

Structural requirements for the biosynthesis of backbone cyclic peptide libraries

Charles P. Scott ^a, Ernesto Abel-Santos ^a, A. Daniel Jones ^{a,b}, Stephen J. Benkovic ^{a,*}

^aDepartment of Chemistry, 414 Wartik Laboratory, The Pennsylvania State University, University Park, PA 16802, USA

^bPenn State Intercollegiate Mass Spectrometry Center, 152 Davey Laboratory, The Pennsylvania State University, University Park, PA 16802, USA

Received 24 April 2001; revisions requested 18 May 2001; revisions received 12 June 2001; accepted 15 June 2001

First published online 29 June 2001

Abstract

Background: Combinatorial methods for the production of molecular libraries are an important source of ligand diversity for chemical biology. Synthetic methods focus on the production of small molecules that must traverse the cell membrane to elicit a response. Genetic methods enable intracellular ligand production, but products must typically be large molecules in order to withstand cellular catabolism. Here we describe an intein-based approach to biosynthesis of backbone cyclic peptide libraries that combines the strengths of synthetic and genetic methods.

Results: Through site-directed mutagenesis we show that the DnaE intein from *Synechocystis* sp. PCC6803 is very promiscuous with respect to peptide substrate composition, and can generate cyclic products ranging from four to nine amino acids. Libraries with five variable amino acids and either one or four fixed residues

were prepared, yielding between 10^7 and 10^8 transformants. The majority of randomly selected clones from each library gave cyclic products.

Conclusions: We have developed a versatile method for producing intracellular libraries of small, stable cyclic peptides. Genetic encoding enables facile manipulation of vast numbers of compounds, while low molecular weight ensures ready pharmacophore identification. The demonstrated flexibility of the method towards both peptide length and composition makes it a valuable addition to existing methods for generating ligand diversity. © 2001 Elsevier Science Ltd. All rights reserved.

Keywords: Combinatorial chemistry; Cyclic peptide; Intein; SICLOPPS; Small molecule

1. Introduction

One of the major objectives of chemical biology is to exploit chemical diversity for the functional interrogation of biological systems. The long-term goal is to identify molecules that modulate the activity of every gene product of the proteome, and use the resulting chemical probes to provide absolute temporal and spatial control over every physiological process. Arranging a one to one (or one to many) correspondence between a gene product and a molecular effector requires the production and screening of libraries of candidate molecules. The discriminating power

of a molecular library is a product of the number of library members and the chemical and structural diversity of the molecular ensemble. The utility of a molecular library for chemical biology depends upon how readily the library can be interfaced with biological screening and the ease with which phenotype can be attributed to a unique compound.

Co-optimization of the chemical and biological properties of a library can be difficult. Both synthetic and genetic methods have been extensively employed to generate molecular libraries, with each having associated strengths and weaknesses. Synthetic methods readily generate libraries of small, drug-like molecules that are rich in chemical diversity. The challenge in using synthetic libraries lies in delivering chemical diversity to intracellular targets without compromising the information content of the library. Membranes are selectively permeable thus the spectrum of compounds that access an intracellular target may not reflect the designed diversity of a synthetic library. The information content of a synthetic library is maintained either by isolating each library member or pool of library

Abbreviations: Ssp, *Synechocystis* sp. PCC6803; I_C, C-intein; I_N, N-intein; SICLOPPS, split intein circular ligation of peptides and proteins; DnaE, gene encoding the replicative polymerase in *Synechocystis* sp. PCC6803; PAGE, polyacrylamide gel electrophoresis; MALDI, matrix assisted laser desorption ionization

* Corresponding author.

E-mail address: sjb1@psu.edu (S.J. Benkovic).

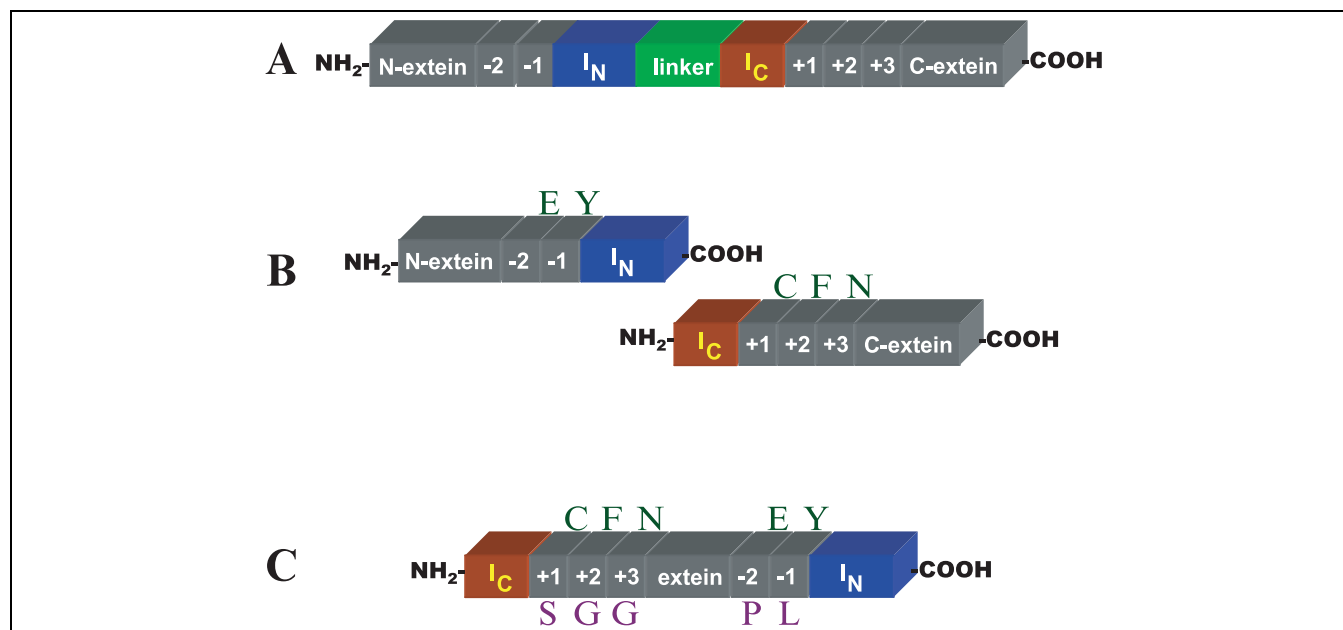


Fig. 1. Intein classes and nomenclature. (A) A canonical (or *cis* acting) intein precursor consists of the amino-terminus (or N-extein) of a gene product, the amino-terminus of an intein (N-intein or I_N), a linker domain (often with homing endonuclease activity), the carboxy-terminus of an intein (C-intein or I_C) and the carboxy-terminus of a gene product (C-extein). The amino acid of the N-extein immediately N-terminal to I_N is referred to as I_N-1, which is preceded by I_N-2, etc. The amino acid of the C-extein immediately carboxy-terminal to I_C is referred to as I_C+1, which is followed by I_C+2 and I_C+3, etc. (B) In a split (or *trans* acting) intein, the I_N and I_C elements of the intein are on separate polypeptides. The one letter code above the graphic represents the amino acids found at the indicated extein positions in the *Ssp* DnaE gene product. (C) In a SICLOPPS construct, the position of the I_N and I_C components is permuted with respect to a canonical intein. In the SICLOPPS configuration the 'extein' is between I_C and I_N. Again, the one letter code above the graphic represents the amino acids found at the indicated extein positions in the *Ssp* DnaE gene product, while the letters beneath the graphic represent the amino acids that are found at the corresponding positions in the SICLOPPS fusion protein precursor to pseudostellarin F.

members in a separate reactor (such as a well in a micro-titer plate) [1,2] or by working with the library at high enough dilution to ensure that an observed phenotype is attributable to a single compound [3–5]. Such spatial addressing strategies typically allow for the facile manipulation of only modest size libraries (10³–10⁵ members) in cell based assays [1–5]. Despite these limitations, synthetic compound libraries have proven useful for the dissection of biological pathways, and have provided a means for simultaneous identification of novel therapeutic targets and small molecules that modulate target function (for example, see [6,7]).

Genetic encoding enables the production of vast biosynthetic libraries that can be readily interfaced with biological selection and screening methodologies, allowing facile manipulation of large numbers (10⁶–10¹⁰) of library members. Functional groups, while limited at the monomer level (amino and nucleic acids), include charged, aromatic, polar, and hydrophobic residues, and colocalization with the intracellular target or pathway of interest prevents loss of chemical diversity. In order to elicit reliable phenotypes, biosynthetic libraries must be significantly stabilized against cellular degradation. To improve stability, variable segments either are large molecules that can fold independently [8–10] or are embedded within or fused to larger biomolecules [11–13]. In either case, a large number of

positions are simultaneously varied (>15), thus library sizes that can be achieved by standard molecular biology methods fall many orders of magnitude short of theoretical library sizes. This complicates identification of the structural determinants of activity.

In an effort to combine some of the useful properties of synthetic and genetic libraries, we have sought to develop methods for the production of genetically encoded libraries of small molecules. While both peptides and oligonucleotides can be genetically encoded, peptide libraries incorporate a much wider array of functional groups and encode five times more information per position. Cyclization is an effective method to stabilize peptides against cellular catabolism, and has the added benefit of restricting conformational freedom, thereby increasing the affinity of cyclic peptides for targeted receptors [14]. Both side chain and backbone cyclization confer significant resistance to proteolysis [15], but side chain cyclization (such as disulfide bond formation) can be compromised by cellular environment [16]. In contrast, backbone cyclic peptides should be stable to a wide range of physiological conditions. Cyclic peptides are an important class of therapeutics with pharmacological functions ranging from immunosuppressants [17] to antineoplastic [18], antibacterial [19], and antiviral agents [20]. The stability and chemical diversity of backbone cyclic peptides coupled with their

broad therapeutic potential make them ideal candidates for the generation of intracellular libraries of small molecules.

In previous work we described a method for intracellular catalysis of peptide backbone cyclization called *split intein circular ligation of peptides and proteins* (SICLOPPS) [21]. In SICLOPPS, an intein is permuted such that its carboxy-terminus (C-intein or I_C) precedes its amino-terminus (N-intein or I_N ; see Fig. 1). In the permuted configuration an amino acid sequence interposed between I_C and I_N is cyclized by the intrinsic protein ligase activity of the intein (Fig. 2). We [21] and others [22] have demonstrated the production of cyclic proteins using SICLOPPS. Moreover, we were able to produce and screen for the eight amino acid cyclic peptide tyrosinase inhibitor pseudostellarin F in bacteria using the same method, and isolate milligram quantities of the cyclic product from the fermentation medium.

Since the SICLOPPS construct is genetically encoded, the tools of molecular biology can be employed to elaborate vast intracellular libraries of cyclic peptides. To be a useful source of ligands, the resulting cyclic peptide libraries must be chemically diverse. The SICLOPPS genetic construct was designed with restriction sites within the intein components to avoid any cloning constraints on

the nucleotide sequence interposed between the I_C and I_N genes [21]. However, the efficiency of intein-mediated protein ligation is known to depend on the identity of the substrate (or extein) amino acid residues immediately adjacent to I_C and I_N (see Fig. 1) [23]. The extent to which similar constraints apply to circular ligation is unknown. In this study we investigate the influence of extein sequence parameters on peptide backbone cyclization by a SICLOPPS construct utilizing the *Synechocystis* sp. PCC6803 (Ssp) DnaE *trans* intein. The results from these structure–activity studies informed the design of a biosynthetic intracellular cyclic peptide library in excess of 10^8 members.

2. Results

2.1. Affinity purification of cyclic products

Study of the amino acid sequence context dependence of intein mediated cyclization requires robust methods to identify cyclic products. While products derived from the intein (N-intein, C-intein, lariat) can be readily visualized by polyacrylamide gel electrophoresis (PAGE) or mass spectroscopy [21], evidence for the production of intein

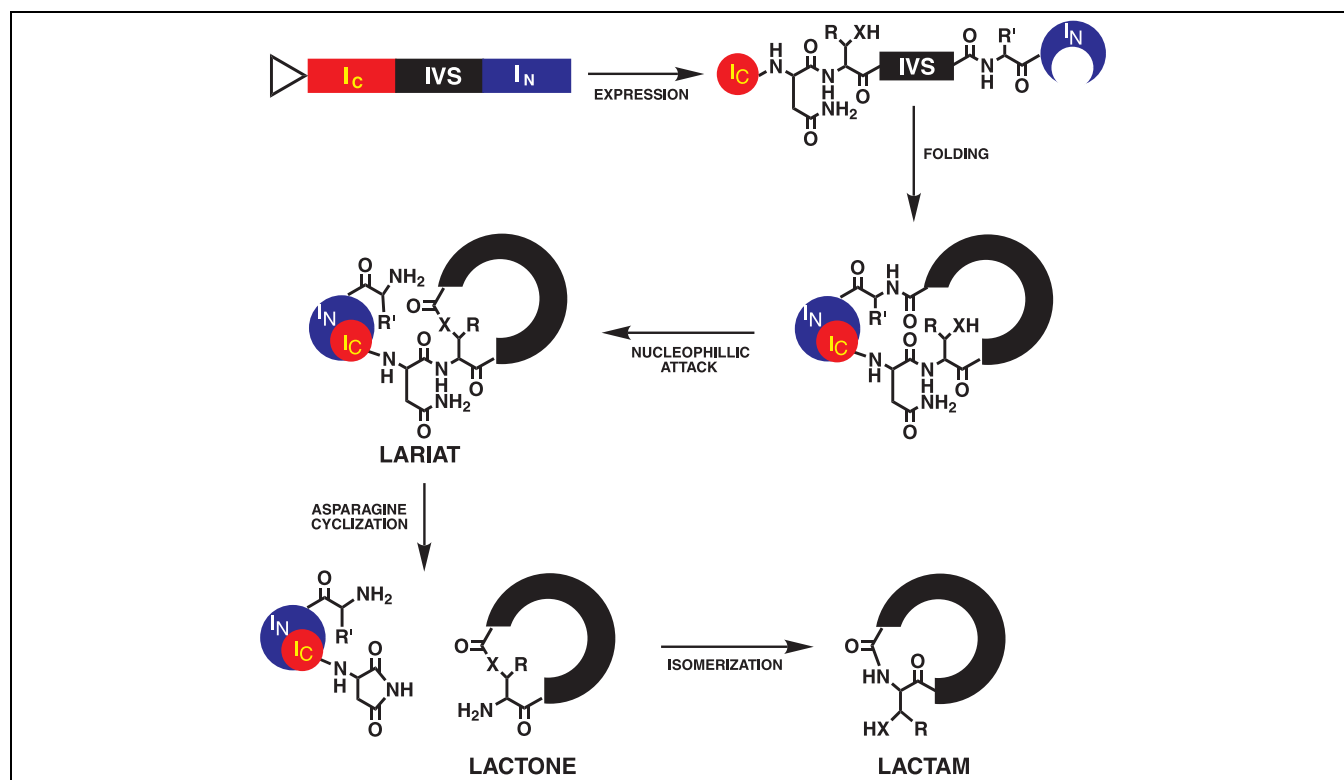


Fig. 2. The SICLOPPS method. A chimeric gene is created that interposes an intervening sequence (IVS) between genes for the C-intein (I_C) and N-intein (I_N) components of an intein. Expression in a suitable host yields a fusion protein (where R is hydrogen or methyl, R' is methyl, hydroxymethyl or thiomethyl and X is oxygen or sulfur) that folds to generate an active intein. The intein catalyzes the attack of a side chain nucleophile from the carboxy-terminus of the IVS upon the carbonyl carbon of the amino acid at the amino-terminus of the IVS to generate a lactone lariat intermediate. The cyclic product is liberated as a lactone by cyclization of the asparagine residue at the carboxy-terminus of I_C . Lactone to lactam isomerization occurs spontaneously.

derived products from a SICLOPPS construct does not constitute sufficient evidence for peptide or protein cyclization. Pseudostellarin F was a convenient model system for SICLOPPS because methods for its identification and purification had been established [24]. As the sequence context flanking the intein is changed, however, the biophysical properties of a cyclic product (particularly a cyclic peptide product) necessary for its identification and purification from the cell lysate are altered. In the present study, a more general procedure to purify peptide products was therefore pursued. We took advantage of a published method [22] (summarized in Fig. 3) to perform SICLOPPS *in vitro* and isolate sufficient quantities of peptide products to enable characterization by mass spectral analysis. In the context of this article, the terms ‘cyclic’, ‘cyclized’ and ‘cyclization’ are only used when molecular ions consistent with the mass of the predicted cyclic product are observed. The terms ‘process’, ‘processed’ and ‘processing’ are used to indicate PAGE or mass spectral evidence for the formation of other products that are diagnostic of intein activity. Optimal detection of the products of the *in vitro* SICLOPPS reaction requires efficient

induction of the fusion protein precursor, but incomplete processing *in vivo* (see Fig. 4, lane 1), because peptide products result from affinity immobilized protein starting materials and/or splicing intermediates that can process *in vitro*. If precursors are incompatible with *in vitro* processing for any reason, no peptide products will be observed. Elution of the chitin affinity column with a high salt buffer affords peptide products of reasonable purity (Fig. 5). Mass spectral analysis reveals that C-intein and lariat intermediates also leach from the column under high salt elution conditions (Fig. 6B). Affinity tagged starting materials and products of the splicing reaction can be harvested from the column by elution with a low salt buffer [21].

2.2. Analysis of mass spectral data

Affinity column eluates were analyzed by matrix assisted laser desorption ionization mass spectrometry (MALDI). To positively identify a peptide product, we required that at least two ionization states be populated (for example $M+H^+$, $M+Na^+$; see Fig. 6A and Tables 1, 2, 5 and 6),

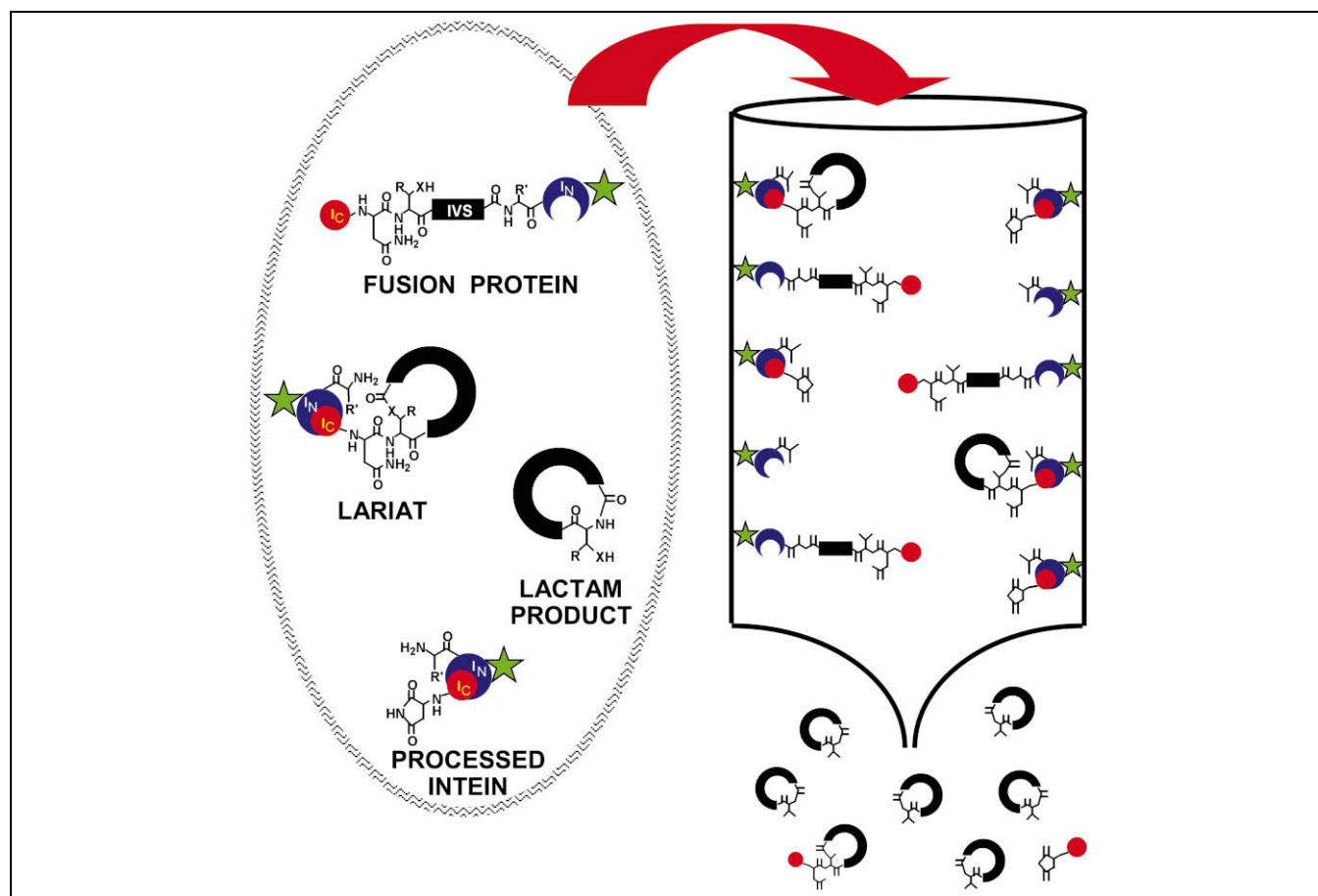


Fig. 3. Affinity purification of peptide products. SICLOPPS products with affinity tags fused either amino-terminal to I_C or carboxy-terminal to I_N (as shown, in green) can be purified away from untagged proteins and concentrated on an affinity column following expression and cell lysis. Peptide products resulting from *in vitro* processing of SICLOPPS fusion protein starting materials and lariat intermediates can be purified away from affinity tagged proteins by a second column wash.

that the signal of each be at least 1.5-fold above background and that the peaks be unique to the sample (not observed in eluates from other clones or attributable to the ionization matrix). Ions corresponding to the mass of the C-intein succinimide and lactone lariat were observed in all samples. These ions are a sensitive, qualitative indicator of intein chemistry, although they do not distinguish between *in vivo* and *in vitro* processing. The difference in mass between molecular ions for the lariat and C-intein peaks corresponds to the mass of the peptide product and thus provides a useful internal standard (see example in Fig. 6B).

2.3. Varying amino acids at position I_C+1

The extein residue immediately adjacent to the C-intein (I_C+1) serves as the nucleophile for the transesterification reaction that generates the branched and lariat intermediates in protein splicing and circular ligation, respectively (see Fig. 2). Only three residues occupy the I_C+1 position in active inteins: cysteine, serine and threonine [25]. Cysteine serves as the transesterification nucleophile for the wild-type Ssp DnaE intein [26]. To evaluate the dependence of circular ligation on the identity of the transesterification nucleophile, the serine residue of pseudostellarin F was mutated either to cysteine (+1C) or threonine (+1T) in the pARCBD-p expression vector [21]. The resulting constructs were overexpressed, and cyclic peptide production was evaluated following *in vitro* incubation and elution from a chitin affinity column [21,22].

All three constructs (pseudostellarin F, +1C and +1T) expressed well and showed evidence for post-translational

processing by PAGE analysis. The extent of post-translational processing for the +1T construct was similar to pseudostellarin F (data not shown), but the +1C construct appeared to process more extensively *in vivo* (compare lanes 1 and 3 or 2 and 4 in Fig. 4). Note that whole cell fractions in Fig. 4 (odd numbered lanes) exclusively reflect *in vivo* processing, while resin bound material (even lanes) reflects the sum of *in vivo* and *in vitro* processing. Further evidence in support of correct post-translational processing of the fusion protein precursors was generated by mass spectral analysis. Molecular ions consistent with the mass of the N-intein were observed from all three constructs in low salt eluates from chitin affinity columns [21] (data not shown), and molecular ions consistent with the mass of the C-intein succinimide and the expected mass of the hydrolyzed lactone lariat were observed in high salt eluates in all cases (see Table 1). Wild type pseudostellarin F could be readily detected in high salt eluates as well, but no molecular ions for either cyclic or linear cysteine or threonine containing derivatives were observed (Table 1). Attempts to isolate cyclic or linear +1C or +1T peptides by *n*-butanol extraction of the cell lysate or growth medium (by analogy with pseudostellarin F [21]) also failed to yield product.

2.4. Varying amino acids at position I_N-1

The identity of the amino acid adjacent to the N-intein (I_N-1) has also been shown to influence the efficiency of protein splicing for several inteins [23]. Unlike the extein amino acid residue at position I_C+1 , the side chain of the amino acid residue in the I_N-1 position is not directly

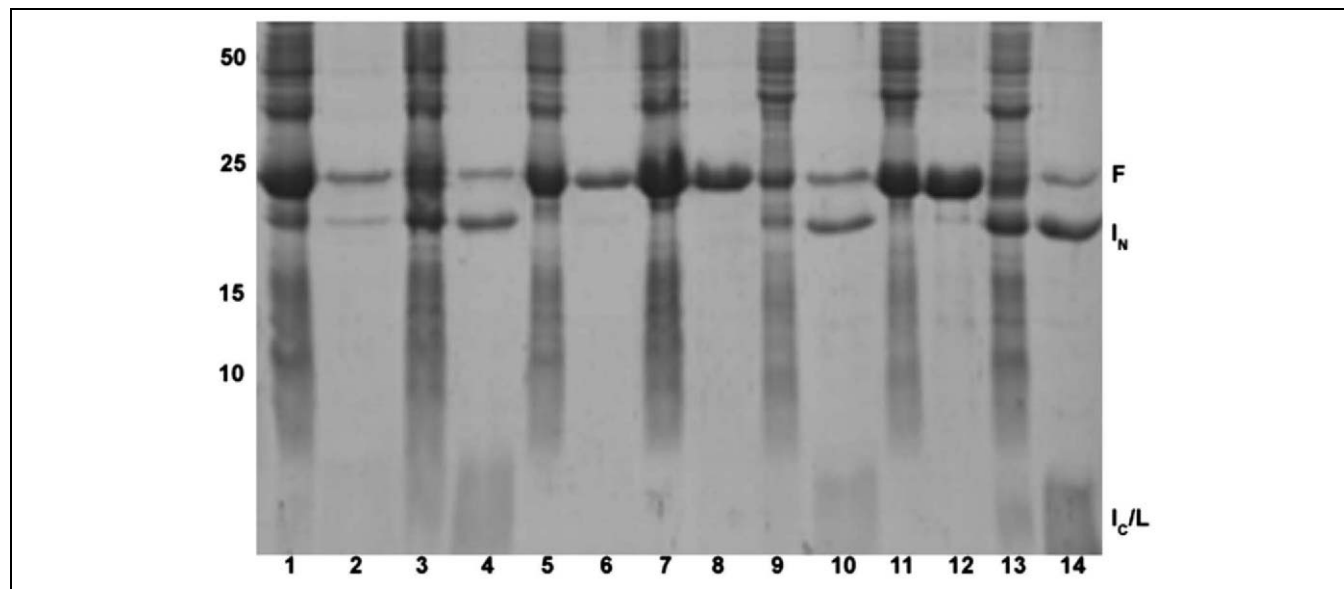


Fig. 4. PAGE analysis of SICLOPPS product expression and post-translational processing. Induced cells (odd numbered lanes) and chitin resin bound proteins (even numbered lanes) were analyzed following expression of SICLOPPS constructs encoding pseudostellarin F (lanes 1 and 2), +1C (lanes 3 and 4), -1E (lanes 5 and 6), -1P (lanes 7 and 8), S+5.3 (lanes 9 and 10), S+5.6 (lanes 11 and 12) and $\Delta 4$ (lanes 13 and 14). Molecular weight markers are shown at the far left (50, 25, 15 and 10 kDa, as indicated). F stands for fusion protein, I_N stands for N-intein and I_C/L stand for C-intein and lariat intermediates.

Table 1
Mass spectral characterization of pseudostellarin F derivatives

Clone	Sequence	Mass cyclic (calc)	Mass linear (calc)	<i>m/z</i> cyclic (obs)	<i>m/z</i> linear (obs)	<i>m/z</i> lariat (obs)	<i>m/z</i> C-intein (obs)	<i>m/z</i> lariat minus <i>m/z</i> C-intein
+1C	CPPYLPPL	800.4	818.4	–	–	4771.5	3953.9	817.6
+1T	TGGYLPPL	798.4	816.4	–	–	4770.3	3953.7	816.6
pseudoF	SGGYLPPL	784.4	802.4	H ⁺ : 785.5 (1000); Na ⁺ : 807.5 (5400)	Na ⁺ : 825.5 (4200)	4756.6	3953.9	802.7
–1S	SGGYLPPL	758.4	776.4	H ⁺ : 759.4 (800); Na ⁺ : 781.4 (5900)	Na ⁺ : 799.4 (900)	4732.3	3955.4	776.9
–1A	SGGYLPPL	742.4	760.4	H ⁺ : 743.4 (2000); Na ⁺ : 765.4 (9000)	Na ⁺ : 783.4 (3000)	4714.7	3954.0	760.7
–1G	SGGYLPPL	728.4	746.4	H ⁺ : 729.7 (3200); Na ⁺ : 751.7 (1500)	–	4703.5	3955.5	748.0
–1K	SGGYLPPL	799.4	817.4	H ⁺ : 800.5 (2700); Na ⁺ : 822.5 (2700)	H ⁺ : 818.5 (1600); Na ⁺ : 840.5 (2800)	4772.8	3954.7	818.1
–1Y	SGGYLPPL	834.4	852.4	H ⁺ : 835.5 (1000); Na ⁺ : 857.5 (3500)	Na ⁺ : 875.4 (3500)	4799.3	3947.8	851.5
–1I	SGGYLPPL	784.4	802.4	H ⁺ : 785.5 (1800); Na ⁺ : 807.4 (4100)	–	4757.6	3954.7	802.9
–1E	SGGYLPPL	800.4	818.4	–	–	4766.2	3948.3	817.9
–1N	SGGYLPPL	785.4	803.4	–	–	4757.6	3954.1	803.5
–1P	SGGYLPPL	768.4	786.4	–	–	4741.9	3955.6	786.3

Numbers in parentheses are background subtracted peak intensities in counts.

involved in the catalytic mechanism, and may tolerate extensive modification. The terminal leucine of the linear pseudostellarin F precursor was therefore replaced with polar, non-polar, large, small, aromatic, charged and beta-branched amino acids to evaluate the influence of the I_N–1 position on circular ligation.

All of the I_N–1 mutant constructs expressed well, and with the exception of the leucine to proline mutation (–1P), showed some evidence for post-translational processing by PAGE. –1P showed excellent expression of the fusion protein precursor but no bands indicating post-translational production of N- or C-inteins (Fig. 4, lanes 7 and 8). Six of the nine I_N–1 mutants gave cyclic products in vitro, including constructs encoding polar (–1S),

non-polar (–1A), large (–1L or pseudostellarin F), small (–1G), charged (–1K), aromatic (–1Y) and β-branched (–1I) amino acids (Table 1). Three mutant constructs, I_N–1 glutamic acid (–1E), asparagine (–1N) and proline (–1P), failed to give cyclic or linear peptide products in vitro. All three constructs gave molecular ions corresponding to the C-intein succinimide and hydrolyzed lactone lariat (Table 1). The apparent discrepancy between the PAGE and mass spectroscopic results for the –1P construct reflects the enhanced sensitivity of MALDI mass spectroscopy compared to Coomassie staining. The difference in mass between the molecular ions for the C-intein and hydrolyzed lactone lariat was consistent with the linear product in all cases.

Table 2
Mass spectral characterization of pseudostellarin F deletion mutants

Clone	Sequence	Mass cyclic (calc)	Mass linear (calc)	<i>m/z</i> cyclic (obs)	<i>m/z</i> linear (obs)
pseudoF	SGGYLPPL	784.4	802.4	H ⁺ : 785.5 (1000); Na ⁺ : 807.5 (5400)	Na ⁺ : 825.5 (4200)
Δ1	SGGYPL	671.3	689.3	H ⁺ : 672.3 (800); Na ⁺ : 694.3 (7100); H ⁺ : 672.1 (2400) ^a ; Na ⁺ : 694.4 (1200) ^a ; K ⁺ : 710.3 (700) ^a	Na ⁺ : 712.3 (2500); H ⁺ : 690.0 (600) ^a ; Na ⁺ : 712.5 (500) ^a ; K ⁺ : 728.5 (800) ^a
Δ2	SGGYPL	574.3	592.3	Na ⁺ : 597.3 (1700)	Na ⁺ : 615.2 (3300)
Δ3	SGYPL	517.3	535.3	Na ⁺ : 540.3 (1100); Na ⁺ : 540.3 (1300) ^a ; K ⁺ : 556.3 (2200) ^a	–
Δ4	SGPL	354.2	372.2	Na ⁺ : 377.2 (400) ^a ; K ⁺ : 393.2 (400) ^a	Na ⁺ : 395.2 (100) ^a

Numbers in parentheses are background subtracted peak intensities in counts.

^aButanol extract.

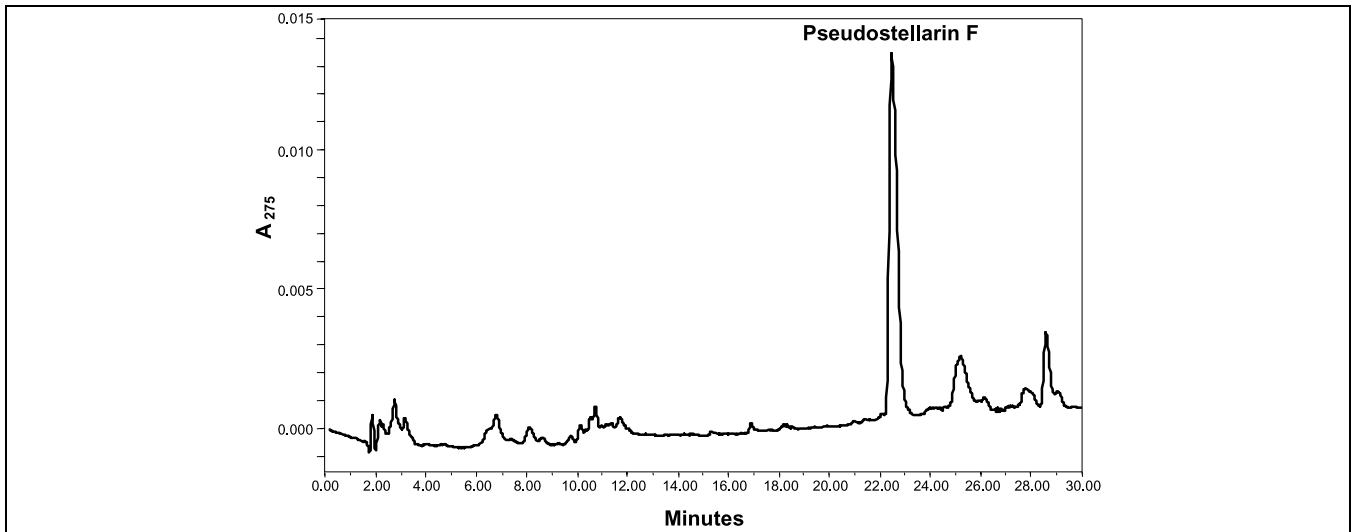


Fig. 5. HPLC characterization of in vitro cyclized pseudostellarin F. Chromatogram of the high salt eluate from a chitin affinity column after overnight incubation of proteins purified following bacterial expression of a SICLOPPS construct encoding pseudostellarin F. Chromatographic conditions and peak identification are as previously described [21].

2.5. Intervening sequence length requirements for peptide cyclization

To determine the minimal product size and investigate the influence of steric constraints on circular ligation, a series of pseudostellarin F derivatives with internal deletions of one to four amino acids ($\Delta 1-4$) were prepared (Table 2). All four constructs expressed well and processed efficiently in vivo. Three of the four deletion constructs

($\Delta 1-3$) gave cyclic products in vitro, but the smallest ($\Delta 4$) processed so well in vivo that insufficient starting material was available for the in vitro production of cyclic product in quantities detectable by mass spectral analysis (lanes 13 and 14 in Fig. 4). Fortunately, the $\Delta 4$ product could be concentrated from the cell lysate and growth medium by *n*-butanol extraction. The concentrated material was sufficiently pure to allow identification of a unique pattern of molecular ions consistent with the molecular

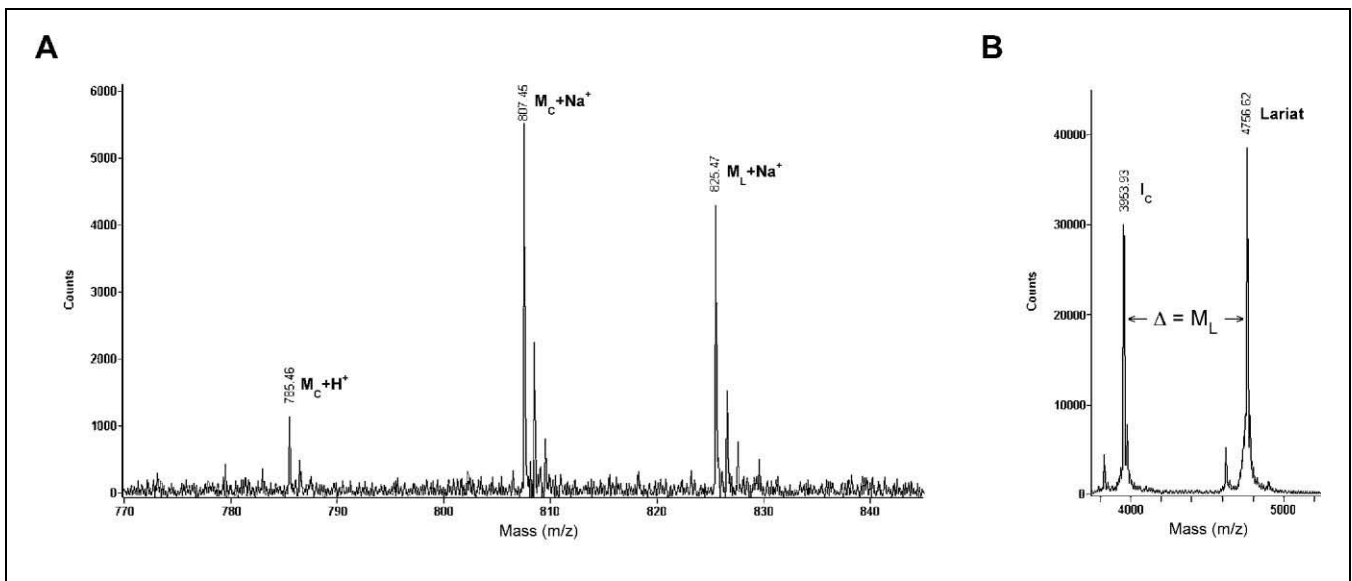


Fig. 6. Mass spectral analysis of in vitro cyclized pseudostellarin F. MALDI mass spectral analysis of the high salt eluate from a chitin affinity column after overnight incubation of proteins purified following bacterial expression of a SICLOPPS construct encoding pseudostellarin F. (A) Reflector mode data collection of peptide product masses. Indicated molecular ions correspond to the mass of the cyclic peptide product (pseudostellarin F) plus protium ($M_C + H^+$) and sodium ($M_C + Na^+$) and the mass of the linear peptide product plus sodium ($M_L + Na^+$). (B) Linear mode data collection of protein product masses. Indicated molecular ions correspond to the mass of the C-intein succinimide plus protium (I_C) and the mass of the hydrolyzed lactone lariat plus protium ('Lariat'). The difference in mass between the C-intein succinimide and the hydrolyzed lactone lariat corresponds to the mass of the linear peptide product ($\Delta = M_L$).

Table 3
Nucleotide usage in SICLOPPS libraries

Position	Library	A	C	G	T
N ₁	S+5	10	7	18	15
	Δ4+5	14	12	13	11
	total	24	19	31	26
N ₂	S+5	11	10	11	18
	Δ4+5	5	13	18	14
	total	16	23	29	32
N _t	S+5	21	17	29	33
	Δ4+5	19	25	31	25
	total	40	42	60	58
S	S+5	—	16	34	—
	Δ4+5	—	29	21	—
	total	—	45	55	—

N₁, N₂ and S refer to position within the N₁N₂S sequence used to encode variable amino acids in the S+5 and Δ4+5 libraries. N_t represents the sum of the N₁ and N₂ data.

weight of Δ4 plus sodium and potassium (Table 2). LC-MS analysis confirmed the production of the Δ4 product.

2.6. Biosynthesis and characterization of a six amino acid cyclic peptide library

A plasmid construct was designed to encode serine at I_C+1 followed by five variable amino acids (SXXXXX) between the C- and N-intein genes. A library of 1.5×10^7 bacterial transformants was prepared, and several colonies were selected at random to evaluate the nucleotide and codon usage in the library, as well as the extent of circular ligation. Induction of the library construct was evaluated by PAGE analysis of induced and uninduced clones. Efficient induction was observed for 13 of 19 colonies selected,

with complete in vivo processing in two cases. Ten library clones displaying efficient induction but incomplete in vivo processing were grown at half-liter scale, induced and purified by chitin affinity chromatography. The nucleotide and amino acid usage in the S+5 library is compiled in Tables 3 and 4, respectively. Eighteen amino acids are represented in the fifty variable positions sequenced (five per clone, 10 clones; see Table 4).

The selected clones displayed post-translational processing ranging from extensive to minimal (compare lanes 9 and 11 or lanes 10 and 12 in Fig. 4) by PAGE analysis. The difference in mass between molecular ions for the C-intein and lariat is consistent with the linear peptide in eight cases and with the cyclic peptide in one case (S+5.8, Table 5). In the remaining clone (S+5.3, Table 5), a plurality of molecular ions was observed beginning at approx-

Table 4
Amino acid usage in SICLOPPS libraries

Amino acid	Number of codons	Frequency S+5	Frequency Δ4+5	% theoretical	% observed
L	3	9	7	9.4	16
R	3	2	6	9.4	8
S	3	1	7	9.4	8
T	2	4	5	6.3	9
A	2	4	4	6.3	8
G	2	3	5	6.3	8
V	2	4	4	6.3	8
P	2	1	2	6.3	3
E	1	6	0	3.1	6
H	1	1	3	3.1	4
M	1	2	2	3.1	4
C	1	3	1	3.1	4
F	1	3	1	3.1	4
W	1	3	1	3.1	4
Y	1	0	2	3.1	2
D	1	1	0	3.1	1
K	1	1	0	3.1	1
N	1	1	0	3.1	1
Q	1	1	0	3.1	1
I	1	0	0	3.1	0
amber	1	0	0	3.1	0
Totals	32	50	50	100	100

Table 5
Mass spectral characterization of selected clones from the S+5 library

Clone	Sequence	Mass cyclic (calc)	Mass linear (calc)	<i>m/z</i> cyclic (obs)	<i>m/z</i> linear (obs)	<i>m/z</i> lariat (obs)	<i>m/z</i> C-intein (obs)	<i>m/z</i> lariat minus <i>m/z</i> C-intein
S+5.1	SFTFLS	682.3	700.3	H ⁺ : 683.0 (3200); Na ⁺ : 704.5 (17 000)	H ⁺ : 701.1 (1800); Na ⁺ : 722.9 (1600)	4650.8	3951.5	699.3
S+5.2	SKMLVA	629.4	647.4	H ⁺ : 630.3 (2200); Na ⁺ : 652.3 (900)	H ⁺ : 648.9 (1300); Na ⁺ : 670.5 (900)	4605.5	3957.9	647.6
S+5.3	SLLNVC	629.3	647.3	–	–	4585.4– 4674.3	3957.3	628.1–717.0
S+5.4	SVRLEV	683.4	701.4	H ⁺ : 684.3 (300); Na ⁺ : 706.2 (200)	H ⁺ : 702.5 (3200); Na ⁺ : 724.3 (400)	4669.5	3967.1	702.4
S+5.5	SLLCPA	584.3	602.3	–	–	4557.5	3956.5	601.0
S+5.6	SMTATE	620.3	638.3	–	–	4593.7	3956.5	637.2
S+5.7	SAEWQG	658.3	676.3	H ⁺ : 659.5 (8600); Na ⁺ : 681.4 (5500)	H ⁺ : 677.4 (1500); Na ⁺ : 699.3 (800)	4630.1	3955.1	675.0
S+5.8	SWTEFC	753.3	771.3	Na ⁺ : 776.5 (1400); H ⁺ : 1507.0 (1100) ^a ; Na ⁺ : 1529.0 (3400) ^a	–	4718.7	3967.4	751.3
S+5.9	SGLELW	685.4	703.4	H ⁺ : 686.5 (200); Na ⁺ : 708.5 (1000)	–	4674.2	3969.8	704.4
S+5.10	SGDRHE	681.3	699.3	H ⁺ : 682.5 (1400); Na ⁺ : 704.5 (700)	H ⁺ : 700.4 (1100)	4668.4	3970	698.4

Numbers in parentheses are background subtracted peak intensities in counts.

^aDisulfide linked dimer.

imately the mass expected for the lactone lariat intermediate and extending an additional 90 Da. No peptide product was observed with this clone. Molecular ions consistent with cyclic peptide products were observed in seven of 10 clones, with ions consistent with linear peptide

products observed in five of 10 cases (Table 5). Three clones failed to give evidence for either linear or cyclic peptide products in vitro. One of these clones (S+5.6) displayed poor processing by PAGE (lanes 11 and 12 in Fig. 4), while the other two (S+5.3 and S+5.5) processed

Table 6
Mass spectral characterization of selected clones from the Δ4+5 library

Clone	Sequence	Mass cyclic (calc)	Mass linear (calc)	<i>m/z</i> cyclic (obs)	<i>m/z</i> linear (obs)	<i>m/z</i> lariat (obs)	<i>m/z</i> C-intein (obs)	<i>m/z</i> lariat minus <i>m/z</i> C-intein
Δ4+5.1	SGHVRHLPL	996.6	1014.6	H ⁺ : 997.6 (4200); Na ⁺ : 1019.4 (1100)	H ⁺ : 1015.6 (2400); Na ⁺ : 1037.6 (500)	4966.2	3951.6	1014.6
Δ4+5.2	SGGRSVLPL	866.5	884.5	H ⁺ : 867.8 (6800); Na ⁺ : 889.7 (3300)	H ⁺ : 885.7 (3200); Na ⁺ : 907.6 (1800)	4828.3	3945.2	883.1
Δ4+5.3	SGRSSTRPL	941.5	959.5	H ⁺ : 942.7 (8400); Na ⁺ : 964.6 (7300)	H ⁺ : 960.7 (1700); Na ⁺ : 983.3 (1100)	4915.6	3955.4	960.2
Δ4+5.4	SGAYVGLPL	857.5	875.5	H ⁺ : 858.9 (800); Na ⁺ : 881.2 (9400)	Na ⁺ : 899.3 (1600)	4824.6	3950.2	874.4
Δ4+5.5	SGARSAMPL	870.4	888.4	H ⁺ : 871.8 (10 600); Na ⁺ : 893.8 (9400)	H ⁺ : 889.9 (7100); Na ⁺ : 911.9 (4600)	4846.4	3956.8	889.6
Δ4+5.6	SGGSCHFPL	885.4	903.4	H ⁺ : 886.3 (1300); Na ⁺ : 908.3 (400); H ⁺ : 1770.3 (2700) ^a ; Na ⁺ : 1792.5 (800) ^a	–	4839.9; 4857.6	3954.4	885.5; 903.2
Δ4+5.7	SGTTTPRPL	910.5	928.5	H ⁺ : 911.8 (10 000); Na ⁺ : 933.8 (7100)	H ⁺ : 930.0 (1900); Na ⁺ : 952.0 (1100)	4889.0	3959.4	929.6
Δ4+5.8	SGMPAGLPL	823.4	841.4	H ⁺ : 825.7 (1600); Na ⁺ : 847.1 (3200)	–	4797.1	3955.0	842.1
Δ4+5.9	SGLSWYGPL	960.5	978.5	H ⁺ : 961.7 (12 600); Na ⁺ : 983.3 (24 300)	H ⁺ : 979.7 (2800); Na ⁺ : 1001.5 (2200)	4929.0	3951.0	978.0
Δ4+5.10	SGVLSLTPL	867.5	885.5	H ⁺ : 868.9 (900); Na ⁺ : 890.5 (1100)	H ⁺ : 886.8 (200); Na ⁺ : 908.4 (300)	4839.6	3954.5	885.1

Numbers in parentheses are background subtracted peak intensities in counts.

^aDisulfide linked dimer.

extensively (see lanes 9 and 10 in Fig. 4 for S+5.3 PAGE data).

2.7. Biosynthesis and characterization of a scaffolded cyclic peptide library

A library was prepared with five variable positions embedded between the glycine and proline residues of the $\Delta 4$ scaffold (SGXXXXXPL). 1.1×10^8 bacterial transformants were obtained following ligation of the digested PCR product into the SICLOPPS construct. Randomly selected colonies were subjected to small-scale induction studies in order to identify clones with appropriate properties for in vitro characterization. From 24 selected colonies, 13 displayed efficient induction but incomplete in vivo processing. Plasmids from 10 selected colonies that expressed SICLOPPS fusion protein in the trial induction study were harvested and sequenced to evaluate nucleotide (Table 3) and amino acid (Table 4) usage. Fourteen amino acids are represented in the 50 variable positions in the selected clones. The selected colonies were grown, induced and subjected to chitin affinity chromatography. All 10 gave cyclic products in vitro (Table 6), with some partitioning to linear products observed in eight of 10 cases.

3. Discussion

We have previously shown that the naturally occurring Ssp DnaE *trans* intein can be used for SICLOPPS to affect the biosynthesis of the eight amino acid backbone cyclic peptide pseudostellarin F in *Escherichia coli* [21]. We expected that the method would be generally applicable for the intracellular biosynthesis of a wide variety of cyclic peptides, or even genetically encoded cyclic peptide libraries. In moving from the production of a single cyclic peptide to a cyclic peptide library, the overriding concern was whether the chemistry of cyclization would be influenced by the composition of the intervening sequence. Any preference displayed by the intein in cyclizing the intervening sequence would translate into a functional bias in resulting libraries. Our objective in this study was to use mutagenesis as a tool to identify the structural determinants of any such bias, evaluate whether these constituted an unacceptable loss in chemical diversity, and, if so, to install invariant residues in the intervening extein sequence to ameliorate or eliminate the problem.

3.1. I_C+1 mutagenesis

We sought first to evaluate whether the serine residue used for cyclization of pseudostellarin F was required for circular ligation, or whether cysteine or threonine could functionally substitute as transesterification nucleophiles. While inclusion of both cysteine and threonine could only increase potential library sizes by a factor of three,

the utility of these residues makes them valuable in any chemical library. The serine to cysteine substitution seemed particularly promising because the transesterification nucleophile for the wild-type Ssp DnaE *trans* intein is cysteine [26], a cysteine nucleophile was used for the cyclization of dihydrofolate reductase [21] and maltose binding protein [22], and a cysteine codon can be readily generated in SICLOPPS genetic constructs with existing restriction sites [21].

Upon substitution of either threonine or cysteine at the I_C+1 position of pseudostellarin F, no linear or cyclic peptide products were observed even after lengthy incubation of in vitro affinity concentrated cyclization precursors. PAGE analysis indicated that the +1T construct was post-translationally processed as efficiently as the pseudostellarin F parent, and that in vivo processing of the +1C construct was even more efficient (compare lanes 1 and 3 or 2 and 4 in Fig. 4). It may be that our difficulty in isolating peptide products following in vitro processing of +1C resulted in part from overprocessing in vivo, which would be expected to deplete the starting materials needed for in vitro peptide product generation.

All of the expected protein intermediates and products for both +1C and +1T were observed in mass spectral analyses (Table 1). Since the difference in mass between the lariat and C-intein molecular ions corresponds to the mass of the expected linear peptide product in both cases, post-translational modification of the intervening peptide sequence could be ruled out as an explanation for our failure to observe peptide products with diagnostic masses. Attempts to isolate cyclic products directly from cell lysate and medium were unsuccessful, as were attempts to alter conditions of the in vitro reaction to promote product formation through the addition of reducing agents or organic solvents. Since PAGE and mass spectral analysis clearly demonstrate that the expected protein products are being generated in vivo (see Fig. 4 and Table 1), peptide products must also be generated. Isolation of the +1C peptide may be complicated by cysteine oxidation (see below), but we have no adequate explanation for our failure to isolate +1T peptide products from the cell lysate or growth medium. It may be that cyclization of +1T and +1C is inefficient, or that significant partitioning to linear products occurs in the expression host. Linear products would be difficult or impossible to isolate from bacterial lysates because the half-life of unconstrained peptides in complex physiological mixtures is short [15]. We envision SICLOPPS libraries as 'inside-out' entities, where preliminary discovery from cyclic peptide libraries takes place in vivo, and biochemical characterization of functional library members is carried out in vitro, so cyclic constructs that are compatible with both in vitro and in vivo synthesis are of the greatest utility. With this in mind, serine appears to be the optimal residue for the I_C+1 position in SICLOPPS library constructs employing the Ssp DnaE intein.

3.2. I_N-1 mutagenesis

Limited studies of the influence of amino acids adjacent to the N-intein on *trans* splicing by the Ssp DnaE intein indicate that mutation at either the I_N-1 position or the I_N-2 position has little effect on the efficiency of protein splicing (C.P. Scott, unpublished). A more comprehensive survey of the influence of the I_N-1 position on circular ligation was nonetheless warranted to guard against systematic bias in cyclic peptide libraries. The results largely corroborate the predictions from the *trans* splicing study and indicate that the Ssp DnaE intein is remarkably tolerant toward a wide variety of amino acids at the I_N-1 position (Table 1). Of the nine mutants studied, we only failed to observe peptide products in three cases. Both $-1E$ and $-1P$ processed poorly if at all (lanes 5–8, Fig. 4), so failure to observe cyclic products with these constructs was not surprising. Glutamate occurred at the I_N-1 position in two of the 10 peptides sequenced from the S+5 library (Table 5: S+5.6, S+5.10), and cyclic peptide was observed with one of these two constructs (S+5.10). It therefore appears that while glutamate at the I_N-1 position is not favorable, it does not preclude *in vitro* cyclization. The difficulty in post-translational processing displayed by the $-1P$ construct may reflect the steric demands imposed by accommodating three consecutive proline residues within an eight amino acid cyclic peptide. By PAGE analysis, the $-1N$ construct appeared to process to a similar extent as the wild-type pseudostellarin F parent (data not shown), so our failure to isolate a peptide product following affinity concentration and *in vitro* incubation cannot be attributed either to overprocessing *in vivo* or underprocessing *in vitro*. Nevertheless, constructs with amide or carboxylate residues at the I_N-1 position may be functional *in vivo*, and would only be expected to occur in the I_N-1 position in 20% of all library peptides, thus the I_N-1 position was fully varied in the preliminary (S+5) library construct.

3.3. Other extein positions

In studies of the extein sequence dependence of protein splicing by the Ssp DnaE intein, we (C.P. Scott, unpublished) and others [22] have determined that three amino acid residues adjacent to the C-intein (I_C+1 , I_C+2 and I_C+3) are important or essential for efficient processing of the two model substrates that have been investigated. However, milligram quantities of pseudostellarin F could be produced by SICLOPPS [21] although this substrate lacks sequence homology in all three positions (see Fig. 1). The influence of the I_C+2 and I_C+3 positions on circular ligation could be evaluated within the context of the deletion and library data without having to resort to exhaustive mutagenesis studies analogous to that performed on the I_N-1 position.

Phenylalanine occurs at the I_C+2 position in the native

DnaE substrate of the Ssp DnaE intein [26] (Fig. 1) and mutation of this residue to lysine (C.P. Scott, unpublished) or threonine [22] eliminates protein splicing. Glycine occurs at I_C+2 in pseudostellarin F and all pseudostellarin F analogues (Tables 1 and 2). Cyclic products were isolated from the S+5 library with valine (S+5.4), alanine (S+5.7) and tryptophan (S+5.8) at the I_C+2 position in addition to the wild-type phenylalanine (S+5.1) and previously observed glycine (S+5.9 and 10) residues. This data clearly indicates that large, medium and small hydrophobic residues at the I_C+2 position are compatible with circular ligation. The observation of a cyclic product with lysine at the I_C+2 position (S+5.2) directly contrasts our *trans* splicing data. Although the apparent dependence of the Ssp DnaE intein upon phenylalanine at the I_C+2 position for efficient *trans* splicing is based on only two amino acid substitutions, our previous [21] and current data with both peptide and protein (DHFR, in [21]) substrates suggests that circular ligation may be more tolerant to substitution at I_C+2 than *trans* splicing. Even in S+5 library constructs where circular ligation is not observed (S+5.3, 5 and 6, where I_C+2 is either leucine or methionine), the amino acid at the I_C+2 position is unlikely to be the determining factor (see $-1E$ discussion, above, and discussion of cysteine containing peptides, below).

Mutation of the asparagine residue that occurs at the I_C+3 position of DnaE to lysine (C.P. Scott, unpublished) or threonine [22] in model substrates almost completely eliminates protein splicing whereas glycine at I_C+3 is well tolerated for the circular ligation of pseudostellarin F and its analogues (Table 1). All 12 of the amino acids that are found in the I_C+3 position in SICLOPPS constructs from the deletion and library studies are compatible with circular ligation (deletion study: G, Y and P; S+5 library: T, M, R, E, L and D; $\Delta 4+5$ library: G, H, R, A, T, M, L and V). In contrast to the results from Evans et al. [22] four library constructs had threonine at the I_C+3 position, and three of the four gave cyclic products. Neither lysine nor the wild-type DnaE I_C+3 asparagine residue were among the 12 amino acids observed in the I_C+3 position in the current study, but arginine occurred in the I_C+3 position in two constructs, and both gave cyclic products. The sum of the data indicates that amino acid residues in the I_C+3 position have little effect on circular ligation.

3.4. Deletion studies

We undertook deletion studies of pseudostellarin F to determine the lower size limit for cyclization and to evaluate the influence of steric constraints imposed by the intervening sequence on cyclic product formation. Based on the estimated copy number of the DnaE ligation product in Ssp [26], the dissociation constant for the heterodimeric Ssp DnaE *trans* intein is likely to be in the nanomolar range. Our ability to copurify each intein

component when the other is affinity tagged [21] is consistent with this estimate. Since cyclization is driven by the association of the intein components, we expected that even somewhat strained cyclic products might be accessible by SICLOPPS. The deletion series represents an idealized case considering that at least one proline and one glycine were present in all of the constructs that were studied. Nevertheless, SICLOPPS proved capable for biosynthesis of cyclic products as small as four amino acids (Table 2), and may be capable of producing even smaller products. The ability of SICLOPPS to generate cyclic products ranging from four to hundreds [21,22] of residues shows the flexibility and wide applicability of this system.

3.5. Design and expression of random cyclic peptide libraries

SICLOPPS libraries were designed by introducing codons for five variable amino acids between the C- and N-intein genes. The variable segment was encoded in the form NNS where N represents any of the four DNA bases (A, C, G or T) and S represents C or G. The NNS sequence generates 32 codons ($4 \times 4 \times 2$) and encodes all 20 amino acids while eliminating both ochre (UAA) and opal (UGA) stop codons from the library. Amber stop codons (UAG) are represented in the library at a frequency of 1/32, but peptide products can be generated from constructs with amber stop codons provided the library is manipulated in a host that expresses suppressor tRNAs. A library of five variable amino acids was chosen because the theoretical library size at the amino acid level (20^5 or 3.2 million members) is comparable in size to the largest reported solution phase libraries of small molecules, while the theoretical library size at the DNA level (32^5 or 34 million members based on the NNS codon usage in the library construct) is well within the number of transformants that can be readily achieved by standard molecular biology techniques.

High level expression of cyclization precursors is essential to ensure detection of cyclic products following in vitro cyclization because the yield of this process is poor. Even with sensitive mass spectral detection, in vitro cyclization is often difficult to detect. Unfortunately, high level expression of SICLOPPS libraries can result in significant toxicity to the expression host (E. Abel-Santos, unpublished). The library was cloned into a pET vector under control of the T7 promoter to ensure efficient transcription with a lysogen encoded T7 RNA polymerase [27]. The failure of 30 (S+5) to 50% ($\Delta 4+5$) of selected clones to express the SICLOPPS construct in induction studies may reflect loss of the DE3 lysogen to avoid toxicity associated with high level expression of certain library constructs. When SICLOPPS libraries are expressed at moderate levels more consistent with library screening, toxicity is significantly reduced.

3.6. S+5 library

As described, the results from mutagenesis studies indicate that serine at the I_C+1 position is the optimal transesterification nucleophile, but that a large number of amino acids at the I_N-1 position are compatible with circular ligation. To generate a library with the maximal chemical diversity and the minimal average molecular weight, our initial construct encoded a library of the form SXXXXX (where X represents any amino acid). Cyclic products from this library have an average molecular weight of less than 650 Da. As such, library members may prove useful as lead compounds for drug discovery, or as therapeutics in their own right. A library of 1.5×10^7 bacterial transformants was generated, which was sufficiently large to characterize the extent of circular ligation of the S+5 construct.

Plasmids from 10 clones that expressed the SICLOPPS fusion protein in trial induction studies were harvested and sequenced to evaluate the nucleotide and codon usage in the library. Cytosine was somewhat under-represented in the library, particularly at the first (N_1) and third (S) positions of each codon (Table 3). Guanine and thymine bases were slightly over-represented, but none of the variation indicated a statistically significant bias at the DNA level. At the protein level, 18 amino acids were represented in the 50 variable positions sequenced (Table 4). Leucine and glutamate were over-represented in the sequenced clones and arginine and serine were correspondingly under-represented, although with such a small sample size it is difficult to determine whether these apparent biases are statistically significant.

Of the 10 selected clones from the S+5 library, two of the three that failed to give cyclic products (S+5.3, S+5.5) contained cysteine residues. In total, five cysteine containing constructs were produced in this study (+1C, S+5.3, S+5.5, S+5.8, $\Delta 4+5.6$). All five processed well in vivo (see +1C and S+5.3 in Fig. 4 lanes 3 and 4 and 9 and 10, respectively), but only two (S+5.8, $\Delta 4+5.6$) gave cyclic products in vitro. Efficient in vivo processing (as exemplified by +1C and S+5.3 in Fig. 4) is incompatible with in vitro peptide production because needed starting materials for in vitro cyclization are consumed in vivo. Indeed, any mechanism through which either starting materials or products are depleted, even slightly, can drop the in vitro peptide product levels below the detection limit. For example, clone S+5.3 (Table 5) displayed a plurality of lariat molecular ions that began at the expected mass and continued for an additional 90 Da. The pattern suggests that either intracellular post-translational modification of the intervening sequence or oxidation of the cysteine residue during in vitro incubation could be responsible for reducing the concentration of peptide products with diagnostic masses. Even with clones S+5.8 and $\Delta 4+5.6$, the most abundant ion corresponded to the mass of the oxidized cyclic product (disulfide dimer) rather than the reduced,

monomeric form (Tables 5 and 6). Moreover, these cysteine containing peptides had low abundance when compared to other cyclic products within the same library. The chemical lability of cysteine may therefore contribute to the difficulty in isolating cysteine containing products following *in vitro* incubation.

The S+5 library highlights the limitations imposed by using *in vitro* cyclization to evaluate intracellular (*in vivo*) cyclization. We were able to produce detectable quantities of pseudostellarin F by allowing affinity immobilized proteins expressed from the SICLOPPS construct to process *in vitro* (Figs. 5 and 6). Together with the results from our previous work describing the purification of pseudostellarin F from the cell lysate and medium [21], these data suggest that the ability of a given construct to cyclize *in vitro* likely reflects its ability to cyclize *in vivo*. We suspect, however, that the failure of a construct to cyclize *in vitro* does not necessarily reflect failure to cyclize *in vivo*. Several of the reported constructs (+1C, +1T, -1E, -1N, S+5.3, S+5.5, S+5.6) show all of the protein products and intermediates in gel (Fig. 4) and mass spectral data (Tables 1 and 5). Since protein products are observed, linear or cyclic peptide products must also be generated. Our failure to observe peptide products of any kind indicates that *in vitro* processing is not an accurate predictor of *in vivo* processing in all cases. Clones that process well *in vivo*, yet fail to give peptide products *in vitro* cannot be evaluated adequately because it is uncertain whether the *in vivo* peptide product is linear or cyclic. Likewise, it is impossible to predict whether the ratio of cyclic to linear product observed *in vitro* reflects the degree of partitioning *in vivo*. For these reasons, our estimates that 70% (seven of 10) of the S+5 library produces cyclic products and that more than 70% of these (five of seven) show some degree of partitioning between cyclic and linear peptides may represent an underestimate of the extent of circular ligation and concentration of cyclic peptide products *in vivo*.

3.7. $\Delta 4+5$ library

Cyclization is likely to be critical for intracellular library screening, so only the most conservative estimates of the useful diversity of the library are justified. Although the majority of randomly selected clones from the S+5 library generated cyclic peptide products, failure to observe cyclization in all cases and significant partitioning to linear peptide products prompted us to investigate whether fixing additional residues might improve the extent of cyclization in the context of a library. The results of the deletion study suggested that the four amino acids of the pseudostellarin F $\Delta 4$ deletion mutant (SGPL) are sufficient for circular ligation (Table 2). A library of five variable residues was therefore inserted between the glycine and proline residues of the $\Delta 4$ sequence so that it could serve as a scaffold to promote cyclization of library constructs. The

resulting constructs (cyclo-[SGXXXXXPL]) are still relatively small molecules, with an average molecular mass of approximately 900 Da. Given the number of transformants generated (in excess of 10^8), there is greater than a 99% chance that every combination of five amino acids is represented at least once in the scaffolded library.

Nucleotide and amino acid codon usage data were determined from the sequences of 10 selected clones. The only apparent biases in the $\Delta 4+5$ library at the nucleotide level (Table 3) involve under-representation of adenine at the central position (N_2) of each variable codon and slight over-representation of cytosine at the wobble base (S). Neither of these apparent biases appears to influence the amino acid distribution, which conforms quite nicely to the distribution predicted from the codon usage in the library (Table 4). When the nucleotide and codon usage data from both the S+5 and $\Delta 4+5$ libraries are compiled (Table 3) the apparent biases observed in the individual libraries are reduced, and 19 amino acids are represented in the 100 variable positions of the accumulated library data.

All 10 of the clones selected from the $\Delta 4+5$ library produced cyclic peptides *in vitro*. The resulting peptides are cleaner coming off the chitin affinity column than peptides resulting from the S+5 library and consequently show more abundant molecular ions in mass spectral analyses (Table 6). Despite the scaffold, some partitioning between cyclic and linear peptide products was observed in eight of 10 cases, but the cyclic peptide was the dominant or predominant product in each case. Direct comparison of the abundance of molecular ions corresponding to the cyclic and linear peptide may overestimate the relative concentration of the linear peptide because the zwitterionic linear peptide should ionize more readily than the corresponding cyclic peptide (see Figs. 5 and 6). Regardless, introduction of the $\Delta 4$ scaffold had the desired effect of enhancing the cyclization of intervening peptide library sequences, and enabled the biosynthesis of a complete multimillion member library of cyclic peptides.

4. Significance

We have described a method, SICLOPPS, that harnesses the biosynthetic machinery of the cell for combinatorial chemistry, creating vast intracellular libraries of small cyclic peptides. Cyclic products ranging from four amino acids to complete proteins have been generated, and intracellular cyclic peptide libraries in excess of 10^8 transformants have been produced. The intein chemistry used for peptide cyclization does not introduce any obvious bias in cyclic products, therefore SICLOPPS libraries can encompass the full chemical diversity of the amino acid monomer pool. Because cyclic peptides are biosynthesized within the cell, libraries can be exploited to inter-

rogate colocalized intracellular targets. Colocalization circumvents loss of chemical diversity at membranes, which plagues chemical genetics approaches that rely on synthetic compound libraries.

Both of the cyclic peptide libraries that were prepared in this study (S+5 and Δ 4+5) gave products with an average molecular mass of less than 1 kDa. As such, these cyclic peptides should be able to penetrate biomolecular complexes and access sites that would be sterically inaccessible to genetically encoded protein probes. Standard molecular biology methodology enables the preparation of complete libraries with up to seven variable positions where every genetic construct is represented at least once in the library. Consequently, thorough structure–activity relationship data can be generated for active cyclic peptide effectors of any physiological process directly out of the primary screen. Moreover, since the method is fully compatible with cell based assays, toxic library members are eliminated early in the screening process. Collectively, these properties make SICLOPPS a valuable new source of ligand diversity that should find wide application in drug discovery and chemical biology.

5. Materials and methods

5.1. Materials

All restriction enzymes and chitin agarose beads were obtained from New England Biolabs (Beverly, MA). T4 DNA ligase was obtained from Gibco-BRL (Gaithersburg, MD). *E. coli* strains DH5 α E and XL1-Blue were obtained from Stratagene (La Jolla, CA). *E. coli* strain Tuner(DE3) was from Novagen (Madison, WI). Oligonucleotides were synthesized on an Expedite 8909 DNA synthesizer (PerSeptive Biosystems, Framingham, MA) and deprotected according to the protocol provided by the manufacturer. PCR primers were passed through Sephadex G25 (Sigma, St. Louis, MO) before use. All other chemicals were of analytical grade or better.

5.2. Methods

5.2.1. Vector construction

Vector pARCBD-p has been reported previously [21]. Vectors pARCBD-p(+1C), pARCBD-p(+1T), pARCBD-p(–1S), pARCBD-p(–1A), pARCBD-p(–1G), pARCBD-p(–1K), pARCBD-p(–1Y), pARCBD-p(–1I), pARCBD-p(–1E), pARCBD-p(–1N) and pARCBD-p(–1P) were derived from pARCBD-p by QuikChange mutagenesis (Stratagene, La Jolla, CA, USA) following the procedure provided by the manufacturer. Deletion vectors pARCBD-p(Δ 1), pARCBD-p(Δ 2), pARCBD-p(Δ 3), pARCBD-p(Δ 4) were similarly constructed by QuikChange mutagenesis. Each construct (Δ X) was prepared by one amino acid deletion of the (Δ X+1) vector. Primer sequences used to construct pseudostellarin F derivatives are available upon request. The identity of all plasmid constructs was confirmed by DNA sequencing. Verified pARCBD-p series plasmids were transformed into XL1-Blue for peptide expression. Expression

was monitored by SDS–PAGE [28] on 15% gels and peptide products were affinity purified as previously described [22].

The expression vector pETCBD-p was constructed by digesting pARCBD-p with *Nco*I and *Hind*III. The smaller fragment was ligated into a similarly digested pET28a(+) plasmid (Novagen, Madison, WI, USA). Libraries were prepared by ligating PCR products encoding variable sequence fused to the N-intein gene into pETCBD-p. Detailed procedures for the construction of the S+5 and Δ 4+5 libraries are described elsewhere [29]. Library plasmids were transformed into the *E. coli* strain Tuner (DE3). Individual colonies were picked, and the peptides were purified as previously described [22].

5.2.2. Mass spectral analysis

Butanol extracts and chitin affinity column eluates were characterized using MALDI mass spectrometry on a Voyager-DE STR time of flight mass spectrometer (PerSeptive Biosystems, Framingham, MA, USA). Sample aliquots (0.5 μ l) were cocrystallized with α -cyano-4-hydroxycinnamic acid (Aldrich, St. Louis, MO, USA) matrix on the MALDI sample plate. Desalting was accomplished by five times washing of the crystallized solid with cold (4°C) 5 μ l volumes of 0.01 M trifluoroacetic acid. Positive ion MALDI mass spectra were acquired in linear mode as the summed signal from 256 shots of 337 nm radiation from a nitrogen laser.

LC/MS analyses were performed using positive mode electrospray ionization on a Mariner mass spectrometer (PerSeptive Biosystems). HPLC separations used a 1 mm \times 50 mm Aquasil C18 (3 μ m) column (Keystone Scientific, Bellefonte, PA, USA) with the following solvents: A=0.15% formic acid in H₂O; B=0.15% formic acid in acetonitrile; C=0.15% formic acid in 2-propanol. Flow rates through the column of approximately 50 μ l/min were achieved using a pump flow rate of 0.5 ml/min and a pre-injection split. The solvent gradient consisted of initial conditions (A/B/C = 5/95/0, hold for 0–1 min) followed by a linear increase to 5/95/0 at 10 min and an additional linear ramp to 5/0/95 at 20 min followed by a 5 min hold.

Acknowledgements

We thank Dr. Deborah S. Grove (Nucleic Acid Facility, Penn State University) for assistance with DNA sequencing. The mass spectrometers used in this study were purchased in part with funds from National Institutes of Health Grant RR11318. C.P.S. was supported in part by National Institutes of Health Grant GM19891.

References

- [1] K. McMillan, M. Adler, D.S. Auld, J.J. Baldwin, E. Blasko, L.J. Browne, D. Chelsky, D. Davey, R.E. Dolle, K.A. Eagen, S. Erickson, R.I. Feldman, C.B. Glaser, C. Mallari, M.M. Morrissey, M.H. Ohlmeyer, G. Pan, J.F. Parkinson, G.B. Phillips, M.A. Polokoff, N.H. Sigal, R. Vergona, M. Whitlow, T.A. Young, J.J. Devlin, Allosteric inhibitors of inducible nitric oxide synthase dimerization discovered via combinatorial chemistry, *Proc. Natl. Acad. Sci. USA* 97 (2000) 1506–1511.
- [2] R.A. Houghton, C. Pinilla, S.E. Blondelle, J.A. Appel, C.T. Dooley,

- J.H. Cuervo, Generation and use of synthetic peptide combinatorial libraries for basic research and drug discovery, *Nature* 354 (1991) 84–86.
- [3] A. Borchardt, S.D. Liberles, S.R. Biggar, G.R. Crabtree, S.L. Schreiber, Small molecule-dependent genetic selection in stochastic nanodroplets as a means of detecting protein-ligand interactions on a large scale, *Chem. Biol.* 4 (1997) 961–968.
- [4] J.L. Silen, A.T. Lu, D.W. Solas, M.A. Gore, D. MacLean, N.H. Shah, J.M. Coffin, N.S. Bhinderwala, Y. Wang, K.T. Tsutsui, G.C. Look, D.A. Campbell, R.L. Hale, M. Navre, C.R. DeLuca-Flaherty, Screening for novel antimicrobials from encoded combinatorial libraries by using a two-dimensional agar format, *Antimicrob. Agents Chemother.* 42 (1998) 1447–1453.
- [5] C.K. Jayawickreme, H. Sauls, N. Bolio, J. Ruan, M. Moyer, W. Burkhardt, B. Marron, T. Rimele, J. Shaffer, Use of a cell-based, lawn format assay to rapidly screen a 442,368 bead-based peptide library, *J. Pharmacol. Toxicol. Methods* 42 (1999) 189–197.
- [6] T.U. Mayer, T.M. Kapoor, S.J. Haggarty, R.W. King, S.L. Schreiber, T.J. Mitchison, Small molecule inhibitor of mitotic spindle bipolarity identified in a phenotype-based screen, *Science* 286 (1999) 971–974.
- [7] N.S. Gray, L. Wodicka, A.M. Thunnissen, T.C. Norman, S. Kwon, F.H. Espinoza, D.O. Morgan, G. Barnes, S. LeClerc, L. Meijer, S.H. Kim, D.J. Lockhart, P.G. Schultz, Exploiting chemical libraries, structure, and genomics in the search for kinase inhibitors, *Science* 281 (1998) 533–538.
- [8] A.R. Davidson, R.T. Sauer, Folded proteins occur frequently in libraries of random amino acid sequences, *Proc. Natl. Acad. Sci. USA* 91 (1994) 2146–2150.
- [9] M. Blind, W. Kolanus, M. Famlok, Cytoplasmic RNA modulators of an inside-out signal-transduction cascade, *Proc. Natl. Acad. Sci. USA* 96 (1999) 3606–3610.
- [10] T.L. Gururaja, S. Narasimhamurthy, D.G. Payan, D.C. Anderson, A novel artificial loop scaffold for the noncovalent constraint of peptides, *Chem. Biol.* 7 (2000) 515–527.
- [11] P. Colas, B. Cohen, T. Jessen, I. Grishina, J. McCoy, R. Brent, Genetic selection of peptide aptamers that recognize and inhibit cyclin-dependent kinase 2, *Nature* 380 (1996) 548–550.
- [12] T.C. Norman, D.L. Smith, P.K. Sorger, B.L. Drees, S.M. O'Rourke, T.R. Hughes, C.J. Roberts, S.H. Friend, S. Fields, A.W. Murray, Genetic selection of peptide inhibitors of biological pathways, *Science* 285 (1999) 591–595.
- [13] G. Caponigro, M.R. Abedi, A.P. Hurlburt, A. Maxfield, W. Judd, A. Kamb, Transdominant genetic analysis of a growth control pathway, *Proc. Natl. Acad. Sci. USA* 95 (1998) 7508–7513.
- [14] A.R. Khan, J.C. Parrish, M.E. Fraser, W.W. Smith, P.A. Bartlett, M.N. James, Lowering the entropic barrier for binding conformationally flexible inhibitors to enzymes, *Biochemistry* 37 (1998) 16839–16845.
- [15] D.F. Veber, R.M. Freidlinger, D.S. Perlow, W.J. Paleveda Jr., F.W. Holly, R.G. Strachan, R.F. Nutt, B.H. Arison, C. Homnick, W.C. Randall, M.S. Glitzer, R. Saperstein, R. Hirschmann, A potent cyclic hexapeptide analogue of somatostatin, *Nature* 292 (1981) 55–58.
- [16] P.H. Bessette, F. Aslund, J. Beckwith, G. Georgiou, Efficient folding of proteins with multiple disulfide bonds in the *Escherichia coli* cytoplasm, *Proc. Natl. Acad. Sci. USA* 96 (1999) 13703–13708.
- [17] S.L. Schreiber, Chemistry and biology of the immunophilins and their immunosuppressive ligands, *Science* 251 (1991) 283–287.
- [18] C. Bousquet, E. Puente, L. Buscail, N. Vaysse, C. Susini, Antiproliferative effect of somatostatin and analogs, *Chemotherapy* 47 (Suppl. 2) (2001) 30–39.
- [19] D.G. McCafferty, P. Cudic, M.K. Yu, D.C. Behenna, R. Kruger, Synergy and duality in peptide antibiotic mechanisms, *Curr. Opin. Chem. Biol.* 3 (1999) 672–680.
- [20] S.R. Bartz, E. Hohenwarter, M.K. Hu, D.H. Rich, M. Malkovsky, Inhibition of human immunodeficiency virus replication by nonimmuno-suppressive analogs of cyclosporin A, *Proc. Natl. Acad. Sci. USA* 92 (1995) 5381–5385.
- [21] C.P. Scott, E. Abel-Santos, M. Wall, D.C. Wahnnon, S.J. Benkovic, Production of cyclic peptides and proteins in vivo, *Proc. Natl. Acad. Sci. USA* 96 (1999) 13638–13643.
- [22] T.C. Evans Jr., D. Martin, R. Kolly, D. Panne, L. Sun, I. Ghosh, L. Chen, J. Benner, X.Q. Liu, M.Q. Xu, Protein trans-splicing and cyclization by a naturally split intein from the dnaE gene of *Synechocystis* species PCC6803, *J. Biol. Chem.* 275 (2000) 9091–9094.
- [23] F.B. Perler, E. Adam, Protein splicing and its applications, *Curr. Opin. Biotechnol.* 11 (2000) 377–383.
- [24] H. Morita, T. Kayashita, H. Kobata, A. Gonda, K. Takeya, H. Itokawa, Pseudostellarins D–F, new tyrosinase inhibitory cyclic peptides from *Pseudostellaria heterophylla*, *Tetrahedron* 50 (1994) 9975–9982.
- [25] F.B. Perler, InBase, the Intein Database, *Nucleic Acids Res.* 28 (2000) 344–345.
- [26] H. Wu, Z. Hu, X.Q. Liu, Protein trans-splicing by a split intein encoded in a split DnaE gene of *Synechocystis* sp. PCC6803, *Proc. Natl. Acad. Sci. USA* 95 (1998) 9226–9231.
- [27] F.W. Studier, B.A. Moffatt, Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes, *J. Mol. Biol.* 189 (1986) 113–130.
- [28] U.K. Laemmli, Cleavage of structural proteins during the assembly of the head of bacteriophage T4, *Nature* 227 (1970) 680.
- [29] E. Abel-Santos, C.P. Scott, S.J. Benkovic, Use of inteins for the in vivo production of stable cyclic peptide libraries in *Escherichia coli*, in: P. Vaillancourt (Ed.), *Methods in Molecular Biology: E. coli Gene Expression Protocols*, Humana Press, Totowa, NJ (in press).