Published online 7 February 2007 *Nucleic Acids Research, 2007, Vol. 35, No. 6* ***e38***
doi:10.1093/nar/gkm017

Motif programming: a microgene-based method for creating synthetic proteins containing multiple functional motifs

Hirohide Saito[1,2], Tamiko Minamisawa[1] and Kiyotaka Shiba[1,2,*]

[1]Department of Protein Engineering, Cancer Institute, Japanese Foundation for Cancer Research, Koto-ku, Tokyo 135-8550, Japan and [2]CREST, Japan Science and Technology Agency (JST), c/o Cancer Institute

Received October 18, 2006; Revised December 20, 2006; Accepted January 2, 2007

ABSTRACT

The presence of peptide motifs within the proteins provides the synthetic biologist with the opportunity to fabricate novel proteins through the programming of these motifs. Here we describe a method that enables one to combine multiple peptide motifs to generate a combinatorial protein library. With this method, a set of sense and antisense oligonucleotide primers were prepared. These primers were mixed and polymerized, so that the resultant DNA consisted of combinatorial polymers of multiple microgenes created from the stochastic assembly of the sense and antisense primers. With this motif-mixing method, we prepared a protein library from the BH1-4 motifs shared among Bcl-2 family proteins. Among the 41 clones created, 70% of clones had a stable, presumably folded expression product in human cells, which was detectable by immunohistochemistry and western blot. The proteins obtained varied with respect to both the number and the order of the four motifs. The method enables homology-independent polymerization of DNA blocks that coded motif sequences, and the frequency of each motif within a library can be adjusted in a tailor-made manner. This motif programming has a potential for creating a library with a large proportion of folded/functional proteins.

INTRODUCTION

Investigators in the emerging field of synthetic biology seek an understanding of biological systems that would be difficult to achieve using more conventional approaches and to construct novel systems and biomacromolecules that exhibit unparalleled behaviors, potentially leading to the development of new technologies (1). The artifacts synthesized include biomolecules (proteins, DNA or RNA) (2,3), replicators (4), gene circuits (5,6) and genomes (7), all of which have complex and dynamic structures composed of rather simpler block units. Natural versions of these entities are known to have been organized to their existing forms over the course of billions of years of evolution. Our knowledge about the principles governing this organizing process is limited, however, which is why the rational design of biological systems is still at a very rudimentary stage, and which is why the act of synthesis can deepen our understanding of the self-organizing principles involved.

In the *in vitro* evolution experiments that were conducted in the early 1990s, synthetic molecules were created by adopting a combinatorial approach (8) in which blocks of nucleic acids or amino acids were randomly assembled to prepare pools of random sequences, and functional molecules were selected from those pools. Although numerous nucleic acids-based enzymes have been created using such random sequence approaches, the emergence of novel proteins has been limited to a few examples (9). In contrast, ‘exon-shuffling type’ or ‘constrained’ DNA libraries constructed using biased sequences have proven to be more effective than random sequences for generating artificial proteins. The numerous approaches that have made use of biased libraries (10–16) can be generally classified into two groups: those requiring short DNA sequences that are commonly shared amongst DNA blocks for recombination and those enabling homology-independent polymerization. Homology-independent recombination of molecular units has proven to be a powerful tool for evolving novel structures and functions (12–16). Ideally, with this approach, the reaction conditions should be simple, a desired number of non-homologous genes should be used to

*To whom correspondence should be addressed. Tel: +81 3 3570 0489; Fax: +81 3 3570 0461; Email: kshiba@jfcr.or.jp
Present address:
Hirohide Saito, Laboratory of Gene Biodynamics, Graduate School of Biostudies, Kyoto University, Oiwake-cho, Kitashirakawa, Sakyo-ku, Kyoto 606-8502, Japan

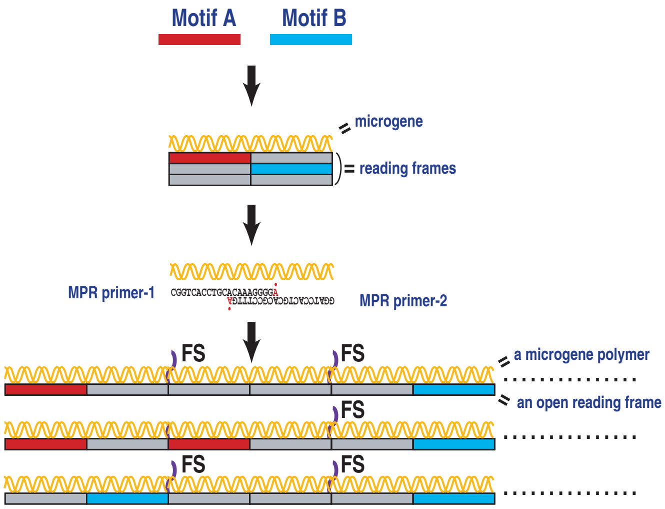

Figure 1. Single microgene-based method. A single microgene block has motif A (red) and motif B (cyan) embedded in different reading frames. The microgene (shown by the double helix) is polymerized in tandem using a microgene polymerization reaction (MPR) in which two overlapping MPR primer pairs having a 3′ mismatched base (shown by red letters with dots) are used. During MPR, insertion and/or deletion mutations randomly occur at junctions between microgene blocks (shown by wavy lines labeled 'FS') causing the reading frame to randomly shift, leading to combinatorial polymerization of motifs A and B.

construct the libraries, and the libraries should be user programmable.

We previously established a simple protein evolution system in which artificial proteins were synthesized from the combinatorial polymerization of peptide motifs (17–20). In our original protocol, motifs to be mixed were first embedded within different reading frames encoded by a single short DNA sequence, a microgene (Figure 1). Next, from the designer microgene, a pair of MPR (microgene polymerization reaction) primers was synthesized such that (i) the pair contained complementary bases in their 3′ region; (ii) the sequence of the primer dimer obtained from the elongation reaction re-created the microgene block; and (iii) the primers had mismatched bases at their 3′-OH ends (shown by red letters with dots in Figure 1). These mismatched nucleotides at the 3′-OH ends of MPR primers are critical for successful polymerization of the microgene (17). During MPR, the primer pairs are reacted under conditions similar to PCR—i.e. a thermal cycle reaction is repeated in the presence of a thermostable archaeal DNA polymerase and dNTP (but without any template DNA). Without the 3′-OH end mismatch, the primer dimer that corresponds to one unit of the microgene would be amplified; including the 3′-OH end mismatch in the MPR primers and using a DNA polymerase having 3′–5′ exonuclease activity enables large DNAs consisting of tandem repeats of the microgene to be synthesized. Moreover, nucleotide insertions and deletions randomly occur at end-joining junctions between microgenes, resulting in synthesis of combinatorial libraries of the three reading frames (motifs) from a single microgene. Although the detailed mechanism of MPR is not yet known, the reaction is apparently related to illegitimate recombination in which double-strand breaks in the DNA are joined. It is noteworthy in that regard that DNA polymerases such as the Klenow fragment of DNA polymerase I and Taq polymerase can serve as 'alignment proteins' that juxtapose the two DNA ends so that DNA synthesis can proceed on discontinuously aligned DNA (21).

Using this MPR method, we have previously reported that translations of microgene polymers, in which α-helix or β-sheet forming peptides were encrypted, produced proteins having secondary structures (18). In addition, we have demonstrated that bifunctional proteins that penetrate through cell membranes and exert a pro-apoptotic effect can be generated by combinatorially polymerizing two short peptide motifs respectively related to induction of apoptosis and protein transduction. Because simple linkage of these motifs was not sufficient to create a bifunctional peptide, and the successful reconstitution was dependent on how these motifs were joined together, the combinatorial polymerization strategy was shown to be important for reconstitution of function from mixtures of short sequence motifs (20). The original MPR method had an inherent limitation, however: the motif number was restricted to the three reading frames, and the creation of any combinatorial library was dependent on a randomly occurring frameshift (Figure 1). Because combinatorics created in this way originate from frame shift mutations that randomly occur at junctions of microgene polymers, the motifs to be mixed are embedded in different reading frames of the microgene. Consequently, the number of motifs that can be mixed is limited to three, the number of reading frames. In addition, random switching between reading frames is indispensable to the creation of combinatorics in this protocol (17–20).

Here we describe a new motif-mixing protocol (outlined in Figure 2) that overcomes these inherent limitations and enables polymerization of more than three motifs. With this method, segments of microgenes (microgenes$_{\mathrm{core}}$) are first designed so that they encode peptide motifs. Thereafter, multiple MPR sense and antisense primers are designed based on these microgenes$_{\mathrm{core}}$. Because these primers have sequences that allow formation of base pairs in their 3′ regions, they can stochastically re-create multiple microgenes, so that when MPR is carried out, combinatorial polymers of multiple microgenes are generated. We applied this method to construct a combinatorial protein library from mixtures of four different peptide motifs (BH1–4). These short peptide sequences are conserved among Bcl-2 family proteins, which constitute a critical checkpoint in the intracellular signaling network regulating the process of mitochondria-dependent apoptosis (22,23). The proteins obtained with different length (68–250 amino acids) varied with respect to both the number and order of the four motifs. The frequency and the ratio of each motif within a library could be controlled in a tailor-made manner. We also demonstrated that these proteins were effectively expressed in human cells, and localized in the mitochondria in some cases.

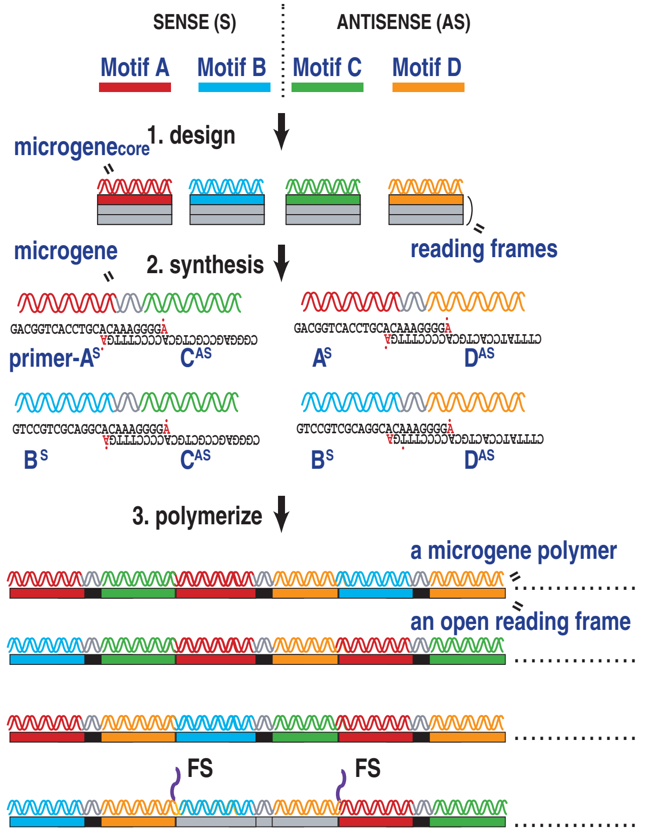

Figure 2. Schematic diagram of a motif-mixing protocol used in this study. Initially, we designed DNA sequences for microgenes$_{core}$ that each encode a peptide motif to be mixed in their first reading frames, after which sense and antisense MPR primers were synthesized based on these microgenes$_{core}$. These primers share 3′ sequences that enable base-pair formation between the sense and antisense primers, but contain mismatched bases at their 3′-OH ends (shown by red letters with dots). In the polymerization step, motifs can be embedded either in the sense or antisense primer. In the figure, motifs A and B are embedded in the sense primers, producing primers A^S and B^S, while motifs C and D are in the antisense primers, producing primers C^{AS} and D^{AS}. The thermal cycle reaction is carried out in the presence of these MPR primers, a thermostable, a DNA polymerase and dNTP. The resultant high molecular weight DNAs are combinatorial polymers of multiple microgenes created by stochastic base paring of the MPR primers. In some clones, nucleotide insertions or deletions allow frame shift mutations (denoted by FS), so that peptide sequences encoded by the second and third reading frames appear in the translated products.

MATERIALS AND METHODS

Library construction

Microgenes were designed so that they fulfilled the following rules: (i) none contained translation termination codons in any of their three reading frames; (ii) codons were chosen so that the microgenes would code for peptides having a propensity to form α helix in the third reading frame; and (iii) all had similar GC contents (55–65%). For the experiments schematically depicted in Figure 3, 0.4 μM sense- and antisense MPR primers, four dNTP (0.35 mM each) and 2.6 units of 3′–5′ *exo*$^+$Vent DNA polymerase (New England Biolabs) were mixed in reaction buffer (10 mM KCl, 10 mM $(NH_4)_2SO_4$, 20 mM Tris-HCl, 2 mM $MgSO_4$, 0.1% TritonX-100, pH 8.8).

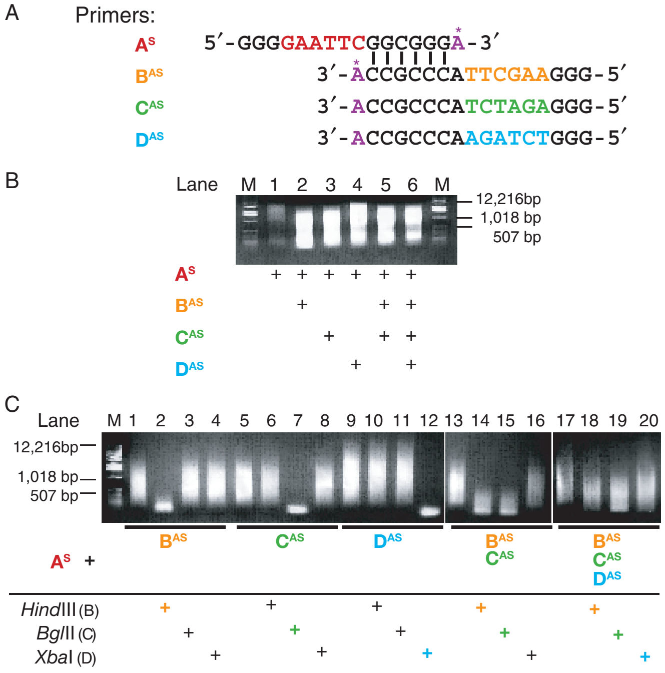

Figure 3. Mixing four short DNAs using the new protocol. (**A**) MPR primers A^S (sense), B^{AS}, C^{AS} and D^{AS} (all antisense) were designed so that they contained a restriction endonuclease recognition site (red, orange, green or cyan). The sense and antisense primers overlap at the 3′-region, forming six base pairs (shown by black bars). All of the primers had a mismatched adenosine residue at the 3′-OH end (represented by a purple A with asterisks). (**B**) Polymerization of four microgenes. The four MPR primers described in (A) were polymerized to yield a large DNA fragment; note that a pair of MPR primers (both sense and antisense) was necessary for the polymerization (compare lane 1 with lanes 2–6). The sense primer A^S (0.4 μM) and antisense primers (B^{AS}, C^{AS} and D^{AS}; 0.4 μM total) were mixed as shown in the figure. (**C**) The MPR products obtained were digested with the indicated restriction enzymes and analyzed by gel electrophoresis (1% agarose). The primers used and their concentrations were same as in (B): lanes 1–4, A^S (0.4 μM) + B^{AS} (0.4 μM); lanes 5–8, A^S (0.4 μM) + C^{AS} (0.4 μM); lanes 9–12, A^S (0.4 μM) + D^{AS} (0.4 μM); lanes 13–16, A^S (0.4 μM) + B^{AS} (0.2 μM) + C^{AS} (0.2 μM); lanes 17–20, A^S (0.4 μM) + B^{AS} (0.134 μM) + C^{AS} (0.134 μM) + D^{AS} (0.134 μM); lane M, size standards.

The thermal program was initiated with 10 min at 94°C; polymerization was accomplished with 40 cycles of 94°C for 10 s and 55°C for 1 min, and terminated with 7 min at 69°C. The resultant microgene polymers were analyzed by agarose (1%) gel electrophoresis. Similar conditions were used for the experiments in Figure 4 and Table 1 except that cycling protocol included 1 min at 72°C instead of 55°C. To monitor the incorporation of motifs into the polymers, MPR products were digested with appropriate restriction enzymes (see the Results section) and analyzed by agarose (1%) gel electrophoresis.

Cloning and expression of microgene polymers

Microgene polymers were directly cloned into the pcDNA3.1 directional TOPO expression vector (Invitrogen), which enables the directional ligation of DNA fragments with CACC tetranucleotides at their 5′ end. This vector was then used to initiate translation

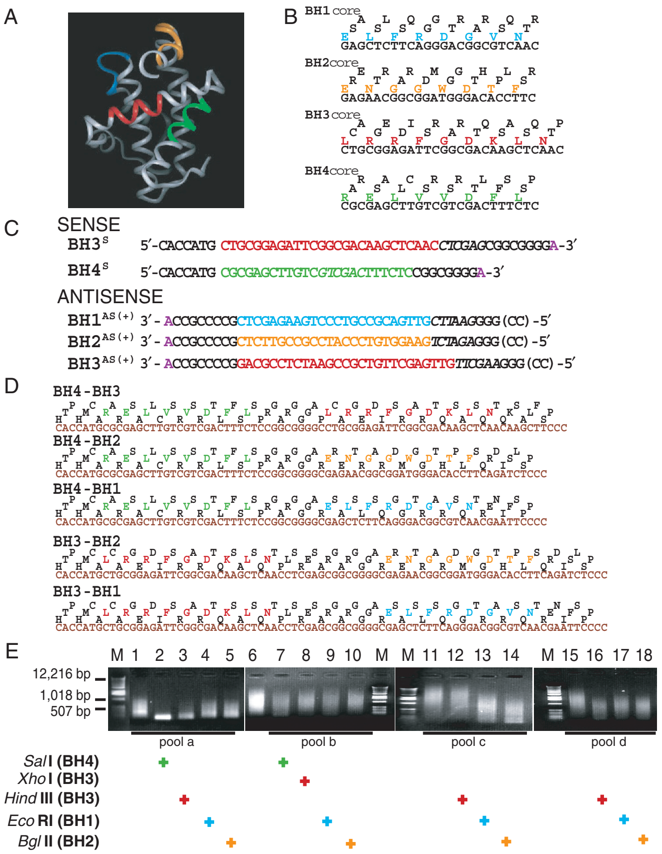
A
B
BH1 core
A L Q G R R Q R
S S S G T A S T
E L F R D G V N
GAGCTCTTCAGGGACGGCGTCAAC
BH2core
E R R M G H L R
R T A D G T P S
E N G G W D T F
GAGAACGGCGGATGGGACACCTTC
BH3 core
A E I R R Q A Q P
C G D S A T S S T
L R R F G D K L N
CTGCGGAGATTCGGCGACAAGCTCAAC
BH4core
R A C R R L S P
A S L S S T F S
R E L V V D F L
CGCGAGCTTGTCGTCGACTTTCTC
C
SENSE
BH3^S 5′-CACCATG CTGCGGAGATTCGGCGACAAGCTCAACCTCGAGCGGCGGGGA-3′
BH4^S 5′-CACCATG CGCGAGCTTGTCGTCGACTTTCTCCGGCGGGGA-3′
ANTISENSE
BH1^AS(+) 3′-ACCGCCCCGCTCGAGAAGTCCCTGCCGCAGTTGCTTAAGGGG(CC)-5′
BH2^AS(+) 3′-ACCGCCCCGCTCTTGCCGCCTACCCTGTGGAAGTCTAGAGGG(CC)-5′
BH3^AS(+) 3′-ACCGCCCCGGACGCCTCTAAGCCGCTGTTCGAGTTGTTCGAAGGG(CC)-5′
D
BH4-BH3
CACCATGCGCGAGCTTGTCGTCGACTTTCTCCGGCGGGGCCTGCGGAGATTCGGCGACAAGCTCAACAAGCTTCCC
BH4-BH2
CACCATGCGCGAGCTTGTCGTCGACTTTCTCCGGCGGGGCGAGAACGGCGGATGGGACACCTTCAGATCTCCC
BH4-BH1
CACCATGCGCGAGCTTGTCGTCGACTTTCTCCGGCGGGGCGAGCTCTTCAGGGACGGCGTCAACGAATTCCCC
BH3-BH2
CACCATGCTGCGGAGATTCGGCGACAAGCTCAACCTCGAGCGGCGGGGCGAGAACGGCGGATGGGACACCTTCAGATCTCCC
BH3-BH1
CACCATGCTGCGGAGATTCGGCGACAAGCTCAACCTCGAGCGGCGGGGCGAGCTCTTCAGGGACGGCGTCAACGAATTCCCC
E
M 1 2 3 4 5 6 7 8 9 10 M M 11 12 13 14 M 15 16 17 18
12,216 bp
1,018 bp
507 bp
pool a
pool b
pool c
pool d
Sal I (BH4)
Xho I (BH3)
Hind III (BH3)
Eco RI (BH1)
Bgl II (BH2)

Figure 4. Mixing four BH motifs. (**A**) Structure of human Bcl-xL, an anti-apoptotic multidomain protein (based on the [1R2D]). The core regions of the BH1, BH2, BH3 and BH4 motifs focused on in this study are shown in cyan, orange, red and green, respectively, though we used the BH3 motif from Noxa instead of Bcl-xL. (**B**) Core microgenes that encode BH peptides in their first reading frames; the color scheme is the same as in (A). (**C**) MPR primers used in this study. CACC tetranucleotides that facilitated directional cloning into pcDNA were added at the 5′ end of the sense primers ($BH3^S$ and $BH4^S$). Six base pairs were formed in the 3′-regions between the sense and antisense primers. They also have mismatched adenosines at their 3′-OH ends (shown in purple). Derivatives of antisense primers ($BH1^{AS+}$, $BH2^{AS+}$ and $BH3^{AS+}$) had an extra CC at their 5′ termini so that the reconstituted microgenes would have lengths that were multiples of three. To monitor incorporation of the corresponding blocks into polymers, the recognition sequences for SalI, XhoI, HindIII, BglII and EcoRI were introduced into $BH4^S$, $BH3^S$, $BH3^{AS(+)}$, $BH2^{AS(+)}$ and $BH1^{AS(+)}$, respectively (shown by italics). (**D**) The designed microgenes used in this study. They generated the $BH1_{core}$–$BH4_{\ core}$ peptide motifs, which consisted of arranged blocks. (**E**) Microgenes polymers prepared from combinations of the primers shown in Table 1. The DNA polymers obtained from pools a–d were digested with the motif-specific restriction enzymes and electrophoresed through 1% agarose.

Table 1. Libraries made from motif programming

Pool	Sense primer(s)	Antisense primers	concentrations (μM)
a*	$BH4^S$	$BH1^{AS}$, $BH2^{AS}$, $BH3^{AS}$	0.4:0.008:0.004:0.4
b	$BH4^S$, $BH3^S$	$BH1^{AS}$, $BH2\ ^{AS}$	0.04:0.4:0.16:0.04
c	$BH4^S$	$BH1^{AS+}$, $BH2^{AS+}$, $BH3^{AS+}$	0.4:0.14:0.14:0.14
d	$BH4^S$	$BH1^{AS+}$, $BH2^{AS+}$, $BH3^{AS+}$	0.4:0.08:0.06:0.4

Four pools were prepared using different combinations of MPR primers.
*See the Materials and methods section also.

at an initiation codon (ATG) located at position 5–7 of the first microgene unit within each polymer. The ligated plasmids were then used to transform *E. coli* TOP-10 cells (Invitrogen), after which the cloned sequences were determined using a CEQ2000XL DNA analyzer (Beckman). We noticed that clones in pool-a unintentionally contained sequences derived from the antisense primer $BH3^{Bcl\text{-}xL}$-AS (5′-GGGAAGCTTGAATTCGTC GCCGGCTTCGCGCAAGCCCCGCCA-3′) (5 out of 13 clones). However, the embedded BH3 motif from Bcl-xL only appeared in clone a11; other clones translated the second and/or third reading frames. The sequenced microgene polymers were then cut from the original vector using BamHI and EcoRV, and sub-cloned into one of the three vectors (pcDNA 3.1/myc-His A, B or C; Invitrogen) to add a myc epitope and a poly-histidine tag at the C-terminal ends of the microgene products, and the resultant plasmids were transfected into MCF-7 cells using lipofectamine 2000, according to the manufacturer's instructions (Invitrogen).

Cell lines

MCF-7 (a human breast cancer line) cells were cultured in RPMI 1640 (GIBCO) supplemented with 10% fetal bovine serum (FBS, Morigate) and antibiotic/antimycotic solution (SIGMA, A5955) at 37°C in humidified air containing 5% CO_2.

Immunohistochemical analysis

For the immunohistochemical analysis in Figure 5C, cells were fixed in methanol for 10 min at room temperature. The fixed cells were washed by PBS and then incubated in blocking solution (PBS with 10% goat serum) for 1 h at room temperature. The cells were then incubated for 1 h at 37°C with anti-penta-his antibody with Alexa Fluor 488 conjugate (1:200, QIAGEN). Stained samples were examined using a confocal laser scanning microscope.

RESULTS

Outline of motif programming

The procedure for mixing more than three motifs explored in this study consists of three processes (Figure 2). In the first two processes $microgenes_{core}$ that each encode a peptide motif (Motif A–D) in its first reading frames were *designed*, after which sense and antisense MPR primers were *synthesized* based on those $microgenes_{core}$ [in Figure 2, Motifs A and B were embedded in the sense primer (A^S and B^S), while motifs C and D were embedded in the antisense primers (C^{AS} and D^{AS})]. These primers also contained additional sequences at their 3′ ends that allowed formation of base pairs between sense and antisense primers, but contained mismatched bases at their 3′-OH ends (indicated by red letters with dots). In the third process, thermal cycling was carried out with the MPR primers, a thermostable DNA polymerase having 3′–5′ exonuclease activity and dNTP, which was shown to efficiently

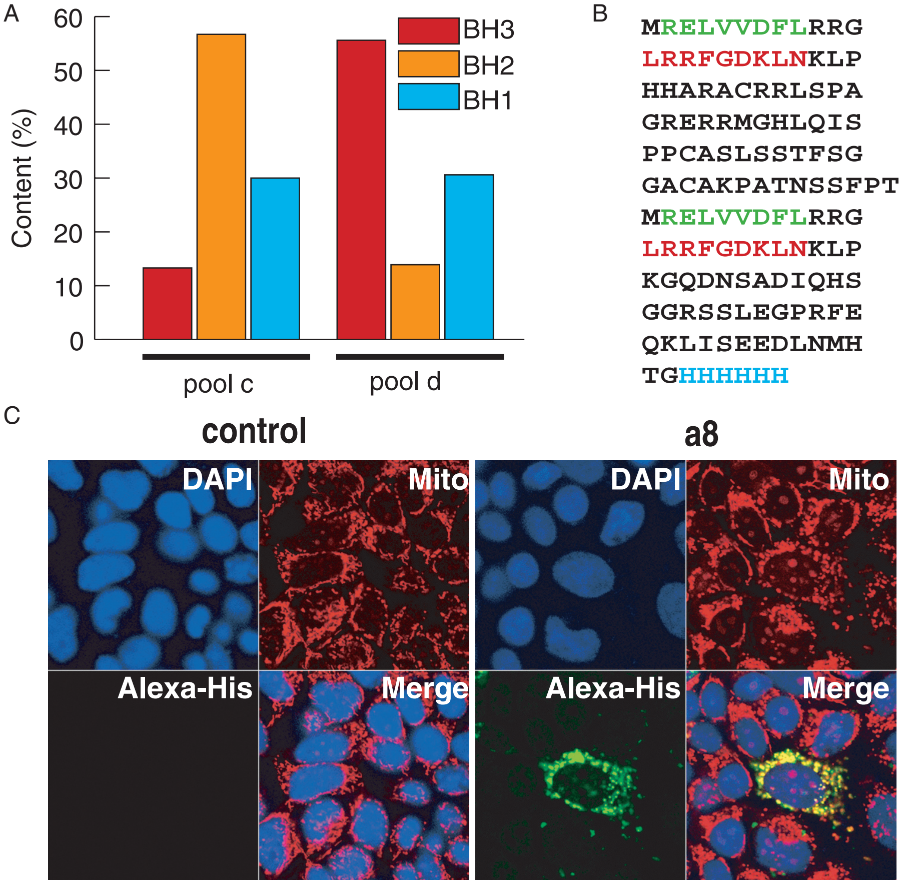

Figure 5. (**A**)The ratio of $BH1_{core}$–$BH4_{core}$ in pool c or pool d was determined from the sequences of 41 randomly selected clones. (**B**) Primary structure of artificial protein, a8, which contained the two $BH4_{core}$ and two $BH3_{core}$ motifs. (**C**) Localization of the a8 protein. The a8 plasmids were transfected into MCF-7 cells using lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions. The cells were incubated for 24 h and then fixed in methanol for 10 min. The localizations of synthetic proteins were analyzed using the penta-his antibody with Alexa Fluor 488 dyes (Alexa-His, QIAGEN). The mitochondria (Mito) and nucleus were probed by Mitotracker Orange (Molecular Probes) and DAPI, respectively. Stained samples were analyzed using a confocal laser scanning microscope.

polymerize a microgene unit created from the sense and antisense MPR primers. With this new protocol, multiple microgenes were created by stochastic base pairing of the MPR primers, enabling their combinatorial polymerization. In the example shown in Figure 2, two sense (A^S and B^S), and two antisense (C^{AS} and D^{AS}) MPR primers yield four distinct microgene units (containing motif-dimers 'A–C', 'A–D', 'B–C' and 'B–D') that are combinatorially polymerized in the MPR process. Alternatively, three microgene units, 'A–B,' 'A–C,' and 'A–D' could be formed by using one sense primer (A^S) and three antisense primers (B^{AS}, C^{AS} and D^{AS}).

We initially assessed the practicality of the new protocol using four short (16–17 nt) MPR primers (Figure 3A). Six bases (GGCGGG) in the 3′ region of the sense primer-A (A^S) were complementary to the 3′ regions of antisense primers-B, C and D (B^{AS}, C^{AS} and D^{AS}) and, therefore, pairwise combination yielding A–B, A–C and A–D should stochastically occur. We then mixed these primers in the combinations shown in Figure 3B and ran the MPR protocol to obtain high molecular weight DNAs (Figure 3B, lanes 1–6). A pair of MPR primers (both sense and antisense) were facilitated to synthesize large DNAs (lanes 2–6). In this pilot experiment, we also introduced restriction endonuclease recognition sequences for EcoRI, HindIII, BglII and XbaI into primers-A^S, -B^{AS}, -C^{AS} and -D^{AS}, respectively, which enabled us to distinguish the DNA units incorporated into the high molecular weight DNAs by simple restriction enzyme digestion. If the DNAs prepared from primers-A^S and -B^{AS} were digested into small fragments by HindIII but not by BglII or XbaI, it would indicate that the polymers contained the primer-B^{AS} unit (Figure 3C lanes 1–4). Similarly, polymers comprising combinations of primer-A^S and -C^{AS} or -D^{AS} were digested only by BglII or XbaI, respectively (lanes 5–12), polymers made from a mixture of primers-A^S, -B^{AS} and -C^{AS} were digested by both HindIII and BglII (lanes 13–16), and polymers made from all four primers were digested by HindIII, BglII and XbaI (lanes 17–20). We then cloned the high molecular weight DNAs into a vector and determined the DNA sequences of some clones, which confirmed that these DNAs were indeed polymers of the three microgenes created from combinations of the four primers used (data not shown). These results indicated that multiple microgenes can be combinatorially polymerized using multiple MPR primers.

BH library construction

We next used the newly developed protocol to construct artificial protein libraries containing mixtures of BH1–4 peptide motifs. Although BH1–4 are unambiguously conserved between Bcl-2 family proteins, there is diversity among the identities of the amino acids in each motif. For BH1, BH2 and BH4, we extracted motif sequences composed of 8 amino acids from human Bcl-xL (24), and for BH3 we extracted a 9-amino-acid sequence from human Noxa (25) ($BH1_{core}$–$BH4_{core}$, Figure 4A and B). We chose the $BH3^{Noxa}$ motif to construct a library because the simple conjugation of $BH3^{Noxa}$ and the protein transduction domain of Tat protein (PTD^{Tat}) has been shown to penetrate into cells but fail to induce apoptosis, whereas a combinatorial library constructed from these motifs contained bifunctional proteins (20). We initially used these natural amino acid sequences to design a set of $microgenes_{core}$ that independently encoded $BH1_{core}$–$BH4_{core}$ in their first reading frames (Figure 4B). Degeneracy in the genetic code allowed us to choose a set of codons such that the other two frames of the genes did not contain any termination codons, and the peptides encoded by the third reading frame had a propensity to form α-helical structures, which we expected would help the structural formation of synthetic proteins (20). To design such microgenes, we have previously developed the microgene design program 'CyberGene' (19). Using CyberGene, appropriate sequences with the above criteria have been selected from all possible sequences that coded the motifs *in silico*. We then synthesized MPR primers based on these core sequences (Figure 4C). The core sequences that coded for the BH motifs were flanked by 3′ association sequences (5′-CGGCGGGGA-3′ and 5′-GCCCCGCCA-3′ for the sense and antisense primers, respectively) and recognition sites for restriction endonucleases, and a CACCATG sequence at the 5′-teminus of the sense primers enabled directional cloning and provided a translation initiation codon, yielding primers $BH3^{S}$, $BH4^{S}$, $BH1^{AS}$, $BH2^{AS}$ and $BH3^{AS}$ (Figure 4C). We also synthesized derivatives of antisense primers that had an extra CC at their 5′ termini ($BH1^{AS+}$, $BH2^{AS+}$ and $BH3^{AS+}$) so that the reconstituted microgenes would have lengths that were multiples of three. Polymerization of such microgenes would maintain the same reading frame unless frameshift mutations randomly occur at junctions of microgene polymers. Therefore, we expected that a library derived from BH^{AS+} primers would contain more peptide motifs compared with that derived from BH^{AS} primers. Addition of these appendix sequences did not create termination codons in the microgenes created (Figure 4D). Thus, the BH_{core}-coding sequences were connected to 3′ association sequences and restriction endonuclease sites, producing a set of sense ($BH3^{S}$ and $BH4^{S}$) and antisense ($BH1^{AS(+)}$, $BH2^{AS(+)}$ and $BH3^{AS(+)}$) primers (Figure 4C).

Using these MPR primers, we tested four conditions for polymerization by changing the ratios and lengths of the primers (Table 1). In the first condition (pool-a), $BH1^{AS}$, $BH2^{AS}$, $BH3^{AS}$ and $BH4^{S}$ primers were mixed at a ratio of 2:1:100:100. With this combination of primers, each re-created microgene should contain the $BH4_{core}$ sequence because each of the three antisense primers (BH1-3 ^{AS}) must associate with the $BH4^{S}$ primer for the MPR polymerization to proceed. Confirming this configuration, digestion of the resultant high molecular weight DNAs using *Sal*I, which cut the BH4 unit, yielded small DNA fragments whose sizes corresponded to single microgene units (Figure 4E, lane 1 versus 2). Because the association of the three antisense primers with $BH4^{S}$ was basically a stochastic event, we expected that the microgene containing $BH3_{core}$ would predominate over those containing $BH1_{core}$ or $BH2_{core}$, as there was 50 and 100-fold more $BH3^{AS}$ present than $BH1^{AS}$ or $BH2^{AS}$, respectively. As expected, digestion of microgene polymers by HindIII (a marker of $BH3^{AS}$) yielded smaller DNA fragments than did EcoRI or BglII (markers of $BH1^{AS}$ and $BH2^{AS}$, respectively) (Figure 4E, lanes 3–5), indicating a high content of $BH3_{core}$ among the polymers.

In the second condition (pool-b), we used two sense primers ($BH3^{S}$ and $BH4^{S}$) and two antisense primers ($BH1^{AS}$ and BH2 ^{AS}), in which stochastic associations of the four primers would yield four types of microgenes: BH3–BH1, BH3–BH2, BH4–BH1 and BH4–BH2 (Figure 4D). Confirming this condition did indeed polymerize all four microgenes, we observed efficient digestion of microgene polymers by SalI ($BH4^{S}$), XhoI ($BH3^{S}$), EcoRI ($BH1^{AS}$) and BglII ($BH2^{AS}$) (Figure 4E, lanes 6–10). In the third (pool-c) and fourth conditions (pool-d), we used three antisense ($BH1^{AS+}$, $BH2^{AS+}$, $BH3^{AS+}$) and one sense primers ($BH4^{S}$). In these conditions, we were able to regulate the frequency of each motif within a library by changing the concentrations of the MPR primers. For instance, when we mixed $BH1^{AS+}$, $BH2^{AS+}$ and $BH3^{AS+}$ at a ratio of 1:1:1 with $BH4^{S}$ (pool-c, Table 1), endonuclease digestion indicated that $BH2_{core}$ (BglII) was most abundant within the polymers, while little $BH3_{core}$ (HindIII) was present (Figure 4E, lanes 12–14). This was confirmed by analyzing the sequences of randomly chosen clones (Figures 6A and 5A, pool c). In contrast, when we mixed $BH1^{AS+}$, $BH2^{AS+}$ and $BH3^{AS+}$ at a ratio of 1:0.75:5 (pool-d, Table 1) to increase the ratio of $BH3_{core}$ and decrease the ratio of $BH2_{core}$ in the pool, both restriction endonuclease digestion (Figure 4E, lanes 16–18) and analysis of the motif content (Figures 6A and 5A, pool d) confirmed $BH3_{core}$ to be the most abundant sequences within the polymers. Thus, our method enables homology-independent polymerization of DNA blocks, in which the frequency of each block within a library can be adjusted by changing the concentrations of MPR primers.

Properties of synthetic proteins

From the prepared libraries, we randomly selected 41 clones (13, 5, 10 and 13 clones from pools-a, -b, -c and -d, respectively) and determined their DNA sequences.

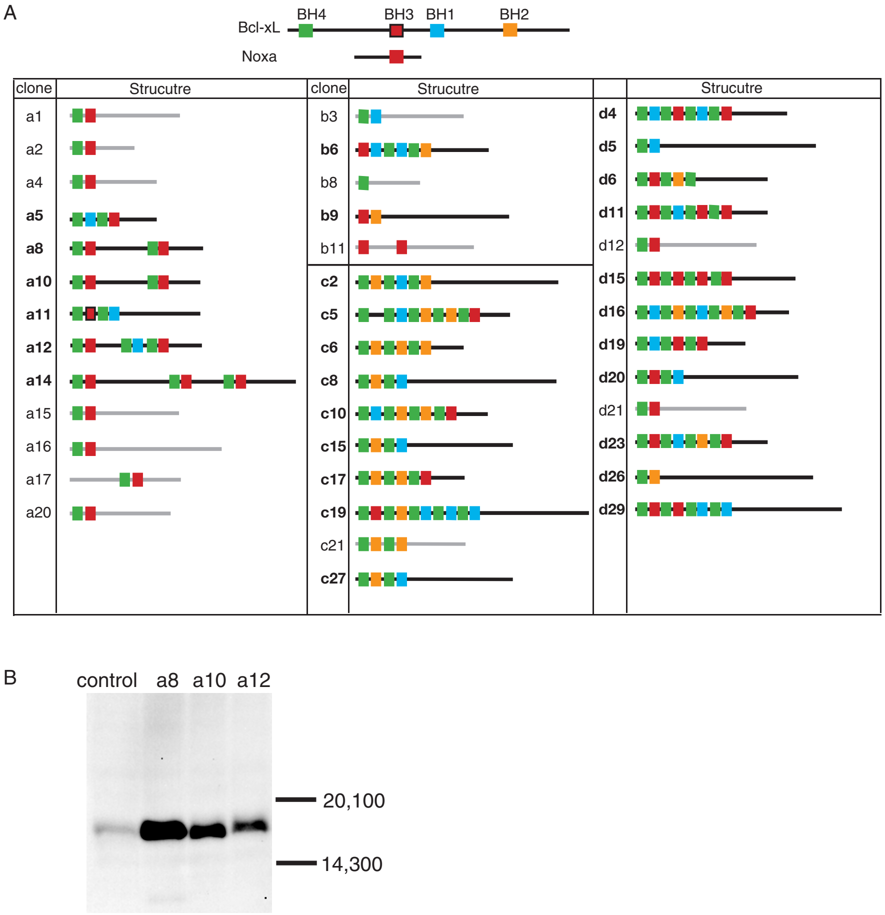

Figure 6. (**A**) Primary structure of synthetic proteins generated from four libraries (pools a–d) prepared under different conditions (see also Table 1). Forty-one clones were sequenced, and their expression was investigated in MCF-7 cells. Transient protein expression in MCF-7 cells was observed for 28 of the 41 clones (68%, shown by bold black bars), which was detected immunohistochemically using anti-c-myc antibody; all of the synthetic proteins contained a myc epitope (EQKLISEEDL) and a poly-histidine tag at their C-terminus. The lengths of the bars correspond to the relative number of amino acids (e.g. d29 has 221 amino acid residues). The order and arrangement of BH1–BH4$_{core}$ in each protein are shown by squares (the color coding of each motif is the same as in Figure 4). In pools-a and -b, MPR primers were designed so that the reconstituted microgenes would have lengths that were *not* multiples of three. The reading frame of the microgene polymers was altered at every junction between the microgene units, unless the junction contained insertion/deletion mutations. In contrast, the microgenes in pools-c and -d were designed to maintain the reading frame throughout the polymers. Reflecting this design, clones obtained from pools-c and -d contained fewer frameshifts than those from pools-a and -b. (**B**) MCF-7 cells were transfected with DNA encoding a8, a10, a12 or empty vector (control), and after 24 h, the clones were analyzed by western blotting using anti-myc antibody. Predicted sizes of synthetic proteins were detected.

The predicated polypeptides contained various combinations of the four BH motifs translated from 68 to 250 codons (Figure 6A). In the experiments in which pools-a and -b were prepared, the reading frame of the polymers was altered at every junction between the microgene units, unless the junction contained insertion/deletion mutations, because the lengths of the microgenes were not multiples of three. By contrast, the microgenes in pools-c and -d were designed to maintain the reading frame throughout the polymers (Figure 4C and Table 1). Reflecting this design, clones obtained from pools-a and -b contained fewer BH motifs (55/18 = 3.1 motifs per clone) than those from pools-c and -d (134/23 = 5.8 motifs per clone) (Figure 6A).

We next transfected each of the 41 clones into MCF-7 cells to investigate the expression profiles of the artificial proteins. After sub-cloning each polymer into an expression vector that added a myc epitope and a poly-histidine tag at its C-terminus, we transfected the polymers into MCF-7 cells and then detected the translated products using an anti-myc antibody. Of the 41 clones, 28 (∼70%) yielded proteins that were immunohistochemically detectable (Figure 6, clones with black bold bars). We also analyzed expressions of randomly selected 10 clones (a8, a10, a12, c17, d5, d11, d16, d19, d26, d29) that were immunohistochemically detectable by western blotting, confirming that predicted polypeptides were expressed in cells (Figure 6B and data not shown). It is noteworthy that clones obtained from pools-c and -d, which contained fewer peptides encoded by the second and third reading frames, were detected at higher rates than peptides obtained from pools-a and -b (44 versus 87%).

The intriguing observation was the mitochondrial localizations of the synthetic a8 protein (Figure 5B and C). 24 h after transfecting the a8 plasmid into

MCF-7 cells, cells stained with the anti-penta-his antibody (green) and Mitotracker^{TM} (red) exhibited overlapping distributions of a8 and mitochondria, indicating that the a8 protein associated with the mitochondria (Figure 5C). The good correlation between a8 and mitochondria was also confirmed by immunofluorescence using the anti-c-myc antibody (data not shown). Because we did not incorporate a mitochondrial targeting sequence into the designed microgenes, it is likely that the a8 have either acquired a synthetic signal sequence or interacted with proteins that target to mitochondria. We also observed that 'aX (a10, a12 and a14)' proteins localized in mitochondria, but some clones from different pools (b–d) localized in different organelles (data not shown). Although further analyses are required to investigate the structural and functional properties of these proteins, it is noteworthy that synthetic proteins made by the motif programming method can be effectively expressed in human cells and localized in the organelle in some cases.

DISCUSSION

Our synthesis of artificial proteins entails a hierarchical approach to *in vitro* protein evolution in which peptide blocks (microgenes) are combinatorially polymerized to construct a protein library. This approach contrasts with the first generation of *in vitro* evolution systems in which a large pool of random sequences prepared by combinatorial polymerization of nucleotide or amino acid units were used as naïve libraries (8). Although the 'selection from naïve sequences' strategy has enabled creation of peptide binders (26), catalytic peptides (27) and even a small protein (9), the larger, modular structures of existing natural proteins suggest that they did not directly arise from random sequences, but developed hierarchically from assemblages of smaller primordial genetic units; that is to say, primordial microgenes endowed with rudimentary activities initially emerged from naïve sequences, after which these microgenes served as building blocks for the larger, more exquisite genes that evolved from their combinatorial assemblage. The 'exon theory of genes', which proposes that polymerization of exons *via* their flanking introns ('exon shuffling') gave rise to sets of genes, is fully compatible with this notion (28).

In our system, we used peptide blocks (motifs) that were preliminarily correlated with a particular function or structure and then combined to make biased libraries. These peptide blocks can include (i) motifs identified from natural proteins; (ii) motifs artificially created by the first generation evolution system; and (iii) motifs rationally designed through protein engineering. We named this operation 'motif programming'. In some ways, motif programming resembles protein engineering in which a rational *de novo* design of a novel protein is sought. However, previous studies by others and ourselves have shown that when it comes to manifesting a desired function, peptide motifs are capricious, as they are strongly influenced by their context within the artificial proteins (20,29). Therefore, with our incomplete knowledge on the structure–function relationships of proteins, the combinatorial or irrational approach is still very important (10). But while motif programming emphasizes the selection from libraries, because it starts with biased libraries, not naïve ones, the size of library to be screened should be small compared to those in first generation evolution systems.

Several methods have been developed to shuffle DNA blocks to make protein libraries. For instance, Stemmer's 'DNA shuffling' has been widely used to improve the properties of existing enzymes, cytokines etc. (11). This method is essentially *in vitro* homologous recombination among a family of genes and DNA blocks that are difficult to combinatorially polymerize—i.e. it can create polymers of A–B′–C from recombination between A—B–C and A′–B′–C, but cannot produce B′–C′–A or A′–C′–C–A. In the years following Stemmer's innovation, several other methods were proposed to enable combinatorial polymerization of DNA blocks. These methods are sketchily classified into two groups: one that requires short DNA sequences commonly shared among DNA blocks for recombination, and one that enables homology-independent polymerization. The latter includes 'SHIPREC' (13), 'ITCHY' (14), 'Y-ligation' (15), 'NRR' (16) and our MPR method (17). The differences between MPR and other methods are (i) the reaction conditions of MPR are rather simple and (ii) the proteins created by MPR are repetitious, which seems to contribute to the emergence of structured proteins (18). Originally, MPR had an inherent restriction, however: the number of motifs that could be embedded was limited to three, the number of reading frames, and the creation of any combinatorial library was dependent on a randomly occurring frameshift (Figure 1). Although we have previously succeeded in synthesizing the functional proteins using MPR method (20), these characteristics of MPR may limit its application potential to generate synthetic proteins with diverse functions in some cases. To overcome that limitation, we developed a new method for microgene polymerization in which a desired number of MPR primers are used to stochastically create two or more microgenes in a single reaction tube. This, in turn, enables polymerization of more than three motifs (Figure 2). Since we have previously demonstrated that β-sheet or copper-binding peptide motifs could be incorporated into synthetic proteins by the original MPR method (18, 19), the scope of the new method is not explicitly restricted to α-helical motifs and proteins. Therefore, in principle, the method can use any peptide sequences for generating a protein library. Moreover, the method can control the frequency of each motif within a library by simply changing the concentration of MPR primers. Such a 'tailor-made library' had not been obtained by the methods reported previously. The new method can also increase the number of different sequences (complexity) in a library compared to the former MPR method. For example, polymers consisting of 10 mer of the single microgene (Figure 1) could have $3^{10} = \sim 0.6 \times 10^5$ molecular diversity. In contrast, polymers consisting of 10 mer of four microgenes (Figure 2) could have

$(4 \times 3)^{10} = \sim 6.2 \times 10^{10}$ diversity. Although it seems that random frameshifts are a rather rare event in this study, it is possible to increase the complexity by using a mixture of antisense primers (BH1-3^{AS} and BH1-3^{AS+}). Because we can incorporate desired number of motifs within a library, this method, by combining with a high-throughput screening, may represent a novel system for synthesizing functional proteins.

Combinatorics of motifs in the new protocol are attributed to (i) the combinations of MPR primers used to create microgenes and (ii) polymerization of microgenes that employs illegitimate recombination by DNA polymerase (Figure 2). With the restriction on motif number removed, we made use of microgene designs in which combinations of BH4–BH1, BH4–BH2 and BH4–BH3 (pools -a, -c and -d) and combinations of BH4–BH1, BH4–BH2, BH3–BH1 and BH3–BH2 (pool-b) were used to create protein libraries (Figure 4D). This means that the polymers obtained were not the result of random shuffling of four BH motifs but of three (pools -a, -c and -d) or four (pool -b) motif heterodimers. If combinatorics of four motifs was needed, one could design microgenes so that (i) each microgene would encode one motif, or (ii) a microgene could be formed from four sense and four antisense MPR primers encoding four sense and antisense motifs, respectively. We are currently working to generate such a library. In conclusion, we invented a new synthetic method for protein creation, in which polymers of multiple peptide motifs are combinatorially assembled. The method enables homology-independent polymerization of DNA blocks, and the frequency of each block within a library can be adjusted in a tailor-made manner. Therefore, the method represents the potential system to create *de novo* folded synthetic proteins with desired functionalities.

ACKNOWLEDGEMENTS

We thank D.B. Murray for his critical proofreading of the manuscript. This work was supported in part by a grant-in-aid scientific research, Ministry of Education, Culture, Sports, Science and Technology. H.S. acknowledges the JSPS Research Fellowships for Young Scientists. Funding to pay the Open Access publication charge was provided by Japan Science and Technology Agency.

Conflict of interest statement. None declared.

REFERENCES

1. Benner,S.A. (2003) Synthetic biology: act natural. *Nature*, **421**, 118.
2. Szostak,J.W. (1992) In vitro genetics. *Trends Biochem. Sci.*, **17**, 89–93.
3. Saito,H., Kourouklis,D. and Suga,H. (2001) An in vitro evolved precursor tRNA with aminoacylation activity. *EMBO J.*, **20**, 1797–1806.
4. Lee,D.H., Granja,J.R., Martinez,J.A., Severin,K. and Ghadri,M.R. (1996) A self-replicating peptide. *Nature*, **382**, 525–528.
5. Elowitz,M.B. and Leibler,S. (2000) A synthetic oscillatory network of transcriptional regulators. *Nature*, **403**, 335–338.
6. Gardner,T.S., Cantor,C.R. and Collins,J.J. (2000) Construction of a genetic toggle switch in Escherichia coli. *Nature*, **403**, 339–342.
7. Smith,H.O., Hutchison,C.A.III, Pfannkoch,C. and Venter,J.C. (2003) Generating a synthetic genome by whole genome assembly: phiX174 bacteriophage from synthetic oligonucleotides. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 15440–15445.
8. Wilson,D.S. and Szostak,J.W. (1999) In vitro selection of functional nucleic acids. *Annu. Rev. Biochem.*, **68**, 611–647.
9. Keefe,A.D. and Szostak,J.W. (2001) Functional proteins from a random-sequence library. *Nature*, **410**, 715–718.
10. Hecht,M.H., Das,A., Go,A., Bradley,L.H. and Wei,Y. (2004) De novo proteins from designed combinatorial libraries. *Protein Sci.*, **13**, 1711–1723.
11. Stemmer,W.P. (1994) Rapid evolution of a protein in vitro by DNA shuffling. *Nature*, **370**, 389–391.
12. Hiraga,K. and Arnold,F.H. (2003) General method for sequence-independent site-directed chimeragenesis. *J. Mol. Biol.*, **330**, 287–296.
13. Udit,A.K., Silberg,J.J. and Sieber,V. (2003) Sequence homology-independent protein recombination (SHIPREC). *Methods Mol. Biol.*, **231**, 153–163.
14. 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.
15. Kitamura,K., Kinoshita,Y., Narasaki,S., Nemoto,N., Husimi,Y. and Nishigaki,K. (2002) Construction of block-shuffled libraries of DNA for evolutionary protein engineering: Y-ligation-based block shuffling. *Protein Eng.*, **15**, 843–853.
16. Bittker,J.A., Le,B.V., Liu,J.M. and Liu,D.R. (2004) Directed evolution of protein enzymes using nonhomologous random recombination. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 7011–7016.
17. Shiba,K., Takahashi,Y. and Noda,T. (1997) Creation of libraries with long ORFs by polymerization of a microgene. *Proc. Natl. Acad. Sci. U.S.A.*, **94**, 3805–3810.
18. Shiba,K., Takahashi,Y. and Noda,T. (2002) On the role of periodism in the origin of proteins. *J. Mol. Biol.*, **320**, 833–840.
19. Shiba,K. (2004) MolCraft: a hierarchical approach to the synthesis of artificial proteins. *J. Mol. Catal. B.*, **28**, 145–153.
20. Saito,H., Honma,T., Minamisawa,T., Yamazaki,K., Noda,T., Yamori,T. and Shiba,K. (2004) Synthesis of functional proteins by mixing peptide motifs. *Chem. Biol.*, **11**, 765–773.
21. King,J.S., Fairley,C.F. and Morgan,W.F. (1996) DNA end joining by the Klenow fragment of DNA polymerase I. *J. Biol. Chem.*, **271**, 20450–20457.
22. Strasser,A. (2005) The role of BH3-only proteins in the immune system. *Nat. Rev. Immunol.*, **5**, 189–200.
23. Opferman,J.T. and Korsmeyer,S.J. (2003) Apoptosis in the development and maintenance of the immune system. *Nat. Immunol.*, **4**, 410–415.
24. Sattler,M., Liang,H., Nettesheim,D., Meadows,R.P., Harlan,J.E., Eberstadt,M., Yoon,H.S., Shuker,S.B., Chang,B.S. *et al.* (1997) Structure of Bcl-xL-Bak peptide complex: recognition between regulators of apoptosis. *Science*, **275**, 983–986.
25. Oda,E., Ohki,R., Murasawa,H., Nemoto,J., Shibue,T., Yamashita,T., Tokino,T., Taniguchi,T. and Tanaka,N. (2000) Noxa, a BH3-only member of the Bcl-2 family and candidate mediator of p53-induced apoptosis. *Science*, **288**, 1053–1058.
26. Smith,G.P. (1991) Surface presentation of protein epitopes using bacteriophage expression systems. *Curr. Opin. Biotechnol.*, **2**, 668–673.
27. Tanaka,F., Fuller,R. and Barbas,C.FIII. (2005) Development of small designer aldolase enzymes: catalytic activity, folding, and substrate specificity. *Biochemistry*, **44**, 7583–7592.
28. Go,M. (1983) Modular structural units, exons, and function in chicken lysozyme. *Proc. Natl. Acad. Sci., U.S.A.*, **80**, 1964–1968.
29. Frugier,M., Giege,R. and Schimmel,P. (2003) RNA recognition by designed peptide fusion creates 'artificial' tRNA synthetase. *Proc. Natl. Acad. Sci., U.S.A.*, **100**, 7471–7475.