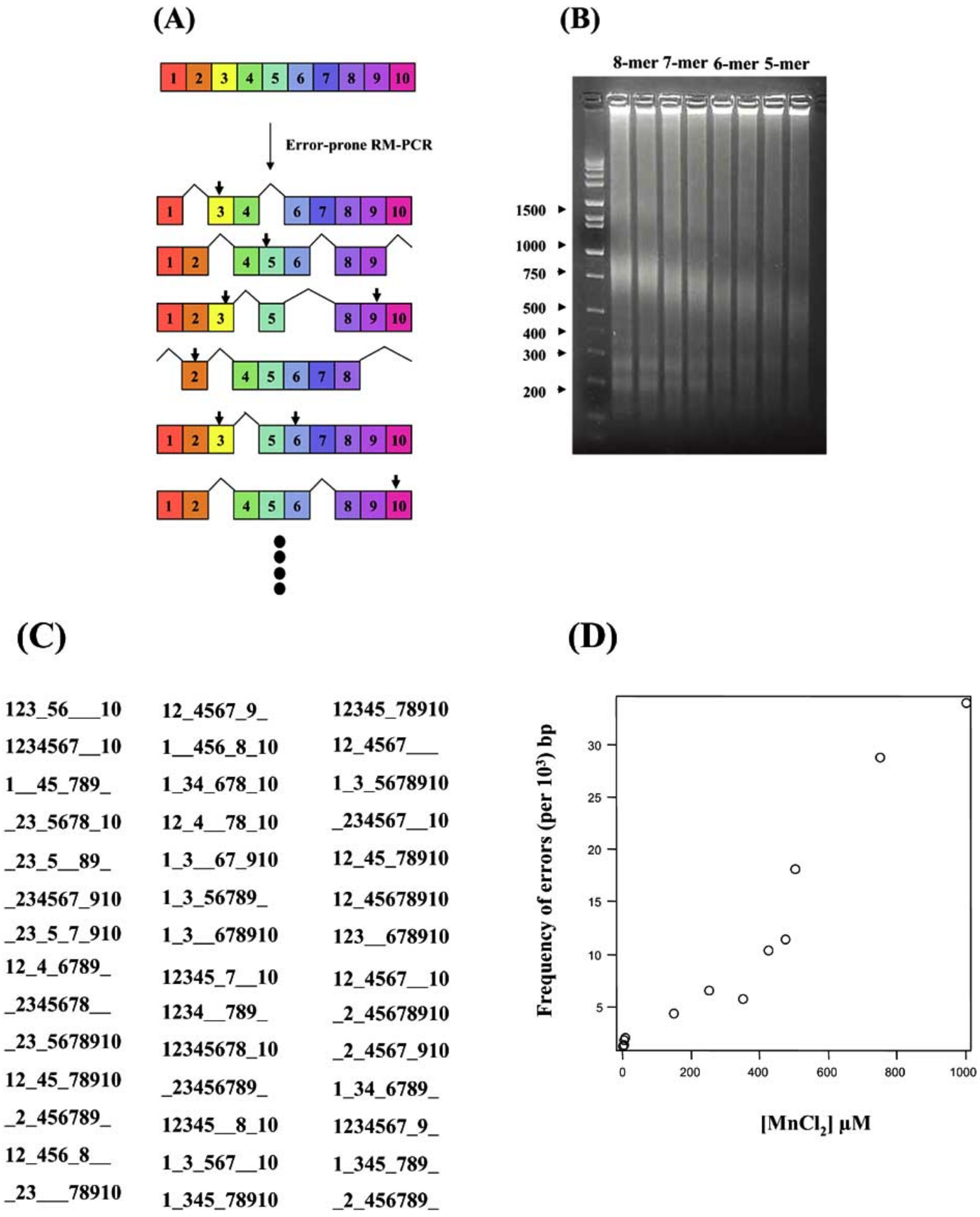

Fig. (9). A schematic diagram for the construction of an artificial alternative splicing library by error-prone RM-PCR. (A) A structural gene consisting of 10 building blocks was used as a parent sequence. (B) RM-PCR products obtained from reaction mixtures in which dimer templates were mixed to yield structural genes consisting of eight, seven, six, and five building blocks were subjected to agarose gel electrophoresis. (C) DNA fragments obtained from the 7-mer library were purified from the gel, cloned, and sequenced. Only a part of the library was analyzed, and the results revealed that many block sequences consisting of five to nine building blocks were obtained. (D) Frequencies of point mutations during error-prone RM-PCR. Details are given in our previous paper [49].

goal, and would allow us to explore global and local sequence space extensively. Alternatively, proteins with higher foldability may be obtained by performing *in vitro* directed evolution experiments in the presence of a denaturant [76].

6. PROSPECTS FOR THE FUTURE

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, which have similar folds, but have low 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 to create combinatorial libraries by block shuffling based on random digestion and blunt-end ligation of parent genes in the presence of appropriate amount 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 pre-selection 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].

Our developed RM-PCR and other methods described above will allow us to search the global sequence space efficiently, but the most exciting challenge is to create proteins with novel functions which have never appeared on the earth. This challenge may already have been met in the field of materials science. Recently many peptides binding to different inorganic materials, including semiconductors, were obtained by phage display or bacterial cell surface display [81]. These peptides were used for designing functional nanostructures [82]. Sano and Shiba identified a hexapeptide motif binding to titanium from a random peptide library, and this would be used for many applications, including implant materials [83]. Shiba and colleagues have also developed a method termed microgene polymerization reaction (MPR), which creates combinatorial libraries by combining functional motifs and other sequences that have the propensity to form secondary structures [84-86]. Because proteins have not encountered such artificial materials in the course of natural protein evolution, an interesting strategy would be to use such artificial peptide motifs as one of the building blocks to create artificial proteins with novel functions by block shuffling. As described in this review article, fundamental strategies for *in vitro* selection and block shuffling have been established. Now, it is possible to combine these strategies to obtain novel functional proteins, and to evolve them to useful mature proteins. Finally, it will be of interest to examine the evolutional trajectories of artificial proteins on the fitness landscape.

ABBREVIATIONS

Designation of barnase mutants is based on the modules and secondary structure units in them: for example, M3245 represents a mutant containing a permutation of the four internal modules in which the six modules of barnase numbered from the N-terminal coding region are rearranged in the order 1, 3, 2, 4, 5, 6. Similarly S2543 represents a mutant containing a permutation of the four internal secondary structure units in which the six structural fragments numbered from the N-terminal coding region are rearranged in the order 1, 2, 5, 4, 3, 6. S2354H102A is a mutant of S2354 in which His-102 is replaced with alanine.

REFERENCES

[1] Gilbert, W.; de Souza, S.J.; Long, M. *Proc. Natl. Acad. Sci. USA*, **1997,** *94,* 7698-7703.

[2] Gilbert, W. *Cold Spring Harb. Symp. Quant. Biol*., **1987,** *52*, 901-905.

[3] Farinas, E.T.; Bulter, T.; Arnold, F.H. *Curr. Opin. Biotechnol*., **2001,** *12,* 545-551.

[4] Doi, N.; Yanagawa, H. *Comb. Chem. High Throughput Screen.,* **2001,** *4,* 497-509.

[5] Arnold, F.H.; Wintrode, P.L.; Miyazaki, K.; Gershenson, A. *Trends Biochem. Sci.,* **2001,** *26,* 100-106.

[6] Stemmer, W.P. *Nature,* **1994,** *370,* 389-391.

[7] Zhao, H; Giver, L; Shao, Z; Affholter, J.A; Arnold, F.H. *Nat. Biotechnol.,* **1998,** *16,* 258-261.

[8] Crameri, A.; Raillard, S.A.; Bermudez, E.; Stemmer, W.P. *Nature,* **1998,** *391,* 288-291.

[9] Keefe, A.D.; Szostak, J.W. *Nature,* **2001,** *410,* 715-718.

[10] Blake, C.C. *Nature,* **1979,** *277,* 598.

[11] Go, M. *Nature,* **1981,** *291,* 90-92.

[12] Go, M. *Proc. Natl. Acad. Sci. USA,* **1983,** *80,* 1964-1968.

[13] Go, M. *Adv. Biophys.,* **1985,** *19,* 91-131.

[14] Go, M; Nosaka, M. *Cold Spring Harb. Symp. Quant. Biol*., **1987,** *52,* 915-924.

[15] Fersht, A.R.; Matouschek, A.; Serrano, L. *J. Mol. Biol.,* **1992,** *224,* 771-782.

[16] Serrano, L.; Kellis, J.T.Jr.; Cann, P.; Matouschek, A.; Fersht, A.R. *J. Mol. Biol.,* **1992,** *224,* 783-804.

[17] Serrano, L.; Matouschek, A.; Fersht, A.R. *J. Mol. Biol.,* **1992,** *224,* 805-818.

[18] Matouschek, A.; Serrano, L.; Fersht, A.R. *J. Mol. Biol.,* **1992,** *224,* 819-835.

[19] Noguti, T.; Sakakibara, H.; Go, M. *Proteins,* **1993,** *16,* 357-363.

[20] Yanagawa, H.; Yoshida, K.; Torigoe, C.; Park, J. S.; Sato, K.; Shirai, T.; Go, M. *J. Biol. Chem.,* **1993,** *268,* 5861-5865.

[21] Yoshida, K.; Shibata, T.; Masai, J.; Sato, K.; Noguti, T.; Go, M.; Yanagawa, H. *Biochemistry,* **1993,** *32,* 2162-2166.

[22] Ikura, T.; Go, N.; Kohda, D.; Inagaki, F.; Yanagawa, H.; Kawabata, M.; Kawabata, S.; Iwanaga, S.; Noguti, T.; Go, M. *Proteins,* **1993,** *16,* 341-356.

[23] Takahashi, K.; Oohashi, M.; Noguti, T.; Go, M. *FEBS Lett.,* **1997,** *405,* 47-54.

[24] Takahashi, K.; Noguti, T.; Hojo, H.; Ohkubo, T.; Go, M. *Biopolymers,* **2001,** *58,* 260-267.

[25] Takahashi, K.; Noguti, T.; Hojo, H.; Yamauchi, K.; Kinoshita, M.; Aimoto, S.; Ohkubo, T.; Go, M. *Protein Eng.,* **1999,** *12,* 673-680.

[26] Wakasugi, K.; Ishimori, K.; Imai, K.; Wada, Y.; Morishima, I. *J. Biol. Chem.,* **1994,** *269,* 18750-18756.

[27] Inaba, K.; Wakasugi, K.; Ishimori, K.; Konno, T.; Kataoka, M.; & Morishima, I. *J. Biol. Chem.,* **1997,** *272,* 30054-30060.

[28] Kaneko, S.; Iwamatsu, S.; Kuno, A.; Fujimoto, Z.; Sato, Y.; Yura, K.; Go, M.; Mizuno, H.; Taira, K.; Hasegawa, T.; Kusakabe, I.; Hayashi, K. *Protein Eng.,* **2000,** *13,* 873-879.

[29] Tsuji, T.; Yoshida, K.; Satoh, A.; Kohno, T.; Kobayashi, K.; Yanagawa, H. *J. Mol. Biol.,* **1999,** *286,* 1581-1596.

[30] Tsuji, T.; Yanagawa, H. *Biochemistry,* **2004,** *43,* 6968-6975.

[31] Tsuji, T.; Kobayashi, K.; Yanagawa, H. *FEBS Lett.,* **1999,** *453,* 145-150.

[32] Meiering, E.M.; Bycroft, M.; Lubienski, M.J.; Fersht, A.R. *Biochemistry,* **1993,** *32,* 10975-10987.

[33] Saito, S.; Sasai, M.; Yomo; T. *Proc. Natl. Acad. Sci. USA,* **1997,** *94,* 11324-11328.

[34] Yomo, T.; Saito, S.; Sasai; M. *Nat. Struct. Biol.,* **1999,** *6,* 743-746.

[35] Miyazaki, K.; Arnold, F.H. *J. Mol. Evol.,* **1999,** *49,* 716-20.

[36] Henke, E.; Bornscheuer, U.T. *Biol. Chem.,* **1999,** *380,* 1029-1033.

[37] Bryngelson, J.D; Onuchic, J.N.; Socci, N.D.; Wolynes, P.G. *Proteins,* **1995,** *21,* 167-195.

[38] Waterhous, D.V.; Johnson, W.C. Jr. *Biochemistry,* **1994,** *33,* 2121-2128.

[39] Shiraki, K.; Nishikawa, K.; Goto, Y. *J. Mol. Biol.,* **1995,** *245,* 180-194.

[40] Minor, D.L. Jr; Kim, P. S. *Nature,* **1996,** *380,* 730-734.

[41] Tabtiang, R.K.; Cezairliyan, B.O.; Grant, R.A.; Cochrane, J.C.; Sauer, R.T. *Proc. Natl. Acad. Sci. USA,* **2005,** *102,* 2305-2309.

[42] Iwakura, M.; Nakamura, T.; Yamane, C.; Maki, K. *Nat. Struct. Biol.,* **2000,** *7,* 580-585.

[43] Freund, S.M.; Wong, K.B.; Fersht, A.R. *Proc. Natl. Acad. Sci. USA,* **1996,** *93,* 10600-10603.

[44] Honda, S.; Kobayashi, N.; Munekata, E. *J. Mol. Biol.,* **2000,** *295,* 269-278.

[45] Hiraga, K.; Yamagishi, A.; Oshima, T. *J. Mol. Biol.,* **2004,** *335,* 1093-1104.

[46] Hartley, R.W. *J. Mol. Biol.,* **1988,** *202,* 913-915.

[47] Bennett, M.J.; Eisenberg, D. *Structure,* **2004,** *12,* 1339-1341.

[48] Tsuji, T.; Onimaru, M.; Yanagawa, H. *Nucleic Acids Res.,* **2001,** *29,* E97.

[49] Tsuji, T.; Onimaru, M.; Kitagawa, M.; Kojoh, K.; Tabata, N.; Yanagawa, H. *Methods Enzymol.,* **2004,** *388,* 61-75.

[50] Horton, R.M.; Hunt, H.D.; Ho, S.N.; Pullen, J.K.; Pease, L.R. *Gene,* **1989,** *77,* 61-68.

[51] Nemoto, N.; Miyamoto-Sato, E.; Hushimi, Y.; Yanagawa, H. *FEBS Lett.,* **1997,** *414,* 405-408.

[52] Roberts, R.W.; Szostak, J.W. *Proc. Natl. Acad. Sci. USA,* **1997,** *94,* 12297-12302.

[53] Liu, R.; Barrick, J.E.; Szostak, J.W.; Roberts, R.W. *Methods Enzymol.,* **2000,** *318,* 268-293.

[54] Miyamoto-Sato, E.; Takashima, H.; Fuse, S.; Sue, K.; Ishizaka, M.; Tateyama, S.; Horisawa, K.; Sawasaki, T.; Endo, Y.; Yanagawa, H. *Nucleic Acids Res.,* **2003,** *31,* E78.

[55] Kurz, M.; Gu, K. ; Lohse, P.A. *Nucleic Acids Res.,* **2000,** *28,* E83.

[56] Tabuchi, I.; Soramoto, S.; Suzuki, M.; Nishigaki, K.; Nemoto, N.; Husimi, Y. *Biol. Proced. Online,* **2002,** *4,* 49-54.

[57] Kurz, M.; Gu, K.; Al-Gawari, A.; Lohse, P.A. *Chembiochem.,* **2001,** *2,* 666-672.

[58] Tabuchi, I.; Soramoto, S.; Nemoto, N.; Husimi, Y. *FEBS Lett.,* **2001,** *508,* 309-312.

[59] Li, S.; Millward, S.; Roberts, R.W. *J. Am. Chem. Soc.,* **2002,** *124,* 9972-9973.

[60] Li, S.; Roberts, R.W. *Chem. Biol.,* **2003,** *10,* 233-239.

[61] Frankel, A.; Roberts, R.W. *RNA,* **2003,** *9,* 780-786.

[62] Frankel, A.; Millward, S.W.; Roberts, R.W. *Chem. Biol.,* **2003,** *10,* 1043-1050.

[63] Xu, L.; Aha, P.; Gu, K.; Kuimelis, R.G.; Kurz, M.; Lam, T.; Lim, A.C.; Liu, H.; Lohse, P.A.; Sun, L.; Weng, S.; Wagner, R.W.; Lipovsek, D. *Chem. Biol.,* **2002,** *9,* 933-942.

[64] Hammond, P.W.; Alpin, J.; Rise, C.E.; Wright, M.; Kreider, B.L. *J. Biol. Chem.,* **2001,** *276,* 20898-20906.

[65] McPherson, M.; Yang, Y.; Hammond, P.W.; Kreider, B.L. *Chem. Biol.,* **2002,** *9,* 691-698.

[66] Lo Surdo, P.; Walsh, M.A.; Sollazzo, M. *Nat. Struct. Mol. Biol.,* **2004,** *11,* 382-383.

[67] Horisawa, K.; Tateyama, S.; Ishizaka, M.; Matsumura, N.; Takashima, H.; Miyamoto-Sato, E.; Doi, N.; Yanagawa, H. *Nucleic Acids Res.,* **2004,** *32,* E169.

[68] Horisawa, K.; Doi, N.; Takashima, H.; Yanagawa, H. *J. Biochem.,* **2005,** *137,* 121-124.

[69] Miyamoto-Sato, E.; Ishizaka, M.; Horisawa, K.; Tateyama, S.; Takashima, H.; Fuse, S.; Sue, K.; Hirai, N.; Masuoka, K.; Yanagawa, H. *Genome Res.,* **2005,** *15,* 710-717.

[70] Nemoto, N.; Miyamoto-Sato, E.; Yanagawa, H. *FEBS Lett.,* **1999,** *462,* 43-46.

[71] Miyamoto-Sato, E.; Nemoto, N.; Kobayashi, K; Yanagawa, H. *Nucleic Acids Res.,* **2000,** *28,* 1176-1182.

[72] Doi, N.; Takashima, H.; Kinjo, M.; Sakata, K.; Kawahashi, Y.; Oishi, Y.; Oyama, R.; Miyamoto-Sato, E.; Sawasaki, T.; Endo, Y.; Yanagawa, H. *Genome Res.,* **2002,** *12,* 487-492.

[73] Kawahashi, Y.; Doi, N.; Takashima, H.; Tsuda, C.; Oishi, Y.; Oyama, R.; Yonezawa, M.; Miyamoto-Sato, E.; Yanagawa, H. *Proteomics,* **2003,** *3,* 1236-1243.

[74] Doi, N.; Yanagawa, H. *FEBS Lett.,* **1999,** *457,* 227-230.

[75] Yonezawa, M.; Doi, N.; Kawahashi, Y.; Higashinakagawa, T.; Yanagawa, H. *Nucleic Acid Res.,* **2003,** *31,* E118.

[76] Chaput, J.C.; Szostak, J.W. *Chem. Biol.,* **2004,** *11,* 865-874.

[77] O'Maille, P.E.; Bakhtina, M.; Tsai, M.D. *J. Mol. Biol.,* **2002,** *321,* 677-691.

[78] Hiraga, K.; Yamagishi, A.; Oshima, T. *J. Mol. Biol.,* **2004,** *335,* 1093-1104.

[79] Bittker, J.A.; Le, B.V.; Liu, J.M.; Liu, D.R. *Proc. Natl. Acad. Sci. USA,* **2004,** *101,* 7011-7016.

[80] Kitamura, K.; Kinoshita, Y.; Narasaki, S.; Nemoto, N.; Husimi, Y.; Nishigaki, K. *Protein Eng.,* **2002,** *15,* 843-853.

[81] Sarikaya, M.; Tamerler, C.; Jen, A.K.; Schulten, K.; Baneyx, F. *Nat. Mater.,* **2003,** *2,* 577-585.

[82] Whaley, S.R.; English, D.S.; Hu, E.L.; Barbara, P.F.; Belcher, A.M. *Nature,* **2000,** *405,* 665-668.

[83] Sano, K.; Shiba, K. *J. Am. Chem. Soc.,* **2003,** *125,* 14234-14235.

[84] Shiba, K.; Takahashi, Y.; Noda, T. *Proc. Natl. Acad. Sci. USA,* **1997,** *94,* 3805-3810.

[85] Shiba, K.; Takahashi, Y.; Noda T. *J. Mol. Biol.,* **2002,** *320,* 833-840.

[86] Saito, H.; Honma, T.; Minamisawa, T.; Yamazaki, K.; Noda, T.; Yamori, T.; Shiba, K. *Chem. Biol.,* **2004,** *11,* 765-773.

Received: July 22, 2005 **Revised: September 12, 2005** **Accepted: September 12, 2005**