Protein Engineering, Design & Selection vol. 18 no. 4 pp. 201–207, 2005
Published online April 22, 2005 doi:10.1093/protein/gzi020

An intein-based genetic selection allows the construction of a high-quality library of binary patterned *de novo* protein sequences

Luke H.Bradley[1], Ralph E.Kleiner[1], Anna F.Wang[1], Michael H.Hecht[1,3] and David W.Wood[2,3]

Departments of [1]Chemistry and [2]Chemical Engineering, Princeton University, Princeton, NJ 08544, USA

[3]To whom correspondence should be addressed.
E-mail: hecht@princeton.edu; dwood@princeton.edu

Combinatorial libraries of synthetic DNA are increasingly being used to identify and evolve proteins with novel folds and functions. An effective strategy for maximizing the diversity of these libraries relies on the assembly of large genes from smaller fragments of synthetic DNA. To optimize library assembly and screening, it is desirable to remove from the synthetic libraries any sequences that contain unintended frameshifts or stop codons. Although genetic selection systems can be used to accomplish this task, the tendency of individual segments to yield misfolded or aggregated products can decrease the effectiveness of these selections. Furthermore, individual protein domains may misfold when removed from their native context. We report the development and characterization of an *in vivo* system to preselect sequences that encode uninterrupted gene segments regardless of the foldedness of the encoded polypeptide. In this system, the inserted synthetic gene segment is separated from an intein/thymidylate synthase (TS) reporter domain by a polyasparagine linker, thereby permitting the TS reporter to fold and function independently of the folding and function of the segment-encoded polypeptide. TS-deficient *Escherichia coli* host cells survive on selective medium only if the insert is uninterrupted and in-frame, thereby allowing selection and amplification of desired sequences. We demonstrate that this system can be used as a highly effective preselection tool for the production of large, diverse and high-quality libraries of *de novo* protein sequences.

Keywords: combinatorial library/ high-quality library/ inteins/preselection/protein design

Introduction

With the development of powerful screening and selection methods, combinatorial libraries of *de novo* genes are emerging as important reagents for the discovery of proteins with novel functions and folds (Hecht *et al*., 2004). The successful isolation of desirable proteins from combinatorial libraries depends not only on the power of the screen (or selection), but also on the quality and diversity of the library. A persistent difficulty has been that the DNA synthesis technologies used to generate *de novo* gene libraries invariably yield unwanted mutations or deletions in a significant fraction of the sequences. Although these aberrant DNA strands typically can be cloned, their translation produces truncated peptides arising from frameshifts or unexpected stop codons. The presence of these non-productive sequences complicates screening strategies by increasing the burden on the selection system, while failing to provide additional valid candidate genes for evaluation.

Our approach for the construction of large combinatorial libraries of *de novo* genes has relied on the assembly of full-length genes from smaller gene segments (Kamtekar *et al*., 1993; West *et al*., 1999). The advantages of this strategy are (i) to minimize the incorporation of unwanted mutations and deletions as a result of DNA synthesis and (ii) to increase the diversity of full-length sequences through the use of combinatorial assembly. Although this strategy has successfully produced libraries that contain full-length assembled genes (Kamtekar *et al*., 1993; West *et al*., 1999), a high-quality library, which is both large and free of frameshifts and/or stop codons, has not yet been produced. It is therefore desirable to develop methods that will remove incorrect sequences from these libraries.

Among large combinatorial libraries of synthetic genes—particularly those encoding subdomains for subsequent assembly into full length *de novo* proteins—it can be assumed that some sequences in the library will misfold and/or aggregate. Therefore, it is important to ensure that *in vivo* systems designed to select open reading frames are not biased by the foldedness and/or aggregation state of the expressed gene segments.

In order to construct a high-quality combinatorial library of *de novo* genes, we have developed a novel system for screening segments of genes. The new system is based on a previously developed pMΔI†T-CM (Maltose binding protein-Intein-Thymidylate synthase-Cleaving Mutant) vector used to screen for intein function (Wood *et al*., 1999). In this system, synthetic DNA from a library of gene segments is inserted upstream from an intein–thymidylate synthase (TS) fusion protein. If the inserted gene segments are in-frame and devoid of stop codons, a tripartite precursor protein is expressed. Subsequent cleavage by the intein releases and activates the TS enzyme, thereby enabling TS-deficient *Escherichia coli* host cells to survive in selective media. Because the inserted gene segments are separated from the intein–TS fusion by a polyasparagine linker, the intein and TS can fold and function independently of the polypeptide encoded by the inserted gene segment. Moreover, because the selectable activity (TS) is cleaved away from the test sequence, the ultimate fate of the test sequence (e.g. misfolding or precipitation) has no bearing on the selection.

Here we describe the use of this conditionally lethal selection system to isolate in-frame, uninterrupted sequences from several synthetic libraries designed to encode segments of binary-patterned amphiphilic α-helices. These libraries of α-helical segments were subsequently assembled combinatorially into full-length genes encoding *de novo* four- helix bundle proteins. This new preselection strategy enabled the construction of a large and high-quality library of uninterrupted full-length genes.

Materials and methods

Reagents

Enzymes were purchased from New England Biolabs (Beverly, MA) and DNA oligonucleotides from Integrated DNA Technologies (Coralville, IA). DNA segments were purified using Qiagen Quickspin kits. All other chemicals were of reagent grade.

Construction of the preselection plasmid pPPV

The plasmid pMΔI†T-CM (Wood *et al*., 1999) was constructed by inserting a gene fusion composed of the fast-cleaving ΔI-CM *Mtu* mini-intein mutant with the T4 TS enzyme into pMal-c2 (New England Biolabs). The pPPV preselection vector was constructed by initially amplifying by polymerase chain reaction (PCR) a region containing the *Nco*I (2444) and *Eco*RI (2795) sites of the pMΔI†T-CM vector. A *Bgl*II site was introduced at the 5′ end of the insert so that reinsertion of this segment into pMΔI†T-CM would place the intein–TS domains in the incorrect (+2) reading frame relative to the pMal-c2 ribosome binding site. The pMΔI†T-CM vector and PCR insert were doubly digested with *Bgl*II and *Eco*RI, gel purified, ligated and transformed into competent D1210Δ*thyA*::Kan^{R} [F^{-}Δ(*gpt-proA*)62 *leuB*6 *supE*44 *ara*-14 *galK*2 *lacY*1 Δ(*mcrC-mrr*) *rpsL*20 (*Str*^{r}) *xyl*-5 *mtl*-1 *recA*13 *lacI*^{q}] cells. The correct insertion was confirmed by restriction analysis, PCR amplification and DNA sequencing.

Construction of individual inserts

A restoration fragment of 17 nucleotides containing *Bsi*WI and *Bgl*II restriction site overhangs was constructed from two synthetic oligonucleotides (forward restoration 5′-GTACGGCAGCAGCAAAA-3′, reverse restoration 5′-GATCTTTTGCTGCTGCC-3′). Oligonucleotides were phosphorylated at the 5′ end by treatment with T4 polynucleotide kinase (Roche) for 1 h at 37°C. The phosphorylated single-stranded oligonucleotides were then mixed, heated at 95°C for 30 min and allowed to cool to room temperature over 1 h. The double-stranded DNA was then ligated to doubly digested pPPV vector and transformed into competent D1210Δ*thyA* cells.

DNA encoding the $A\beta_{1-42}$ peptide was introduced by primer extension PCR of a synthetic $A\beta_{1-42}$ gene (Wurth *et al*., 2002). The primers were forward 5′-CACCACCACGCGTACGAATGGATGCGGAATTTCGC-3′ and reverse 5′-AGCCACCACGAGATCTTTCCCGCAATCACCACGCC-3′.

Primers were also used to disrupt the reading frame at the N-terminus (forward 5′-CACCACCACGCGTACGATGGATGCGGAATTTCGC-3′) and the C-terminus (reverse 5′-AGCCACCACGAGATCTTCCCGCAATCACCACGCC-3′) of the peptide. For each PCR product, restriction digests were performed with *Bsi*WI and *Bgl*II. The desired fragment was then purified, ligated into digested pPPV and subsequently transformed into D1210Δ*thyA* cells.

The full-length *de novo* S-824 four-helix bundle protein (Wei *et al*., 2003b) was amplified from its parental vector by primer extension PCR and then introduced into pPPV. The primers used were forward 5′-GAGGAGGAGGCGTACGAATGTATGGCAAGTTGAACG-3′ and reverse 5′-GACGACGACCCGGCCGACGGTGGACGAGCTCTTCC-3′.

Primers were also used to disrupt the reading frame at the N-terminus (forward 5′-GAGGAGGAGGCGTACGAAATGTATGGCAAGTTGAACG-3′) and the C-terminus (reverse 5′-GACGACGACCCGGCCGTACGGTGGACGAGCTCTTCC-3′). For each PCR product, restriction digests were performed with *Bsi*WI and *Eag*I. DNA was purified, ligated into doubly digested pPPV and subsequently transformed into D1210Δ*thyA* cells.

Construction of a library of inserts

Four test libraries (H1, H2, H3 and H4) of *de novo* sequences encoding binary-patterned amphipathic α-helices were constructed from four synthetic PAGE-purified oligonucleotides using an approach described previously (Kamtekar *et al*., 1993). For each library, the 5′ end insert contained a *Bsi*WI site and the 3′ end contained either a *Bgl*II or *Eag*I site for insertion into pPPV. For each library, restriction digests were performed with *Bsi*WI and either *Bgl*II or *Eag*I. The library segments were then purified to remove enzymes and nucleotides by Qiagen Quickspin kits, ligated into pPPV and subsequently transformed into D1210Δ*thyA* cells.

Growth conditions

After 1 h of incubation in SOC medium, transformed cells were pelleted and resuspended in ~200 μl of –THY medium (i.e. medium lacking thymine) (Belfort *et al*., 1990) and then plated on rich, non-selective medium (2×YT), –THY or –THY supplemented with 50 μg/ml thymine (+THY) agar containing 100 μg/ml ampicillin. Plates were incubated at 37°C in all experiments unless noted otherwise. DNA was obtained from colony PCR-amplified segments from isolated *E.coli* colonies. Sequencing was determined using the following sequencing primers: forward 5′-TTTTCACGAGCACTTCAC-3′ and reverse 5′-ACGACATGAATAGGCTTG-3′.

Results

Construction of the preselection vector pPPV

The pPPV preselection vector is based on a previously developed genetic selection system for monitoring and modulating intein function (Wood *et al*., 1999). This system includes the pMΔI†T-CM vector, which was constructed by fusing a rapidly cleaving engineered intein to TS and inserting this fusion downstream from the maltose binding protein (MBP) in the commercially available pMal-c2 vector. A linker encoding 10 asparagine residues separates the MBP from the intein–TS domain, allowing the intein to function independently of MBP. Upon translation of the MIT precursor (maltose binding protein–intein–TS), the intein cleaves the TS domain from the fusion protein, leading to production of active TS enzyme. Therefore, a functional pMΔI†T-CM plasmid enables *E.coli* strain D1210Δ*thyA* cells, which contains a chromosomal TS knockout, to grow on medium lacking thymine (Wood *et al*., 1999; this work, Figure 1).

The pPPV preselection vector was constructed by removing part of the MBP domain from the pMΔI†T-CM vector and replacing it with a multiple cloning site (MCS). (Previous findings demonstrated that a functional MBP is not required for the preselection.) The MCS includes unique restriction sites (*Bsi*WI, *Bgl*II and *Eag*I) and shifts the intein–TS fusion sequence to be out-of-frame with the ribosome binding site and start codon of the maltose binding protein (Figure 1). Therefore, for a cell to have a positive TS phenotype (i.e. grow on selective plates lacking thymine), a DNA fragment inserted

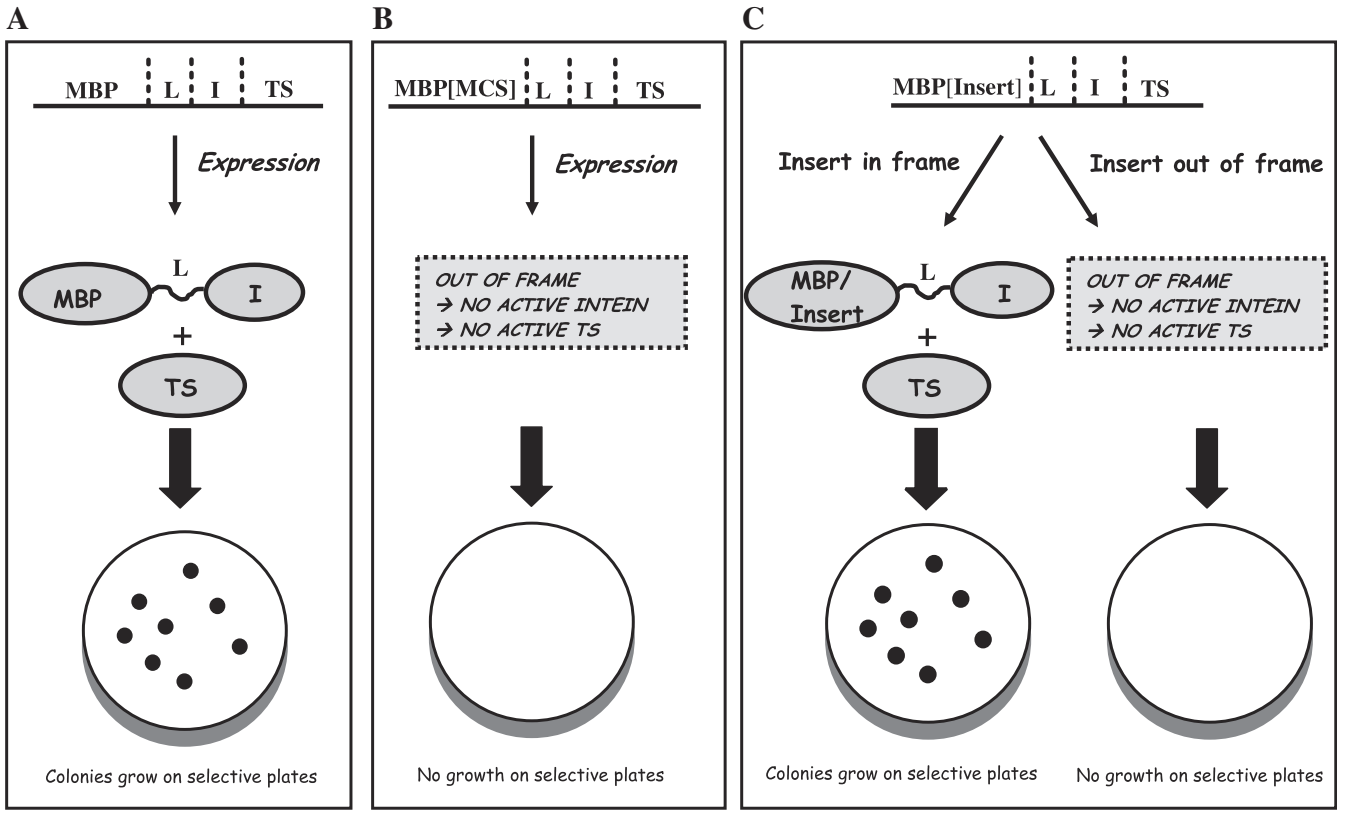

Fig. 1. Overview of the preselection system. (**A**) The pMΔI†T-CM vector contains a maltose binding protein domain (MBP) linked to an intein (I)–thymidylate synthase (TS) fusion by a polyasparagine linker (L). Upon expression, TS is activated by successful intein cleavage, thereby allowing transformed D1210Δ*thyA* cells to produce colonies on –THY plates. (**B**) The pPPV vector contains a multiple cloning site (MCS) within the MBP region. The I–TS fusion is not in-frame with the MBP start codon, hence TS is not activated upon translation. Consequently, no colonies grow. (**C**) Only an in-frame insert (left) will restore the I–TS reading frame, resulting in TS activation and thereby allow colonies to grow on –THY medium. Any insert that is not in-frame or contains a stop codon (right) would result in cell death on –THY medium.

into the MCS must restore the intein–TS reading frame (Figure 1). The new pPPV construct was verified by sequence analysis and used for the subsequent preselection studies.

Characterization of the preselection system

Ideally, this preselection system should allow the isolation of in-frame gene segments, which encode protein subdomains, from libraries of sequences that also contain frameshifts and/or stop codons. The pPPV system was initially tested for this capability by insertion of a small 17 bp double-stranded fragment designed to restore the reading frame between MBP and the intein. As expected, cells transformed with the restored pPPV grew on selective plates, whereas cells transformed with unmodified pPPV did not grow (Figure 2, Table I).

Next, the preselection system was challenged by insertion of a gene encoding a well-folded, full-length protein. For this study, the *de novo* binary-patterned four-helix bundle protein S-824 was chosen. This protein has been studied extensively and has been shown to be highly soluble and stable (Wei *et al*., 2003b). Recently, we determined its three-dimensional structure by NMR spectroscopy and demonstrated that it formed a four-helix bundle consistent with its design (Wei *et al*., 2003a). The gene encoding S-824 was inserted into pPPV to form an uninterrupted tripartite fusion protein. In separate experiments, variants of the S-824 gene that disrupt the reading frame at either the 5′ (N*) or 3′ (C*) ends of the gene were also inserted into pPPV. As with the restoration fragment, only cells transformed with the preselection vector containing the in-frame uninterrupted gene grew on selective plates (Figure 2, Table I). Cells transformed with either of the disrupting S-824 insertions, N* or C*, were unable to grow on selective plates (Table I). Hence the preselection system successfully distinguished between a full-length gene encoding a 102 amino acid protein and incorrect genes containing frameshifts.

Next, the preselection system was tested for its ability to tolerate inserted sequences that are in-frame, but prone to misfold or aggregate. The $A\beta_{1-42}$ peptide, which is the major component of amyloid plaque in Alzheimer's disease, was chosen for this test. A synthetic gene encoding $A\beta_{1-42}$ was inserted into pPPV. In addition, two negative controls were made with frameshifts at either the 5′ (N*) or 3′ (C*) ends of the $A\beta_{1-42}$ insert. As with the previous inserted sequences, only D1210Δ*thyA* cells expressing the $A\beta_{1-42}$ insert in the correct reading frame grew on the selective plates (Figure 2, Table I). The N* and C* frame-disrupting $A\beta_{1-42}$ inserts, and also unmodified pPPV produced no growth on selective plates (Figure 2, Table I). Hence the preselection system is able to select in-frame sequences even for sequences with a high propensity to misfold or aggregate.

Preselection of libraries of sequences

The pPPV preselection system was then used to isolate in-frame sequences from libraries of synthetic DNA encoding *de novo* protein sequences. Four test libraries were screened. These libraries encode four combinatorial collections of

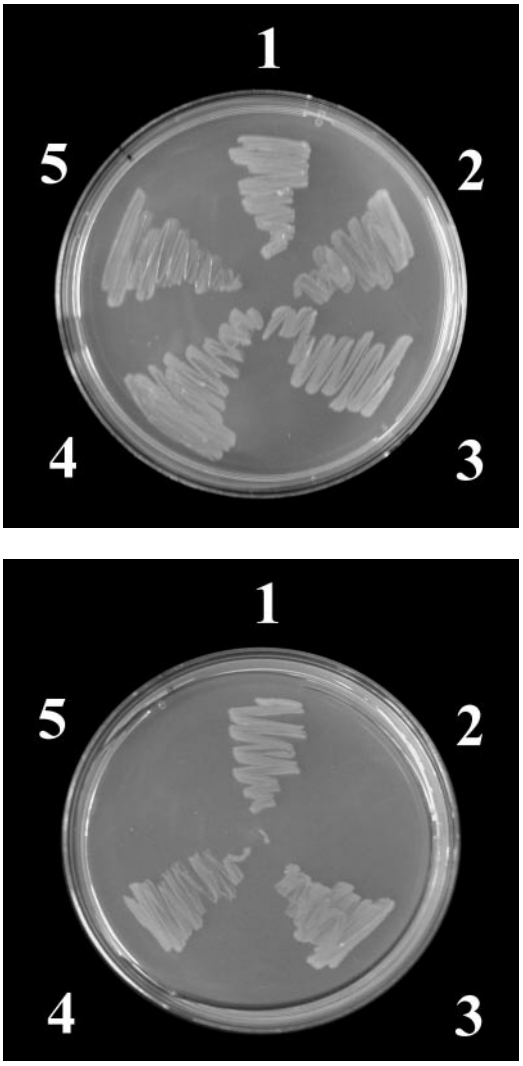

Fig. 2. Viability of transformed D1210Δ*thyA* cells depends on the inserted sequence. Cells were streaked on rich (2×YT, top) and minimal (–THY, bottom) plates containing ampicillin. The streaked colonies contain pPPV with the following inserts: (1) restoration fragment; (2) $A\beta_{1–42}$ N* frameshift; (3) $A\beta_{1–42}$; (4) S-824; and (5) pPPV with no insert. Cells lacking insert (5) or containing a frameshifted insert (2) did not survive on –THY plates. Inserts encoding a full-length well-folded protein (4) or the amyloid forming $A\beta_{1–42}$ peptide (3) permit cell growth.

Table I. Viability of D1210Δ*thyA* on selective (–THY) or non-selective (2×YT or +THY) media for cells transformed with pMΔI†T-CM or with pPPV containing various inserts; cells were incubated for 20 h at 37°C

Variant	Description	2×YT plates	Minimal (–THY)	Supplemented (+THY)
pMΔI†T-CM	Parent vector	+++	+++	++
pPPV	Preselection vector	+++	–	++
Restoration	17 bp restoration fragment	+++	+++	++
S-824	Full-length *de novo* S-824	+++	+++	++
S-824 N*	S-824 with 5′ frameshift	+++	–	++
S-824 C*	S-824 with 3′ frameshift	+++	–	++
$A\beta_{1–42}$	$A\beta_{1–42}$ peptide insert	+++	+++	++
$A\beta_{1–42}$ N*	$A\beta_{1–42}$ with 5′ frameshift	+++	–	++
$A\beta_{1–42}$ C*	$A\beta_{1–42}$ with 3′ frameshift	+++	–	++

Table II. Unselected sequences: analysis of four unselected libraries of synthetic sequences. DNA was ligated into pPPV, transformed into D1210Δ*thyA* cells and plated on non-selective (2×YT) medium containing ampicillin

Library	Total sequenced	In-frame	Frameshift	Self-closed	Internal stop codon	% in-frame
H1	33	30	2	1	–	91
H2	25	23	2	–	–	92
H3	24	21	3	–	–	88
H4	26	22	3	–	1	85
Total	108	96	10	1	1	89

binary-patterned amphipathic α-helices, which were designed for subsequent assembly into full-length *de novo* genes encoding 102 amino acid, four-helix bundle proteins. The libraries, designated H1, H2, H3 and H4, are 128, 122, 122 and 98 bp in length, respectively. The sequences in each library are combinatorially diverse, with polar amino acids (Lys, His, Glu, Gln, Asp and Asn) encoded by the degenerate DNA codon VAN and non-polar amino acids (Met, Leu, Ile, Val and Phe) encoded by the degenerate DNA codon NTN (where V = A, G or C and N = A, G, C or T).

The libraries were designed to be free of stop codons; however, libraries of synthetic DNA invariably contain some number of incorrect sequences with deletions, stop codons and/or frameshifts. Because libraries containing an abundance of incorrect sequences would be inappropriate for further studies, the preselection method is important to ensure that all, or nearly all, sequences in our libraries are free of errors and accurately encode the designed binary pattern of polar and non-polar amino acids required for amphipathic α-helices. Moreover, because the four libraries encode single amphipathic α-helices rather than full-length four-helix bundles, the expressed sequences are prone to oligomerize or aggregate (Xiong *et al.*, 1995). Therefore, it is essential that the preselection effectively screens for correct gene segment sequences, irrespective of whether the encoded amino acid sequences fold or aggregate.

pPPV plasmids with inserted test libraries were transformed into D1210Δ*thyA*. In all cases several thousand colonies were observed on non-selective rich medium. Sequence data from a total of 108 colonies from the four libraries plated on non-selective medium indicated that ~10–15% of the non-selected sequences were out-of-frame (Table II). Similar results were observed with the non-selective *E.coli* strain XL-1 Blue (data not shown). To assess whether our preselection system could 'weed out' these frameshifted sequences, the library was replica plated on to selective –THY medium plates. A total of 253 colonies, transformed with the H4 library, were replica plated from rich medium plates on to selective –THY medium. Approximately 78% of the colonies grew on selective –THY plates. This value is similar to the 85% correct gene segment sequences determined by direct DNA sequencing of the unselected H4 library (Table II).

For each library, 10^4 colonies were collected from the selective –THY plates (Table III). DNA sequence analysis of arbitrarily chosen colonies showed that all of the selected sequences were in-frame and devoid of stop codons (Table III). These preselected libraries of individual α-helices were then assembled combinatorially to produce a library of

Table III. Selected sequences: analysis of the preselection of four libraries of sequences. Library DNA was ligated into pPPV, transformed into D1210Δ*thyA* cells and plated on selective –THY medium plates

Library	Colonies selected	Sequenced	In-frame/ no stop codons
H1	10 300	5	5 (100%)
H2	10 000	7	7 (100%)
H3	13 600	6	6 (100%)
H4	11 100	8	8 (100%)
Total		26	26 (100%)

Table IV. Time dependence of the preselection: results of replica plating experiments on –THY plates

Incubation period at 37°C (h)	Colonies	Colonies sequenced	Sequences in-frame
18	13	4	4 (100%)
24	45	4	4 (100%)
30	158	8	8 (100%)
36	171	8	8 (100%)
42	197	8	8 (100%)

Replica plates were made from a 2×YT plate containing 253 colonies, incubated at 37°C for 18 h with the H4 library. Each of the sequences was unique and of the expected size.

full-length genes encoding four-helix bundle proteins (L.H.Bradley and M.H.Hecht, unpublished work). [With 10^4 correct sequences from each of the four libraries, combinatorial assembly of genes encoding full-length 102-residue four-helix bundles could, in theory, generate a library with a diversity of 10^{16}, i.e. $(10^4)^4$.] From this library of combinatorially assembled full-length genes, 21 sequences were determined. All 21 were found to be in-frame and devoid of stop codons through the full length of the gene. Hence the preselection system permitted the construction of a high-quality, highly diverse library of *de novo* protein sequences.

Optimization of the preselection system

To exploit the large combinatorial diversity of our libraries, we sought conditions that would enable the largest number of correct sequences to survive on –THY medium plates, while simultaneously removing all of the incorrect (e.g. frameshifted) sequences. To establish these conditions, 253 colonies containing the H4 library in pPPV were replica plated from rich 2×YT plates on to selective –THY plates and allowed to incubate for various lengths of time. The number of colonies increased with incubation time. After 18 h, 13 colonies were observed and this number grew to 197 colonies after 42 h (Table IV). Thus, after extended incubation, ~78% of the H4 library yielded colonies on selective –THY medium. Four colonies from the 18 and 24 h incubations and eight colonies from each of the 30, 36 and 42 h incubations were chosen for sequence analysis. All of the isolated sequences were found to be in-frame and free of stop codons (Table IV). Similar results were observed for libraries H1, H2 and H3 (data not shown). These findings demonstrate that extended incubation is required to ensure a high level of representation of the library on selective plates, with incubation times of 30–42 h yielding the greatest number of correct sequences. Furthermore, the system was shown to be insensitive to temperature over the range 23–37°C, as expected for a preselection system based on reading frame, as opposed to foldedness or aggregation state of the expressed polypeptide (data not shown).

Discussion

Our long-term goal is to use high-throughput screens and selections to find novel proteins with desirable activities from amidst large and diverse libraries of *de novo* sequences. To enhance the effectiveness of these downstream screens and selections, it is important to use libraries of genes that are free of frameshifts and/or stop codons. Our method for constructing full-length *de novo* gene libraries relies on combinatorial assembly of libraries of shorter fragments (Kamtekar *et al*., 1993; West *et al*., 1999). Among the advantages of this approach is an increase in the diversity of the full-length library. For example, four libraries containing only 10^4 sequences per library can be combined to yield a theoretical diversity of 10^{16}. Therefore, in order to take full advantage of the true diversity of a screened library, it is important to isolate successfully in-frame, uninterrupted segment sequences from several synthetic libraries.

Tests of the newly devised pPPV system demonstrated that only in-frame sequences of (i) a short restoration fragment, (ii) the Alzheimer's peptide, $A\beta_{1–42}$, or (iii) a well-folded, binary-patterned protein (S-824) yield viable cells on –THY plates. Control experiments using inserts with disrupted reading frames produced cells that cannot survive on selective plates. Hence the system readily differentiates between open reading frames and non-coding inserts.

To compare our newly developed system with previous work, it is important to note that previously described systems for the selection of high-quality sequences from expression libraries can be divided into two classes: (i) those that select (or screen) for folding and solubility (Kristensen and Winter, 1998; Sieber *et al*., 1998, 2001; Finucane *et al*., 1999; Maxwell *et al*., 1999; Waldo *et al*., 1999) and (ii) those that aim to isolated open reading frames (Daugelat and Jacobs, 1999; Cho *et al*., 2000; Lutz *et al*., 2002; Gerth *et al*., 2004). As described below, selections for folding and solubility are not suitable for screening gene fragments prior to assembly into full-length genes: appropriate screening of libraries of gene segments requires that the selection be based on the reading frame of the DNA and not the structure of the expressed polypeptide.

A number of previously described systems for the selection of uninterrupted sequences have shown limited tolerance for long sequences (Daugelat and Jacobs, 1999) or sequences with insoluble translation products (Maxwell *et al*., 1999; Waldo *et al*., 1999; Sieber *et al*., 2001; Lutz *et al*., 2002). In particular, systems in which there is intimate contact between the library insert and selectable functional domains may be less effective for screening long or misfolded/insoluble sequences (Cho *et al*., 2000). More recent systems for the selection of correct reading frames incorporate flexible linkers and/or inteins to uncouple the properties of test sequences from those of the selectable reporter proteins (Gerth *et al*., 2004; this work).

Because isolated segments of binary-patterned sequences have a propensity to misfold and/or aggregate (Xiong *et al*., 1995), it was essential for us to use a system in which the selection of gene segments from our libraries is not biased

by the foldedness and/or aggregation state of the encoded polypeptide segments. To test the tolerance of the pPPV preselection system for aggregated or misfolded sequences of approximately the same length as those in our libraries of gene segments, we inserted into pPPV a synthetic gene encoding the amyloidogenic peptide $A\beta_{1-42}$. Although $A\beta_{1-42}$ is known to prevent the proper folding of C-terminally fused GFP (Wurth *et al*., 2002), insertion of $A\beta_{1-42}$ into pPPV did not prevent intein function and consequently allowed the production of active TS (Table I, Figure 2).

The pPPV vector has four important features that might prevent $A\beta_{1-42}$ (or other misfolded sequences) from disrupting the activity of the TS reporter. First, the MCS is located within a truncated MBP domain containing the first 121 amino acids of MBP. These amino acids include most of domain I of MBP (Spurlino *et al*., 1991; Quiocho *et al*., 1997), which may enhance the solubility of sequences inserted into the MCS. Second, the pPPV vector was designed with a poly-Asn linker separating the MBP/MCS region from the C-terminal intein–TS reporter. This linker has been shown to allow significant changes to be made in the MBP/MCS domain without transmitting the affects of these changes to the intein–TS reporter (D.W.Wood, unpublished results). Thus, misfolded sequences cloned into the MCS are isolated from the C-terminal intein–TS reporter. Third, because selectable levels of TS activity occur without overexpression of the TS reporter fusion, the likelihood of aggregation is diminished. Finally, if a test sequence is nonetheless prone to aggregate, cleavage of TS enzyme from the test sequence ensures that subsequent accumulation of such aggregates does not interfere with the selection.

Although it is possible that cryptic ribosome binding sites might lead to false positives during preselection (Sieber *et al*., 2001), such occurrences will be extremely rare in our libraries. An intact ribosome binding site of seven base pairs at a proper distance from an ATG start codon has an expected frequency of 1 in 1 048 576. ($4^7 \times 4^3 = 1\,048\,576$). The shorter the sequences of DNA, the smaller is the number of 'windows' in which this sequence can occur. The DNA sequences in our libraries encoding segments of α-helices are slightly over 100 bp long (see above) and thus contain ~100 windows per sequence. Therefore, one can expect one cryptic ribosome binding site per 10 000 sequences. ($\sim 100 \times 1/1\,048\,576 \approx 1/10\,000$). As described above, we selected 10^4 colonies from each of the four libraries of α-helical segments. Therefore, on average, we expect one false positive per library. We consider this an acceptable level of false positives.

It is also possible that an occasional fragment of synthetic DNA might contain the deletion of an entire codon (or even several codons). Since such sequences leave the intein–TS fusion in the correct reading frame, they would not be excluded by the pPPV preselection system. Because such sequences arise only through the deletion of three bases (or six or nine, etc.), such false positives will be extremely rare.

We note that the libraries synthesized for this study were relatively free of errors and contained a surprisingly high frequency of open reading frames even prior to selection (Table II). When challenged with these high-quality libraries, the pPPV preselection strategy successfully removed those few sequences that were out-of-frame. For synthetic libraries that are not of such high quality, the impact of the new preselection system would be even more significant.

In summary, when the pPPV preselection system was challenged with four libraries of synthetic sequences of varying length, only in-frame sequences of the expected length were recovered (Tables III and IV). The four libraries of gene segments were then assembled combinatorially to yield a library of full-length genes encoding binary patterned four-helix bundles. All of the resulting full-length genes were found to be in-frame and free of stop codons. By preselecting libraries of smaller fragments before assembling the full-length gene, we were able to (i) create a high-quality library without frameshifts and/or stop codons, (ii) minimize the potential for false positives due to cryptic ribosome binding sites and (iii) enhance the combinatorial diversity of our library by assembling full-length genes (>300 base pairs) from preselected parts.

As combinatorial libraries of genes develop as important reagents for the discovery of proteins with new folds and activities, it is becoming increasingly important to devise methods to optimize the quality and diversity of these libraries. Here we have demonstrated that the pPPV preselection system selects sequences based on reading frame, while allowing the diversity of the sequences to be maintained. Hence this system provides an important tool that can be used towards the discovery of *de novo* proteins with novel functions.

Acknowledgements

The authors thank Georgios Skretas, Judy Hsii and Jesse Platt for assistance. L.H.B. was supported in part by a Postdoctoral Fellowship from the Council on Science and Technology (Princeton University). D.W.W. was supported by Princeton University Chemical Engineering Department startup funds. R.E.K. and A.F.W. received support from the Department of Chemistry Summer Undergraduate Research Program. Other support was from NIH grant R01 GM062869 (to M.H.H.).

References

Belfort,M., Ehrenman,K. and Chandry,P.S. (1990) *Methods Enzymol*., **181**, 521–539.

Cho,G., Keefe,A.D., Liu,R., Wilson,D.S. and Szostak,J.W. (2000) *J. Mol. Biol*., **297**, 309–319.

Daugelat,S. and Jacobs,W.R.Jr (1999) *Protein Sci*., **8**, 644–653.

Finucane,M.D., Tuna,M., Lees,J.H. and Woolfson,D.N. (1999) *Biochemistry*, **38**, 11604–11612.

Gerth,M.L., Patrick,W.M. and Lutz,S. (2004) *Protein Eng. Des. Sel*., **17**, 595–602.

Hecht,M.H., Das,A., Go,A., Bradley,L.H. and Wei,Y. (2004) *Protein Sci*., **13**, 1711–1723.

Kamtekar,S., Schiffer,J.M., Xiong,H., Babik,J.M. and Hecht,M.H. (1993) *Science*, **262**, 1680–1685.

Kristensen,P. and Winter,G. (1998) *Fold. Des.*, **3**, 321–328.

Lutz,S., Fast,W. and Benkovic,S.J. (2002) *Protein Eng*., **15**, 1025–1030.

Maxwell,K.L., Mittermaier,A.K., Forman-Kay,J.D. and Davidson,A.R. (1999) *Protein Sci*., **8**, 1908–1911.

Quiocho,F.A., Spurlino,J.C. and Rodseth,L.E. (1997) *Structure*, **5**, 997–1015.

Sieber,V., Pluckthun,A. and Schmid,F.X. (1998) *Nat. Biotechnol*., **16**, 955–960.

Sieber,V., Martinez,C.A. and Arnold,F.H. (2001) *Nat. Biotechnol*., **19**, 456–460.

Spurlino,J.C., Lu,G.Y. and Quiocho,F.A. (1991) *J. Biol. Chem*., **266**, 5202–5219.

Waldo,G.S., Standish,B.M., Berendzen,J. and Terwilliger,T.C. (1999) *Nat. Biotechnol*., **17**, 691–695.

Wei,Y., Kim,S., Fela,D., Baum,J. and Hecht,M.H. (2003a) *Proc. Natl Acad. Sci. USA*, **100**, 13270–13273.

Wei,Y., Liu T., Sazinsky,S.L., Moffet,D.A., Pelczer,I. and Hecht,M.H. (2003b) *Protein Sci*., **12**, 92–102.

West,M.W., Wang,W., Patterson,J., Mancias,J.D., Beasley,J.R. and Hecht,M.H. (1999) *Proc. Natl Acad. Sci. USA*, **96**, 11211–11216.

Wood,D.W., Wu,W., Belfort,G., Derbyshire,V. and Belfort,M. (1999) *Nat. Biotechnol.*, **17**, 889–892.
Wurth,C., Guimard,N.K. and Hecht,M.H. (2002) *J. Mol. Biol.*, **319**, 1279–1290.
Xiong,H., Buckwalter,B.L., Shieh,H.M. and Hecht,M.H. (1995) *Proc. Natl Acad. Sci. USA*, **92**, 6349–6353.

Received March 7, 2005; accepted March 9, 2005

Edited by Andrew Griffiths