Random multi-recombinant PCR for the construction of combinatorial protein libraries

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

[1]Mitsubishi Kagaku Institute of Life Sciences, 11 Minamiooya, Machida, Tokyo 194-8511, Japan and
[2]Department of Applied Chemistry, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama 223-8522, Japan

Received May 21, 2001; Received August 1, 2001; Accepted August 12, 2001

ABSTRACT

Development of a new methodology to create protein libraries, which enable the exploration of global protein space, is an exciting challenge. In this study we have developed random multi-recombinant PCR (RM-PCR), which permits the shuffling of several DNA fragments without homologous sequences. In order to evaluate this methodology, we applied it to create two different combinatorial DNA libraries. For the construction of a 'random shuffling library', RM-PCR was used to shuffle six DNA fragments each encoding 25 amino acids; this affords many different fragment sequences whose every position has an equal probability to encode any of the six blocks. For the construction of the 'alternative splicing library', RM-PCR was used to perform different alternative splicings at the DNA level, which also yields different block sequences. DNA sequencing of the RM-PCR products in both libraries revealed that most of the sequences were quite different, and had a long open reading frame without a frame shift or stop codon. Furthermore, no distinct bias among blocks was observed. Here we describe how to use RM-PCR for the construction of combinatorial DNA libraries, which encode protein libraries that would be suitable for selection experiments in the global protein space.

INTRODUCTION

Exploring global protein space by directed evolution is an exciting challenge, because only limited sequence spaces around natural proteins have so far been explored. It is very likely that new functional or foldable proteins are present in the global protein space (1,2). Recent techniques, such as directed evolution performed completely *in vitro* (3–6), would allow us to explore the global protein space if appropriate protein libraries are employed. Our laboratory has developed two types of directed *in vitro* evolution systems termed '*in vitro* virus' (3) and 'STABLE' (6). In the '*in vitro* virus' method, polypeptides obtained *in vitro* are attached to their mRNAs with a covalent bond through a puromycin derivative that is synthetically coupled to the 3′ end of the mRNA (7). In the 'STABLE', a library of biotinylated DNAs, which encode polypeptides fused to streptavidin is transcribed and translated *in vitro* in water-in-oil emulsions. In both systems, the sequences of functional proteins can be obtained from the nucleic acid portion of the nucleic acid–protein fusion molecules. These techniques would allow us to find rare functional proteins among the ~10^{13} independent members. Therefore, it is timely to develop a novel strategy for the construction of protein libraries suitable for exploring global protein space.

Several different types of protein library can be used for this purpose. They include a random amino acid library encoded by chemically synthesized DNAs (8–10) and a binary patterned library favoring secondary structure formation (11,12). Another is a combinatorial protein library consisting of different building blocks such as modules (13–15), secondary structures, functional motifs and so on. The latter strategy is based on the 'exon shuffling theory' of 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 (16,17). However, some difficulties are associated with the construction of such a DNA library, because several DNA fragments without homologous sequences must be combined in an appropriate direction. DNA shuffling (18), which can thoroughly explore local sequence space around natural proteins, cannot be applied because the building blocks one wants to shuffle do not always have homologous sequences. A method termed incremental truncation for creating hybrid enzymes (ITCHY) can combine two parent sequences in a homology-independent manner (19), but this method cannot create structural genes consisting of more than two building blocks per experiment. Several DNA fragments flanked by non-palindromic restriction enzyme sites can be used for the construction of a library containing different sequences with different building blocks connected in an appropriate direction (10), but the sites specify certain kinds of amino acids at every junction between building blocks, which will cause a distinct bias in the protein space that we explore.

To address such problems, in this study we have developed random multi-recombinant PCR (RM-PCR), which permits the shuffling of plural DNA fragments without homologous sequences in a single PCR. This method is based on the 'multi-recombinant PCR' reported by us previously (20). Figure 1A shows a schematic diagram of multi-recombinant PCR, in which three building blocks are combined. T7 (an artificial sequence encoding a T7 promoter sequence) and Ex are 5′ and

*To whom correspondence should be addressed at: Department of Applied Chemistry, 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; Email: hyana@applc.keio.ac.jp

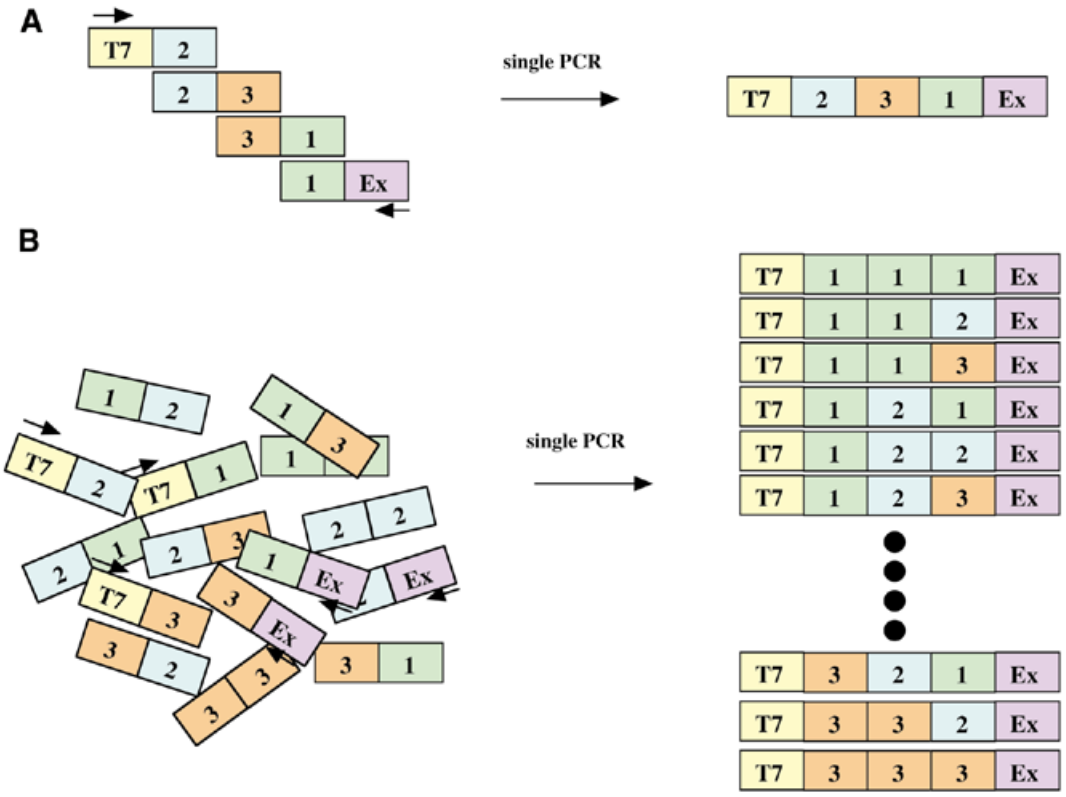


Figure 1. Schematic diagrams of multi-recombinant PCR (**A**) and RM-PCR (**B**). In the RM-PCR, different sequences consisting of several building blocks arranged in different orders may be synthesized in a single PCR.

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 overlapped segments, are combined by a single PCR. Therefore, if many more dimer templates with overlapped segments (for example 1-1, 1-2, . . . , 3-3) 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 may be obtained by a single PCR (Fig. 1B).

In order to test this idea we have applied RM-PCR to create two different types of a combinatorial DNA library. For the construction of the 'random shuffling library', RM-PCR was used to shuffle and combine six DNA fragments each encoding 25 amino acids without homologous sequences, which would result in many different fragment sequences whose every position has an equal probability to encode any of the six building blocks (Fig. 2A). For the construction of the second library termed the 'alternative splicing library', RM-PCR was used to perform different alternative splicings at the DNA level, i.e. DNA fragments encoding 10 building blocks with different chain lengths from human estrogen receptor α ligand binding domain (hERα LBD) were spliced alternatively, resulting in different block sequences (Fig. 2B). Here we describe how to use RM-PCR for the construction of combinatorial DNA libraries, which encode protein libraries that would be suitable for selection experiments in the global protein space.

MATERIALS AND METHODS

Preparation of dimer templates

Plus chains of six building blocks from *Escherichia coli* glutaminyl-tRNA synthetase (21,22), T7, Ex, T7OM and CBPHis were purchased from ESPEC Oligo Service (Japan) or DATE Concept (Japan). The sequences were 5′-acc aca gta cac acc cgt ttc ccg ccg gag ccg aat ggc tat ctg cat att ggc cat gcg aaa tct atc tgc ctg-3′ (block 1), 5′-aac ttc ggg atc gcc cag gac tat aaa ggc cag tgc aac ctg cgt ttc gac gac act aac ccg gta aaa gaa gat-3′ (block 2), 5′-atc gag tat gtt gag tcg atc aaa aac gac gta gag tgg tta ggt ttt cac tgg tct ggt aac gtc cgt tac tcc-3′ (block 3), 5′-atg cgc gat ccg gtg ctg tac cgt att aaa ttt gct gaa cac cac cag act ggc aac aag tgg tgc atc tac ccg-3′ (block 4), 5′-atg tac gac ttc acc cac tgc atc agc gat gcc ctg gaa ggt att acg cac tct ctg tgt acg ctt gag ttc cag-3′ (block 5), 5′-gac aac cgt cgt ctg tac gac tgg gta ctg gac aac atc acg att cct gtt cac ccg cgc cag tat gag ttc tcg-3′ (block 6), 5′-gat ccc gcg aaa tta ata cga ctc act ata ggg aga cca caa cgg ttt ccc tct aga aat aat ttt gtt att ctt taa gaa gga gat gcc acc atg-3′ (T7), 5′-atc tcg atc ccg cga taa tac gac tca cta tag gga caa tta cta ttt aca att aca atg gac tac aaa gat gac gac gat aag-3′ (T7OM), 5′-ggc ggg gcc gct gcg ctg tct ggt gcc ctg tcc atc agc gct gtc ggt tct ctg tcc ttg atc ggc gtg atc ctc ggc gct gga ggc-3′ (Ex) and 5′-aag cga cga tgg aaa aag aat ttc ata gcc gtc tca gca gcc aac cgc ttt aag aaa atc tca tcc tcc ggg gca ctt cat cac cat cac cat cac-3′ (CBPHis). The letters underlined indicate the sequences and complementary sequences of the forward and

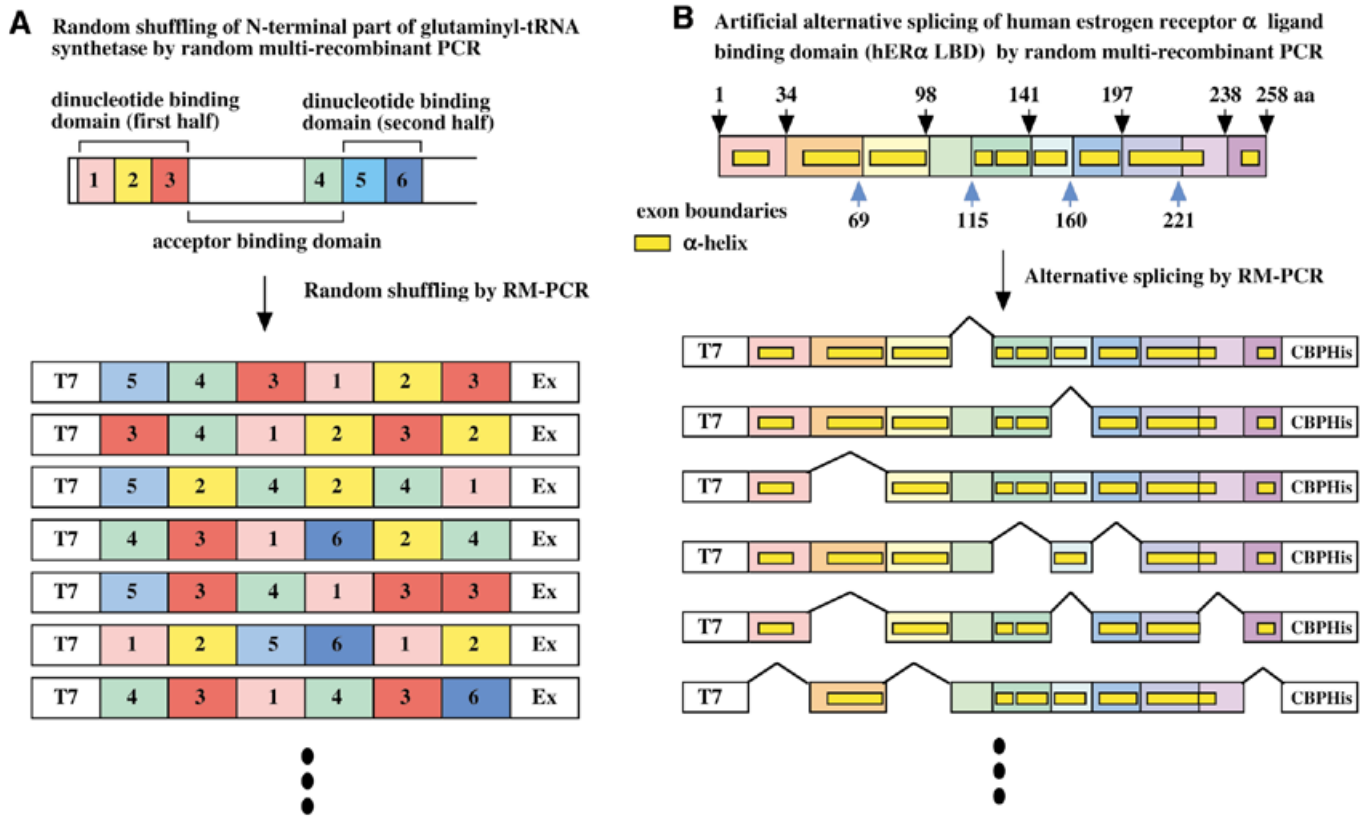


Figure 2. (**A**) Construction of a random shuffling library by RM-PCR. The N-terminal part of *E.coli* glutaminyl-tRNA synthetase consists of the first half of the dinucleotide binding domain, the acceptor binding domain, and the second half of the dinucleotide binding domain (21). The six building blocks each encoding 25 amino acids are from these domains. These six building blocks are shuffled and combined semi-randomly to yield many different structural genes in RM-PCR. In this library, every position on the block sequences should have an equal probability of encoding each building block. Therefore, there are 6^6 possibilities for a sequence composed of six building blocks. (**B**) Construction of an alternative splicing library by RM-PCR. hERα LBD consisting of 258 amino acid residues was divided into 10 building blocks at the boundaries of the exons (27) and secondary structures (28). Ten building blocks from hERα LBD were spliced alternatively at the DNA level to yield different structural genes in RM-PCR. Blue and black arrows indicate boundaries of exons and secondary structures, respectively. Yellow bars indicate α-helices.

reverse primers (DATE Concept). Double-stranded DNAs of these sequences were obtained by PCR. The PCR mixtures (50 µl) contained 20 pmol of each phosphorylated primer, 200 µM of each dNTP, 10 ng of DNA template, 2.5 U of *Pfu* DNA polymerase (Stratagene) and 5 µl of 10× *Pfu* DNA polymerase buffer. The program for the PCR was one cycle at 95°C for 5 min, followed by 30 cycles consisting of 95°C for 30 s and 56°C for 30 s. Building blocks from hERα LBD were amplified by PCR with the same program as described above using appropriate sets of phosphorylated primers and the plasmid containing structural genes of the domain. This plasmid was a gift from Dr G. L. Greene (The University of Chicago, IL) (23). The forward and reverse primers (DATE Concept) used were 5′-atgatcaaacgctctaagaagaacag-3′ (block 1f), 5′-ctcggaatagagtatgggggg-3′ (block 1r), 5′-tatgatcctaccagacccttcag-3′ (block 2f), 5′-tggcaccctcttcgccc-3′ (block 2r), 5′-ggctttgtggatttgaccctcc-3′ (block 3f), 5′-gcgccagacgagaccaatcat-3′ (block 3r), 5′-tccatggagcacccagtgaag-3′ (block 4f), 5′-gtccaagagcaagttaggagc-3′ (block 4r), 5′-aggaaccagggaaaatgtgtagag-3′ (block 5f), 5′-catgcggaaccgagatgatgta-3′ (block 5r), 5′-atgaatctgcagggagaggagttt-3′ (block 6f), 5′-agaattaagcaaaataatagatttgaggcac-3′ (block 6r), 5′-ggagtgtacacatttctgtccag-3′ (block 7f), 5′-agaattaagcaaaataatagatttgaggcac-3′ (block 6r), 5′-tgccttggccatcaggtggat-3′ (block 7r), 5′-ggcctgaccctgcagc-3′ (block 8f), 5′-catgtgcctgatgtgggagag-3′ (block 8r), 5′-agtaacaaaggcatggagcatctg-3′ (block 9f), 5′-caccacgttcttgcacttcatg-3′ (block 9r), 5′-cccctctatgacctgctgct-3′ (block 10f) and 5′-gctagtgggcgcatgtagg-3′ (block 10r).

The DNA fragments obtained were purified using Wizard PCR Preps (Promega), and then ligated to each other using T4 DNA ligase (New England Biolabs). The dimer templates were amplified by PCR using ligation products and appropriate sets of primers. The PCR mixtures (50 µl) contained 20 pmol of each primer, 200 µM of each dNTP, 10 ng of DNA template, 3 µl of 25 mM of $MgCl_2$, 2.5 U of *rTaq* DNA polymerase (TOYOBO) and 5 µl of 10× *rTaq* DNA polymerase buffer. The program for the PCR was one cycle at 95°C for 5 min, followed by 30 cycles consisting of 95°C for 30 s and 56°C for 30 s. The dimer templates amplified were purified using Wizard PCR Preps (Promega), cloned using a TOPO TA cloning kit with TOP10F′ cells (Invitrogen), and sequenced using a CEQ2000 DNA analysis system (Beckman Coulter). Dimer templates used for RM-PCR were amplified by PCR from the plasmids with correct dimer template, and recovered

from low melting agarose gels (Sigma) by phenol extraction and ethanol precipitation. The PCR mixtures (50 µl) contained 30 pmol of each primer, 200 µM of each dNTP, 10 ng of DNA template, 3.75 U of *Pfu* DNA polymerase (Stratagene) and 5 µl of 10× *Pfu* DNA polymerase buffer. The program for the PCR was one cycle at 95°C for 5 min, followed by 30 cycles consisting of 95°C for 30 s and 56°C for 30 s. Apparent concentrations of all dimer templates were estimated from the intensity of ethidium bromide-stained bands loaded on agarose gels and from the absorbance at 260 nm.

Random multi-recombinant PCR

The program of the RM-PCR for the construction of a random shuffling library was one cycle at 95°C for 5 min, followed by 20 cycles consisting of 95°C for 30 s, 60°C for 30 s and 72°C for 3 min. The reaction mixture (50 µl) contained 20 pmol of each primer, 750 fmol of dimer templates of the six building blocks, 15 fmol of dimer templates containing T7 or Ex, 5 µl of 10× Vent DNA polymerase buffer and 1 U Vent DNA polymerase (New England Biolabs). The program of the RM-PCR for the construction of alternative splicing libraries was one cycle at 95°C for 5 min, followed by 20 cycles consisting of 95°C for 30 s, 54°C for 30 s and 72°C for 1 min. The reaction mixture (50 µl) contained 20 pmol of each primer, 630 or 1260 fmol of total dimer templates, 5 µl of 10× KOD Dash DNA polymerase buffer and 1.25 U of KOD Dash DNA polymerase (TOYOBO). The RM-PCR products were concentrated by ethanol precipitation, and loaded on low melting agarose gels (Sigma). DNA bands were purified using Wizard PCR Preps (Promega), incubated with 2.5 U of *Taq* DNA polymerase (Perkin Elmer) and 200 µM dATP at 72°C for 15 min, cloned using a TOPO TA cloning KIT with TOP10F′ cells (Invitrogen) without blue-white selection to avoid bias arising from expression of cloned genes, and sequenced using a CEQ2000 DNA analysis system (Beckman Coulter).

RESULTS

Random shuffling library

Building blocks. The six building blocks used are parts of the dinucleotide binding domain and the acceptor-binding domain of *E.coli* glutaminyl-tRNA synthetase (Fig. 2A; 21,22). These building blocks consist of 25 amino acids. T7 is a 5′ consensus sequence, and has a T7 promoter sequence, a Kozak sequence (24) and an initiation codon at the 3′ end for protein *in vitro* directed evolution systems, such as the *in vitro* virus method (3) or STABLE (6). Ex is a 3′ consensus sequence designed for encoding a flexible peptide chain. Chemically synthesized DNA fragments encoding these building blocks, T7 and Ex were amplified by PCR using appropriate sets of primers.

Construction of dimer templates. Building blocks in dimer templates must be ligated in an appropriate direction. We constructed dimer templates as follows: (i) eight kinds of monomer fragments were ligated with blunt ends, resulting in 21 mixtures; (ii) 48 dimer templates were amplified by PCR from the mixtures using appropriate sets of primers. The dimer templates amplified were T7-1, T7-2, T7-3, T7-4, T7-5, T7-6, 1-1, 1-2, 1-3, 1-4, 1-5, 1-6, 2-1, 2-2, 2-3, 2-4, 2-5, 2-6, 3-1, 3-2, 3-3, 3-4, 3-5, 3-6, 4-1, 4-2, 4-3, 4-4, 4-5, 4-6, 5-1, 5-2, 5-3, 5-4, 5-5, 5-6, 6-1, 6-2, 6-3, 6-4, 6-5, 6-6, 1-Ex, 2-Ex, 3-Ex, 4-Ex, 5-Ex and 6-Ex. Because the shortest fragment flanked by two primers is amplified by PCR preferentially, dimers encoding different peptides with identical direction were obtained easily. Although dimers encoding identical peptides with opposite direction are possibly amplified by an identical primer, this did not occur under our experimental conditions. Six dimers encoding a single peptide sequence in the same direction (1-1 to 6-6) were obtained together with by-products such as monomer, trimer and/or tetramer (data not shown). The DNA bands corresponding to the dimer templates were purified from agarose gels.

Deletions or insertions resulting in frame shifts may occur at the ligation junctions, especially when blunt-end ligation is performed. Indeed, such mutations were often observed in genes generated by RM-PCR (on average, 0.2 mutations per junction). We selected the 48 dimers without mutations as templates for the RM-PCR after DNA sequencing. The 36 dimer templates of building blocks and 12 dimer templates of consensus sequences without mutations were mixed, respectively, based on the intensity of ethidium bromide-stained bands loaded on agarose gels and the absorbance at 260 nm, so that the final concentrations of all dimers in mixtures were approximately equal.

RM-PCR for the construction of a random shuffling library

RM-PCR was performed for the construction of a random shuffling library in which every position of the block sequences generated should have an equal probability of encoding each of the six building blocks. Therefore, two mixtures containing equal amounts of dimer templates of six building blocks and consensus sequences, respectively, were mixed in different ratios, and an appropriate concentration of each dimer template in each reaction mixture was obtained. Figure 3 shows the PCR products obtained under the best conditions examined, where the reaction mixture (50 µl) contained 750 fmol of 36 dimers of six building blocks, 15 fmol of 12 dimers of T7 or Ex and 1 U of Vent DNA polymerase (for details see Materials and Methods). There were >15 DNA bands on the gel, indicating the presence of PCR products with different chain lengths. The spans of these bands appeared to be 75 bp, suggesting that these bands are T7-$(block)_1$-Ex, T7-$(block)_2$-Ex, . . . , T7-$(block)_{15}$-Ex, . . . , from the smallest band. These ladder DNA bands were obtained reproducibly as shown in three lanes in Figure 3, where PCR products obtained from three independent reaction mixtures were loaded.

For further characterization of the random shuffling library obtained by the RM-PCR, the DNA band corresponding to T7-$(block)_6$-Ex was purified, cloned and sequenced. All DNA sequences analyzed are shown in Table 1. Sixty-four of 66 sequences were quite different; most of the sequences had several building blocks flanked by T7 and Ex, and the blocks were arranged in different orders. Many different multi-recombinant PCRs had thus proceeded simultaneously, and different structural genes had been constructed in single RM-PCR. The frequencies of the six building blocks were 37, 49, 40, 37, 51 and 38, so that no distinct bias among the building blocks was observed.

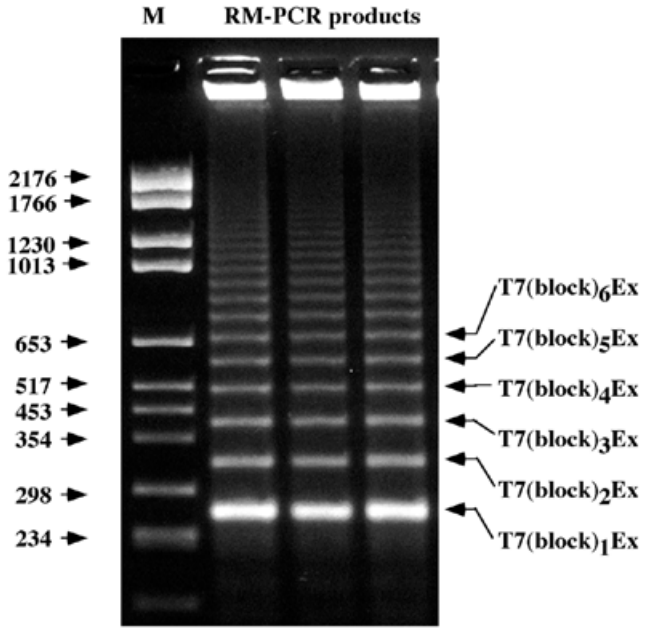


Figure 3. RM-PCR products for the random shuffling library. Each lane contained 50 µl of PCR products. M indicates a DNA size marker.

Table 1. DNA sequences given as block numbers analyzed in the random shuffling library constructed by RM-PCR

112114	245	341514	432335	525155	645465[a]
113134	245256	342	446262	526655	646161
12156	254136	343411	45325	532	646362
12465	256541	345265	455124	5336155[b]	646551
134526	261136	346452	465222	542526	651656
136564	26212	3465125[c]	466325	563656	652636
145412	2641	35354	511533	565145	6552
146521	313542	354531	514231	614231	665115
2116	3212	363262	521314	623652	665145625
216431	324355	412351	521351	625362	
213531	332612[a]	4263152[b]	523515	631234	

[a]Sequences obtained from two colonies.
[b]Sequences without Ex at the 3′ end.
[c]Sequence having a deletion between blocks 6 and 5.

Some sequences were longer or shorter than expected. This must be because the band fraction corresponding to T7-(block)$_6$-Ex contained sequences with different lengths. These sequences also had several building blocks flanked by T7 and Ex, and the blocks were arranged in different orders. Two sequences did not have Ex at the 3′ terminus, indicating that dimer templates had primed the extension by themselves, instead of the reverse primer. Nine point mutations were found among 40 299 bp. It is possible that point mutations give rise to some stop codons in genes generated by RM-PCR, but the frequency of errors (2×10^{-4}) under our experimental conditions was quite low, and would not present a problem in further selection experiments (any point mutations in the sequences analyzed did not give rise to stop codons). Only one sequence among those analyzed had a deletion, which occurred between blocks five and six. This was due to recombination with the identical sequence 'gtacgact' in these two building blocks. As found here, most of the sequences obtained by RM-PCR should not contain any insertion or deletion resulting in a frame shift or stop codon, and should have a long open reading frame suitable for directed evolution of proteins.

Alternative splicing library

Building blocks. hERα LBD (25,26) was divided into 10 building blocks at the boundaries of exons and secondary structures (27,28). The 10 building blocks encoding 34, 35, 29, 17, 26, 19, 37, 24, 17 and 20 amino acid residues (Fig. 2B) were amplified by PCR from the plasmid containing hERα LBD structural gene (23). DNA fragments termed T7OM and CBPHis were used as the 5′ and 3′ consensus sequences, and were amplified by PCR from chemically synthesized DNA. T7OM encodes T7 promoter sequence, an omega-like sequence (29), an initiation codon, and the FLAG tag (10), and CBPHis encodes the calmodulin binding peptide tag (30) and the hexa-histidine tag (31).

Dimer templates. The dimer templates required for the RM-PCR to create an alternative splicing library are only dimers encoding block i – block j (i < j), because sequences with permutations, insertions and duplications of building blocks must be excluded in an alternative splicing library.

RM-PCR for the construction of the alternative splicing library

For the initial attempt to create an alternative splicing library, two mixtures containing equal amounts of dimer templates encoding two building blocks (i.e. 1-2, 1-3, . . . , 9-10) and dimer templates encoding a consensus sequence and a building block (i.e. T7OM-1, T7OM-2, . . . , 9-CBPHis, 10-CBPHis) were mixed in different ratios (mixture A, ratio of dimer templates encoding two building blocks:dimer templates encoding a consensus sequence and a building block of 100:1; mixture B, 10:1; mixture C, 1:1) and then RM-PCR was performed. The PCR products obtained from these reaction mixtures were loaded on an agarose gel for electrophoresis, in which they showed smear bands due to the heterogeneity in the length of the building blocks (Fig. 4A). PCR products obtained from reaction mixture A seemed to encode somewhat longer open reading frames than those obtained from the other reaction mixtures. DNA bands corresponding to 500–1000 bp, which should be structural genes with more than five building blocks flanked by T7OM and CBPHis, were recovered from the gel, cloned and sequenced. DNA sequencing revealed that PCR products obtained from the reaction mixture A were 4 or 5mers lacking the 5′ or 3′ consensus sequence (the sequences were T7OM-12369, T7OM-23579, T7OM-23679, T7OM-1237, T7OM-23467, T7OM-23479, T7OM-23510 and 127810-CBPHis). DNA sequencing of the products from the mixtures B and C was subsequently performed, and six DNA sequences analyzed were 3 or 4mer structural genes flanked by 5′ and 3′ consensus sequences (the sequences were T7OM-138-CBPHis, T7OM-15710-CBPHis, T7OM-2589-CBPHis,

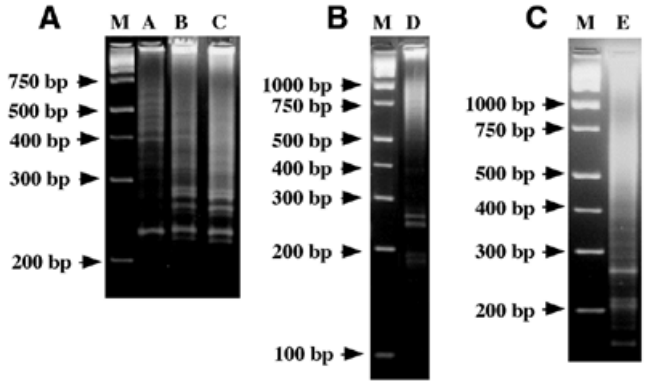
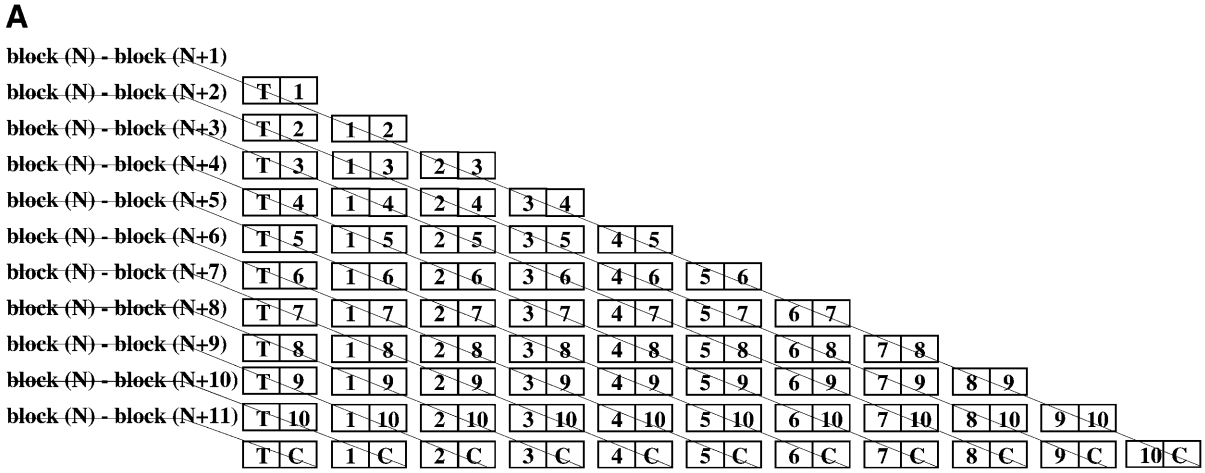


Figure 4. (**A**) RM-PCR products from reaction mixtures containing equal amounts of dimer templates (mixture A, ratio of dimer templates encoding two building blocks:dimer templates encoding a consensus sequence and a building block of 100:1; mixture B, 10:1; mixture C, 1:1). RM-PCR products from reaction mixtures where dimer templates were mixed to yield structural genes with eight (lane D in **B**) or five (lane E in **C**) building blocks. Each lane contained 50 µl of PCR products. M indicates a DNA size marker.

T7OM-126-CBPHis, T7OM-147-CBPHis and T7OM-1510-CBPHis). Thus, alternative splicing proceeded successfully at the DNA level, although most of the PCR products obtained under these conditions would be short molecules, which are probably not suitable for selection experiments. These results indicate that reaction mixtures containing equal amounts of dimer templates are not suitable for RM-PCR to create alternative splicing libraries.

All dimer templates required for the construction of the alternative splicing library can be classified into 11 classes [Fig. 5A, i.e. block (N)–block (N + 1) to block (N)–block (N + 11)]. The frequency of each dimer template in all possible sequences with a certain number of building blocks is shown in Figure 5B. For example, there are 10 kinds of structural genes encoding nine building blocks, and they are composed of dimer templates classified into block (N)–block (N + 1) and block (N)–block (N + 2). The ratio between these dimer templates is 9 (= $_9C_1$):1 (= $_8C_0$). Similarly, there are 45 kinds of structural genes with eight building blocks, and they are composed of dimer templates classified into block (N)–block (N + 1), block (N)–block (N + 2) and block (N)–block (N + 3), and the ratio between these dimer templates is 36 (= $_9C_2$):8 (= $_8C_1$):1 (= $_7C_0$). Therefore, an alternative splicing library consisting of structural genes with a certain number of building blocks, suitable for selection experiments, would be obtained by RM-PCR from



B

	10 mer	9 mer	8 mer	7 mer	6 mer	5 mer	4 mer	3 mer	2 mer
block (N) - block (N+1)	$_9C_0$	$_9C_1$	$_9C_2$	$_9C_3$	$_9C_4$	$_9C_5$	$_9C_6$	$_9C_7$	$_9C_8$
block (N) - block (N+2)	—	$_8C_0$	$_8C_1$	$_8C_2$	$_8C_3$	$_8C_4$	$_8C_5$	$_8C_6$	$_8C_7$
block (N) - block (N+3)	—	—	$_7C_0$	$_7C_1$	$_7C_2$	$_7C_3$	$_7C_4$	$_7C_5$	$_7C_6$
block (N) - block (N+4)	—	—	—	$_6C_0$	$_6C_1$	$_6C_2$	$_6C_3$	$_6C_4$	$_6C_5$
block (N) - block (N+5)	—	—	—	—	$_5C_0$	$_5C_1$	$_5C_2$	$_5C_3$	$_5C_4$
block (N) - block (N+6)	—	—	—	—	—	$_4C_0$	$_4C_1$	$_4C_2$	$_4C_3$
block (N) - block (N+7)	—	—	—	—	—	—	$_3C_0$	$_3C_1$	$_3C_2$
block (N) - block (N+8)	—	—	—	—	—	—	—	$_2C_0$	$_2C_1$
block (N) - block (N+9)	—	—	—	—	—	—	—	—	$_1C_0$

Figure 5. (**A**) Classification of dimer templates required for the construction of the alternative splicing library. Ten building blocks from hERα LBD are numbered from the N-terminal coding region. T and C represent T7OM and CBPHis, respectively. They are numbered as 0 and 11, because they are the 5′ and 3′ ends of the RM-PCR products. All dimer templates were classified into 11 classes depending on the building blocks which they encode [block (N)–block (N + 1) to block (N)–block (N + 11)]. Each diagonal line covers each type of dimer template. (**B**) Relative frequencies of different dimer templates in all possible sequences with a certain number of building blocks, are calculated as binomial coefficient ($_nC_r = n!/(n-r)!/r!$).

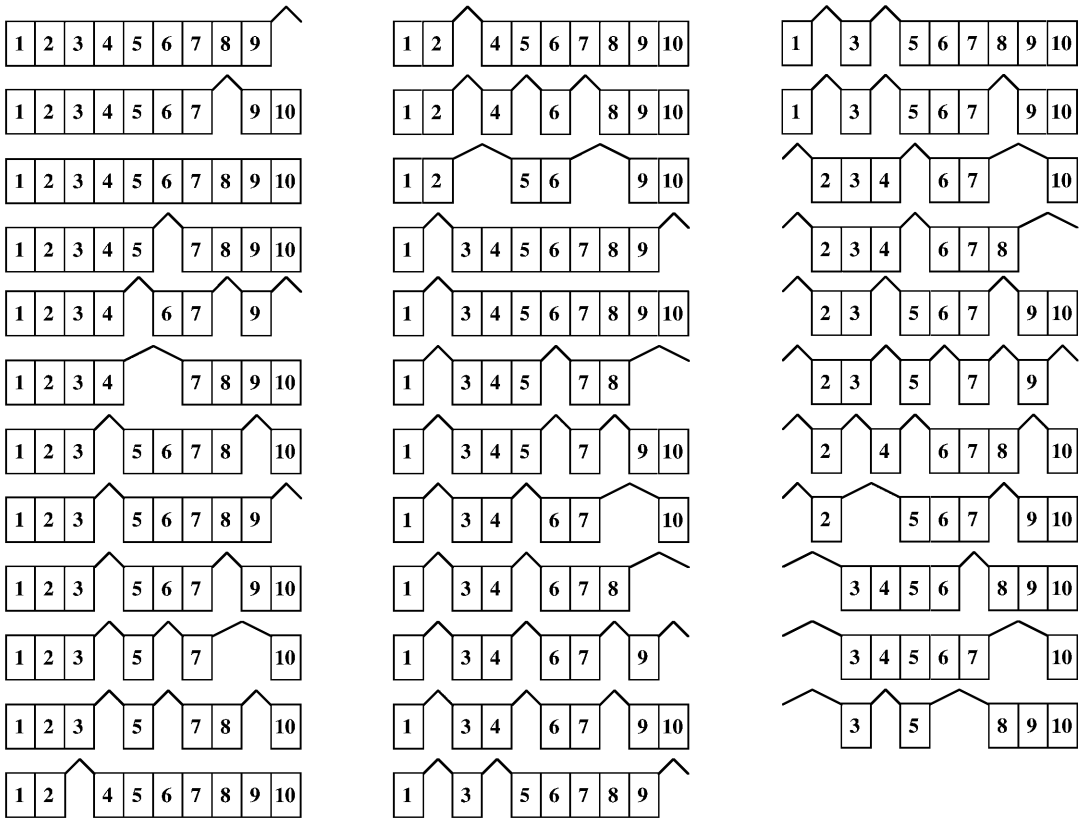


Figure 6. DNA sequences given as block numbers analyzed in the alternative splicing library constructed by RM-PCR. DNA sequences shown were obtained from reaction mixtures where dimer templates were mixed to yield structural genes with eight building blocks flanked by T7OM and CBPHis.

reaction mixtures containing appropriate classes of dimer templates in the proper ratios.

We constructed 5 and 8mer libraries in an initial attempt to examine whether libraries consisting of different sizes of open reading frames can be created by RM-PCR. Figure 4B and C show PCR products obtained from reaction mixtures where dimer templates were mixed to yield structural genes with eight or five building blocks, respectively. Strong DNA bands appeared in the regions expected to contain structural genes with the intended numbers of building blocks. DNA bands in these regions were purified, cloned and sequenced. DNA sequences obtained are shown in Figures 6 and 7, respectively. The DNA sequences shown in Figure 6 consist of five to 10 building blocks flanked by T7OM and CBPHis. The numbers of the structural genes with five, six, seven, eight, nine and 10 building blocks were 2, 11, 9, 6, 6 and 1, respectively, indicating that different structural genes with relatively long open reading frames had been obtained. Thirty-seven of 38 sequences analyzed were quite different, and only two sequences among those analyzed had unexpected recombination sites, which occurred between building blocks two and one and between building blocks five and four. These are due to recombination with the identical sequence 'atgatc' in building blocks one and two, and the similar sequences 'ccaggga' and 'ccagtga' in building blocks five and four, respectively. The DNA sequences shown in Figure 7 consist of two to five building blocks flanked by T7OM and CBPHis. The numbers of structural genes consisting of two, three, four, five and six building blocks were 1, 2, 9, 17 and 4, respectively, indicating that many structural genes of the intended length were obtained. All 33 sequences analyzed were quite different, and no products with unexpected recombination sites or mutations resulting in frame shifts were obtained.

The frequencies of the 10 building blocks in all sequences analyzed in these two libraries were 39, 30, 52, 34, 40, 43, 56, 32, 43 and 40, indicating that there was no marked bias of usage among these building blocks. Sixty point mutations were found among 47 533 bp. Again, this frequency of errors (1.3×10^{-3}) is sufficiently low that it will not be a problem for protein selection experiments (only two point mutations in all the sequences analyzed changed the codons into stop codons). As already mentioned, most of the sequences analyzed (67 of 71) did not have deletions or insertions resulting in frame shifts. Thus, different structural genes with relatively long open reading frames were successfully obtained by RM-PCR.

DISCUSSION

At the beginning of this study, we expected that several factors would bias the frequencies of the building blocks in DNA libraries created by RM-PCR. For example, they include the differences in length, melting temperature and higher-order structure of the DNA sequences, and the difficulty of estimating the actual concentration of each dimer template. However, under our experimental conditions, no distinct bias was observed among six building blocks with identical chain

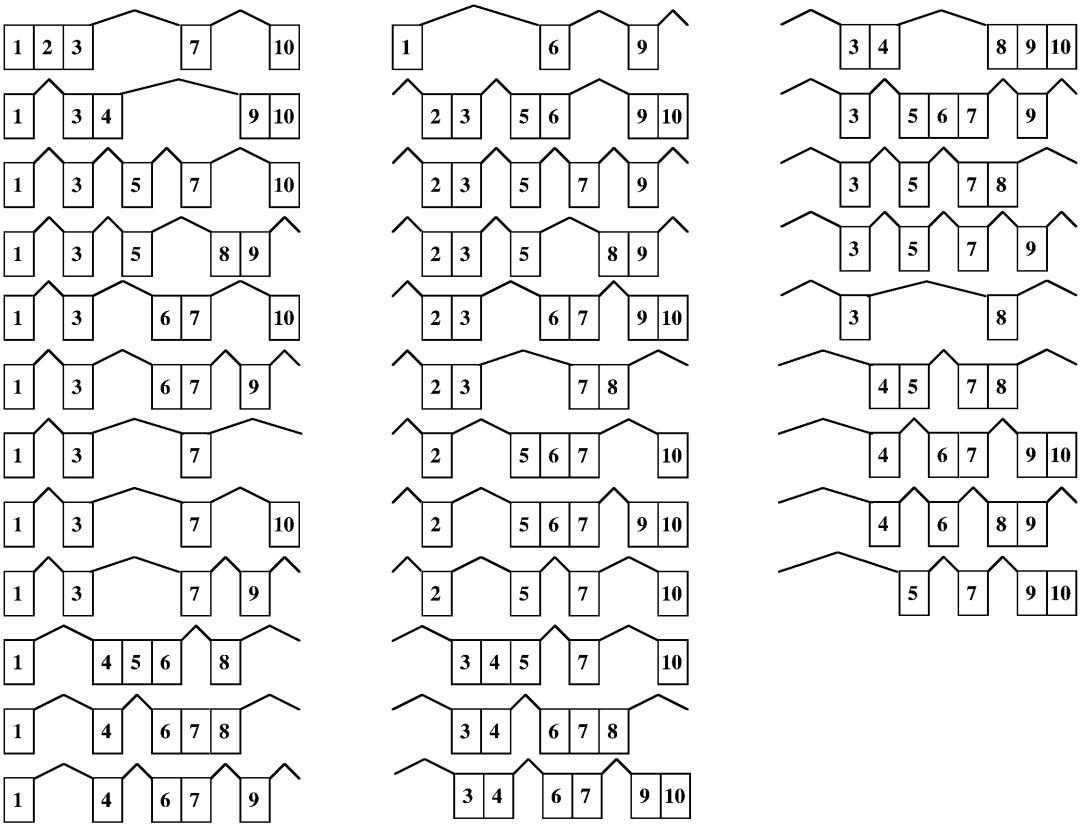


Figure 7. DNA sequences given as block numbers analyzed in the alternative splicing library constructed by RM-PCR. DNA sequences shown were obtained from reaction mixtures where dimer templates were mixed to yield structural genes with five building blocks flanked by T7OM and CBPHis.

length in the random shuffling library or among 10 building blocks with different chain lengths in two alternative splicing libraries. Some possible explanations for this result are proposed below. The splicings in RM-PCR are based on the overlapped segments that are building blocks themselves. If the melting temperature of each building block is higher than the annealing temperature in PCR, and the extension time is long enough, the bias arising from the length or melting temperature of each building block would be weakened. Next, we need to know only the relative concentrations of dimer templates, but not the actual concentrations. Although it is difficult to measure actual concentrations of dimer templates by UV absorbance because of contamination, especially when low melting agarose gel is used, it is not very difficult to determine the relative concentration of each dimer template, because all dimer templates are purified from the gels by the same procedures. Indeed, the apparent concentrations of the dimer templates estimated from the intensity of the ethidium bromide-stained bands loaded on agarose gels and the absorbance at 260 nm corresponded very well (data not shown).

The random shuffling covers global protein space that would not be accessible through point mutations alone. The sequences in the random shuffling library would have block substitutions, block duplications and/or block permutations and, therefore, every position of the block sequences generated should have an equal probability of encoding each of the six building blocks. This results in 6^6 possibilities for a sequence composed of six building blocks in the library (Fig. 2A). In addition, because longer or shorter sequences are generated simultaneously by random shuffling (Fig. 3), the theoretical maximum library size would be significantly $>6^6$. If 10 building blocks are used for the random shuffling (this requires 120 dimer templates, which could be prepared by one person in a month), the theoretical maximum library size is $>10^{10}$. However, the number of independent sequences obtained experimentally will be smaller than the number of all possible sequences obtained theoretically. If RM-PCR is performed in two separate processes, it is possible to estimate the maximum number of independent sequences that can be created experimentally. In the ideal case, cycles consisting of random annealing of dimer templates and overlap extension are repeated until every sequence is flanked by T7 and Ex, and these sequences are only amplified in the next step with primers. Thus, the complexity of the library is determined at the first step. Because all sequences are flanked by the 5′ and 3′ consensus sequences at the first step, the maximum number of independent sequences obtainable experimentally is less than the number of molecules of dimer templates encoding T7 and Ex. As 15 fmol of dimer templates of T7 and Ex are appropriate for 50 µl of RM-PCR, the maximum number of independent sequences generated is 9×10^9/50 µl under the conditions described here. Sequences encoding too long or short open reading frames that would not yield functional proteins are present in the library, diminishing the number of

independent members with appropriate lengths. However, it is difficult to estimate the number of independent sequences generated in this experiment because three reactions, random annealing of the dimer templates, overlap extension and amplification, proceed simultaneously. Although random shuffling allows us to search the global protein space, the probability of finding functional molecules from the library would be rather low. In addition, the number of independent sequences generated experimentally is very low relative to the number of molecules tested by the directed *in vitro* evolution systems (~10^{13}). Therefore, we would like to use error-prone PCR (32) to create a library with further complexity. An alternative splicing library contains ~10^2 independent candidate functional proteins. This library, together with mutagenesis of individual amino acids, could be used to create downsized proteins. Another use of RM-PCR is to construct '*in vitro* exon shuffling' libraries, as described by Kolkman and Stemmer (33). They proposed an alternative method for combinatorial reassembly of gene fragments without sequence homology. They used chimeric oligonucleotides encoding parts of two domains as primers to amplify targeted sequences, and the amplified sequences (termed 'pre-made PCR fragments') were then mixed and used as both primers and templates in a PCR-like reaction without the primers that reassemble the full length genes. The pre-made PCR fragments seem to be more conveniently obtainable than the dimer templates prepared by our protocol. However, the usefulness of their method cannot be compared with that of RM-PCR, because experimental data and detailed protocols relating to their experimentally created library have not yet been reported.

In our previous paper (20,34), the foldability and RNase activity of 45 barnase mutants obtained by the permutation of modules or secondary structure units were investigated. This study suggests that amino acids not involved directly in catalysis in foldable proteins have been selected as scaffolds to stabilize the active site with appropriate conformations in the course of globular protein evolution. The mutant having a nascent active site, but not a stable scaffold, can be considered as globular protein in an 'evolutionarily intermediate state' (34). Theoretical studies have successfully simulated the trajectory in which evolution of stability around the active site leads to a sequence which folds globally into a conformation (35,36).

Several proteins in evolutionarily intermediate states should be present in a protein library constructed by RM-PCR using different building blocks from identical or different proteins. Although these premature proteins would contain unfavorable interactions between building blocks, mutagenesis at the level of individual amino acids by error-prone PCR or DNA shuffling (18) could be used to modify such interactions and to evolve proteins in evolutionarily intermediate states into mature proteins. This strategy is very different from conventional strategies for protein engineering such as site-directed mutagenesis or error-prone PCR, because they alter only fine surface structures of natural foldable proteins, but not the global folds of the proteins. To our knowledge, only one experimental study has observed the evolution of both foldability and function simultaneously (37). In that study, a targeted enzyme with a functional loop substituted by a loop from another enzyme with similar fold was aggregation-prone at an initial stage, and was successfully evolved to a soluble and functional enzyme by further directed evolution using DNA shuffling.

A recent theoretical study by Bogarad and Deem (38) suggests that point mutations and/or DNA shuffling are incapable of evolving new protein fold, whereas non-homologous DNA 'swapping' of low-energy structures is a possible strategy to find new protein folds in the global protein space. RM-PCR can be used to achieve non-homologous DNA swapping among several building blocks. However, it is not clear whether swapping of building blocks with low-energy structures is the best way for the construction of new foldable and/or functional proteins, because parts of folded proteins are stabilized by both local and non-local interactions, and the relative contributions of these interactions are different from protein to protein. Therefore, we wished to investigate experimentally what types of building blocks are appropriate to create new proteins. RM-PCR can be used for shuffling of many different building blocks, such as modules, secondary structures with a low-energy conformation, secondary structures capable of taking different conformations by adapting themselves to the surrounding environment (39), functional motifs and so on. RM-PCR, which permits the shuffling of different types of building blocks without homologous sequences, will be a useful method for protein engineering to create new proteins and as a tool in protein science to understand relationships between folding, function and evolution of globular proteins.

ACKNOWLEDGEMENTS

The authors thank Prof. G. L. Greene (The University of Chicago, IL) for providing the plasmid with hERα LBD and Dr Tsuge (Mitsubishi Kagaku Institute of Life Sciences, Tokyo, Japan) for helpful discussions. This research was supported by the New Energy and Industrial Technology Development Organization (NEDO).

REFERENCES

1. Keefe,A.D. and Szostak,J.W. (2001) Functional proteins from a random-sequence library. *Nature*, **410**, 715–718.
2. Wilson,D.S., Keefe,A.D. and Szostak,J.W. (2001) The use of mRNA display to select high-affinity protein-binding peptides. *Proc. Natl Acad. Sci. USA*, **98**, 3750–3755.
3. Nemoto,N., Miyamoto-Sato,E., Hushimi,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.
4. Roberts,R.W. and Szostak,J.W. (1997) RNA-peptide fusions for the *in vitro* selection of peptides and proteins. *Proc. Natl Acad. Sci. USA*, **94**, 12297–12302.
5. Hanes,J. and Pluckthun,A. (1997) *In vitro* selection and evolution of functional proteins by using ribosome display. *Proc. Natl Acad. Sci. USA*, **94**, 4937–4942.
6. Doi,N. and Yanagawa,H. (1999) STABLE: protein–DNA fusion system for screening of combinatorial protein libraries *in vitro*. *FEBS Lett.*, **457**, 227–230.
7. 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.
8. Mandecki,W. (1990) A method for construction of long randomized open reading frames and polypeptides. *Protein Eng.*, **3**, 221–226.
9. 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.

10. Cho,G., Keefe,A.D., Liu,R., Wilson,D.S. and Szostak,J.W. (2000) Constructing high complexity synthetic libraries of long ORFs using *in vitro* selection. *J. Mol. Biol.*, **297**, 309–319.
11. Kamtekar,S., Schiffer,J.M., Xiong,H., Babik,J.M. and Hecht,M.H. (1993) Protein design by binary patterning of polar and nonpolar amino acids. *Science*, **262**, 1680–1685.
12. Xu,G., Wang,W., Groves,J.T. and Hecht,M.H. (2001) Self-assembled monolayers from a designed combinatorial library of *de novo* beta-sheet proteins. *Proc. Natl Acad. Sci. USA*, **98**, 3652–3657.
13. Go,M. and Nosaka,M. (1987) Protein architecture and the origin of introns. *Cold Spring Harb. Symp. Quant. Biol.*, **52**, 915–924.
14. Yanagawa,H., Yoshida,K., Torigoe,C., Park,J.S., Sato,K., Shirai,T. and Go,M. (1993) Protein anatomy: functional roles of barnase module. *J. Biol. Chem.*, **268**, 5861–5865.
15. Yoshida,K., Shibata,T., Masai,J., Sato,K., Noguti,T., Go,M. and Yanagawa,H. (1993) Protein anatomy: spontaneous formation of filamentous helical structures from the N-terminal module of barnase. *Biochemistry*, **32**, 2162–2166.
16. Gilbert,W. (1978) Why genes in pieces? *Nature*, **271**, 501.
17. de Souza,S.J., Long,M., Schoenbach,L., Roy,S.W. and Gilbert,W. (1996) Intron positions correlate with module boundaries in ancient proteins. *Proc. Natl Acad. Sci. USA*, **93**, 14632–14636.
18. Stemmer,W.P.C. (1994) DNA shuffling by random fragmentation and reassembly: *in vitro* recombination for molecular evolution. *Proc. Natl Acad. Sci. USA*, **91**, 10747–10751.
19. Ostermeier,M., Shim,J.H. and Benkovic,S.J. (1999) A combinatorial approach to hybrid enzymes independent of DNA homology. *Nat. Biotechnol.*, **17**, 1205–1209.
20. 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.
21. Rould,M.A., Perona,J.J., Soll,D. and Steitz,T.A. (1989) Structure of *E.coli* glutaminyl-tRNA synthetase complexed with tRNA (Gln) and ATP at 2.8 Å resolution. *Science*, **246**, 1135–1142.
22. Yamao,F., Inokuchi,H., Cheung,A., Ozeki,H. and Soll,D. (1982) *Escherichia coli* glutaminyl-tRNA synthetase. I. Isolation and DNA sequence of the glnS gene. *J. Biol. Chem.*, **257**, 11639–11643.
23. Shiau,A.K., Barstad,D., Loria,P.M., Cheng,L., Kushner,P.J., Agard,D.A. and Greene,G.L. (1998) The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell*, **95**, 927–937.
24. Kozak,M. (1983) Comparison of initiation of protein synthesis in procaryotes, eucaryotes and organelles. *Microbiol. Rev.*, **47**, 1–45.
25. Greene,G.L., Gilna,P., Waterfield,M., Baker,A., Hort,Y. and Shine,J. (1986) Sequence and expression of human estrogen receptor complementary DNA. *Science*, **231**, 1150–1154.
26. Brzozowski,A.M., Pike,A.C., Dauter,Z., Hubbard,R.E., Bonn,T., Engstrom,O., Ohman,L., Greene,G.L., Gustafsson,J.A. and Carlquist,M. (1997) Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature*, **389**, 753–758.
27. Ponglikitmongkol,M., Green,S. and Chambon,P. (1988) Genomic organization of the human oestrogen receptor gene. *EMBO J.*, **7**, 3385–3388.
28. Tanenbaum,D.M., Wang,Y., Williams,S.P. and Sigler,P.B. (1998) Crystallographic comparison of the estrogen and progesterone receptor's ligand binding domains. *Proc. Natl Acad. Sci. USA*, **95**, 5998–6003.
29. Gallie,D.R and Walbot,V. (1992) Identification of the motifs within the tobacco mosaic virus 5′-leader responsible for enhancing translation. *Nucleic Acids Res.*, **20**, 4631–4638.
30. Zheng,C.F., Simcox,T., Xu,L. and Vaillancourt,P. (1997) A new expression vector for high level protein production, one step purification and direct isotopic labeling of calmodulin-binding peptide fusion proteins. *Gene*, **186**, 55–60.
31. Porath,J., Carlsson,J., Olsson,I. and Belfrage,G. (1975) Metal chelate affinity chromatography, a new approach to protein fractionation. *Nature*, **258**, 598–599.
32. Matsumura,I., Wallingford,J.B., Surana,N.K., Vize,P.D. and Ellington,A.D. (1999) Directed evolution of the surface chemistry of the reporter enzyme β-glucuronidase. *Nat. Biotechnol.*, **17**, 696–701.
33. Kolkman,J.A. and Stemmer,W.P.C. (2001) Directed evolution of proteins by exon shuffling. *Nat. Biotechnol.*, **19**, 423–428.
34. Tsuji,T., Kobayashi,K. and Yanagawa,H. (1999) Permutation of modules or secondary structure units creates proteins with basal enzymatic properties. *FEBS Lett.*, **453**, 145–150.
35. Saito,S., Sasai,M. and Yomo,T. (1997) Evolution of the folding ability of proteins through functional selection. *Proc. Natl Acad. Sci. USA*, **94**, 11324–11328.
36. Yomo,T., Saito,S. and Sasai,M. (1999) Gradual development of protein-like global structures through functional selection. *Nature Struct. Biol.*, **6**, 743–746.
37. Altamirano,M.M., Blackburn,J.M., Aguayo,C. and Fersht,A.R. (2000) Directed evolution of new catalytic activity using the α/β-barrel scaffold. *Nature*, **403**, 617–622.
38. Bogarad,L.D. and Deem,M.W. (1999) A hierarchical approach to protein molecular evolution. *Proc. Natl Acad. Sci. USA*, **96**, 2591–2595.
39. Minor,D.L.,Jr and Kim,P.S. (1996) Context-dependent secondary structure formation of a designed protein sequence. *Nature*, **380**, 730–734.