Protein Engineering, Design & Selection vol. 17 no. 7 pp. 595–602, 2004
Published online August 25, 2004 doi:10.1093/protein/gzh068

A second-generation system for unbiased reading frame selection

Monica L.Gerth[1], Wayne M.Patrick[1] and Stefan Lutz[2]

Department of Chemistry and Center for Fundamental and Applied Molecular Evolution, Emory University, 1515 Dickey Drive, Atlanta, GA 30322, USA

[1]These authors contributed equally to this work

[2]To whom correspondence should be addressed.
E-mail: sal2@emory.edu

Reading frame selection of nucleic acids has important implications for protein engineering and genomics. Current methods are limited because selection of the gene of interest inevitably depends on the solubility of its translated product. Here we report the construction of the pInSALect vector, which provides strict reading frame selection without concomitant selection for protein solubility or folding. This plasmid incorporates the *cis*-splicing VMA intein sequence from *Saccharomyces cerevisiae* to facilitate the post-translational self-excision of the protein of interest, thereby eliminating potential aggregation problems. Results from two libraries of chimeric glycinamide ribonucleotide formyltransferases confirm the superior performance of pInSALect over existing reading frame selection systems.

Keywords: directed evolution/intein/methodology/protein engineering/selection

Introduction

Eliminating out-of-frame DNA sequences has become desirable, and sometimes even essential, for an increasing number of applications in protein engineering and genomics (Waldo, 2003; Yang *et al*., 2003; Zacchi *et al*., 2003). With the advent of methods for homology-independent *in vitro* recombination, reading frame selection has also become an important component of many directed evolution experiments (Cho *et al*., 2000; Lutz *et al*., 2001b; Sieber *et al*., 2001; Bittker *et al*., 2004). Prescreening a library to remove frame-shifted variants is an effective way to reduce the amount of molecular diversity to be assessed in subsequent steps, while simultaneously enriching for useful diversity and therefore increasing the probability of identifying a clone with the desired function.

In recent years, a variety of methods for identifying and removing frame-shifted DNA sequences have been developed. Initial approaches involved expressing the protein of interest as an N-terminal fusion to green fluorescent protein (Waldo *et al*., 1999) or chloramphenicol acetyltransferase (Maxwell *et al*., 1999) (Figure 1A) to provide an observable phenotype (i.e. fluorescence or antibiotic resistance). While these methods were developed as intracellular assays for folding, primarily in response to the requirement for improved protein yields and solubility for structural genomics initiatives, the assumption *a priori* is that the reading frame must also be conserved. However, these systems are prone to identifying false positives due to initiation of translation at internal ribosomal binding sites and this lessens their utility in the context of directed evolution (Sieber *et al*., 2001; Lutz *et al*., 2002).

A number of bipartite selection systems have been developed in an attempt to overcome the limitations of using simple fusion proteins to select for the maintenance of reading frame. In these systems, the DNA sequence of interest is positioned between 'head' and 'tail' reporters, both of which are required to give an observable phenotype (Figure 1B). Szostak and co-workers employed dual N- and C-terminal epitope tags to pre-screen mRNA display libraries that were constructed entirely *in vitro* (Cho *et al*., 2000). In contrast, the plasmid-based, *in vivo* systems pLAB (Seehaus *et al*., 1992), ORFTRAP (Daugelat and Jacobs, 1999) and pSALect (Lutz *et al*., 2002) require an in-frame, correctly oriented gene of interest to render a host cell antibiotic resistant. In pSALect, the head reporter—the Tat signal sequence—directs the post-translational export of the fusion protein to the periplasm (Yahr and Wickner, 2001; Palmer and Berks, 2003), whereas the tail reporter—β-lactamase—requires periplasmic export for function. As a result, only completely translated, in-frame fusion proteins (head + protein of interest + tail) can be exported to confer ampicillin resistance (Figure 1C). Two recent reports describe conceptually similar methods for reading frame selection in the phage display of genomic DNA libraries (Ansuini *et al*., 2002; Zacchi *et al*., 2003).

While the dual reporter systems have proven efficient at removing frame-shifted library members, they also inevitably impart selection pressure for the folding and solubility of the protein of interest (Sieber *et al*., 2001; Lutz *et al*., 2002; Zacchi *et al*., 2003; Bittker *et al*., 2004). For example, in pSALect an in-frame but aggregation-prone fusion protein would be unable to confer antibiotic resistance, especially as it has recently been shown that the Tat signal sequence does not export misfolded proteins (DeLisa *et al*., 2003). Selection against misfolded or insoluble variants may be helpful in many directed evolution applications, as these clones are unlikely to be those that display the evolved property of interest. However, it is noteworthy that many functional variants identified by directed evolution suffer from an impaired ability to fold (Lee *et al*., 2003). As it is impossible to quantify the degree of selection pressure for folding imparted by the existing systems, the risk is that functional variants will be eliminated in the course of pre-selection for reading frame. This is an especially pertinent concern for the SCRATCHY methodology (Lutz *et al*., 2001b), in which reading frame selection is carried out after generating single-crossover hybrids, but before these variants are recombined to yield multiple-crossover clones (Figure 2). Clearly, single-crossover hybrids that misfold may still be able to contribute useful diversity to the final, multiple-crossover library; it would therefore be undesirable to eliminate them at the pre-selection stage.

Here we describe an improved plasmid vector, pInSALect, for reading frame selection without a prerequisite for the

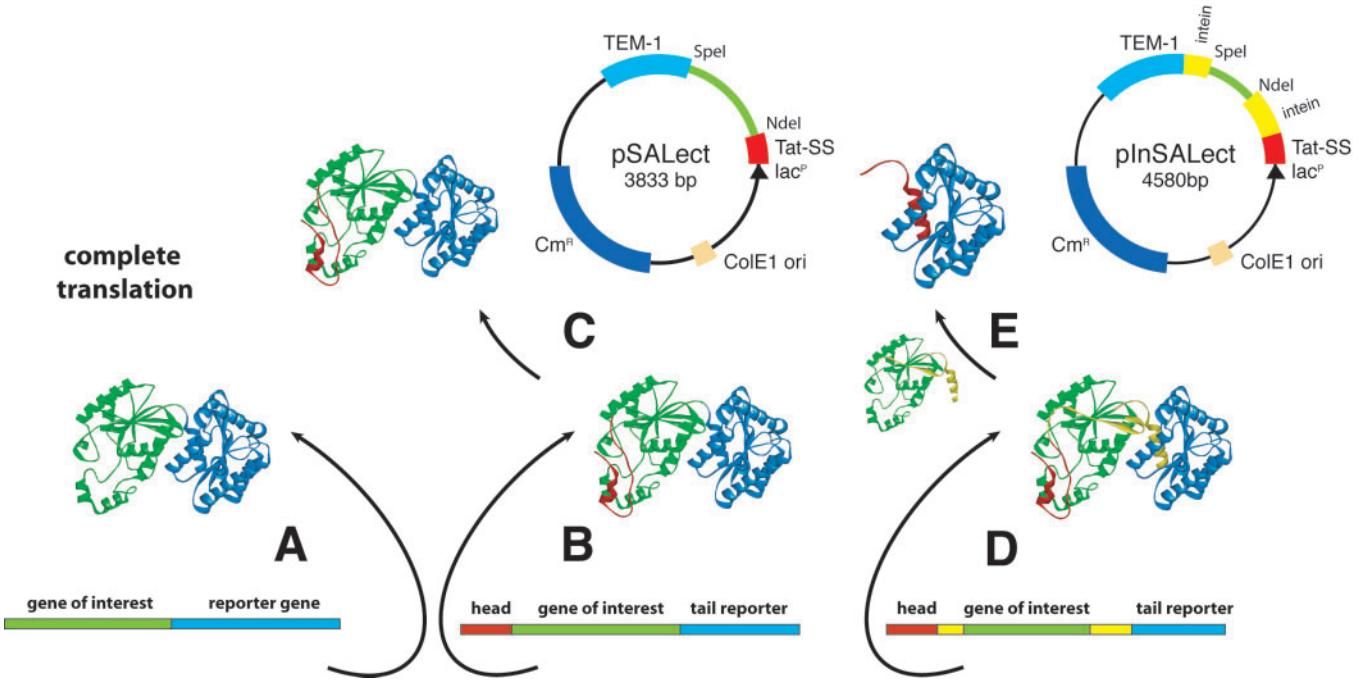

Fig. 1. Overview of reading frame selection systems. (**A**) Fusion of a gene of interest (green) to a C-terminal reporter gene (blue). (**B** and **C**) pSALect: the gene of interest is inserted between head (red) and tail (blue) reporters. The dual selection system eliminates false positives arising from internal initiation of translation, but is limited to genes of interest that encode non-aggregating fusion proteins. (**D** and **E**) pInSALect: the introduction of intein sequences (yellow) flanking the gene of interest facilitates the post-translational self-excision of the region linking the head and tail reporters. The concomitant ligation of the exteins (i.e. the head and tail sections) results in a stable protein and confers a selectable phenotype.

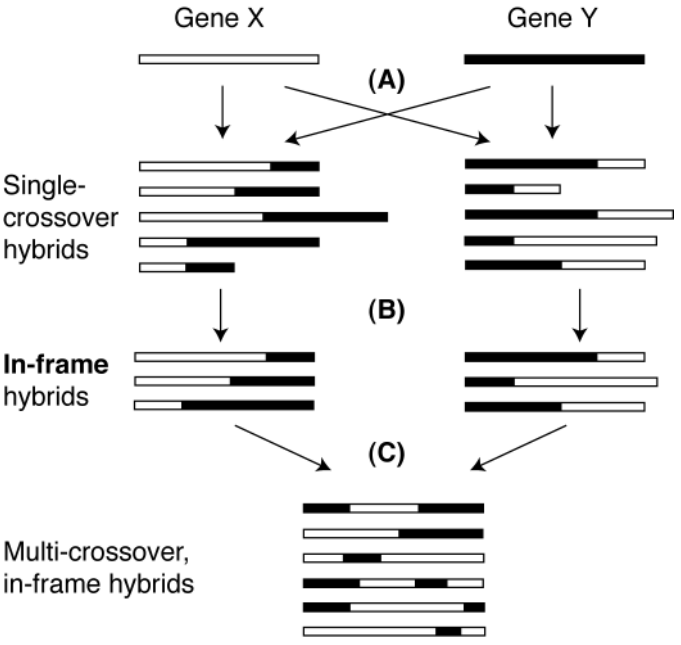

Fig. 2. The role of reading frame selection in SCRATCHY library construction. (**A**) Single-crossover hybrids of genes X and Y are constructed using incremental truncation (Ostermeier *et al*., 1999; Lutz *et al*., 2001a). (**B**) The single-crossover library is subjected to size selection and then sub-cloned into a suitable vector for reading frame selection. (**C**) The in-frame hybrids are re-amplified and recombined by DNA shuffling, yielding a library of multi-crossover hybrids in which all clones must retain a correct reading frame.

folding, solubility or function of the translated hybrid protein. This was accomplished by incorporating the gene encoding the VMA intein from *Saccharomyces cerevisiae* between the Tat 'head' and β-lactamase 'tail' of pSALect and re-engineering the cloning site for genes of interest into the intein sequence (Figure 1D). *Cis*-splicing inteins such as VMA catalyze their own post-translational excision through a series of nucleophilic displacements, concomitantly ligating their flanking sequences (exteins) via a peptide bond (Gogarten *et al*., 2002). Preceding the possible aggregation of the fusion protein, the splicing reaction takes place in the cytoplasm, excising the intein–protein of interest polypeptide and resulting in the ligation of the exteins, the Tat signal sequence and β-lactamase. This soluble antibiotic resistance marker is then exported to the periplasm regardless of the folding state of the excised protein portion (Figure 1E). The intein–protein of interest splicing product might accumulate in the cytoplasm or aggregate as insoluble cytoplasmic inclusion bodies; its folding status is no longer relevant to the selection. The construction of pInSALect ensures that self-catalyzed excision of the intein and the protein of interest will only occur if the latter maintains the correct reading frame. If the gene of interest contains a frame-shift or a premature stop codon, the C-terminal intein domain and β-lactamase reporter will be out of frame and antibiotic resistance will not be observed.

We have evaluated pInSALect by expressing in-frame and frame-shifted genes and by directly detecting the products of the splicing reaction. Additionally, libraries of hybrid genes created by incremental truncation (Lutz *et al*., 2001a) were subjected to reading frame selection in both pSALect and pInSALect and were subsequently assessed for biases in the crossover distributions of the selected variants that would reflect an underlying selection pressure for folding. The results demonstrated that pInSALect maintains the same, stringent selection for reading frame as pSALect, while eliminating undesired selection pressure for the folding of hybrid proteins.

Materials and methods

Materials

Enzymes were purchased from New England Biolabs (Beverly, MA) unless otherwise indicated. *Taq* DNA polymerase was obtained from Promega (Madison, WI). DNA samples were purified using the QIAquick and QIAprep purification kits (Qiagen, Valencia, CA) according to the manufacturer's protocols. The protease inhibitor cocktail for use in the purification of histidine-tagged proteins and all antibodies for immunoblotting were purchased from Sigma (St Louis, MO). Primers were purchased from Integrated DNA Technologies (Coralville, IA). All plasmid modifications were confirmed by DNA sequencing.

Construction of pInSALect

The gene encoding the VMA intein (GenBank No. AB093499) was amplified by polymerase chain reaction (PCR) from *S.cerevisiae* (commercial sample of baker's yeast) using oligonucleotides ScIN.for (5′-GCG ATT AAT GGG TGC TTT GCC AAG GGT ACC-3′) and ScIC.rev (5′-CGC GCT AGC GCA ATT ATG GAC GAC AAC CTG-3′), which also introduced restriction sites for *Ase*I and *Nhe*I (underlined). The resulting 1386 bp fragment was digested with *Ase*I and *Nhe*I and cloned into the *Nde*I/*Spe*I sites of pSALect (Lutz *et al*., 2002), which have compatible cohesive ends but result in a ligation product that cannot be recleaved. Next, overlap extension PCR (Horton *et al*., 1990) was used to introduce unique restriction sites for *Nde*I and *Spe*I into the endonuclease domain of the intein, between amino acid resides Gly273 and Gly275 (numbered according to amino acid position in the *S.cerevisiae* intein, beginning with Cys1 and ending with Asn454). Primary amplifications were performed with the following primer pairs: ScIN.for with 5′-ACC ACT AGT ACC ACC CAT ATG ACC TCT GAC AAC TTT AGA GTA-3′ (5′ fragment) and 5′-CAT ATG GGT GGT ACT AGT GGT ATT CGC AAT AAT CTT AAT ACT GAG-3′ with ScIC.rev (3′ fragment). The primer-encoded restriction sites are underlined. The full-length intein was reassembled by PCR with primers ScIN.for and ScIC.rev and the product was re-inserted into pSALect after digestion with *Ase*I and *Nhe*I. This strategy resulted in the replacement of Asn274 of the parental intein sequence with the linker His–Met–Gly–Gly–Thr–Ser.

Construction of test plasmids

The *Drosophila melanogaster* deoxyribonucleoside kinase (*Dm*dNK) gene was amplified by reverse transcription of embryonic mRNA (kindly provided by Professor Susan Abmayr, Pennsylvania State University, University Park, PA) using Qiagen's OneStep RT-PCR kit with primers 5′-CGC CAT ATG AAG TAC GCC GAG GGC ACC CAG-3′ and 5′-GCG GAA TTC CTC GAG ACT AGT TCA GGG CTG TTG GTT ACT TGA-3′. The resulting PCR product was digested with *Nde*I and *Xho*I and ligated with pET-16b (Novagen, Madison, WI). The reverse primer also introduced an *Spe*I restriction site (underlined) into the pET-16b vector.

The gene encoding human mitochondrial thymidine kinase 2 (hTK2) was isolated from a human pancreas cDNA library (Marathon-Ready, Clontech, Palo Alto, CA) by PCR with primers 5′-CGC CAT ATG TCA GTG ATC TGT GT CGA GGG C-3′ and 5′-CGC ACT AGT TCA TGG GCA ATG CTT CCG ATT CTC TGG-3′. The underlined *Nde*I and *Spe*I restriction sites were used to sub-clone *hTK2* into pET-16b. The hTK2 gene was corrected for codon usage in *Escherichia coli*; the rare arginine codons AGA and AGG were mutated to CGC at amino acid positions R79, R93, R99, R107, R129, R151, R155 and R157.

For sub-cloning into pSALect and pInSALect, the stop codon of each gene was replaced by a GGA triplet (underlined below), creating a glycine linker between the test protein and β-lactamase, using the forward primer T7 (5′-TAA TAC GAC TCA CTA TAG GG-3′) and the reverse primers 5′-CGC ACT AGT TCC GGG CTG TTG GTT ACT TGA GAT-3′ (*DmdNK*) or 5′-CGC ACT AGT TCC TGG GCA ATG CTT CCG ATT CTC-3′ (*hTK2*). Frame-shifted variants of *DmdNK* and *hTK2* were constructed by a single nucleotide insertion (underlined) in the reverse primers (*DmdNK*, 5′-CGC ACT AGT TCC GTG GCT GTT GGT TAC TTG AGA T-3′; *hTK2*, 5′-CGC ACT AGT TCC TCG GGC AAT GCT TCC GAT TCT C-3′). The PCR products were digested with *Nde*I and *Spe*I and ligated into pSALect and pInSALect.

Construction of pBlaFla

For western immunoblotting analyses, the FLAG tag sequence (DYKDDDDK) was introduced at the C-terminus of the pInSALect-encoded fusion protein by PCR amplification with primers 5′-CAT ATG GGT GGT ACT AGT GGT ATT CGC AAT AAT CTT AAT ACT GAG-3′ and 5′-CGC GAA TTC TCA CTT GTC GTC ATC GTC CTT GTA GTC TCC CCA ATG CTT AAT CAG TGA GGC-3′ (FLAG tag sequence underlined). The 1395 bp product was digested with *Spe*I and *Eco*RI and cloned into the corresponding *Spe*I/*Eco*RI sites of pInSALect.

Restreaking test

Escherichia coli DH5α-E (Invitrogen, Carlsbad, CA) were transformed with each plasmid, plated on LB–agar plates containing chloramphenicol (Cm; 50 μg/ml) and incubated at 37°C. Single colonies from each transformation were used to inoculate 2 ml cultures of LB–Cm (50 μg/ml). After overnight growth at 30°C, LB–Cm medium was used to dilute each culture to A_{600} = 0.2. An aliquot (4 μl) of each normalized culture was plated on LB–agar containing carbenicillin (Carb; 100 μg/ml) and, as a control, on Cm-containing plates. Plates were incubated at 22, 30 and 37°C.

Solubility test

Escherichia coli BL21(DE3)pLysS cells (Novagen) transformed with pET-16b(*Dm*dNK) and pET-16b(hTK2) were grown at 37°C in 50 ml LB broth containing Carb (100 μg/ml) and Cm (50 μg/ml) to an A_{600} of ∼ 0.6. Induction was for 3 h at 30°C in the presence of 1 mM isopropyl-β-D-thiogalactopyranoside. The cells were harvested by centrifugation and the resulting cell pellets were each resuspended in 5 ml of lysis buffer (50 mM potassium phosphate pH 8, 300 mM NaCl, 10 mM imidazole, 25 μl protease inhibitor cocktail). An aliquot of the resuspended cells was collected for SDS–PAGE analysis of the total cell fraction. Lysozyme was added to a final concentration of 0.1 mg/ml and the bacteria were lysed by sonication on ice. Benzonase nuclease (Novagen) was added to reduce the viscosity of the lysate and the sample was centrifuged for 30 min at 4500 *g*. The clarified supernatant was collected for the analysis of soluble cellular protein. The

remaining pellet was resuspended in 5 ml of lysis buffer for analysis of the insoluble fraction. Samples were analyzed by SDS–PAGE.

In vivo protein splicing assays

Transformed *E.coli* DH5α-E were grown to mid-log phase in LB medium containing Carb (100 μg/ml) and harvested by centrifugation. The soluble periplasmic fraction was obtained by the method of Neu and Heppel (1965) except that an ice-cold solution of $MgSO_4$ (5 mM) was used in place of pure water to resuspend the cell pellet. After collecting the cold osmotic shock fluid, it was buffered by the addition of Tris–HCl (pH 7.5; final concentration 20 mM) and concentrated using Amicon Ultra-4 spin filters (10 kDa molecular weight cut-off; Millipore, Billerica, MA). Immunoblots utilized the anti-FLAG M2 monoclonal antibody and were developed colorimetrically using a peroxidase-conjugated anti-mouse IgG with the SigmaFast peroxidase substrate tablet set.

Library construction

This study employed the library of *purN* and *hGART* single-crossover hybrids (the 'PGX' library) previously constructed for validating the pSALect system (Lutz *et al*., 2002). Hybrid genes were excised from pSALect by digestion with *Nde*I and *Spe*I and ligated directly into pInSALect that had been treated with the same enzymes. The products were desalted by ethanol precipitation and used to transform *E.coli* DH5α-E by electroporation. Cells were plated on LB–agar medium containing Cm (50 μg/ml) and, after 12 h at 37°C, colonies were recovered in 2×YT medium supplemented with glucose (2%, w/v) and glycerol (15%, v/v). Reading frame selection was carried out by replating the pSALect and pInSALect libraries on LB–ampicillin (100 μg/ml) and incubating at 21°C. Colonies appeared after 72–96 h and were analyzed by DNA sequencing.

The inverse incremental truncation library, consisting of *hGART–purN* (GPX) hybrid genes, was also investigated in the two reading frame selection systems. This library had been characterized previously (Lutz *et al*., 2001b). The GPX library was removed from pDIM (Lutz *et al*., 2001a) by digestion with *Not*I and *Spe*I and the reaction products were separated on a 2.5% (w/v) agarose gel. Hybrid genes of approximately parental size were excised from the gel and recovered using QIAquick spin columns. Library amplification was by PCR using the forward primer 5′-ATA GAT TTC AAG GAG ACA GTC CAT ATG-3′ and the reverse primer 5′-GCA CTA GTT CCC TCG TCG GCA GC-3′, with the underlined sequence replacing the stop codon from pDIM-GPX with a codon for glycine. The PCR products were restricted with *Nde*I and *Spe*I and ligated into the corresponding sites in pSALect and pInSALect as described above. Reading frame selection for both GPX libraries was performed as described for the PGX libraries.

Results and discussion

Vector construction

To construct a reading frame selection system without protein folding requirements, we redesigned pSALect through the introduction of the *cis*-splicing VMA intein, producing pInSALect. The *VMA* sequence was isolated with gene-specific primers from commercially available baker's yeast. The full-length intein includes a homing endonuclease domain (Chong *et al*., 1996), which we initially removed in order to minimize the size of the fusion protein encoded by pInSALect. While this domain is not required for protein splicing (Chong and Xu, 1997), its absence led to significantly impaired cell growth under selection conditions. We therefore incorporated the entire intein sequence in the pInSALect vector. The restriction sites for cloning genes of interest were engineered into the same loop of the endonuclease domain that had previously been shown to tolerate insertions without affecting splicing (Chong *et al*., 1998a,b).

An important element in the design of pInSALect was the incorporation of the Tat signal sequence. While the use of this signal sequence has been discussed in the context of pSALect (Lutz *et al*., 2002), it offers additional advantages for pInSALect. In requiring the completion of translation and protein folding prior to export (Yahr and Wickner, 2001; DeLisa *et al*., 2003; Palmer and Berks, 2003), the Tat pathway allows splicing to take place in the cytoplasm. In contrast, translocation by the Sec pathway can be either co- or post-translational, depending on the nature of the protein (Lee and Beckwith, 1986). Unbiased reading frame selection using the latter pathway could then rely on splicing in the comparatively oxidizing environment of the periplasm, the feasibility of which remains untested. Furthermore, the comparatively poor efficiency of protein translocation by the Tat pathway (DeLisa *et al*., 2004) is advantageous, as it ensures that fusion proteins remain in the cytoplasm long enough for the splicing reaction to occur.

Vector validation

The primary consideration with our new system was to ensure that it imparted a reliable selection for maintenance of reading frame. To test this, the *D.melanogaster* deoxynucleoside kinase (*Dm*dNK) gene (Johansson *et al*., 1999) was inserted into the *Nde*I/*Spe*I cloning site of pInSALect, generating pInSALect(*Dm*dNK). A frame-shifted version with a single nucleotide insertion in the last codon of the gene was constructed as a negative control. When *E.coli* harboring pInSALect(*Dm*dNK) were plated on LB–carbenicillin medium, colony growth was observed within 24 h at temperatures of 22 or 30°C. In contrast, no growth was observed under any conditions (22, 30 or 37°C; incubation times of up to 10 days) on plates with the frame-shifted gene in pInSALect. The results indicated that, as expected, disruption of the reading frame in a gene of interest is not compatible with survival of the host cell. These data were consistent with those previously reported for pSALect (Lutz *et al*., 2002). More importantly, the experiment demonstrated that the bacterial host could express a functional version of the extended fusion protein (763 amino acids plus the target protein), provided the correct reading frame was maintained.

The human mitochondrial thymidine kinase (hTK2) gene was cloned into pInSALect to assess whether it truly provided reading frame selection without a requirement for protein folding. This protein had previously been reported to form inclusion bodies when it was over-expressed in *E.coli*, unless the GroEL/ES system was co-expressed and a heat-shock response was induced (Barroso *et al*., 2003). Our own over-expression experiments confirmed that hTK2 partitioned into the insoluble fraction, in contrast to *Dm*dNK, which was expressed solubly (Figure 3A). We hypothesized that intein-catalyzed removal of insoluble hTK2 from the pInSALect-encoded fusion protein would enable growth under selection conditions, while the fusion protein expressed from pSALect(hTK2) would not confer antibiotic resistance. Plating the two clones on

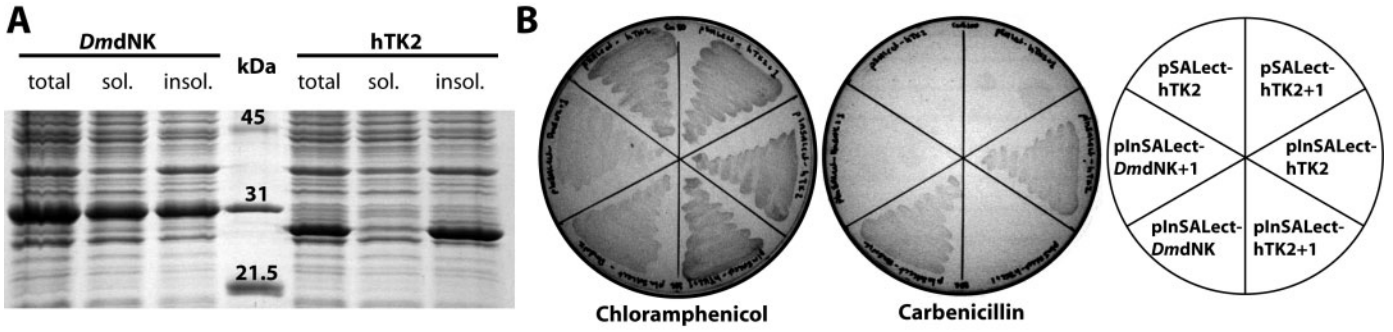

Fig. 3. Protein solubility and reading frame selection. (**A**) SDS–PAGE analysis comparing the total, soluble and insoluble cellular fractions when the deoxyribonucleoside kinases from *D.melanogaster* (*Dm*dNK) and human (hTK2) are over-expressed in *E.coli*. While expression of *Dm*dNK results in significant amounts of soluble protein, hTK2 is found exclusively in the insoluble fraction. (**B**) Reading frame selection of the two kinases in pSALect and pInSALect at 30°C. Growth on LB–chloramphenicol (left) is comparable among all samples. On the LB–carbenicillin selection plate (right), the insolubility of hTK2 does not affect growth of *E.coli* harboring pInSALect(hTK2); however, no colonies are detected with the pSALect clone. The *Dm*dNK-expressing clone serves as a positive control. The frame-shifted versions of both kinases (*Dm*dNK + 1 and hTK2 + 1) serve as negative controls.

LB–carbenicillin confirmed that this was the case (Figure 3B). A variant that incorporated a +1 frame-shift in *hTK2* was not viable under selection conditions in either pSALect or pInSALect (Figure 3B). Critically, then, it was shown that misfolding proteins such as hTK2 could be positively selected by the pInSALect system provided they maintained a correct reading frame, while pSALect would eliminate such clones owing to their insolubility.

Cellular and periplasmic fractions of *E.coli* were analyzed to confirm that the fusion protein was undergoing intein-catalyzed splicing. Although fusion protein expression in pInSALect is under control of the *lac* promoter, typical experiments rely on the natural leakiness of the promoter rather than the addition of an inducer, in order to minimize the impact on the host cells and to avoid saturating the Tat export pathway. Consequently, the concentration of fusion protein is generally low, complicating the detection of splicing products by standard staining techniques. A derivative of pInSALect named pBlaFla was therefore constructed, in which the FLAG epitope sequence was added to the C-terminus of β-lactamase to facilitate detection by immunoblotting. In growth experiments with the in-frame and frame-shifted *Dm*dNK genes, pBlaFla conferred an identical phenotype to pInSALect. The analysis of total cellular protein fractions with anti-FLAG antibody M2 was inconclusive, owing to the high non-specific background (data not shown). The soluble periplasmic fractions from *E.coli* carrying pBlaFla, or with no plasmid (negative control), were subsequently analyzed separately. In each case the quantity of cells harvested was normalized using A_{600} readings; SDS–PAGE of the periplasmic fractions verified that the total protein loading was equal (Figure 4A). Immunoblotting of the positive control yielded a single reactive band close to the 31 kDa marker (Figure 4B). This size is consistent with that predicted for FLAG-tagged β-lactamase, implying that intein self-excision had occurred in the cytoplasm. No full-length fusion protein (~85 kDa) was observed. Although export of unspliced material may be possible, these data indicate that the majority undergoes post-translational splicing prior to export into the periplasm.

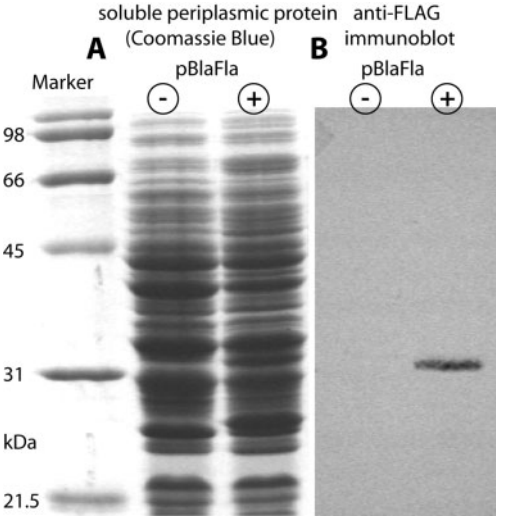

Fig. 4. SDS–PAGE analysis of soluble periplasmic fractions from *E.coli* grown with or without pBlaFla, a pInSALect derivative encoding a C-terminal FLAG tag. (**A**) Coomassie Brilliant Blue staining confirms an equal protein loading for each sample. (**B**) Immunoblotting the same gel with the anti-FLAG antibody M2 identifies a single band, corresponding in size to spliced β-lactamase, in the culture harboring pBlaFla.

Selection for reading frame from combinatorial libraries

Libraries of single-crossover hybrids of the glycinamide ribonucleotide formyltransferases from *E.coli* (*purN*) and human (*hGART*) (Lutz *et al*., 2001a,b, 2002) were used to evaluate the ability of pInSALect to identify in-frame clones in the context of a directed evolution experiment. The *purN-hGART* (PGX) and *hGART-purN* (GPX) libraries were subcloned into both pInSALect and pSALect. Based on colony counts when grown on LB–chloramphenicol medium, the sizes of the four libraries were estimated to be 9.7×10^4 clones for pInSALect-PGX, 1.0×10^6 for pInSALect-GPX, 6.4×10^4 for pSALect-PGX (Lutz *et al*., 2002) and 9.1×10^4 for pSALect-GPX.

After reading frame selection on LB–ampicillin medium, sequences from a total of 167 colonies (GPX, $n = 78$; PGX, $n = 89$) were analyzed. Only one sequence (in the pSALect-PGX sample) was out of frame at the crossover point. We hypothesized that this false positive was a result of ampicillin hydrolysis over the extended selection period (up to 96 h); the use of carbenicillin is therefore preferable. These results demonstated that each vector efficiently removed out-of-frame sequences from the libraries; however, they offered

no insights into whether in-frame sequences were also being excluded.

Unbiased reading frame selection

With pInSALect, we aimed to design a system that was strictly selective for reading frame, but that was not influenced by whether the protein of interest could fold. In such a system, one would expect the percentage of clones surviving reading frame selection to equal the percentage of the naïve (i.e. pre-selection) library that was in frame. Sequence analysis of colonies from the naïve libraries (GPX, $n = 63$; PGX, $n = 66$) showed that 14% (GPX) and 38% (PGX) respectively were in frame (Table I). These percentages serve as baseline values for comparison with survival rates on reading frame selection. In the case of pSALect, plating the libraries under selection conditions led to a dramatically reduced survival rate (Table I). These data suggested that only one in five in-frame hybrids survived reading frame selection from the pSALect-PGX library; the figure was closer to one in 13 for the pSALect-GPX library. The implication is that the pSALect system applies strong selection pressure beyond a requirement for maintenance of reading frame, resulting in the removal of up to 90% of in-frame hybrids because of their misfolding or limited solubility. In contrast, the proportions of cells surviving selection in the pInSALect libraries closely matched the percentages of in-frame clones in the naïve libraries. Selection in pInSALect therefore appears to be limited to reading frame alone.

Unbiased reading frame selection should not affect the distribution of crossover positions in any given library. This was investigated by sequencing members of the PGX and GPX libraries before and after reading frame selection and plotting the positions of the crossovers (e.g. for GPX in Figure 5). The naïve sequences from the pSALect and pInSALect libraries were pooled into a single dataset, as they originated from the same sample of digested GPX fragments and showed no significant difference in their crossover distributions. The average sizes of the PGX and GPX library clones were shown to remain constant over the course of reading frame selection (Table II).

Figure 5 provides a qualitative measure of the level of conservation in crossover distributions after reading frame selection. In pSALect, the hybrids that survived selection show a crossover distribution that has shifted considerably towards crossover points in the C-terminal region of the protein (amino acid positions 110–150). In contrast, the distribution of crossover points after selection with pInSALect closely matches that of the naïve library. A similar although less distinct result was obtained with the PGX library (data not shown), consistent with the apparent lack of bias observed previously for this library (Lutz *et al*., 2002) when a much smaller set of sequences was considered.

Table I. Cell survival rates upon reading frame selection

Library	Naïve (% in frame)	pSALect (% survival)	pInSALect (% survival)
PGX	38 (25 of 66)	8.0	38
GPX	14 (9 of 63)	1.1	15

The crossover positions in each library are clearly not normally distributed. We therefore employed the Kolmogorov–Smirnov (KS) test to compare the distributions of crossovers in naïve, pSALect and pInSALect libraries. The KS test compares

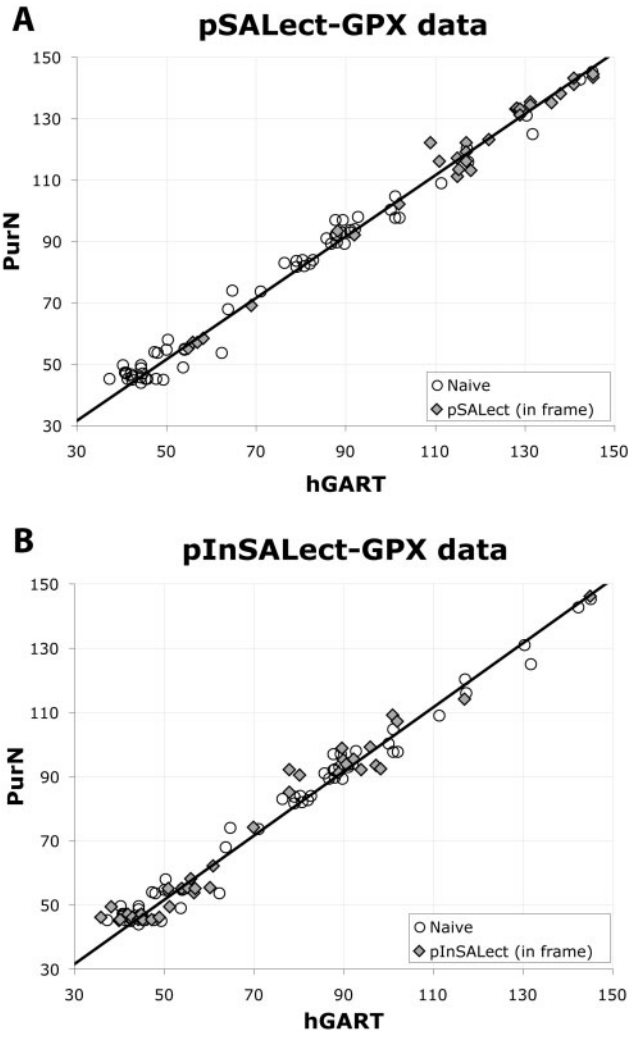

Fig. 5. Crossover distributions of GPX library members in pSALect (**A**) and pInSALect (**B**). Each circle (naïve library) and diamond (selected library) is plotted such that (x, y) represents ('last residue of hGART', 'first residue of PurN'). The diagonal ($y = x + 1$) therefore represents parental size fusion proteins, with points above the line indicating hybrid proteins with amino acid deletions and points below the line indicating insertions. The hybrid crossover distribution after selection in pInSALect accurately replicates the scattering in the naïve library. In contrast, additional selection for hybrid folding and solubility in pSALect results in a significant shift in the crossover distribution.

Table II. Library size distributions before and after reading frame selection

Library	Insert size (bp)[a]	Sample size (n)
PGX		
Naïve	608 ± 4	$n = 66$
pSALect	602 ± 6	$n = 58$
pInSALect	604 ± 6	$n = 31$
GPX		
Naïve	636 ± 3	$n = 63$
pSALect	638 ± 4	$n = 31$
pInSALect	635 ± 4	$n = 43$

[a]Mean ± 95% confidence interval.

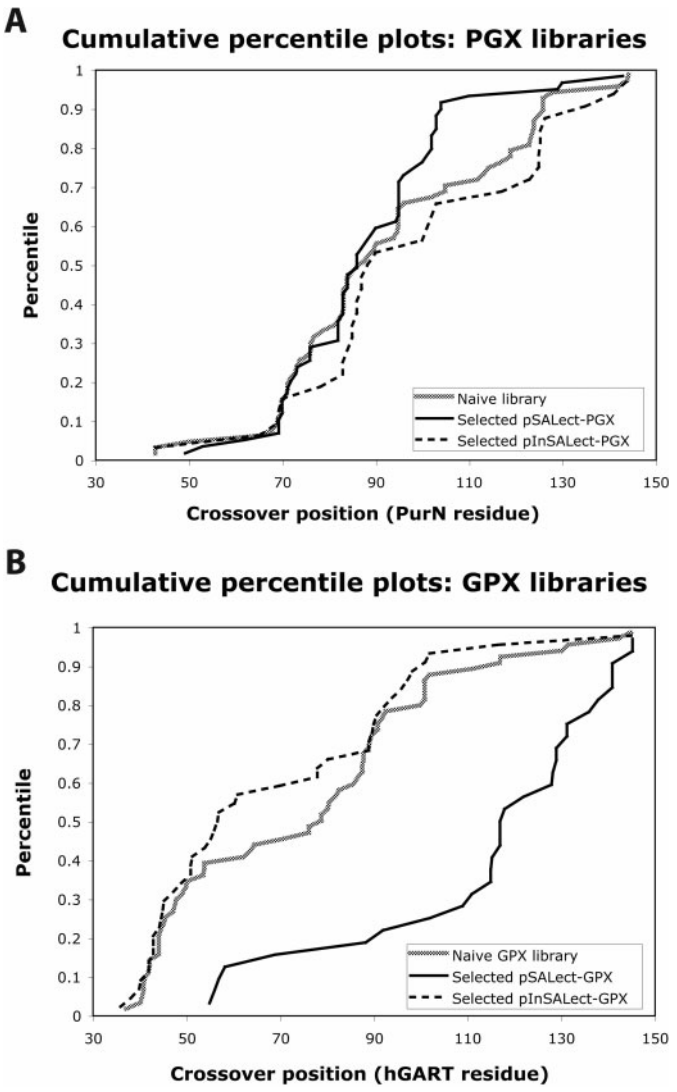

Fig. 6. Cumulative percentile plots of crossover positions for the PGX and GPX libraries. The plots for the PGX library (**A**) and the GPX library (**B**) highlight the similarities and differences in crossover distributions.

Table III. Summary of Kolmogorov–Smirnov statistics

Library	*D*	*P*
PGX		
Naïve vs pSALect	0.2492	0.036
Naïve vs pInSALect	0.1813	0.452
GPX		
Naïve vs pSALect	0.6472	$<5\times10^{-4}$
Naïve vs pInSALect	0.1846	0.316

cumulative percentile plots (Figure 6) and generates a test statistic, *D*, corresponding to the maximum difference between two cumulative distributions. This in turn allows calculation of *P*, the significance of the difference (Press *et al*., 2002). The *D* and *P* values, when each of the GPX and PGX libraries were compared, are summarized in Table III. The statistics confirmed that using pSALect for reading frame selection imparts a significant bias on the distribution of crossovers in the resulting library. This is most evident for the GPX library, where a *P* value of $<5\times10^{-4}$ corresponds to >99.95% confidence that the naïve and pSALect libraries are different. In contrast, both pInSALect libraries show crossover distributions that are statistically indistinguishable from the naïve datasets. Notably, crossovers in the naïve GPX library are biased towards the N-terminal half of the hybrid protein. The pInSALect system faithfully maintains this skew in the process of eliminating frame-shifted clones.

Conclusion

The qualitative and quantitative data presented here confirm that the pInSALect system reliably selects genes with the desired reading frame, while exerting no additional selection pressure for the folding or solubility of the resulting protein. By encoding a self-excising intein that also removes the protein of interest, export of β-lactamase to the periplasm becomes solely dependent on translation of an in-frame gene. Our experiments confirm that splicing does occur prior to export and that pInSALect is superior to pSALect for unbiased reading frame selection of combinatorial libraries.

We envision that pInSALect will find a number of applications in protein engineering and functional genomics. In the SCRATCHY methodology, the premature removal of a subset of in-frame hybrid genes is undesirable because it reduces diversity for the subsequent DNA shuffling step; pInSALect has therefore become an integral component of our protein engineering strategies. Although not intended for protein over-expression, additional applications of pInSALect could include the analysis of genomic or cDNA libraries, providing a convenient pre-screen for codon-optimized open reading frames. Moreover, pInSALect can be used to maintain folding-compromised protein variants in *E.coli*, effectively functioning as a pre-screen for reading frame alone. These variants may then be transferred to a non-bacterial host for protein expression or may become new targets for directed evolution approaches that improve heterologous expression and/or solubility. Finally, pInSALect offers a means to begin the directed evolution of proteins that have traditionally proven troublesome, such as membrane proteins and those that are insoluble in the absence of oligomerization partners.

Acknowledgements

The authors thank Dr Andrew Firth (University of Otago, New Zealand) for his advice and guidance on the statistical analysis of libraries. We would also like to acknowledge financial support from the National Science Foundation (NSF-MRI 0320786). Sequencing was performed at the Center for Fundamental and Applied Molecular Evolution (FAME) at Emory University.

References

Ansuini,H., Cicchini,C., Nicosia,A., Tripodi,M., Cortese,R. and Luzzago,A. (2002) *Nucleic Acids Res.*, **30**, e78.
Barroso,J.F., Elholm,M. and Flatmark,T. (2003) *Biochemistry*, **42**, 15158–15169.
Bittker,J.A., Le,B.V., Liu,J.M. and Liu,D.R. (2004) *Proc. Natl Acad. Sci. USA*, **101**, 7011–7016.
Cho,G., Keefe,A.D., Liu,R., Wilson,D.S. and Szostak,J.W. (2000) *J. Mol. Biol.*, **297**, 309–319.
Chong,S., Shao,Y., Paulus,H., Benner,J., Perler,F.B. and Xu,M.Q. (1996) *J. Biol. Chem.*, **271**, 22159–22168.
Chong,S., Montello,G.E., Zhang,A., Cantor,E.J., Liao,W., Xu,M.Q. and Benner,J. (1998a) *Nucleic Acids Res.*, **26**, 5109–5115.
Chong,S., Williams,K.S., Wotkowicz,C. and Xu,M.Q. (1998b) *J. Biol. Chem.*, **273**, 10567–10577.
Chong,S. and Xu,M.Q. (1997) *J. Biol. Chem.*, **272**, 15587–15590.
Daugelat,S. and Jacobs,W.R.,Jr (1999) *Protein Sci.*, **8**, 644–653.

DeLisa,M.P., Tullman,D. and Georgiou,G. (2003) *Proc. Natl Acad. Sci. USA*, **100**, 6115–6120.
DeLisa,M.P., Lee,P., Palmer,T. and Georgiou,G. (2004) *J. Bacteriol.*, **186**, 366–373.
Gogarten,J.P., Senejani,A.G., Zhaxybayeva,O., Olendzenski,L. and Hilario,E. (2002) *Annu. Rev. Microbiol.*, **56**, 263–287.
Horton,R.M., Cai,Z.L., Ho,S.N. and Pease,L.R. (1990) *Biotechniques*, **8**, 528–535.
Johansson,M., van Rompay,A.R., Degreve,B., Balzarini,J. and Karlsson,A. (1999) *J. Biol. Chem.*, **274**, 23814–23819.
Lee,C. and Beckwith,J. (1986) *Annu. Rev. Cell Biol.*, **2**, 315–336.
Lee,S.G., Lutz,S. and Benkovic,S.J. (2003) *Protein Sci.*, **12**, 2206–2214.
Lutz,S., Ostermeier,M. and Benkovic,S.J. (2001a) *Nucleic Acids Res.*, **29**, E16.
Lutz,S., Ostermeier,M., Moore,G.L., Maranas,C.D. and Benkovic,S.J. (2001b) *Proc. Natl Acad. Sci. USA*, **98**, 11248–11253.
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.
Neu,H.C. and Heppel,L.A. (1965) *J. Biol. Chem.*, **240**, 3685–3692.
Ostermeier,M., Shim,J.H. and Benkovic,S.J. (1999) *Nat. Biotechnol.*, **17**, 1205–1209.
Palmer,T. and Berks,B.C. (2003) *Microbiology*, **149**, 547–556.
Press,W.H., Teukolsky,S.A., Vetterling,W.T. and Flannery,B.P. (eds) (2002) *Numerical Recipes in C*. Cambridge University Press, Cambridge, pp. 620–628.
Seehaus,T., Breitling,F., Dübel,S., Klewinghaus,I. and Little,M. (1992) *Gene*, **114**, 235–237.
Sieber,V., Martinez,C.A. and Arnold,F.H. (2001) *Nat. Biotechnol.*, **19**, 456–460.
Waldo,G.S. (2003) *Methods Mol. Biol.*, **230**, 343–359.
Waldo,G.S., Standish,B.M., Berendzen,J. and Terwilliger,T.C. (1999) *Nat. Biotechnol.*, **17**, 691–695.
Yahr,T.L. and Wickner,W.T. (2001) *EMBO J.*, **20**, 2472–2479.
Yang,J.K., Park,M.S., Waldo,G.S. and Suh,S.W. (2003) *Proc. Natl Acad. Sci. USA*, **100**, 455–460.
Zacchi,P., Sblattero,D., Florian,F., Marzari,R. and Bradbury,A.R. (2003) *Genome Res.*, **13**, 980–990.

Received June 23, 2004; revised August 3, 2004; accepted August 13, 2004

Edited by Andreas Plueckthun