A universal, vector-based system for nucleic acid reading-frame selection

Stefan Lutz1, Walter Fast2 and Stephen J.Benkovic3

Department of Chemistry, The Pennsylvania State University, 414 Wartik Laboratory, University Park, PA 16802, USA
1Present address: Department of Chemistry, Emory University, 1515 Pierce Drive, Atlanta, GA 30322, USA
2Present address: Medicinal Chemistry Division, PHR 4.220, College of Pharmacy, University of Texas at Austin, Austin, TX 78712, USA
3To whom correspondence should be addressed. E-mail: sjb1@psu.edu

The identification of a nucleic acid sequence’s correct reading frame has important implications for homology-independent protein engineering techniques such as incremental truncation and SCRATCHY. We report the development and experimental implementation of a general in-frame selection system, pSALect, a plasmid vector that utilizes two marker sequences flanking the DNA of interest. This dual selection approach overcomes inconsistencies observed with traditional C-terminally fused reporter proteins. In the pSALect vector, sequences of interest are positioned between an N-terminal Tat-signal sequence and a C-terminal β-lactamase reporter. In-frame selection of the resulting three-domain protein is performed by growing colonies on ampicillin-containing plates, requiring full-length translation in order to link covalently the signal sequence to the lactamase for export into the periplasm. This dual selection scheme has been validated successfully using defined sequences of glycinamide ribonucleotide formyltransferases (GARTs) from Escherichia coli and human and, in contrast to C-terminal fusion systems, proved effective when applied towards the selection of in-frame constructs in an incremental truncation library.

Keywords: nucleic acids/pSALect vector/reading-frame selection

Introduction

During translation of genetic information, groups of three nucleotides form a codon. Defining the reading frame, these codons subsequently dictate the type and order of assembly of amino acid building blocks into polypeptides and proteins. A shift in the reading frame by one or two nucleotide positions causes a change in the translated information and will, at best, produce a polypeptide strand of non-sense amino acid sequences, but more likely will lead to early termination of translation.

For a series of combinatorial homology-independent protein engineering methods introduced in recent years, the separation of library members with the desired reading frame from nonsense sequences has become an important aspect. When applied to fusion constructs generated by methods such as incremental truncation for the creation of hybrid enzymes (ITCHY) (Ostermeier et al., 1999a; Lutz et al., 2001a) and sequence homology-independent protein recombination (SHIPREC) (Sieber et al., 2001), the selection for constructs with the correct reading frame can significantly expedite the subsequent screening of the hybrid libraries. For twin techniques such as SCRATCHY, a combination of incremental truncation and DNA shuffling (Lutz et al., 2001b) the removal of out-of-frame library members prior to homologous recombination is important to maximize the formation of multi-crossover hybrid genes with the correct reading frame.

Previous approaches towards the identification of genetic constructs with the desired reading frame have focused on the construction of C-terminal fusion proteins (Figure 1A) (Maxwell et al., 1999; Waldo et al., 1999). However, the screening or selection of target nucleic acid sequences fused to the reporter gene has been shown to be unreliable, potentially producing false positives (Sieber et al., 2001). It is believed that internal ribosomal binding sites (RBSs) in the gene of interest lead to productive translation of the reporter gene with only a fragment of the original target sequence being fused to the N-terminus of the reporter, thereby circumventing the original selection scheme (Figure 1B). Alternatively, a genetic selection based on the hybrid enzyme’s function can be employed to isolate in-frame constructs. Such an approach was used in our previous experiments with mixed success, providing sequences with the correct reading frame but also generating biased libraries (Lutz et al., 2001b).

In order to overcome all these limitations, we have constructed a vector, pSALect, that utilizes a dual selection system (‘head + tail’) to identify nucleic acid sequences with the desired reading frame (Figure 2). This design is similar to a previously reported vector used for deletion mutant selection in antibody libraries (Seehaus et al., 1992) but distinguishes itself by the nature of the signal sequence and the additional antibiotic selection marker.

The fundamental idea of the approach is to apply two functionally distinct, yet ultimately dependent, selection criteria on the gene of interest or more precisely on its reading frame. As outlined in Figure 1C, only a fusion protein with an intact head section will be transported to the periplasm whereas N-terminally truncated translation products, originating for example from internal RBSs, are ‘trapped’ in the cytoplasm (Figure 1D). Upon successful export of a fusion protein, constructs with a properly translated tail section, whose function is limited to a periplasmic environment, then develop a selectable phenotype. This concept was experimentally implemented by fusing the target gene sequence with an N-terminal Tat signal sequence, directing periplasmic transport of the fusion construct and with a C-terminal β-lactamase, providing antibiotic resistance to the host cell upon proper translation and export of the three-domain protein.

Materials and methods

All enzymes used were purchased from New England Biolabs (Beverly, MA) unless indicated otherwise. Taq DNA polymerase was obtained from Promega (Madison, WI). DNA
samples were purified using the QIAquick Gel and PCR purification kit (Qiagen, Valencia, CA) and eluted from the spin columns using the provided EB buffer (10 mM Tris, pH 8.5).

Construction of pSALect
Initially, a BamHI restriction site was introduced near position 816 in the starting vector pBC-SK+ (Stratagene, LaJolla, CA) by primer overlap extension, using the following mutagenic primers: 5’-CGGATAACAATTTCACACAGGATCCAGCTATGACCATGATTACGC-3’ and 5’-GCGTAATCATGGTCATAGCTGGATCCTGTGTGAAATTGTTATCCG-3’ and the corresponding outside primers. The resulting construct was named pBC-SK+ (BamHI).

Separately, the DNA sequence encoding for the C-terminal portion (amino acids 24–287) of β-lactamase (bla) was amplified off pBR322, using gene-specific primers. The gene fragment was cloned into pBC-SK+ via SpeI and EcoRI, creating pBC-SK+ (bla). The DNA sequence for the Tat signal sequence was isolated by PCR of torA (NCBI accession No. X73888) from Escherichia coli genomic DNA. Using primer overlap extension, the short fragment encoding the signal sequence portion of the gene was fused N-terminally to purN (NCBI accession No. M13747, isolated from *E.coli* genomic DNA) and the entire construct was ligated into pBC-SK+ (bla) via the NotI and SpeI restriction sites, creating pBC-SK+ (tat-purN-bla).

Next, pBC-SK+ (BamHI) was digested with NotI and XmnI and pBC-SK+ (tat-purN-bla) was digested with NotI and EcoRV. The vector portion of the former and the tat-purN-bla sequence of the latter were ligated, deleting an 800 basepair stretch including the *fl*-origin and creating pBC-SK+ (BamHI, tat-purN-bla). In a final step, the α-complementation fragment of β-galactosidase was removed by digestion of pBC-SK+ (BamHI, tat-purN-bla) with BamHI and SalI (introduced downstream of the *NotI* site). Following Klenow treatment to blunt-end the restriction sites, the plasmid was ligated intramolecularly, generating pSALect. The final construct was characterized by restriction analysis and DNA sequencing.

Vector validation
The wild-type genes for PurN and human GART were PCR-amplified from genomic DNA (*E.coli*) and cDNA (human), using gene-specific primers with flanking *NdeI* and *SpeI* restriction sites. In addition, two frame-shifted constructs were

---

**Fig. 1.** Conceptual overview of the traditional and the dual selection system to identify gene sequences with the correct reading frame. (A) Traditional C-terminal fusion between a gene of interest (green) and the reporter (blue). Complete translation results in a two-domain complex with an intact reporter protein with selectable features. (B) Internal ribosomal binding sites in the gene of interest, however, can generate N-terminally truncated versions of the fusion protein. These complexes also possess a functional reporter but circumvent the desired selection, producing false positives that contaminate the pool of selected sequences. (C) Overcoming these problems, the dual selection scheme fuses the gene of interest between a head (red) and tail (blue) selection marker. Each is based on a different selection criterion, yet the two are functionally linked. As outlined, the head section encodes for a peptide signal sequence, directing periplasmic transport of the C-terminal fusion construct. The tail marker in this example is a reporter that only functions in the periplasmic environment, for example a β-lactamase. (D) Although the dual selection system cannot prevent the translation of C-terminal protein fragments from internal ribosomal binding sites, the lack of the signal sequence in those constructs prevents allocation of the protein product into the periplasm and therefore prevents the manifestation of selectable phenotype by the truncated product.
generated with primers, carrying single and double nucleotide extensions as shown in Figure 3. In all constructs, the TGA stop codon was replaced by a GGA triplet, encoding for a glycine spacer between the target protein and the lactamase. The four test genes were cloned into pSALect via NdeI and SpeI, creating pSALect(purN), pSALect(hGART), pSALect(purN+1) and pSALect(hGART+2). The correct assembly of the constructs was confirmed by DNA sequencing.

Restreaking test
Following transformation of the individual plasmid constructs into DH5α-E (Invitrogen, Carlsbad, CA), cultures were plated on LB-agar plates containing chloramphenicol (Cm; 50 µg/ml) and incubated at 37°C. Individual colonies from all four transformations were restreaked on LB-agar plates containing ampicillin (Amp; 100 µg/ml) and as a control on Cm-containing plates. Plates were incubated at ambient temperature and 37°C.

Transformation test
A 1:1 mixture of the plasmids pSALect(purN) and pSALect(purN+1) was transformed into DH5α-E and plated on LB-agar plates (Cm; 50 µg/ml). After overnight incubation at 37°C, transformants were recovered by washing the plate with LB medium. Following a dilution to adjust the cell density, an aliquot of the suspension was replated on LB-agar containing either Cm (50 µg/ml; control plates) or Amp (100 µg/ml) and incubated at either 37°C or room temperature.

Library test
Plasmid DNA, containing a combinatorial library of hybrid genes of purN and hGART (2) was digested with NdeI/SpeI. The reaction mixture was loaded on a high-resolution agarose gel [TAE buffer with 2.5% NuSieve agarose (FMC Bioprod-
et al., 2001). These methods provide the means to break new ground in the study of protein properties and their applications. However, a central problem with these homology-independent approaches is the conservation of the correct reading frame. Generally, the reading frame of a gene of interest is randomized during the blunt-end fusion of the hybrid’s fragments, rendering two-thirds of the members of a fusion library useless and even potentially poisonous to the remaining sequence pool. The latter case is significant for dual-randomization protocols such as SCRATCHY where homology-independent methods (e.g., incremental truncation) are combined with homology-dependent approaches (DNA shuffling). Non-sense sequences from the former will rapidly decrease the probability of the latter generating hybrid enzymes with multiple crossovers while maintaining the correct reading frame.

Earlier attempts at in-frame selection of hybrid enzymes using a C-terminal fusion of a reporter protein have met with mixed success. Using the same model system as for pSALect, our laboratory initially designed GART gene fusions with a C-terminal neomycin-resistance marker to confer kanamycin resistance to in-frame constructs. While the total number of colonies per plate rapidly decreased above 5 µg/ml kanamycin, sequence analysis of the gene fusions of survivors (~10–25% of original colony count) indicated no enrichment of in-frame constructs. Sequence analysis of the two parental genes revealed more than a dozen internal start codons, many flanked by potential Shine–Dalgarno sequences, that could function as RBSs. A possible cause of the mixed success of the C-terminal fusion approach can be traced to the presence or absence of additional RBSs within a target DNA sequence. Similar observations with internal and external RBSs have been reported elsewhere (Brueck and Mayfield, 1998; Hrzenjak et al., 2001; Sieber et al., 2001). As outlined in Figure 1B, the internal RBSs can generate N-terminally truncated fusion proteins. These are fully functional with respect to the C-terminal selection marker, yet no longer require expression of the full-length fusion protein to achieve the selected properties and thus do not accurately reflect the reading frame of the hybrid.

Two possible approaches to minimize the generation of false positives by internal RBSs can be imagined. By careful analysis of the parental sequences, potential RBSs could be identified and eliminated by side-directed or cassette mutagenesis. However, either approach is tedious and prone to failure if potential binding sites are missed. Alternatively, a dual selection system (Figure 1C) can be applied, fusing the gene of interest sandwich-like between two functionally distinct but dependent selection markers. In such a three-domain construct, the head section can, for example, be represented by a signal sequence while the tail is formed by a β-lactamase, only functional upon export into the periplasm. The protein in the center, encoded by the gene of interest, functions solely as an extend linker between the two markers and is responsible for maintaining the correct reading frame between the two flanking sequences. No functional constraints other than solubility apply to this linker. An N-terminally truncated version of that fusion construct (Figure 1D) as generated by internal RBSs can no longer produce a false positive owing to the absence of an address tag that directs protein transport to the correct cellular location. β-Lactamase in turn cannot confer ampicillin resistance unless it is exported to the periplasm (Kadonaga et al., 1984). Moreover, the translation of a construct with an out-of-frame gene insert will encounter a stop codon in both reading frames shortly into the bla sequence, preventing the expression of the antibiotic-resistance portion of the fusion protein and therefore growth on selection plates. The pSALect vector was designed to implement experimentally the concept of dual selection.

pSALect construction

The pSALect vector (Figure 2) was constructed into the multiple cloning site of pBC-SK+. Maintaining the original lac-promoter site, the α-complementation fragment of the β-galatosidase in pBC-SK+ was substituted with the tat-signal sequence of torA. The Tat-signal peptide directs transport of the fusion protein into the periplasmic space. This particular sequence was chosen owing to its ability to accommodate the export of a polypeptide strand independent of its folding status (Santini et al., 1998). Alternative signal sequences such as pelB can only export polypeptides that are in at least a partly unfolded state (Schatz and Dobberstein, 1996). This dependence on the folding status would add undesirable selection pressure: spontaneously folding hybrid proteins may trap the entire fusion construct in the cytoplasm. In turn, cytoplasmic hybrid proteins when exported in a partly folded or unfolded state would face the challenge to fold properly in the periplasmic environment. Concerns regarding the lower transport efficiency of Tat in comparison with the alternative export signal sequences have recently been address by Georgiou and co-workers (DeLisa et al., 2002).

Past the unique NdeISpeI restriction sites used for cloning of the gene of interest, the β-lactamase gene bla minus its natural signal sequence (amino acids 1–23) was introduced, providing a suitable periplasmic selection marker. In the process of constructing pSALect, portions of the downstream region including the fl-origin were deleted to minimize the vector size.

Validation of pSALect

Initially, the performance of pSALect was tested on the four rational constructs, pSALect(purN), pSALect(hGART), pSALect(purN+1) and pSALect(hGART+2) (Figure 3). The pSALect(purN) and pSALect(hGART) encode for the wild-type enzymes and served as positive controls. The pSALect(purN+1) and pSALect(hGART+2) constructs contain a single and double nucleotide insertion on the C-termini of GART, resulting in a reading-frame shift which causes translational termination early on in the β-lactamase gene. As expected, when inoculated on agar plates containing Cm as a selection marker, all four vectors conferred growth to their hosts. However, restreaking of the same colonies on plates containing Amp resulted in selective growth of only the two wild-type constructs. Extended incubation at room temperature, and also 37°C, did not result in any observed colony formation of the two out-of-frame vectors.

In a secondary validation test, equal amounts of pSALect(purN) and pSALect(purN+1) were mixed and transformed into E.coli. Colonies that grew on agar plates with Cm were harvested and small aliquots thereof were replated on Amp-containing and, for control purposes, on Cm-containing plates. As expected, the absolute number of colonies on the Amp plates was ~50% lower in comparison with the controls. Five individual colonies from both selection plates were picked and further analyzed by DNA sequencing. While three of the five plasmids from the control plate contained pSALect(purN+1), all five sequences isolated from the Amp-containing plate originated from pSALect(purN).

The data from both validation experiments confirm the
functionality of the dual selection system. Sequences that upon translation do not generate the entire three-domain complex will fall in the selection stage and mixtures of in-frame and out-of-frame constructs can be separated, exclusively allowing growth for the former.

In-frame selection of combinatorial libraries

The pSALect vector was applied to the selection of in-frame constructs of an incremental truncation library, containing hybrid genes of GART from E.coli and human. Plasmid DNA containing the hybrid gene library was obtained as described earlier (Lutz et al., 2001a). Since gene constructs of approximately parental size are believed to be the most promising candidates to yield functional enzymes, plasmids were digested at restriction sites flanking the hybrid gene and a size selection was performed by running the reaction mixture on an agarose gel and excising fragments of 609±50 basepairs. DNA sequence analysis of the resulting sub-library confirmed a size distribution of 605±20 basepairs, resulting in a theoretical library size of 2.4×10⁴ constructs.

Following reamplification, the PCR product was ligated into pSALect and transformed into DH5α-E. When inoculated on LB-agar plates containing Cm, a library of ~6.4×10⁴ colonies was obtained, providing 2.6-fold coverage of the theoretical library size. After recovery, ~1×10⁶ cells were replated on LB-agar containing Amp as the selection marker. Colony counts on control plates indicated that approximately two-thirds of the colonies, growing in the presence of Cm also survive on the Amp-containing plates. Although this experimentally found percentage is higher than the expected one-third, the numbers are probably within the experimental error. For detailed analysis of the in-frame versus out-of-frame distribution of these hybrid genes, the plasmid DNA from 10 members of either the Cm-selected (naive) or the Amp-selected plate (selected) was isolated and their sequence determined by DNA sequencing (Figure 4).

In the naive library, four out of 10 sequences have the correct reading frame across the entire hybrid gene, in line with the predicted 33% as expected for a random distribution. Upon selection in the presence of ampicillin, this distribution shifts to 100%: all 10 sequences show the correct reading frame. As important as the analysis of the reading frame of the selected members is the distribution or diversity of the selected hybrid genes in sequence space. Ideally, pSALect should not apply any selection pressure other than for the correct reading frame of the gene of interest. In other words, the selection should not remove or favor hybrid enzymes with crossovers in particular regions in the gene. As outlined in Figure 4, the plot of sequence data indicates that neither the naive nor the selected library shows biases towards particular regions in the sampled sequence space.

The absence of such biases is particularly important for incremental truncation libraries that will subsequently be used in conjunction with DNA shuffling to create SCRATCHY libraries. In previous experiments on GART, the selection of size- and in-frame constructs by functional selection clearly demonstrated that crossover predispositions towards certain regions will be carried through the entire protocol, biasing the distribution of the final SCRATCHY library (Lutz et al., 2001b). Preventing such regional biases by the use of pSALect, the more balanced crossover distribution of the incremental truncation library will certainly be beneficial to the number and allocation of fusion points in the final SCRATCHY library.

More specifically, computational models for ideal SCRATCHY libraries predict a 1.5-fold increase in the average number of crossovers over regular DNA shuffling (Lutz et al., 2001b). The pSALect selection system now provides the means to test these calculations experimentally.

One aspect that has only been mentioned briefly but is an unavoidable side-effect in the presented selection scheme is protein folding. While the primary objective of pSALect is the isolation of hybrid genes with proper reading frame, the experiment automatically also selects for foldability and solubility of the fusion protein. In order to be functionally selected, the fusion constructs must fold into the proper structure and stay soluble (avoid aggregation) over the time-frame of the experiment. Although protein folding and aggregation of an overexpressed protein can be controlled by variations in growth temperature and induction level, one should keep in mind that the absence of certain constructs post-selection may hint at regions of significance to protein folding.

Future perspectives

The dual-selection scheme, presented here with the combination of the Tat-signal sequence and a β-lactamase, is expandable towards various alternative selection pairs. Particularly interesting in that respect could be the application of protein fragment complementation to achieve dual selection. A variety of split proteins (e.g. green fluorescence protein, dihydrofolate reductase, β-lactamase, GART), used for two-hybrid systems, have been reported (Michnick et al., 2000). These same systems could be used for in-frame selection of hybrid genes although they may be limited to hybrids with proximal N- and C-termini. In addition, the substitution of the tail portion of pSALect with alternative periplasmic selection or screening markers can be envisioned.

In summary, the pSALect vector system provides a solid framework for the rapid selection of genes with a desired
reading frame independent of their function. The system was developed to overcome limitations of previous procedures by requiring not only a C-terminally fused reporter but by sandwiching the gene of interest between N- and C-terminal selection markers. In this role, it provides a compelling alternative for sequences problematic to selection by the traditional fusion methods. Addressing the universal nature of the pSALect vector, we have so far successfully applied this selection system to eight additional proteins from both bacterial and mammalian sources, identifying gene hybrids that maintain the desired reading frame.

References


Received May 22, 2002; revised August 25, 2002; accepted September 25, 2002