PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Combining the polymerase incomplete primer extension method for cloning and mutagenesis with microscreening to accelerate structural genomics efforts

Heath E. Klock,[1,2] Eric J. Koesema,[1] Mark W. Knuth,[1,2] and Scott A. Lesley[1,2]*

[1] The Genomics Institute of the Novartis Research Foundation, San Diego, California 92121

[2] The Joint Center for Structural Genomics at the Genomics Institute of the Novartis Research Foundation, San Diego, California 92121

ABSTRACT

Successful protein expression, purification, and crystallization for challenging targets typically requires evaluation of a multitude of expression constructs. Often many iterations of truncations and point mutations are required to identify a suitable derivative for recombinant expression. Making and characterizing these variants is a significant barrier to success. We have developed a rapid and efficient cloning process and combined it with a protein microscreening approach to characterize protein suitability for structural studies. The Polymerase Incomplete Primer Extension (PIPE) cloning method was used to rapidly clone 448 protein targets and then to generate 2143 truncations from 96 targets with minimal effort. Proteins were expressed, purified, and characterized via a microscreening protocol, which incorporates protein quantification, liquid chromatography mass spectrometry and analytical size exclusion chromatography (AnSEC) to evaluate suitability of the protein products for X-ray crystallography. The results suggest that selecting expression constructs for crystal trials based primarily on expression solubility is insufficient. Instead, AnSEC scoring as a measure of protein polydispersity was found to be predictive of ultimate structure determination success and essential for identifying appropriate boundaries for truncation series. Overall structure determination success was increased by at least 38% by applying this combined PIPE cloning and microscreening approach to recalcitrant targets.

Proteins 2008; 71:982–994.

Key words: ligase-independent; construct optimization; high-throughput; sequence-independent; site-directed; PCR.

INTRODUCTION

The Protein Structure Initiative (PSI, www.nigms.nih.gov/Initiatives/PSI/) seeks to expand structural coverage of large protein families by mining genomes for likely targets of novel structures while integrating innovative technologies, which reduce the time and costs per structure. As part of this effort, The Joint Center for Structural Genomics (JCSG, www.jcsg.org) has created a high-throughput protein expression, purification, and crystallization platform[1,2] and used it to study the entire proteome of *Thermotoga maritima*[1] as well as over 7000 representative protein targets from over 100 additional bacterial genomes. The throughput of this pipeline is driven by the availability of expression-ready clones. To access and translate the ever-increasing amount of genomic and metagenomic information available (www.genomesonline.org)[3] efficiently, rapid methods for cloning and characterizing gene products are critical to protein production. With protein structures deposited in the protein data bank for over 454 targets to date (http://www.jcsg.org/prod/scripts/public_targets/structure.html), the power of a high-throughput pipeline approach to structural genomics is evident.

In general, structural genomics efforts generate large clone collections from which protein crystallization varies widely.[4–6] For problematic targets, mutations, and truncations are often used to improve crystallization. Surface mutations have been shown to be an effective means of improving crystallization.[7] Also, analytical techniques such as partial proteolysis or deuterium-exchange mass spectrometry[8] can be used to define construct boundaries for successful protein crystallization and structure determination.[9] However, prediction of exact boundaries leading to improved crystallization *a priori* is rarely successful when working on targets without known structural homology. Empirically determining correct boundaries for difficult targets typi-

This work was performed at The Genomics Institute of the Novartis Research Foundation.
Grant sponsor: National Institute of General Medical Sciences, Protein Structure Initiative; Grant number: U54 GM074898.
*Correspondence to: Scott A. Lesley, 10675 John J Hopkins Drive, San Diego, CA 92121.
E-mail: slesley@gnf.org
Received 30 May 2007; Revised 8 August 2007; Accepted 14 August 2007
Published online 14 November 2007 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.21786

cally requires cloning, expression and crystallization trials of many truncated gene constructs. The Polymerase Incomplete Primer Extension (PIPE) method for cloning and mutagenesis combined with protein characterization by microscreening is an effective way to create either initial clones or mutant and truncation arrays and then to screen constructs for the highest potential of crystallization success.

The literature is full of diverse cloning strategies for creating expression clones of predicted genes from newly sequenced genomes.[10–21] Popular ligase-independent methods are generally faster and exhibit higher cloning efficiencies than earlier, more traditional protocols. However, most of these protocols suffer from sequence requirements, which can encode unwanted amino acids or significantly limit the flexibility of the cloning strategy. "Enzyme-free cloning" eliminates sequence constraints of other ligase-independent methods by annealing "hetero-staggered" PCR products but suggests multiple rounds of PCR amplifications, removal of residual PCR reactants and a denaturation/hybridization step.[19,20] Most recently, the SLIC method describes a RecA-mediated homologous recombination approach, which mentions the utility of incomplete PCR (iPCR) products.[21] These iPCR products are equivalent to the PCR products generated by PIPE, which are the basis of the method described here. The PIPE method is a facile cloning and mutagenesis approach, which retains both high cloning efficiencies and accelerated throughput without any inherent sequence, strain and recombination limitations or additional steps and reactions of other ligase-independent methods. Simply stated, unpurified PCR products, comprised of vector and insert DNA fragments, are directly transformed into bacteria to create clones without any extra manipulations.

In this study, we used a combination of PIPE cloning methods and protein characterization to evaluate a set of 448 full-length targets as well as an array of 2143 truncation clones derived from a diverse subset of 96 soluble targets. While most truncation studies focus on the production of soluble protein as the initial endpoint prior to crystallization trials,[22,23] our results indicate that solubility alone is insufficient as a selection criterion and that aggregation must also be considered.

METHODS

PIPE overview

The PIPE method takes advantage of the observation that normal PCR reactions generate a population of partially single-stranded DNA fragments[24] resulting from incomplete primer extension. These PCR products contain consistent 5′-ends with sequences defined by the amplification primers, while each 3′-end varies in length as a function of primer extension. Using a simple primer design rule, complementary 5′-ends are created to function as annealing templates for combining PCR fragments when either cloning or generating mutants. This primer design rule proposes the incorporation of 14–17 bases of sequence onto the 5′ end of each primer. This sequence must be reverse-complementary to the 5′ end of an opposite strand primer where annealing is desired, but the overlapping sequence can be composed of any sequence so long as this complementarity is maintained. Since the cloned product is from PCR amplifications, template DNA concentrations can be reduced to the point where background colonies are negligible. The simplicity, flexibility and efficiency of the PIPE method makes it ideal for generating large numbers of expression clones, including protein variants to enhance biophysical properties, quickly and easily.

The SpeedET vector

Genes and mutants were cloned into vector pSpeedET/*ccdB* (~4 kb). This plasmid encodes the N-terminal expression and purification tag MGSDKIHHHHHHEN-LYFQG and was constructed by replacing ampicillin resistance with kanamycin resistance on the pBAD/thio (Invitrogen)-derived vector pMH4[1] and also adding the TEV protease cleavage site sequence. PCR primers for this vector are designed to amplify the entire vector except for the lethal *ccdB* gene such that any vector template contamination in the transformations cannot propagate. The N-terminal tag can be removed by treatment with TEV protease, which results in the addition of only a single glycine residue to the cloned sequence. Protein expression is inducible using either the arabinose or T7 promoter. It is important to note that the PIPE method does not require this vector. Any vector, which can be PCR-amplified should work with this method and we have utilized PIPE with a number of such plasmids. In fact, other ligase-independent cloning vectors are immediately amenable to the PIPE method simply by PCR amplification across the ligation-independent cloning (LIC) sites of both the vector and insert(s).

Observing incomplete primer extension in PCR reactions

Oligonucleotides (desalted, Integrated DNA Technologies) 5′SpeedETreverse (5′-ctggaagtacaggttttcgtgatgatgat gatgatg-3′) and 3′SpeedETforward (5′-cgcgacttaattaactc gtttaaacggtctccagc-3′) were phosphorylated by T4 Polynucleotide Kinase (New England Biolabs). A 50-μL PCR reaction was set up containing 1 μM each of these primers, 1× Cloned *Pfu* DNA Polymerase Reaction Buffer, 2.5 units of *Pfu* Turbo DNA polymerase (Stratagene), 200 μM of each dNTP (Invitrogen) and 1 ng of pSpeedET template. The reaction was treated as follows: Initial denaturation for 2 min at 95°C, then 25 cycles of 95°C for 30 s, 55°C for 45 s, and 68°C for 14 min followed by a cool down to 4°C.

To observe the extent of incomplete primer extension leading to single-stranded 5′-ends, the PCR product above was diluted 1:2 into 50 m*M* NaOAc, 30 m*M* NaCl, 1 m*M* $ZnSO_4$, pH 5.0 containing 10 units of Mung Bean Nuclease (New England Biolabs) and incubated 1 h at 30°C (Mung Bean Nuclease degrades single-stranded DNA leaving ligatable, blunt ends). The resulting blunt-end products were ligated for 1 h at room temperature using 1 unit of T4 DNA Ligase (Invitrogen) in 1X T4 Ligase Buffer and transformed into GeneHogs cells (Invitrogen). Plasmids from resulting colonies were sequenced across the ligation site to determine the lengths of the single-stranded sequences removed by nuclease treatment as well as the frequency of such events.

PIPE cloning of 448 targets

Oligonucleotides (desalted, Integrated DNA Technologies) used for the PIPE method were designed to include extensions of at least 15 bases complementary to the paired oligonucleotide for hybridization. Successful PCR amplification was achieved for 480 of 516 gene targets from 30 different bacterial genomic DNA templates. The gene inserts were amplified using 40–45mer primer pairs, which included extensions complementary to the desired vector cloning site as follows: a 5′ Insert forward primer (5′-*aacctgtacttccag*ggc-plus gene-specific sequence-3′) and a 3′ Insert reverse primer (5′-gagttaattaagtcgcgtta-plus gene-specific sequence-3′). SpeedET was amplified using a 5′ Vector reverse primer (5′-*ctggaagtacaggttt*tcgtgatgatgatgatgatg-3′) and a 3′ Vector forward primer (5′-cgcgacttaattaactcgtttaaacggtctccagc-3′). The italicized and underlined bases highlight the two distinct complementary regions between primers where annealing occurs. Insert PCR was performed as above except using dephosphorylated oligonucleotides, 20 ng of genomic DNA template and extending at 72°C for 1 min per kilobase. Vector PCR was performed exactly as above (the mung bean nuclease treatment was not done here and is not part of the normal PIPE cloning method). Each unpurified insert PCR product was mixed 1:1 (v/v) with the unpurified vector PCR product. Immediately after mixing, a 2-μL aliquot of the mixture was incubated on ice with 20 μL of GeneHogs competent cells (homemade by $CaCl_2$ method, transformation efficiency ~10^6 cfu/μg) for 15–25 min. Cells were transformed by heat shock then plated on selective media. The transformants were sequenced across the annealing junctions.

PIPE mutagenesis to create 2143 truncation clones

Truncation mutants were created by PIPE using primer pairs consisting of a universal vector-specific primer and a truncation-specific primer for each truncation. Miniprepped plasmids from the initial clones were used as the template DNA (1 ng per reaction) for PCR. For N-terminal truncations, the vector-specific, 5′ reverse primer (5′-cgattgaaagtacaggttttcgtgatgatgatgatgatg-3′) and the insert-specific, 5′ forward primer (5′-ctgtactttcaatcg-plus truncated gene sequence-3′) were used. The C-terminal truncations were created by amplifying the same DNA templates with a universal vector-specific, 3′ forward primer (5′-*taggtgacgaccatt*gggtaagtttaaacggtctccagc-3′) and an insert-specific, 3′ reverse primer (5′-*aatggtcgtcaccta*-plus truncated gene sequence-3′). Again, the underlined and italicized bases highlight the complementary regions across the primers (only one annealing site for single-point mutagenesis). PCR reactions were performed as described above for Vector PCR. Unpurified PCR products were directly transformed and resulting clones were validated using LC-MS confirmation of masses of the purified proteins and/or partial DNA sequencing.

Microscreening—expression/purification

Overnight cultures of confirmed clones were diluted 1:25 into 1.2 ml of Modified-TB media (24 g yeast extract, 16 g tryptone, 10 g bacto-casamino acid, 2% glycerol and 100 m*M* Tris-HCl, pH 8.0 per liter) in deep 96-well blocks (Corning Costar® No. 3961), sealed with AirPore Tape Sheets (QIAGEN No. 19571) and incubated at 37°C, 900RPM (Glas-Col incubator). After about 2 h of growth, cultures were induced at an OD_{600} of 1.3–1.7 by adding 0.02% arabinose for 5 h (final OD_{600} of 8–12) and harvested by 15 min of centrifugation at 3000*g*. Pellets were frozen at −20°C overnight, thawed, then resuspended with 600 μL of Buffer A (50 m*M* Tris-HCl, 50 m*M* sucrose and 1 m*M* EDTA, pH 7.5) containing 500 units of Ready-lyse (Epicentre). After 15 min, room temperature incubation, 600 μL of Buffer B (10 m*M* Tris-HCl, 50 m*M* NaCl, 1 m*M* EDTA, and 10 m*M* $MgCl_2$, pH 7.5) containing 60 units of benzonase (Novagen) was added and lysates were incubated an additional 15 min at room temperature. Crude lysates were cleared by 40 min of centrifugation at 7000*g*. Supernatants were applied to 50 μL of Ni-NTA resin (QIAGEN) in a 96-well filter plate (Thomson Instrument Company No. 931919), pre-equilibrated with 50 m*M* HEPES, 50 m*M* NaCl and 10 m*M* imidazole, pH 8.0. The resin was washed with 700 μL of 50 m*M* HEPES, 10% glycerol, pH 8.0. The proteins were eluted with 100 μL of 50 m*M* HEPES, 300 m*M* imidazole and 10% glycerol, pH 8.0. Eluate protein concentrations were measured by protein assay using Coomassie Plus Reagent (Pierce Biotechnology) per manufacturer's recommendation.

Microscreening—liquid chromatography-mass spectrometry and analysis

A 15-μL aliquot from each eluate was analyzed by LC-MS in TFA/H_2O/CH_3CN on a ZORBAX Poroshell

300SB-C18 desalt column (Agilent) using a Micromass QTOFII coupled with a Capillary LC (Waters Corporation) to confirm expected masses. Note: TEV site cleavage efficiency could be assessed by a 2-h room-temperature incubation with TEV protease (the catalytic domain of the NIa protein from tobacco etch virus) followed by LC-MS analysis.

Microscreening—analytical size-exclusion chromatography

Fifteen microliter samples were injected onto a Shodex Protein KW-803 column with a KW-G guard column (Thomson Instrument Company) and eluted at 1.5 mL/min. in 20 m*M* Tris, 200 m*M* NaCl, 0.25 m*M* TCEP and 3 m*M* NaN_3, pH 7.9. The run time was 13.3 min per sample and detection was at 280 nm. The HPLC used was an HP1100 LC with a refrigerated, 384-well autosampler (Agilent Technologies).

Preparative-scale protein production and crystallization

Full-length and truncated proteins were expressed and purified at preparative scale using the GNFermentor/GNFuge expression/purification system.[1] Nickel-affinity purified proteins were buffer exchanged using PD10 columns (GE Healthcare), digested with histidine-tagged TEV protease to remove the N-terminal expression/purification tag and subjected to subtractive IMAC to separate the TEV protease-cleaved and uncleaved fractions before being concentrated to ~15 mg/mL.

Full-length proteins and select truncations were subjected to crystallization trials using a selection of 384 commercial sparse matrix conditions, which were previously sorted by past crystallization success.[6] Protein drops of 250 nL were mixed with 250 nL of each crystallization solution and incubated under vapor diffusion conditions at 4°C. Crystallization was scored after 7, 14, and 28 days.

RESULTS

PIPE cloning and mutagenesis

Observation of incomplete primer extension during PCR amplification

Primer extension heterogeneity was determined by measuring the length distribution of synthesized strands created in normal PCR reactions by treating PCR products with mung bean nuclease to remove any single-stranded regions present. Clones, generated by ligating the nuclease-treated ends of the PCR amplified plasmids, were sequenced across the cloning junction of the PCR products. Sequence analysis of 272 clones showed that deletions from 1 to 171 bases occurred in 254 (93%) of the clones and only 7% of the PCR products were fully double-stranded. The average deletion was 76 ± 43 bases and the median deletion was also 76 bases. This correlated well with Olsen and Eckstein[24] who reported a wide distribution of synthesized strand lengths. This observation was likely a result of incomplete primer extension and was found in parallel experiments to be common in all PCR amplifications attempted regardless of the polymerase used. The thermostable DNA Polymerases from *Pyrococcus furiosus* (*Pfu*Turbo, Stratagene), *Thermus aquaticus* (*Taq,* Invitrogen), *Thermococcus kodakaraensis* (*KOD,* EMD Biosciences) and Phusion™ (Finnzymes) were tested (data shown is for *Pfu*Turbo only). Although complete extension promoted amplification in early cycle PCR, incomplete extension became favored and promoted heterogeneous, partially single-stranded products in late cycle PCR. The sequences on the 5′-ends of these PCR products were controlled by the incorporated oligonucleotide primer, while the extent of single-strandedness was a function of PIPE [Fig. 1(a)].

Application of PIPE for cloning and mutagenesis

Given this observation, it became clear that primers could be designed such that populations of PCR products directly amenable to ligase-independent annealing would automatically result during PCR amplification. By incorporating complementary sequences on the 5′-ends of PCR primers and subsequent PCR products, the populations of DNA fragments containing single-stranded, 5′ ends could readily anneal across these sequences either intermolecularly for PIPE cloning [Fig. 1(b)] or intramolecularly for PIPE mutagenesis [Fig. 1(c)]. For PIPE cloning, primers were designed with 15 bases of complementary sequence on the 5′-ends targeted for assembly. After PCR amplification, unpurified aliquots from the vector and insert reactions were mixed and immediately transformed into recipient cells. Any remaining single-stranded gaps or nicks within the annealed vector/insert hybrid were repaired *in vivo*. With this mechanism, homologous recombination is not required and the method is simplified. In this study, sequence positive clones were obtained for 448 of 516 (87%) genes attempted in this way. Table I summarizes the points at which failures occurred and shows a negligible difference between the average gene size of clones attempted versus clones obtained. Additionally, Table II shows that 71% of the sequence positive clones could be obtained by screening just 1 or 2 colonies. Here, we described amplifying insert and vector DNA fragments separately before mixing and transforming cells. However, it is also possible to amplify the two (or more) fragments in the same reaction, then, directly transform cells to remove the mixing step (data not shown). Although the results shown here are specifically from PIPE cloning into the SpeedET vector, we have used this method for cloning into 15 bacterial expression vectors (4.0–4.7 kb), 14

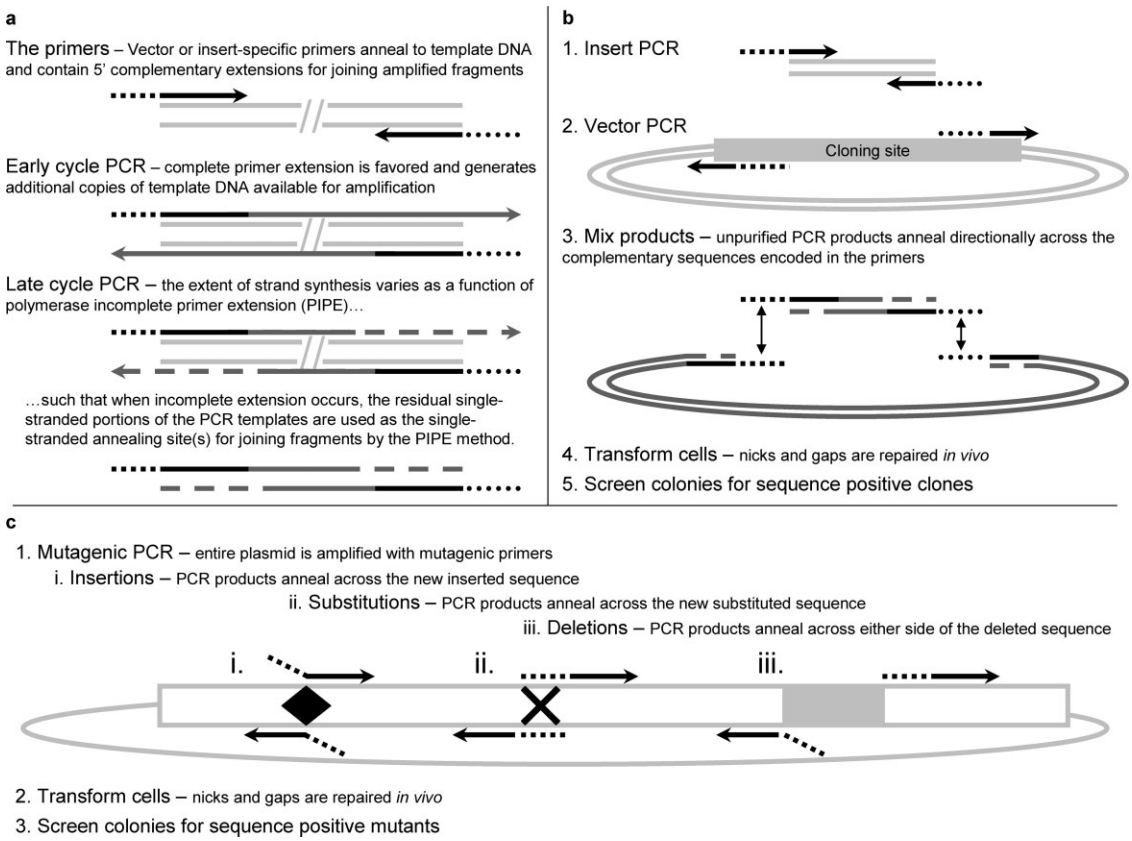

Figure 1

Schematics of proposed PIPE mechanism and methods. The light gray lines represent template DNA. The black lines with dashed or dotted ends represent the primers with 5′ complementary extensions. The black square dashes represent sequences complementary to each other as do the black circles. The straight dark gray lines represent complete strand synthesis. The dashed dark gray lines represent heterogeneous primer extension resulting from PIPE. (a) Progression of PIPE during normal PCR amplification. (b) The PIPE Entry Cloning method. (c) Primer design guidelines and method for PIPE Mutagenic Cloning for creating insertion, substitution and deletion mutants.

baculovirus expression vectors (4.6–5.0 kb) and five mammalian expression vectors (6.3–7.0 kb) with similar success (data not shown). Other, independent labs have had success with PIPE cloning using their own preferred vectors as well. Additionally, while time and resources are insufficient to sequence entire plasmids routinely, deleterious PCR-derived mutations (i.e., occurring within critical expression control features) within the vectors have not been observed to be a major concern. However, the products of vector amplifications can be sequenced across these important features in batch as a general means of quality assurance.

Early expression data for the entire 448 clone collection showed that 114 full-length targets produced soluble

Table I

PCR Amplification, Initial Cloning and Mutagenesis Efficiency Results

	Attempts		Failures		Successes		
	No.	Insert size (bp)	PCR[a]	Cloning[b]	No.	Insert size (bp)	% Success
Initial cloning	516	811 ± 323	36	32	448	795 ± 319	87
Mutagenesis	2304	—	161	0	2143	—	93

[a]Target did not amplify or amplified at incorrect size from genomic DNA.
[b]No insert, wrong insert, mutation(s) observed and/or sequence data was not definitive.

Table II
Number of Colonies Screened to Obtain Sequence-Positve Clones

		Failures		Successes				Totals	
	Attempts	1 colony	1–18	1 colony	1–2	3–4	5–18	1–18 colonies	%
Initial cloning	480[a]	—	32	—	316	99	33	448	93
Mutagenesis	2304	161	—	2143	—	—	—	2143	93

[a]Although there were 516 initial targets, 36 failed PCR leaving only 480, where successful cloning was possible.

protein. Nine of those targets quickly led to structure solutions. From the remaining 105 soluble targets, 96 were selected at random for analysis via truncations arrays. Twenty-four deletion constructs, comprised of 12 N-terminal and 12 C-terminal truncations at four codon increments, were attempted for each of the 96 targets (2304 total). These truncation arrays were made using the PIPE mutagenesis method for deletions [Fig. 1(c)(iii)]. The plasmid DNAs from these 96 initial clones were used as the templates for PCR amplification using primers defining the desired truncation boundaries. After PCR, these PCR products were simply transformed into the recipient host without additional manipulation. Background colonies due to the presence of template DNA in the PCR reactions could be effectively eliminated either by simply using a low template concentration for PCR (we found 50 pg/μl to be sufficiently low, data not shown) or by adding a methylation-dependent restriction enzyme such as *Dpn* I as used in common mutagenesis reactions.[25,26] Either variation provided high mutagenic efficiencies without significant background contamination, although the *Dpn* I treatment adds cost, another step and an hour to the protocol. By screening only one colony per transformation (no *Dpn I* treatment was used), 2143 (93%) truncation clones were created (Tables I–II).

Microscreening structured truncation arrays

Suitability of a target for protein crystallography is dependent on multiple criteria. Sufficient protein yields, which are uniform in integrity and monodispersity and are unaggregated must be produced for crystallographic screens. Regions of disorder should be avoided as they can interfere with crystallization or crystal packing, which may lead to poor diffraction. When working with targets without known structural homology, prediction of disordered regions is difficult. Often, small disordered regions are present at the N- or C-termini, which can be removed by screening truncation series. Evaluating 2143 deletion constructs and the corresponding 96 full-length proteins for crystallographic suitability would normally take too much time and effort to be feasible. However, our microscreening effort offers a way to quickly profile the various constructs.

First, a standard Bradford protein assay was used to determine the soluble yields of all constructs. Of the 96 full-length targets, 67 (70%) had at least one truncation construct, which retained a minimal level of solubility. In this study, minimal solubility was defined as a microscreened, soluble fraction eluate having a protein concentration of at least 100 μg/mL. work has suggested that microscreened protein eluates with soluble expression levels under 100 μg/mL correlate to insufficient protein yields for crystallization trials (data not shown). Soluble proteins were submitted for LC-MS analysis to confirm protein identity. LC-MS was found to be a simple, fast and useful quality control measure. A total of 327 truncations, at least minimally soluble and LC-MS verified, were generated from the 67 truncation-friendly targets (Fig. 2). Within this large subset, the median length of truncation where solubility was retained was 16 amino acids (n = 60, range: 4–48 residues) from the N-terminus and 28 amino acids (n = 37, range: 4–48 residues) from the C-terminus with wide variation observed. Interestingly, several of the targets showed "islands" of solubility within the truncation series. Presumably these islands represent regions between dispensable secondary elements. Nearby truncations, which disrupt, rather than eliminate, these secondary elements likely promote aggregation or result in insoluble expression.

Second, each protein with at least minimal solubility was submitted for analytical size-exclusion chromatography (AnSEC) analysis using a rapid protocol allowing 96 proteins to be screened for monodispersity per day. Our previous work has shown a strong correlation between structure determination success and AnSEC scores. Full-length and truncated proteins were assigned these qualitative AnSEC scores (1 [completely aggregated and/or polydispersed] to 4 [no aggregation and monodispersed]) based on their AnSEC chromatogram profiles [Fig. 3(a)]. In our experience, an AnSEC score of 3 or 4 is necessary but not sufficient for structure determination success. For example, 83% of a diverse group of past structures of full-length proteins gave AnSEC scores of 4 and 6% came from proteins scoring a 3, while the other 11% were not scored for various reasons. Although we do not solve every protein scoring a 3 or 4, no structures have come from proteins scoring a nonoptimal 1 or 2 to date (data not shown). In other words, proteins with AnSEC

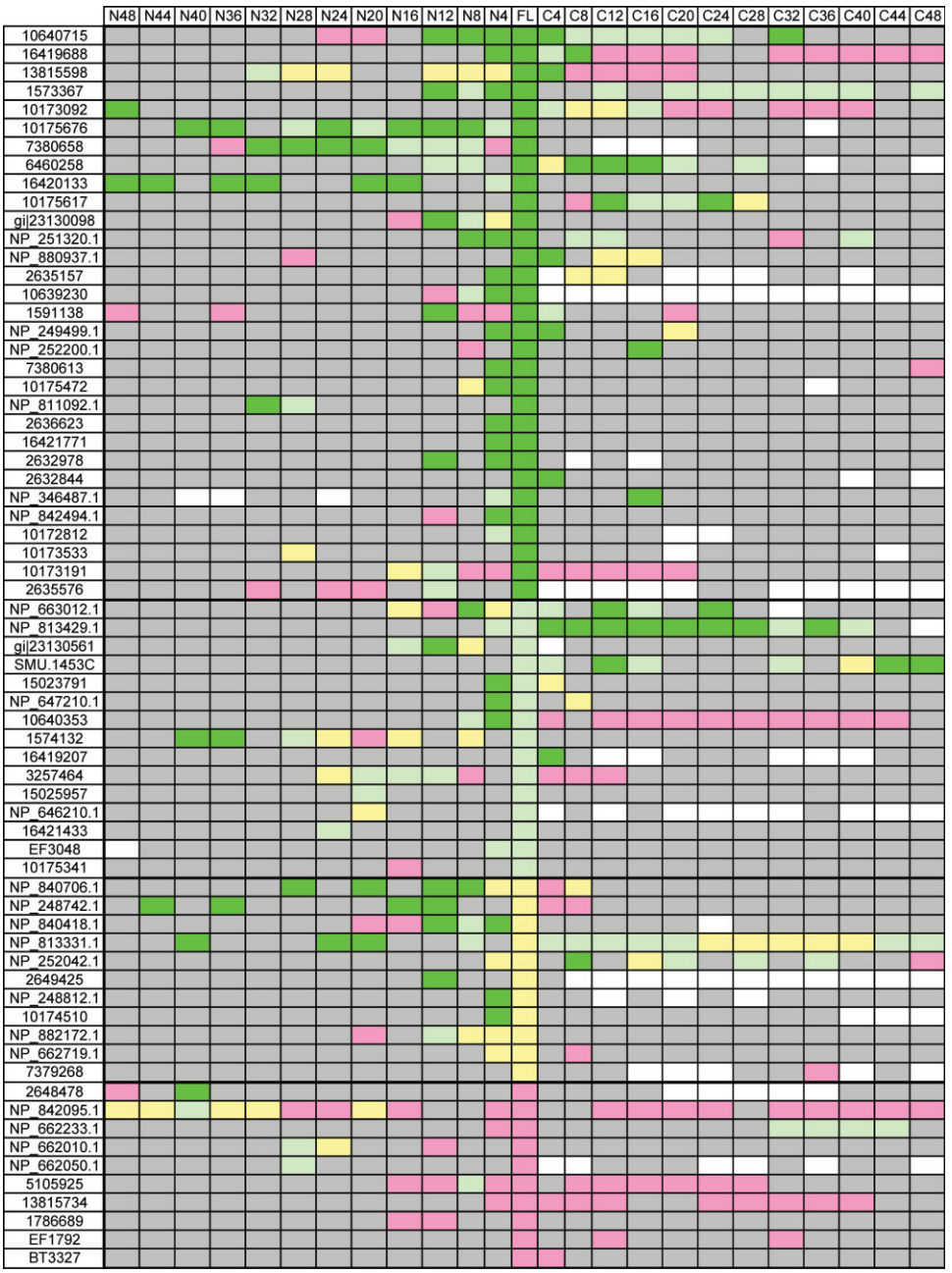

Figure 2

Compilation of solubility and AnSEC scoring results for protein targets with at least one minimally soluble truncation. "FL" indicates the full-length version of the target. Truncations are denoted by "N" for N-terminal or "C" for C-terminal followed by the number of terminal residues removed from the full-length sequence for the given truncation. Bright green squares represent soluble constructs with an AnSEC score of 4. Light green: AnSEC score is 3. Yellow: AnSEC score is 2. Pink: AnSEC score is 1. Light gray squares did not express soluble protein. Clones were not obtained for blank squares.

scores of a 3 or 4 are the optimal candidates for crystallization trials. The AnSEC scoring results for all 67 targets allowing soluble truncations are found in Figure 2.

Of the 67 truncation-friendly, full-length targets, 46 had optimal AnSEC scores (3–4) for crystallization trials. In 43 (93%) instances, at least one (3 on average) additional truncated construct having an optimal AnSEC score could also be generated. There were also 21 of the 67 truncation-friendly, full-length targets that had non-optimal AnSEC scores (1 or 2). Here, optimally scoring truncation constructs could be generated for 15 (71%) targets (2.5 constructs on average), clearly demonstrating

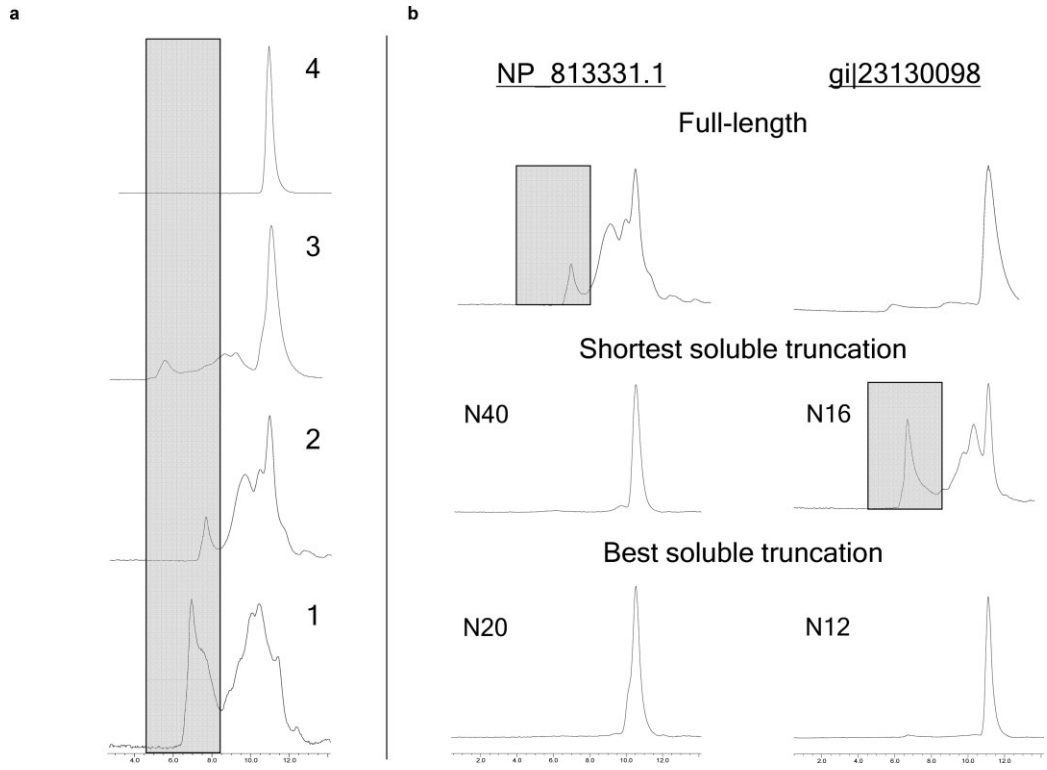

Figure 3

Using AnSEC chromatogram scoring to characterize full-length and truncated protein targets. The shaded areas outline the range where aggregation is normally observed. (a) Examples of protein chromatograms for the four AnSEC scoring categories. A "4" score describes profiles of sharp, monodispersed peaks, which identify well-behaved proteins. A "3" score describes predominately monodispersed peaks with low levels of aggregation and/or polydispersity. A "2" score shows increased aggregation and/or polydispersity, in which obtaining a monodispersed protein by preparative size-exclusion chromatography would be difficult. A "1" score is indicative of a completely aggregated sample and/or a highly polydispersed sample where obtaining an appropriate sample for successful crystal trials is virtually impossible. (b) AnSEC profiles are shown for full-length, shortest truncation and best truncation for NP_83331.1 and gi\|23130098. NP_83331.1 shows substantial improvement in monodispersity when truncated. The target gi\|23130098 shows an island of solubility and monodispersity as seen by its sensitivity to minor truncation at the N-terminus.

that truncation arrays were an effective route to improving target polydispersity. The total (327) minimally soluble truncations were evenly distributed between N-terminal (163) and C-terminal (164) deletions. However, 60% (97 of 163) of the N-terminal truncations yielded optimally scoring proteins, while only 45% (74 of 164) of the C-terminal truncations yielded optimal protein for crystal trials.

Figure 4 shows a full distribution of the AnSEC scores comparing full-length and truncated proteins with 66 of 171 (39%) truncation constructs having better scores than the original full-length constructs where improvement was possible (targets with full-length AnSEC scores of 4 were excluded in this calculation). Truncation AnSEC scores were no worse than full-length scores in another 81 of 273 (30%) possible instances indicating that nonessential sequence could be removed in these cases as well (targets with full-length AnSEC scores of 1 were excluded in this calculation). In few cases, both N- and C-terminal truncations were beneficial (Fig. 2, NP_663012.1 and NP_813331.1). Importantly, however, many of the soluble truncations produced aggregated protein when characterized by AnSEC. The results showed that even four residue changes could result in drastic changes in aggregation state despite adequate soluble expression levels. For example, the protein NP_840418.1 shows a region of good behavior amongst the soluble N-terminal deletions. Removing 4, 8, and 12 amino acids resulted in improvements in aggregation state (AnSEC 2 to AnSEC 3 or 4) over full-length. However, protein quality deteriorated significantly (AnSEC 4 to AnSEC 1) with 16 and 20 residue deletions, while further truncations of this target did not result in minimally soluble proteins.

Crystallization trials and structure determination

Crystallization trials were initiated for all 96 original, full-length targets and a subset of the 327 soluble trun-

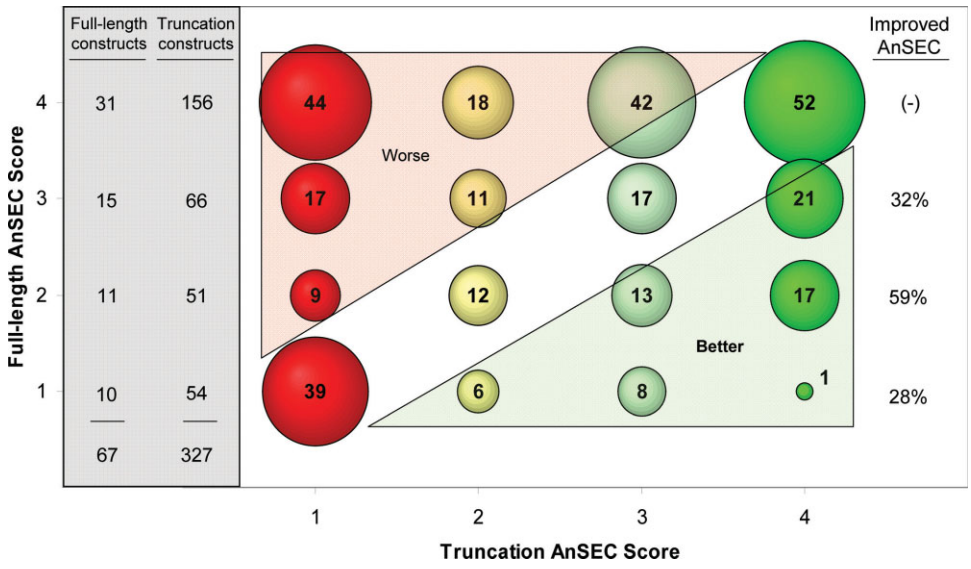

Figure 4

Effect of truncation arrays on AnSEC profiles. The y-axis gives the AnSEC scores of the full-length versions of the 67 protein targets that tolerated at least one truncation. The x-axis gives the AnSEC scores of the 327 minimally-soluble truncation constructs sorted by the parental full-length score. The green shaded triangle highlights the numbers of truncation constructs, which improved on the full-length AnSEC scores ("Better"). The red shaded triangle highlights the numbers of truncation constructs, which worsened the full-length AnSEC scores ("Worse"). The bubbles corresponding to the truncation constructs scoring "4" and "3" are shaded bright and light green, respectively, since they are more likely to lead to structures than the "1" (red) and "2" (yellow) scoring constructs.

cated versions. Proteins for these trials were expressed at large-scale from defined media containing selenomethionine. Purified proteins were screened for crystallization by scoring crystal growth within a sparse matrix of 384 crystallization conditions at 4°C over 1 month. Simply comparing protein constructs by the number of conditions yielding harvestable crystals can be misleading since propensity to crystallize does not always track with structure determination success. Therefore, the level of crystallization success was defined by the furthest stage a construct progressed.

Table III shows the furthest stage reached as either attempted but had "insufficient yield" for crystal trials, setup for crystal trials but yielded "no crystals," setup for crystal trials and yielded "marginal crystals," setup for crystal trials and yielded "harvestable crystals," or resulted in a "**solved structure**". The targets with bold "X"s in the "harvestable crystals" column are cases of very high structure solution potential. There were instances where the truncation series yielded "no soluble truncation" to test and cases where a result for a soluble truncation was "not determined" because the full-length version had already resulted in a structure. Harvestable crystals were obtained for 56 of the 96 targets and structures were determined for 18 of those 56 targets to date. Targets with harvestable crystals were grouped into three classes: targets where only full-length versions yielded harvestable crystals or structures (27 targets, eight structures), targets where only truncated versions yielded harvestable crystals or structures (12 targets, three structures) and targets where both full-length and truncated versions yielded harvestable crystals or structures (17 targets, two truncated and five full-length structures). Within the first class (only full-length versions yielded harvestable crystals or better), there were many targets where soluble truncated versions either did not exist or did not have sufficient yields for crystal trials. In a few cases, the full-length version was solved before the truncation(s) became available. Including truncated protein constructs in the crystallization efforts increased the number of targets with at least harvestable crystals from 44 to 56 targets (a 27% improvement), while adding five truncated structures to the 13 full-length structures (a 38% increase) determined in this test set. Additionally, 5 other truncated targets (bold "**X**"s in the "harvestable crystals" column) have had harvestable crystals diffract to between 1.6 and 3.7 Å (average 2.85 Å). With continued effort, these are likely to lead to four more solved structures. In total, target to structure success rate might be improved by as much as 69% from using this series of truncated constructs.

Generally, those with optimal AnSEC scores of 3 and 4 performed substantially better than those scoring a non-optimal 1 or 2. A few targets with soluble aggregate did result in harvestable crystals. However, few of these crystals were of sufficient quality to screen for diffraction

and none diffracted sufficiently for structure determination. Several targets had multiple truncations put into crystallization trials. In these cases, the shortest truncation attempted was not always the best crystallizing construct. The differences in the AnSEC profiles between these truncations ranged from subtle (NP_813331.1) to stark (gil23130098) [Fig. 3(b)]. Again, structures were determined for 13 full-length targets and 5 truncated targets (32% of the 56 targets with harvestable crystals, 19% overall) and the AnSEC scores of these solved proteins were either "3"s (4 structures, 22%) or "4"s (14 structures, 78%).

DISCUSSION

The PIPE method for cloning and mutagenesis offers significant advantages for high-throughput molecular biology. Unpurified PCR reactions are simply transformed into competent cells without any additional manipulations, which eliminates the additional enzyme costs and steps common with other methods. Clones, substitutions, insertions and deletions can all be made with slight variations of the same protocol without the sequence limitations or added steps needed when using recombinatorial, restriction enzyme-based or other ligase-independent cloning methods. In this study, PIPE cloning manually generated 448 full-length clones in about a week while an array of 2143 nested deletions from 96 of these initial clones was manually created in about 2 weeks using PIPE mutagenesis. No significant bias in cloning efficiencies due to the size of the inserts tested here was observed. Also, most initial clones and all truncation constructs created could be obtained by screening just 1 or 2 colonies. Primer-based mutations were not observed and PCR-based mutations were observed in only 5% of the sequences. The speed, ease and high success rates of the various PIPE protocols have tremendously increased our ability to rapidly produce large numbers of constructs using minimal resources or automation and thereby minimized the traditional bottlenecks normally associated with molecular biology in large-scale structural genomics efforts.

Combining the throughput of PIPE-generated constructs with effective protein microscreening has further enabled the evaluation and optimization of targets for structural genomics. Our results demonstrate that while having a construct that produces soluble recombinant protein is obviously necessary for crystallization success, it is not sufficient. Many soluble expression constructs produce aggregated protein, which is unsuitable for structural studies. Additional biophysical parameters, namely AnSEC scoring, must be measured to properly select, which targets to take to crystal trials to improve the efficiency within our structural genomics platform.

We have chosen to use AnSEC scoring as the method for measurement of monodispersity mainly because of the resolution and direct quantitative readout of the abundances of multiple species present as well as the ready availability of suitable HPLCs with autosampling capability and low sample consumption. It is probable that a similar correlation between crystallization success and aggregation state parameters as determined by other methods, such as light scattering, might also be established. Light scattering might actually be preferred for some specific cases. Such cases might include labs where light scattering plate readers or readers with autosamplers are available, where users have lower throughput requirements, where high degrees of accuracy as to molecular weight are imperative, or where the protein does not behave well with respect to chromatographic behavior or does not absorb at a suitable wavelength.

Both past and present results support screening proteins by AnSEC scores (only proteins with AnSEC scores of 3 and 4 have been successful), however, it is also clear that this score alone is not sufficient to predict crystallization success. The value to structural biology efforts of rapidly generating and characterizing truncations is to find several high AnSEC scoring protein variants that can be put into crystal trials simultaneously to improve the chances of individual target success. This is done empirically using the PIPE method and microscreening in a timeframe that is difficult to attain with other methods. To date, truncated structures have been obtained only for targets where the full-length proteins had optimal AnSEC scores but could not be solved, however, several nonoptimal full-length proteins now appear promising in crystallization trials as optimally scoring truncated versions (5105925, NP_840418, 2648478, etc.) as well.

Although structures were determined for 32% of the targets, which produced harvestable crystals, not all high AnSEC scoring truncations were subjected to crystal trials. Also, additional truncations outside the initial design and predicted surface entropy reduction mutants have not been attempted. However, several targets in this set have truncated constructs, which are already diffracting well and appear close to solution. With additional work, the success shown here should further increase. In this set, N-terminal truncations yielded more optimal candidates than C-terminal truncations. Future studies will examine whether extending the truncation range, changing the truncation granularity or combining N- and C-terminal truncations have additional effects. It is also possible that once enough systematic data of this sort has been generated on a standard protein set, bioinformatic predictions may be able to model a better path for defining truncations. This might further increase efficiency since, on average, only 5 of the 24 truncations resulted in soluble protein. It is evident from these data, however, that the average accuracy of current domain boundary predictions (up to 70% accurate to within 20 resi-

Table III
Crystallization and Structure Determination Results

	Full-length versions						Best truncated versions							
Target	AnSEC score	Insufficient yield	No crystals	Marginal crystals	Harvestable crystals	**Solved structure**	AnSEC score	No soluble truncation	Not determined	Insufficient yield	Marginal crystals	Harvestable crystals[a]	Solved structure	Version tested
Targets where only full-length versions yielded harvestable crystals or structures														
2633731	4					X	—	X						—
NP_880937	4					X	—		X					—
NP_811092	4					X	—		X					—
10175472	4					X	4				X			N4
NP_346487	4					X	3				X			N4
10173191	4					X	2				X			N16
13814777	3					X	—	X						—
EF3048	3					X	—		X					—
1591138	4				X		—			X				—
NP_252200	4				X		—			X				—
NP_345866	4				X		—	X						—
16411388	4				X		—	X						—
NP_249499	4				X		4				X			N4
1573367	4				X		4				X			N12
16421771	4				X		4				X			N4
10175341	3				X		—			X				—
3257464	3				X		—			X				—
6968024	3				X		—	X						—
15024582	3				X		—	X						—
NP_812667	3				X		—	X						—
NP_811426	3				X		—	X						—
16419207	3				X		4				X			C4
NP_647210	3				X		4				X			N4
15023791	3				X		2				X			C4
13815734	2				X		—			X				—
EF1792	1				X		—			X				—
NP_253523	1				X		—	X						—
Targets where only truncated versions yielded harvestable crystals or structures														
10173092	4	X					4						X	N48
2632844	4			X			4						X	C4
NP_663012	3			X			3						X	C16
5105925	1	X					3					**X**		N8
1574132	3			X			3					**X**		N28
10639230	4			X			3					**X**		N8
2648478	1	X					4					X		N40
NP_842095	1		X				2					X		N48
NP_840418	2		X				4					X		N12
NP_813331	2	X					4					X		N20
SMU.1453C	3	X					4					X		C12
10175676	4		X				3					X		N20

Table III
(Continued)

	Full-length versions						Best truncated versions							
Target	AnSEC score	Insufficient yield	No crystals	Marginal crystals	Harvestable crystals	**Solved structure**	AnSEC score	No soluble truncation	Not determined	Insufficient yield	Marginal crystals	Harvestable crystals[a]	Solved structure	Version tested
Targets where both full-length and truncated versions yielded harvestable crystals or structures														
10640715	4					X	4					**X**		N12
13815598	4					X	3					X		N32
10172812	4					X	3					X		N4
16420133	4					X	3					X		N4
2636623	4					X	4					X		N4
NP_813429	3				X		4						X	C28
2635576	4				X		3						X	N12
gil23130098	4				X		4					**X**		N12
10175617	4				X		4					X		C12
NP_251320	4				X		4					X		N8
2632978	4				X		4					X		N12
NP_842494	4				X		4					X		N4
6460258	4				X		3					X		C20
16421433	3				X		3					X		N24
10640353	3				X		4					X		N4
gil23130561	3				X		3					X		N16
NP_646210	3				X		2					X		N20

[a]Bold Xs denote constructs, where quality data sets have been collected.

dues)[27,28] is not high enough in many cases to reliably support production of soluble, monodispersed protein from only a single construct.

Clearly, the ability to rapidly create large numbers of targets and target variants and screen them for biophysical properties is a key step for structural genomics efforts to increase the overall success rate for individual targets, both in present salvage efforts and in providing datasets for better predictive methods. As demonstrated here, PIPE cloning and mutagenesis in coordination with microscreening is a highly practical and routine approach that can facilitate these efforts.

ACKNOWLEDGMENTS

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health. We thank Julie Feuerhelm, Joanna Hale, and Jessica Paulsen for generating expression and purification data, Dennis Carlton and Ylva Elias for generating analytical SEC results, Scott Brittain, and Michelle Stettler-Gill for generating LC-MS data and Eileen Ambing and Linda Okach for crystallization data. We also thank Christian Ostermeier and Roger Benoit for helpful discussions.

REFERENCES

1. Lesley SA, Kuhn P, Godzik A, Deacon AM, Mathews I, Kreusch A, Spraggon G, Klock HE, McMullan D, Shin T, Vincent J, Robb A, Brinen LS, Miller MD, McPhillips TM, Miller MA, Scheibe D, Canaves JM, Guda C, Jaroszewski L, Selby TL, Elsliger MA, Wooley J, Taylor SS, Hodgson KO, Wilson IA, Schultz PG, Stevens RC. Structural genomics of the *Thermotoga maritima* proteome implemented in a high-throughput structure determination pipeline. Proc Natl Acad Sci USA 2002;99:11664–11669.
2. Klock HE, White A, Koesema E, Lesley SA. Methods and results for semi-automated cloning using integrated robotics. J Struct Funct Genomics 2005;6:89–94.
3. Liolios K, Tavernarakis N, Hugenholtz P, Kyrpides NC. The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. Nucleic Acids Res 2006;34:D332–D334.
4. Lesley SA, Wilson IA. Protein production and crystallization at the Joint Center for Structural Genomics. J Struct Funct Genomics 2005;6:71–79.
5. McPherson A. Protein crystallization in the structural genomics era. J Struct Funct Genomics 2004;5:3–12.
6. Page R, Grzechnik SK, Canaves JM, Spraggon G, Kreusch A, Kuhn P, Stevens RC, Lesley SA. Shotgun crystallization strategy for structural genomics: an optimized two-tiered crystallization screen against the *Thermotoga maritima* proteome. Acta Crystallogr D 2003;59:1028–1037.
7. Derewenda ZS. Rational protein crystallization by mutational surface engineering. Structure 2004;4:529–535.
8. Pantazatos D, Kim JS, Klock HE, Stevens RC, Wilson IA, Lesley SA, Woods VL Jr. Rapid refinement of crystallographic protein construct definition employing enhanced hydrogen/deuterium exchange MS. Proc Natl Acad Sci USA 2004;101:751–756.
9. Spraggon G, Pantazatos D, Klock HE, Wilson IA, Woods VL Jr, Lesley SA. On the use of DXMS to produce more crystallizable proteins: structures of the *T. maritima* proteins TM0160 and TM1171. Protein Sci 2004;13:3187–3199.
10. Scharf SJ, Horn GT, Erlich HA. Direct cloning and sequence analysis of enzymatically amplified genomic sequences. Science 1986;233:1076–1078.
11. Costa GL, Grafsky A, Weiner MP. Cloning and analysis of PCR-generated DNA fragments. PCR Methods Appl 1994;6:338–345.
12. Aslanidis C, de Jong PJ. Ligation-independent cloning of PCR products (LIC-PCR). Nucleic Acids Res 1990;18:6069–6074.
13. Hsiao K. Exonuclease III induced ligase-free directional subcloning of PCR products. Nucleic Acids Res 1993;21:5528–5529.
14. Boyd AC. Turbo cloning: a fast, efficient method for cloning PCR products and other blunt-ended DNA fragments into plasmids. Nucleic Acids Res 1993;21:817–821.
15. Bubeck P, Winkler M, Bautsch W. Rapid cloning by homologous recombination in vivo. Nucleic Acids Res 1993;21:3601–3602.
16. Liu Q, Li MZ, Leibham D, Cortez D, Elledge SJ. The univector plasmid-fusion system, a method for rapid construction of recombinant DNA without restriction enzymes. Curr Biol 1998;8:1300–1309.
17. Oliner JD, Kinzler KW, Vogelstein B. *In vivo* cloning of PCR products in *E. coli*. Nucleic Acids Res 1993;21:5192–5197.
18. Hartley JL, Temple GF, Brasch MA. DNA cloning using *in vitro* site-specific recombination. Genome Res 2000;10:1788–1795.
19. Tillett D, Neilan BA. Enzyme-free cloning: a rapid method to clone PCR products independent of vector restriction enzyme sites. Nucleic Acids Res 1999;27:e26.
20. Chiu J, March PE, Lee R, Tillett D. Site-directed, Ligase-Independent Mutagenesis (SLIM): a single-tube methodology approaching 100% efficiency in 4 h. Nucleic Acids Res 2004;32:e174.
21. Li MZ, Elledge SJ. Harnessing homologous recombination *in vitro* to generate recombinant DNA via SLIC. Nat Methods 2007;3:251–256.
22. Cornvik T, Dahlroth SL, Magnusdottir A, Herman MD, Knaust R, Ekberg M, Nordlund P. Colony filtration blot: a new screening method for soluble protein expression in *Escherichia coli*. Nat Methods 2005;2:507–509.
23. Nguyen H, Martinez B, Oganesyan N, Kim R. An automated small-scale protein expression and purification screening provides beneficial information for protein production. J Struct Funct Genomics 2004;5:23–27.
24. Olsen DB, Eckstein F. Incomplete primer extension during *in vitro* DNA amplification catalyzed by Taq polymerase; exploitation for DNA sequencing. Nucleic Acids Res 1989;23:9613–9620.
25. Kirsch RD, Joly E. An improved PCR-mutagenesis strategy for two-site mutagenesis or sequence swapping between related genes. Nucleic Acids Res 1998;26:1848–1850.
26. Sawano A, Miyawaki A. Directed evolution of green fluorescent protein by a new versatile PCR strategy for site-directed and semi-random mutagenesis. Nucleic Acids Res 2000;28:e78.
27. Kong L, Ranganathan S. Delineation of modular proteins: domain boundary prediction from sequence information. Brief Bioinform 2004;5:179–192.
28. Sikder AR, Zomaya AY. Improving the performance of DomainDiscovery of protein domain boundary assignment using inter-domain linker index. BMC Bioinformatics 2006;7(Suppl 5):S6.