doi:10.1006/jmbi.2000.3571 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2000) 297, 309–319

JMB

Constructing High Complexity Synthetic Libraries of Long ORFs Using *In Vitro* Selection

Glen Cho, Anthony D. Keefe, Rihe Liu, David S. Wilson and Jack W. Szostak*

Howard Hughes Medical Institute, Department of Molecular Biology Massachusetts General Hospital, Boston, MA 02114, USA

*Corresponding author

We present a method that can significantly increase the complexity of protein libraries used for *in vitro* or *in vivo* protein selection experiments. Protein libraries are often encoded by chemically synthesized DNA, in which part of the open reading frame is randomized. There are, however, major obstacles associated with the chemical synthesis of long open reading frames, especially those containing random segments. Insertions and deletions that occur during chemical synthesis cause frameshifts, and stop codons in the random region will cause premature termination. These problems can together greatly reduce the number of full-length synthetic genes in the library. We describe a strategy in which smaller segments of the synthetic open reading frame are selected *in vitro* using mRNA display for the absence of frameshifts and stop codons. These smaller segments are then ligated together to form combinatorial libraries of long uninterrupted open reading frames. This process can increase the number of full-length open reading frames in libraries by up to two orders of magnitude, resulting in protein libraries with complexities of greater than 10^{13}. We have used this methodology to generate three types of displayed protein library: a completely random sequence library, a library of concatemerized oligopeptide cassettes with a propensity for forming amphipathic α-helical or β-strand structures, and a library based on one of the most common enzymatic scaffolds, the α/β (TIM) barrel.

Keywords: in vitro selection; mRNA display; mRNA-protein fusions; synthetic library; TIM barrel

Introduction

In vitro selection and directed evolution are powerful tools for discovering new functional molecules. Rare molecules can be isolated from large libraries of random and semi-random sequences through iterative cycles of selection and amplification (Szostak, 1992; Wilson & Szostak, 1999). Recent advances have made it possible to perform *in vitro* selection experiments on protein libraries (Roberts & Ja, 1999). Unlike nucleic acids, proteins cannot be amplified directly. This problem has been addressed by expressing protein libraries on the surface of phage or microorganisms, and even using stalled ribosomes (Boder & Wittrup, 1997; Forrer *et al.*, 1999; Georgiou *et al.*, 1997; Hanes & Pluckthun, 1997; Mattheakis *et al.*, 1994; Smith & Petrenko, 1997). Recently our laboratory has developed a technique that links phenotype and genotype *in vitro* through the covalent attachment of proteins to their own mRNAs (Roberts & Szostak, 1997), a process that we call mRNA display. This technique allows a much larger sampling of random sequences (>10^{13}) than has been previously possible.

Here, we describe a method for the design and synthesis of large synthetic DNA libraries coding for protein sequences suitable for use in selection experiments. The entire length of the DNA library

G.C., A.D.K., R.L., D.S.W. contributed to this paper equally.

Present address: D. S. Wilson, Zyomyx, 3911 Trust Way, Hayward, CA 94545, USA.

Abbreviations used: ORF, open reading frame; RT-PCR, reverse transcription polymerase chain reaction; P, polar residue; N, non-polar residue; IGPS, indole-3-glycerol phosphate synthase.

0022-2836/00/020309–11 $35.00/0

is derived from degenerate DNA oligonucleotides. There are, however, a number of problems associated with creating long synthetic protein libraries of high complexity. Deletions that occur due to imperfect coupling and capping efficiencies during solid phase DNA synthesis will cause frameshifts (Hecker & Rill, 1998). Another problem is that stop codons that are encoded within the randomized region will cause premature termination. The proportion of ''correct'' or intended sequences in the library will be reduced significant]y if these problems are not addressed. For a typical deletion rate of 0.5% per coupling, only 22% of a synthetic library that encodes 100 amino acids will be completely in frame. Stop codons within random regions will further decrease this number depending on the distribution of codons used. For example, only 0.8% of sequences will not have stop codons for a library encoding 100 amino acid residues in which each codon position has an equal mixture of each nucleotide. The use of appropriate ratios of nucleotides in the three positions of the codon can reduce the occurrence of stop codons, but this severely alters the density with which different regions of sequence space are sampled.

To address these problems, we have chosen the following general strategy: (1) synthesize the library in small cassettes; (2) perform an *in vitro* selection that enriches for sequences lacking stop codons and frameshifts; (3) assemble the final library through restriction and ligation of the selected cassettes. Essentially, we perform an *in vitro* selection using the mRNA display technology to enrich for sequences that code for intact open reading frames. This method will selectively enrich the library in sequences that lack insertions, deletions and stop codons. We refer to this procedure as ''pre-selection''.

We have used this method to create three different types of random libraries. The first type of library has all amino acid positions randomized with the exception of constant regions containing affinity tags at both ends of the sequence. The second is a ''patterned'' library made up of segments that have a propensity for forming amphipathic α-helices and β-strands. A third type of library has been created in which random amino acid residues are displayed in the context of a known protein fold. Each of these libraries will be useful in studying specific questions in protein evolution.

Results

General strategy

DNA oligonucleotides corresponding to segments of the designed library are synthesized and amplified by PCR. These templates are then amplified such that flanking affinity tags (FLAG (Chiang & Roeder, 1993) and His_6 (Porath *et al*., 1975)) are encoded at the 5′ and 3′ ends. When translated, these tags correspond to the N and C termini of the expressed polypeptide (Figure 1). The preselection is then performed as follows: RNA is generated from the DNA library and then modified at the 3′ end to add a short DNA segment that is terminated with puromycin. Translation of these constructs *in vitro* results in the formation of a covalent bond between the RNA and the protein it encodes; this produces a protein displayed upon its mRNA (Roberts & Szostak, 1997). These mRNA displayed proteins are then successively purified on the basis of each of the affinity tags. We found it beneficial to include the N-terminal purification tag because we had observed internal initiation events. The translation reaction in which the mRNA displayed proteins are made will contain a heterogeneous mix of species (Figure 1). The majority of the RNA will not display protein. Some sequences have an intact open reading frame (ORF) and contain both N and C-terminal tags. Some contain deletions and go out of frame and lack the C-terminal tag. Internally initiated transcripts lack the N-terminal tag. After purification on Ni-NTA and anti-FLAG resins, those sequences with both tags and thus a contiguous ORF will be enriched. The RNA component of the displayed proteins can then be amplified using reverse transcription and polymerase chain reaction (RT-PCR) to generate double-stranded DNA segments that can be cut and ligated together to form the complete library. Naturally, the number of unique sequences in each segment is decreased after preselection, but the ligation of these pieces will regenerate a high complexity final library through the combinatorial assembly of different cassettes.

Random library

Presumably the first proteins arose from pools of random sequences. The density of functional or even folded molecules in protein sequence space is not known. To address these issues, we constructed a random library in which every position has an equal probability of encoding each amino acid. This random library (R_4) was constructed using a novel strategy. A single cassette (R) was synthesized that encoded 20 contiguous random amino acids flanked by FLAG and His_6 tags. A nucleotide distribution (Table 1) encoding an amino acid composition (Table 2) resembling that of contemporary proteins was arrived at by using an iterative computer program (Michael New, personal communication), but similar programs are available on the internet, for example, at http://gaiberg.wi.mit.edu/cgi-bin/CombinatorialCodons (Wolf & Kim, 1999). This program adjusts the nine variables that define the proportions of each of the four nucleotides at each of the three positions in the codon in an attempt to match a desired amino acid composition of the resultant protein library. Because the selection step removes sequences con-

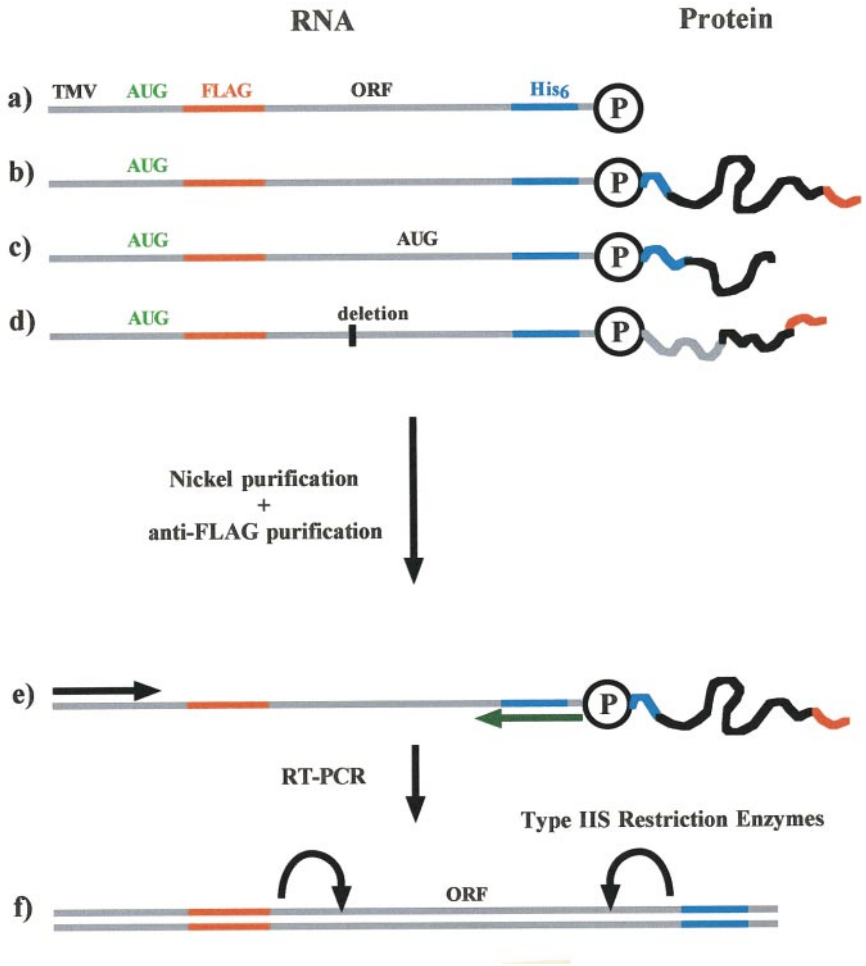

Figure 1. The pre-selection. A heterogeneous mixture of mRNA display templates, some of which display protein, are present in the pre-selection translation mixture immediately before affinity purification. (a) An mRNA display template terminating in puromycin with the parts of the open reading frame coding for the two protein affinity tags indicated. The 5′ end of this template contains the tobacco mosaic virus translation enhancer sequence (TMV) followed by the initiating methionine codon (AUG). (b) An mRNA display template displaying a full-length protein free of frameshifts and stop codons with both protein affinity tags present. (c) An mRNA display template that has initiated internally and displays the corresponding truncated protein lacking the N-terminal FLAG tag. (d) An mRNA display template with a deletion in its open reading frame, displaying the corresponding frameshifted protein lacking the C-terminal His_6 tag. (e) The only one of the constructs described above that is enriched by the pre-selection process, the mRNA display template displaying a full-length protein free of frame shifts with both protein affinity tags present, that corresponds to (b) above; this is amplified using RT-PCR as indicated. (f) The double-stranded DNA cassette that results from the amplification of the pre-selected mRNA display templates with the location of the type IIS restriction enzyme sites indicated.

taining stop codons from the library, it is not necessary to minimize the occurrence of stop codons by adjusting the nucleotide mix.

The random region is flanked by two different type IIS restriction sites, *Bbs*I and *Bbv*I. These enzymes cut outside of their recognition sequences (Figure 2). The position of the recognition sequences with respect to the cut sites was arranged so that the recognition sequences and any other auxiliary sequences such as the affinity tags are excised. The constructs were also designed so that both enzymes created the same four nucleotide non-palindromic overhang. We were able to avoid completely fixed amino acid residues at the ligation junction because the four base overhang symmetrically spans two codons. The first of these codons specifies a mixture of Asp, His, Tyr, and Asn residues, and the second a mixture of Cys and Trp residues. These codons were chosen for the ligation junctions to increase the representation of amino acid residues that were under-represented in the random sequence and could be important for chemical catalysis. Pre-selection was performed, as described below, upon these constructs and double-stranded DNA was generated from the selected material.

Table 1. Percentage of each nucleotide at each position of the designed codons

Library	Codon position	% T	% C	% A	% G
Random	1	20	18	35	27
	2	29	17	33	21
	3	22	29	0	49
Patterned					
α-P	1	0	8	49	43
	2	4.6	26	51	18
	3	0	31	0	69
α-N	1	58	17	16	8.9
	2	89	11	0	0
	3	0	31	0	69
β-P	1	34	0	47	19
	2	0	33	45	22
	3	0	60	0	40
β-N	1	21	18	32	29
	2	100	0	0	0
	3	0	51	0	49
γ	1	12	22	30	36
	2	32	16	32	20
	3	0	72	0	28
Structure-based	1	14	18	32	36
	2	20	23	35	22
	3	35	32	0.61	33

The values shown were determined by sequencing individual cloned cassettes, except in the case of the random library, in which case cassettes were sequenced in the context of the full-length library. These frequencies closely matched the intended ratios (see http://xanadu.mgh.harvard.edu/szostakweb/orf.html internet table for details). For the number of codons sequenced to determine these percentages, refer to Table 2.

The random region of the R cassette encodes 20 amino acid residues, and the final R_4 library was assembled from four cassettes to give a random region 80 amino acid residues long and flanked by FLAG and His_6 tags. The assembly proceeded by dividing the double-stranded DNA library into two aliquots, one of which was cut with *Bbs*I and the other with *Bbv*I. The purified fragments were

Table 2. Percentage of each amino acid encoded by the designed codons

Amino acid	Random library	α-Cassette polar[a]	α-Cassette non-polar[a]	β-Cassette polar[a]	β-Cassette non-polar[a]	γ-Cassette[a]	Patterned library	Structure-based library[b]
Ala	4.1	11	0.98	6.8	0	5.8	4.3	8.7
Arg	6.8	7.6	0	4.2	0	6.1	3.0	5.9
Asn	7.5	7.7	0	14	0	6.9	5.2	7.9
Asp	5.5	6.8	0	5.4	0	8.3	8.8	7.8
Cys	4.6	0	0	4.7	0	1.7	4.4	1.8
Gln	2.6	2.8	0	0	0	2.0	0.8	1.8
Glu	4.0	15	0	3.6	0	3.2	4.1	5.2
Gly	5.3	7.8	0	4.4	0	7.2	3.1	8.2
His	3.8	1.3	0	0	0	5.1	3.8	4.1
Ile	4.8	0.70	4.4	0	16	6.9	4.1	4.1
Leu	7.4	0.37	51	0	28	8.1	14	6.1
Lys	5.1	17	0	9.0	0	2.7	5.5	5.0
Met	4.5	1.6	9.9	0	15	2.7	4.7	2.1
Phe	2.8	0	16	0	11	2.8	4.6	1.7
Pro	2.8	2.1	1.9	0	0	3.5	1.1	4.0
Ser	6.6	2.8	6.3	18	0	6.2	5.9	9.3
Thr	5.4	13	1.8	17	0	4.8	6.7	5.9
Trp	4.4	0	0	3.1	0	0.67	6.5	1.1
Tyr	5.0	0	0	9.8	0	2.8	2.5	3.5
Val	7.1	2.0	7.9	0	29	12	7.5	5.8
Codons sequenced	1200	395	316	400	320	315	2134	657

The values shown were determined by sequencing individual cloned cassettes, except in the case of the random library, in which case cassettes were sequenced in the context of the full-length library.

[a] For individual cassettes, the contribution of positions at the bridging restriction sites is omitted; this contribution is included for the full-length libraries (random and patterned).

[b] For the structure-based library, only the degenerate residues were considered.

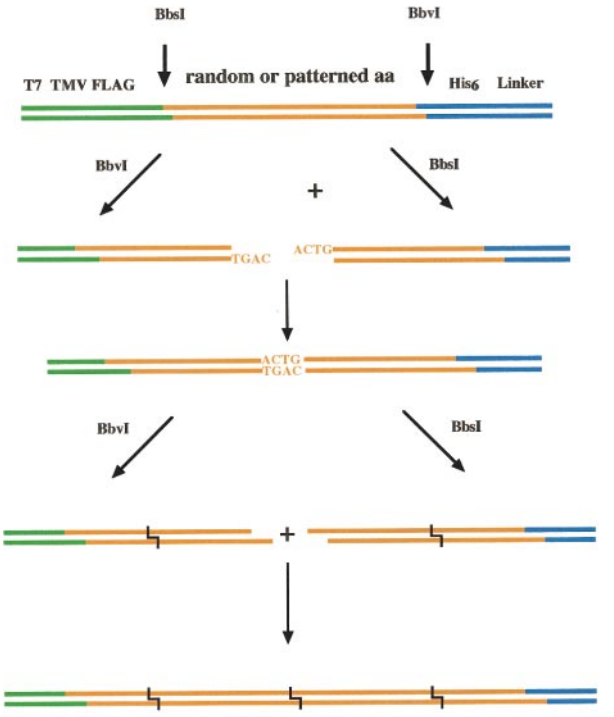

Figure 2. Assembly of the full-length library from the DNA cassettes that result from the amplification of the pre-selected mRNA display templates. The cassettes are divided into two aliquots that are restricted with either *Bbv*I or *Bbs*I; subsequent ligation with T4 DNA ligase gives a new cassette in which the DNA between the restriction sites doubles in length while the flanking regions remain the same. The doubling of the length of this region may be repeated any number of times by repeating the restriction and ligation process.

then ligated together with T4 DNA ligase. Repeating this procedure yielded the final library (Figure 2). The diversity of the pre-selected cassette was 4×10^{12} sequences. This value represents the number of mRNA display template cassettes that were recovered from the pre-selection, and assumes that each cassette is unique. This assumption is justified because the complexity of the mRNA display templates (3×10^{14}) is much larger than the number of mRNA displayed proteins recovered from the pre-selection (4×10^{12}). After ligating the four cassettes together, the diversity of the library was combinatorially increased to 4×10^{14}, assuming that each full-length library member results from the ligation of different combinations of cassettes. After the final ligation, the library was amplified by PCR, which increases the copy number but does not affect the diversity or complexity of the library.

After pre-selection, 60 % of the final library sequences are completely free of deletions, insertions and stop codons (Table 3). Removing stop codons and frameshifts increased the proportion of ''desired'' molecules in the library by a factor of approximately 50. Had the library been assembled from the cassettes without pre-selection, then the proportion that would have been completely free of deletions and stop codons would have been approximately 1 %, with a concomitant reduction in diversity. The random library can be used to generate mRNA display templates that display the proteins they encode when translated *in vitro* (Figure 3).

Patterned library

Only a small fraction of random polypeptide sequences would be expected to fold into unique, globular structures typical of natural proteins. For this reason, we also constructed a second library in which polar (P) and non-polar (N) residues were patterned so as to form peptide stretches capable of forming amphipathic secondary structure elements. Our intention was that this patterning would increase the fraction of library molecules capable of forming a well-defined hydrophobic core (Kamtekar *et al.*, 1993).

The design was approached by first making three cassettes encoding 11 amino acid residues, each displaying a particular pattern of polar and non-polar residues. Cassette α (''α-helix'') encodes sequences with the pattern NNPPNPPNNPP. The nucleotide distributions (Table 1) of the codons were designed so that several different amino acid residues could be encoded at each position, while respecting the P/N distinction (see Table 2 and Materials and Methods for details). A sequence with this P/N pattern can form an α-helix in which all the non-polar residues lie on one face (Eisenberg *et al.*, 1984). Cassette β (''β-strand'') encodes a sequence with the pattern NPNPNPNPNPP. This peptide pattern can form β-strands with all of the non-polar residues on one side. The periodicity of polar and non-polar residues in a linear sequence has been shown to be sufficient to promote folding into α-helices and β-strands (Xiong *et al.*, 1995). Cassette γ encodes a random amino acid sequence (Table 2), similar to that described for the R cassette above without any designed P/N pattern. The goal was then to ligate eight cassettes so as to generate a library of polypeptides with a random ordering of amphipathic α-helices and β-strands, as well as some unpatterned peptide units.

Each of these cassettes has the same constant flanking regions, which are similar to those described for the R cassette. The same restriction sites, *Bbs*I and *Bbv*I, allowed for a strategy similar to that described above. Each cassette was divided into two aliquots, one being restricted with *Bbs*I and the other with *Bbv*I. These six fragments were then mixed together and ligated, as above, so as to give cassette-dimers of the nine possible combinations (αα, αβ, αγ, βα, ββ, βγ, γα, γβ, and γγ). These cassette-dimers were then treated exactly as were the R cassette for making the R_4 library (Figure 2). After two successive ligations, the final

Table 3. Proportions of synthetic DNA cassettes with various kinds of imperfections, both before and after the pre-selection process was applied

	Before pre-selection			After pre-selection			
Library	Without frameshifts	Fraction of library Without stop codons	Perfect	Without frameshifts	Fraction of library Without stop codons	Perfect	Enrichment factor in perfect full-length libraries
R	0.81 (16)	0.41	0.34	0.96 (96)	0.92	0.88	
Random (R_4)			0.013			0.60	47
α	1.00 (16)	1.00	1.00				
β	0.92 (60)	0.59	0.54	0.96 (28)	0.93	0.90	
γ	1.00 (15)	0.95	0.94				
Patterned			0.10			0.65	6.5
AB	0.32 (22)	0.69	0.16	0.92 (14)	1.00	0.71[a]	
CD	0.79 (14)	0.64	0.51	1.00 (17)	1.00	1.00	
EF	0.47 (17)	0.69	0.33	0.94 (18)	1.00	0.94	
GH	0.49 (36)	0.62	0.30	1.00 (15)	1.00	0.54[b]	
IJ	0.63 (11)	0.87	0.55	1.00 (16)	1.00	1.00	
Structure-based			0.0043			0.36	84

The values shown were determined by sequencing individual cloned cassettes, except in the case of the random library, in which case cassettes were sequenced in the context of the full-length library. Columns 2-4 correspond to the unselected cassettes: column 2 shows the fraction of cassettes without frameshifts (deletions and insertions); column 3 shows the fraction of cassettes without stop codons; and column 4 shows the fraction of cassettes without either frameshifts or stop codons, and the same for the full length DNA library members that would have resulted were they to have been assembled without the benefit of the pre-selection process. Columns 5-7 correspond to the cassettes that were selected by the pre-selection process and are exactly analogous to columns 2-4. In the case of the full-length libraries, these numbers are extrapolated from the fraction of perfect cassettes within sequenced full-length library members. Column 8 shows the factor by which the proportion of library members completely free of frameshifts and stop codons has been multiplied by the use of the pre-selection process. This factor is equal to the factor by which the initial library diversity is multiplied in a selection using one of these libraries as compared to the same library constructed without the use of this method. The numbers in parentheses are the numbers of cassettes sequenced in the derivation of these data.

[a] Of the sequenced clones, 92% were free of frameshifts (deletions or insertions) and stop codons: however, only 71% were also free of three, six or nine-base deletions and these sequences were not considered perfect.

[b] This number is aberrantly low due to a technical error, see the experimental protocol for the structure-based library.

library consisted of eight consecutive cassettes. The number of possible combinations of α, β and γ units is 6561. The actual complexity of the library is much higher because each cassette is itself a library of sequences. The number of unique full-length library sequences after the final ligation was 2.4×10^{14}. The composition of the library with respect to the cassettes was 44% α, 45% β and 11% γ.

Three measures were taken to reduce the number of library molecules that were either out of frame or contained in-frame stop codons. First, a small amount of each library was purified by high-resolution denaturing PAGE. After this gel purification, the deletion rate per nucleotide was only 0.2%. Second, the codons in the α and γ libraries were designed such that in-frame stop codons occurred at zero and 1%, respectively. Nearly all of the sequenced α and γ cassettes were completely free from both frameshifts and in-frame stop codons. Third, the pre-selection technique, mentioned above, was performed on the β cassette, since no attempt was made in this case to reduce the frequency of in-frame stop codons. The pre-selection decreased the fraction of defective cassette β fragments from 45% to 10% (Table 3). Since the β fragment was incorporated into the library at 44% on average, this increased the number of completely in-frame molecules without stop codons from 10% to 65%. The patterned library can be used to generate mRNA display templates that display the protein they encode when translated *in vitro* (Figure 3).

Structure-based library

A third library was constructed in which random amino acid positions were placed in the context of a known protein fold. We chose indole-3-glycerol phosphate synthase (IGPS) (Tutino *et al.*, 1993) from the hyperthermophilic archaeon *Sulfolobus solfatiricus* to serve as the structural model for building this library. The IGPS is part of the tryptophan biosynthesis pathway and has the canonical α/β barrel, or TIM barrel fold (Creighton & Yanofsky, 1966). It is highly stable to thermal and chaotropic denaturation (Andreotti *et al.*, 1997). The structure has been solved to 2.0 Å and the protein is active as a monomer (Hennig *et al.*, 1995; Knochel *et al.*, 1996). We have created a library in which 49 random residues have been inserted in place of the eight structurally constrained loops that connect the β-sheets and α-helices at one end of the barrel (Figure 4). The random sequences will be presented from eight different structural positions, enabling the proteins to form large cavities or interaction surfaces.

The complete library was assembled from five separate cassettes (AB, CD, EF, GH, and IJ) that correspond to different regions in the gene. Each cassette was produced by the mutually primed extension of two oligonucleotides. The oligonucleo-

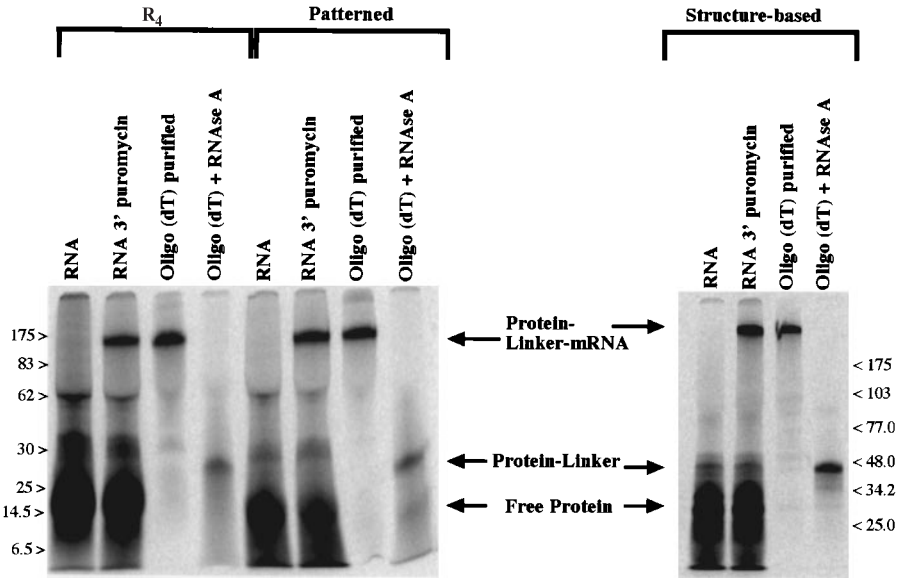

Figure 3. The mRNA displayed proteins of the assembled libraries. *In vitro* translated samples labeled with [^{35}S]-methionine were analyzed by SDS-PAGE (6 % or 8 %) and exposed to a phosphorimager plate. Linker refers to a short oligo(dA) stretch that is terminated by puromycin at the 3′ end. For each of the libraries, the first lane shows the translation product of the unmodified mRNA template. The second lane shows the translation product of the mRNA template modified at the 3′ end with puromycin (the mRNA display template). Lane 3 shows the elution fraction from an oligo(dT) purification performed on the sample shown in lane 2. Lane 4 shows the same oligo(dT)-purified sample subsequently treated with RNase A.

tides have constant sequences that correspond to the amino acids specifying the β-strands or α-helices of the scaffold and random sequences that correspond to the loops. These cassettes were then amplified to add the T7 promoter at the 5′ end and a His_6 tag at the 3′ end. Each cassette was selected for the absence of deletions and stop codons using the pre-selection technique described above. The enzyme Deep Vent polymerase (NEB) was used for PCR amplification to reduce the frequency of mutations introduced into the cassettes subsequent to pre-selection. The design at the C-terminal His_6 tag included two out-of-frame stop codons, that terminate translation for sequences that are not in the correct frame. The results of the pre-selection are summarized in Table 3. After pre-selection and subsequent RT-PCR amplification, *Fok*I and *Ban*I were used in generating asymmetric ligation junctions. AB, GH, and IJ were cut with *Fok*I; CD and EF were cut with both *Fok*I and *Ban*I to generate complementary sticky ends. The ligation junctions were designed so that the overhangs will only be complementary between intended adjacent cassettes to generate the complete library in which the cassettes are arranged AB-CD-EF-GH-IJ.

The random region was designed so that the amino acid distribution would reflect the distribution normally found in the loops of α/β barrel family members. A structural alignment for all of the known enzymes with α/β barrel folds from the FSSP database was used to obtain the frequency of the various amino acids in the loops (Holm & Sander, 1994). An appropriate nucleotide mixture was calculated as described above. The probability of a stop codon occurring within the random region was initially 3.6 %, (20 out of 555 codons) per position before selection as determined by sequencing the random regions of each of the cassettes. This would have resulted in only 16 % of the sequences in the final library having no stop codons. However, after pre-selection, sequencing revealed no stop codons were encountered in the random region (out of 657 codons sequenced). Overall, the pre-selection increased the diversity approximately 80-fold due to the reduction in frameshifts and stop codons. For the final library, 36 % of the members are calculated to be completely intact. If we had attempted to generate the library without pre-selecting for ORFs, then the proportion of intact sequences would be 0.43 %. The final library is 849 nucleotides in length and codes for 269 amino acid residues, 49 of which are in random loops (Figure 4). The diversity of the library is 1.2×10^{14} different sequences. The structure-based library can be used to generate mRNA display templates that display the protein they encode when translated *in vitro* (Figure 3).

Discussion

We would like to compare the frequency and distribution of functional molecules within each of the three different libraries. How many successful molecules will emerge from each library if the

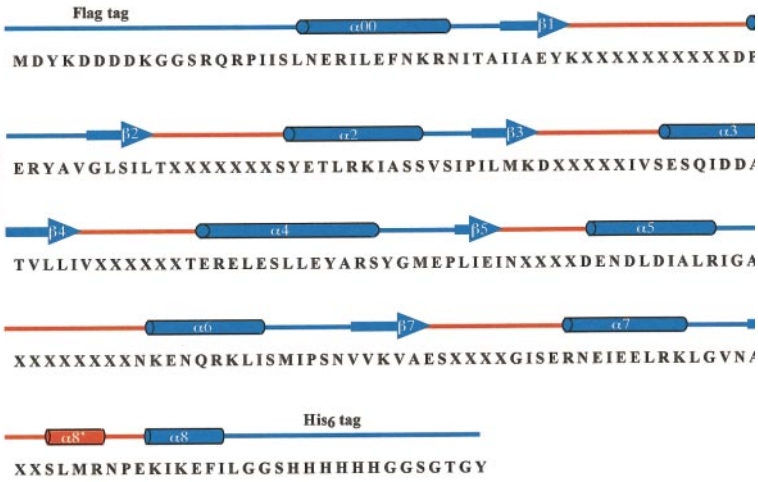

Figure 4. Secondary structure representation of the structure-based library. The red regions correspond to the loops at the randomized end of the barrel. Degenerate positions within the sequence are denoted by X. The first 25 residues of the wild-type indole-3-glycerol phosphate synthase, which correspond to α-helix 0 in the wild-type sequence, were not included in the construction of the library and affinity tags were added at the N and C termini.

same target is used for selection? Also, will there be a difference in the range of activities of these successful molecules with respect to the library from which they were derived? If we perform a selection for specific binding to a small molecule target, then active proteins presumably will have to specify in their sequence: (1) regions of secondary structure, (2) a hydrophobic core, and (3) an appropriate arrangement of residues at the binding site. It is not known how frequently these properties occur in sequence space. The patterned library will have a predisposition for forming secondary structural units, and thus less additional information may be required to form a functional protein than in the case of the totally random library. The α/β barrel library, which has the highest proportion of fixed residues of the three libraries, will need to additionally specify only the correct active site or binding site residues, and thus functional molecules may be even more abundant compared with the other two libraries.

The selection for ORFs described in this study can also be used to create other types of protein libraries, and for increasing the complexity of libraries used in other types of selections such as phage display and ribosome display. For example, if the diversity in a phage display experiment is limited to 10^8 and only 1 % of the sequences is in the correct frame and has no stop codons, then effectively only 10^6 correct sequences will be sampled. However, if a library has been pre-selected for ORFs, then the complexity can be increased to approach the experimental limit. It should be noted that the pre-selection protocol itself could be performed using phage display.

Another possible application of this technique is to select ORFs from genomic libraries. Random genomic DNA clones could be modified so as to possess tags at their 5′ and 3′ ends and then pre-selected for the absence of stop codons. Sequences selected by such an approach will be enriched in fragments of ORFs. These sequences could then be used for selections or could be sequenced to give a database enriched in ORFs.

Materials and Methods

Construction of the random library

Exact nucleotide mixtures and all oligonucleotide sequences for the three libraries can be downloaded from http://xanadu.mgh.harvard.edu/szostakweb/orf.html Further information on library construction is also present on this site. Oligonucleotides were ordered from the Keck Oligo Facility at Yale University. The random cassette was synthesized on a DNA synthesizer using standard phosphoramidite chemistry. The synthesized DNA encoded 72 random nucleotides that comprised 18 consecutive random codons (XYZ) with the distribution of bases shown in Table 1. These random codons were flanked by constant sequence that, after nested PCR, gave the whole cassette as follows: phage T7 RNA polymerase promoter, tobacco mosaic virus translation enhancer sequence (Sleat *et al.*, 1987), initiating methionine (start of ORF), FLAG tag, *Bbs*I site, random region *Bvs*I site, His_6 tag, flexible linker (MGMSGS), (TTC TAATACGACTCACTATAGGGACAATTACTATTTAC AATTACAATGGACTACAAAGACGACGACGATAAG AAGACTYACTGZ$(XYZ)_{18}$YACTGGTCAGCGAGCTGC-CATCATCATCATCATCATATGGGAATGTCTGGATC T). The RNA was produced from the cassette with T7 RNA polymerase and mRNA displayed proteins were made as previously described (Liu *et al.*, 2000). A 10 ml translation yielded 6×10^{13} mRNA displayed proteins. The translation mixture was then diluted tenfold into oligo(dT)-cellulose binding buffer (1 M KCl, 100 mM Tris(hydroxymethyl) amino methane, 0.25 % w/v Triton X-100 (pH 8.0)) and this mixture was incubated with 2 mg/ml oligo(dT)-cellulose (Pharmacia, Piscataway, NJ) for 15 minutes at 4 °C with rotation. The oligo(dT)-cellulose was washed on a chromatography column (Bio-Rad, Hercules, CA) with the oligo(dT)-cellulose binding buffer and then eluted with deionized water. We did not observe problems with insolubility of the mRNA displayed protein fragments, either with this or the other libraries, presumably due to the low concentration of the mRNA displayed proteins during the operations, and from the solubilizing effect of the nucleic acid portion.

The eluate was mixed with 2× Ni-NTA binding buffer (1× is 6 M guanidinium chloride, 0.5 M NaCl, 100 mM sodium phosphate, 10 mM Tris(hydroxymethyl)amino methane, 10 mM 2-mercaptoethanol, 0.25% Triton X-100 (pH 8.0)) and then incubated with Ni-NTA agarose (Qiagen, Valencia, CA) for one hour at 4 °C with rotation. The Ni-NTA agarose was then washed with Ni-NTA first wash buffer (8 M urea, 0.5 M NaCl, 100 mM sodium phosphate, 10 mM Tris(hydroxymethyl)amino methane, 10 mM 2-mercaptoethanol, 0.25% Triton X-100 (pH 6.3)) and then with a gradient of increasing amounts of Ni-NTA second wash buffer (0.5 M NaCl, 10 mM Tris(hydroxymethyl)amino methane, 10 mM 2-mercaptoethanol, 0.25% Triton X-100 (pH 8.0)), and then was eluted with Ni-NTA elution buffer (0.25 M imidazole, 0.5 M NaCl, 10 mM Tris(hydroxymethyl)amino methane, 10 mM 2-mercaptoethanol, 0.25% Triton X-100 (pH 8.0)) for one hour at 4 °C with rotation. EDTA was added to the eluate to give 5 mM.

The buffer was exchanged into FLAG binding buffer (150 mM NaCl, 50 mM Hepes, 0.25% Triton X-100 (pH 8.3)) on a gel filtration column (Pharmacia, Piscataway, NJ), and the solution was then incubated with FLAG M2 agarose (Sigma, St Louis, MO) for one hour at 4 °C with rotation. The mRNA displayed proteins were then eluted with the same buffer containing five molar equivalents of FLAG elution peptide (Sigma, St. Louis, MO) to the FLAG M2 agarose for one hour at 4 °C with rotation.

The volume of the eluate was reduced by lyophilization and the buffer was exchanged to RT buffer (50 mM Tris(hydroxymethyl)amino methane, 75 mM KCl, 3 mM $MgCl_2$ (pH 8.3)) on a gel filtration column (Pharmacia, Piscataway, NJ). The mRNA displayed proteins were then reverse transcribed with Superscript II (Gibco BRL, Rockville, MD) to give a cDNA library of 4×10^{12} different members, which were then amplified using PCR. The resulting dsDNA was gel purified using native PAGE in 1 × TBE with an additional 50 mM NaCl and split into two equal parts that were restricted with either *Bbs*I or *Bbv*I. The resulting fragments were purified using native PAGE and the larger of the resulting fragments was collected and ligated together using phage T4 DNA ligase and the product was again purified using native PAGE. Repeating the restriction and ligation upon the resulting product yielded the final library with four contiguous random regions, each derived from a different combination of cassettes. This final library was then translated in a 10 ml reaction to give 7×10^{13} unique mRNA displayed proteins.

Construction of the patterned library

Three different cassettes were synthesized, each of which contains a random sequence region flanked by constant elements similar to those described for the R cassette, except for the absence of the FLAG tag sequence. The only differences between the α, β and γ cassettes are in the random regions. Cassette α, which encodes peptides with P/N periodicity consistent with amphipathic α-helix formation, has the following codon sequence in its random region: 12332332234, where one encodes a mixture of Cys and Trp, two encodes a mixture of non-polar residues, three encodes a mixture of polar residues, and 4 encodes a mixture of Asn, Asp, His and Tyr. Therefore, the sequence of the entire cassette is: 5′TTCTAATACGACTCACTATAGGGACAATTACTATT TACAATTACAATGGACGAGAAGACCCACTG(29 random nucleotides)ACTGGTCTGGCGGCTGCCACCATC ACCACCATCACAGCAGCGCC. The codon sequence of the β cassette is: 15656565654, where 5 and 6 encode mixtures of polar and non-polar amino acid residues, respectively. The codon sequence of the γ cassette is: 17777777774, where 7 encodes a mixture of all 20 amino acid residues. Each cassette was initially synthesized as a 69 nucleotide long oligonucleotide (encoding the random regions and some flanking constant sequence), and the full-length, double-stranded cassette was then generated by PCR with primers that added the remaining sequence.

The pre-selection was performed on the β cassette using the same protocol as described for cassette R, with the following exceptions: after binding the displayed peptides to the Ni-NTA agarose, the column was washed with Ni-NTA binding buffer and then eluted with the same buffer with 100 mM imidazole (pH 8). The imidazole was then removed by gel filtration and the Ni-NTA purification was repeated a second time. The imidazole was again removed by gel filtration and the RT step was performed as described above. The number of displayed proteins synthesized in this protocol was 4×10^{12} and the number of output molecules was 1.2×10^{11}. The FLAG purification was also omitted. The α and γ cassettes were not pre-selected, as they had a low incidence of frameshift errors (since their relatively small size, 69 nucleotides, allowed for efficient removal of shorter products on denaturing PAGE) and stop codons (due to codon design).

Each of the three cassettes was amplified by PCR, gel-purified using native PAGE, and separated into two equal aliquots for separate restriction by either *Bbs*I or *Bbv*I, as described for cassette R. The six fragments were then gel purified (native PAGE), mixed together and ligated. The resulting cassette-dimers were then treated exactly as was the R cassette, such that after two more successive cutting and ligation cycles, each library molecule contained eight cassettes. The ratio of the cassettes in the final product was $\alpha{:}\beta{:}\gamma = 44{:}45{:}11$. This final library was then translated in a 10 ml reaction to give 1.2×10^{14} unique mRNA displayed proteins.

Construction of the structure-based library

Each segment (AB, CD, EF, GH, and IJ) was made through a mutually primed extension of 1×10^{14} molecules using Deep Vent polymerase (New England Biolabs, Beverly, MA). The codons that correspond to the regions that will remain constant were optimized for translation in rabbit reticulocyte lysate (Wada *et al.*, 1992). These segments were then amplified with their respective primers with auxiliary segments such that the T7 promoter and TMV translational enhancer directly precedes the initiating ATG, and a His_6 tag sequence is added at the 3′ end (the FLAG epitope was encoded in the 5′ extending oligonucleotide). RNA was produced from these libraries and translation of the modified constructs yielded approximately 10^{13} mRNA displayed proteins in 1.5 ml *in vitro* translation reactions for each cassette as described above.

The pre-selection was performed as described for the R cassette of the random library with these exceptions: 8 M urea was used in place of 6 M guanidinium chloride during the Ni-NTA agarose purification. Two successive Ni-NTA purifications were performed by removing the imidazole by gel filtration as described for the patterned library. After precipitation of the Ni-NTA eluate, the

mRNA displayed proteins were reverse transcribed and the FLAG purification was performed as described for the R cassette. The final yield for each cassette was around 10 % of the starting amount of mRNA displayed proteins giving a diversity of molecules ranging from 1×10^{11} to 3×10^{12}.

The cDNA from the FLAG eluate was then PCR amplified using Deep Vent polymerase to generate $>10^{14}$ molecules. Cassettes AB, GH, and IJ were cut with *Fok*I (New England Biolabs, Beverly, MA); cassettes CD and EF were cut with *Fok*I and *Ban*I (New England Biolabs, Beverly, MA) in a double digestion to produce asymmetric ligation junctions. The fragments were purified and ligated as described above to give a final complexity of 1.2×10^{14}. This final library was then translated in a 10 ml reaction to generate approximately 1.3×10^{13} unique mRNA displayed proteins before purification.

One of the cassettes (GH) was anomalous in that the proportion of intact sequences did not seem to increase dramatically upon selection for ORFs (Table 3). Upon careful examination, it appears that one of the primers used to amplify library GH had an extra nucleotide. This unfortunate error caused the selection to enrich those sequences that had a one base deletion to compensate for the one base insertion in order to remain in frame. The results from sequencing show that the deletions occur within a short window of sequence that is flanked by two out-of-frame stop codons. Fortunately, the region is separated by the *Fok*I cleavage site so half of the deletions were cut away and half were retained in the coding region. This had the effect of reducing the overall library complexity of the structure-based library twofold.

Acknowledgments

The authors thank the members of the Szostak Laboratory, in particular Jonathan Davis, Jack Pollard and Jonathan Urbach, for helpful suggestions and Pamela Svec for assistance with cloning and sequencing. The authors thank Michael New of the NASA Ames Research Center for the amino acid composition iteration code. We acknowledge the Cancer Research Fund of the Damon Runyon-Walter Winchell Foundation, Hoechst AG, the NASA Astrobiology Institute and the NIH for grant support.

References

Andreotti, G., Cubellis, M. V., Palo, M. D., Fessas, D., Sannia, G. & Marino, G. (1997). Stability of a thermophilic TIM-barrel enzyme: indole-3-glycerol phosphate synthase from the thermophilic archaeon *Sulfolobus solfataricus*. *Biochem. J.* **323**, 259-264.

Boder, E. T. & Wittrup, K. D. (1997). Yeast surface display for screening combinatorial polypeptide libraries. *Nature Biotechnol.* **15**, 553-557.

Chiang, C. M. & Roeder, R. G. (1993). Expression and purification of general transcription factors by FLAG epitope-tagging and peptide elution. *Pept. Res.* **6**, 62-64.

Creighton, T. E. & Yanofsky, C. (1966). Indole-3-glycerol phosphate synthetase of *Escherichia coli*, an enzyme of the tryptophan operon. *J. Biol. Chem.* **241**, 4616-4624.

Eisenberg, D., Weiss, R. M. & Terwilliger, T. C. (1984). The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl Acad. Sci. USA,* **81**, 140-144.

Forrer, P., Jung, S. & Pluckthun, A. (1999). Beyond binding: using phage display to select for structure, folding and enzymatic activity in proteins. *Curr. Opin. Struct. Biol.* **9**, 514-520.

Georgiou, G., Stathopoulos, C., Daugherty, P. S., Nayak, A. R., Iverson, B. L. & Curtiss, R., III (1997). Display of heterologous proteins on the surface of microorganisms: from the screening of combinatorial libraries to live recombinant vaccines. *Nature Biotechnol.* **15**, 29-34.

Hanes, J. & Pluckthun, A. (1997). *In vitro* selection and evolution of functional proteins by using ribosome display. *Proc. Natl Acad. Sci. USA,* **94**, 4937-4942.

Hecker, K. H. & Rill, R. L. (1998). Error analysis of chemically synthesized polynucleotides. *Biotechniques,* **24**, 256-260.

Hennig, M., Darimont, B., Sterner, R., Kirschner, K. & Jansonius, J. N. (1995). 2.0 Å structure of indole-3-glycerol phosphate synthase from the hyperthermophile *Sulfolobus solfataricus*: possible determinants of protein stability. *Structure,* **3**, 1295-1306.

Holm, L. & Sander, C. (1994). The FSSP database of structurally aligned protein fold families. *Nucl. Acids Res.* **22**, 3600.

Kamtekar, S., Schiffer, J. M., Xiong, H., Babik, J. M. & Hecht, M. H. (1993). Protein design by binary patterning of polar and non-polar amino acids. *Science,* **262**, 1680-1685.

Knochel, T. R., Hennig, M., Merz, A., Darimont, B., Kirschner, K. & Jansonius, J. N. (1996). The crystal structure of indole-3-glycerol phosphate synthase from the hyperthermophilic archaeon *Sulfolobus solfataricus* in three different crystal forms: effects of ionic strength. *J. Mol. Biol.* **262**, 502-515.

Liu, R., Barrick, J., Szostak, J. W. & Roberts, R. W. (2000). Optimized synthesis of RNA-protein fusions for *in vitro* protein selection. *Methods Enzymol.* **318**, 268-293.

Mattheakis, L. C., Bhatt, R. R. & Dower, W. J. (1994). An *in vitro* polysome display system for identifying ligands from very large peptide libraries. *Proc. Natl Acad. Sci. USA,* **91**, 9022-9026.

Porath, J., Carlsson, J., Olsson, I. & Belfrage, G. (1975). Metal chelate affinity chromatography, a new approach to protein fractionation. *Nature,* **258**, 598-599.

Roberts, R. W. & Ja, W. W. (1999). *In vitro* selection of nucleic acids and proteins: what are we learning? *Curr. Opin. Struct. Biol.* **9**, 521-529.

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

Sleat, D. E., Gallie, D. R., Jefferson, R. A., Bevan, M. W., Turner, P. C. & Wilson, T. M. (1987). Characterisation of the 5′-leader sequence of tobacco mosaic virus RNA as a general enhancer of translation *in vitro*. *Gene,* **60**, 217-225.

Smith, G. P. & Petrenko, V. A. (1997). Phage display. *Chem. Rev.* **97**, 391-410.

Szostak, J. W. (1992). *In vitro* genetics. *Trends Biochem. Sci.* **17**, 89-93.

Tutino, M. L., Scarano, G., Marino, G., Sannia, G. & Cubellis, M. V. (1993). Tryptophan biosynthesis genes trpEGC in the thermoacidophilic archaebac-

terium *Sulfolobus solfataricus*. *J. Bacteriol.* **175**, 299-302.

Wada, K., Wada, Y., Ishibashi, F., Gojobori, T. & Ikemura, T. (1992). Codon usage tabulated from the GenBank genetic sequence data. *Nucl. Acids Res.* **20**, 2111-2118.

Wilson, D. S. & Szostak, J. W. (1999). *In vitro* selection of functional nucleic acids. *Annu. Rev. Biochem.* **68**, 611-647.

Wolf, E. & Kim, P. S. (1999). Combinatorial Codons: a computer program to approximate amino acid probabilities with biased nucleotide usage. *Protein Sci.* **8**, 680-688.

Xiong, H., Buckwalter, B. L., Shieh, H.-M. & Hecht, M. H. (1995). Periodicity of polar and non-polar amino acids is the major determinant of secondary structure in self-assembling oligomeric peptides. *Proc. Natl Acad. Sci. USA,* **92**, 6349-6353.

Edited by P. E. Wright

(Received 8 November 1999; received in revised form 27 January 2000; accepted 28 January 2000)