Combinatorial Domain Hunting: An effective approach for the identification of soluble protein domains adaptable to high-throughput applications

STEFANIE REICH,[1,5] LORETTO H. PUCKEY,[1,5] CAROLINE L. CHEETHAM,[2,5]
RICHARD HARRIS,[2] AMMAR A.E. ALI,[4] UMA BHATTACHARYYA,[4]
KATE MACLAGAN,[2] KEITH A. POWELL,[4] CHRISOSTOMOS PRODROMOU,[3,4]
LAURENCE H. PEARL,[3,4] PAUL C. DRISCOLL,[2,4] AND RENOS SAVVA[1,4]

[1]School of Crystallography, Birkbeck College, London WC1E 7HX, United Kingdom
[2]Department of Biochemistry and Molecular Biology, University College London, London WC1E 6BT, United Kingdom
[3]Section of Structural Biology, Institute of Cancer Research, Chester Beatty Laboratories, London SW3 6JB, United Kingdom
[4]Domainex Ltd., London SW7 3RP, United Kingdom

(RECEIVED January 6, 2006; FINAL REVISION June 20, 2006; ACCEPTED July 25, 2006)

Abstract

Exploitation of potential new targets for drug and vaccine development has an absolute requirement for multimilligram quantities of soluble protein. While recombinant expression of full-length proteins is frequently problematic, high-yield soluble expression of functional subconstructs is an effective alternative, so long as appropriate termini can be identified. Bioinformatics localizes domains, but doesn't predict boundaries with sufficient accuracy, so that subconstructs are typically found by trial and error. Combinatorial Domain Hunting (CDH) is a technology for discovering soluble, highly expressed constructs of target proteins. CDH combines unbiased, finely sampled gene-fragment libraries, with a screening protocol that provides ‘‘holistic’’ readout of solubility and yield for thousands of protein fragments. CDH is free of the ‘‘passenger solubilization’’ and out-of-frame translational start artifacts of fusion-protein systems, and hits are ready for scale-up expression. As a proof of principle, we applied CDH to p85α, successfully identifying soluble and highly expressed constructs encapsulating all the known globular domains, and immediately suitable for downstream applications.

Keywords: protein structure/folding; structure; new methods; expression systems

The ability to produce multimilligram quantities of a target protein in a stable and soluble form underpins modern techniques of high-throughput biochemical assays and structure-based drug development (Blundell et al. 2002; Rowlands et al. 2004). Complex multidomain human proteins that constitute many targets present particular problems for expression in simple recombinant systems such as *Escherichia coli*, and soluble expression of full-length gene products is often impossible. In contrast, subconstructs of the target gene can often be expressed to give soluble protein at good yield, so long as the subconstruct encodes a segment of the protein that is capable of folding to a thermodynamically stable three-dimensional structure. Identification of such constructs, which frequently encapsulate one or more domains, commonly uses bioinformatics analysis of the

[5]These authors contributed equally to this work.

Reprint requests to: Laurence H. Pearl, Section of Structural Biology, Institute of Cancer Research, Chester Beatty Laboratories, 237 Fulham Road, London SW3 6JB, UK; e-mail: laurence.pearl@icr.ac.uk; fax: 44-20-7153-5457.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.062082606.

protein sequence, and/or partial proteolytic digestion of the full-length protein or at least a larger construct. Bioinformatics can be effective in localizing domains within the overall sequence and defining conserved core residues, but is poor at predicting boundaries and identifying globular structures formed by multiple domains. For example, knowledge of the true boundaries of the structured domains of TERT protein (Jacobs et al. 2006) and Sir3 (King et al. 2006) has remained elusive until recently, even though their sequences have been known for some time. Precise definition of boundaries is very important, and experience shows that variation by two or three residues can significantly alter the behavior of the protein product—underestimates can lead to burial of a charged N or C terminus, giving an unstable construct, while overestimates generate disordered additional segments that may promote aggregation or prevent crystallization. Furthermore, bioinformatics cannot predict if a defined domain will be expressed in a soluble form in multimilligram amounts. Limited proteolysis is a technique with a good track record in structural biology that relies on folded segments being less accessible to a protease, such as trypsin, than the ''linkers'' that connect them. Thus, readily cleaved sites define the boundaries of folded regions, although accessible loops within folded regions can give misleading results. However, it is an experimental technique with an absolute requirement for some folded soluble protein at the outset, and cannot be applied ab initio. Clearly, a method identifying clones expressing soluble domains from a higher throughput, low-information screen to a lower throughput, high-information screen, while eliminating false positives, is desirable. Diverse attempts, which have been successful, have recently been made to address this (Cabantous et al. 2005a; Cornvik et al. 2005; Jacobs et al. 2006; King et al. 2006). Experience over two decades of the problem of producing protein for structural study has led us to develop a technique that directly addresses the problem of identifying constructs, from a library of DNA fragments, that express soluble, stable protein that is produced at multimilligram levels in *Escherichia coli*. We have developed a combinatorial approach, which generates a random library of contiguous fragments of the target gene in a one-pot reaction, with a defined fragment size distribution and random positional distribution over the parent DNA sequence. We have combined this approach with a holistic screen that identifies stable, soluble, and highly expressed protein segments, free from false positives introduced by ''passenger solubilization'' with fusion proteins, and amenable to scale-up production without further genetic manipulation. Here we report a proof-of-principle study in which this combinatorial domain hunting (CDH) method is applied to a multidomain target protein, the p85α subunit of class 1_A phosphoinositide 3-kinase. Work over many years by ourselves and others has defined the domain architecture of this protein empirically, and elucidated three-dimensional structures for most of its folded regions (Booker et al. 1992, 1993; Liang et al. 1996; Musacchio et al. 1996; Nolte et al. 1996; Siegal et al. 1998; Hoedemaeker et al. 1999). In contrast, it took CDH only months to successfully identify stable, soluble, and highly expressed protein segments encapsulating the known globular BCR, N-SH2, and C-SH2 domains individually, in addition to a new construct expressing the tandem SH3-BCR segment. We show CDH to be a rapid and effective method applicable ab initio to discovery and production of highly expressed soluble constructs from protein targets.

Results and Discussion

Generation of gene fragment library

The first stage in the CDH methodology requires generation of a fragment library of the target gene (Fig. 1A).

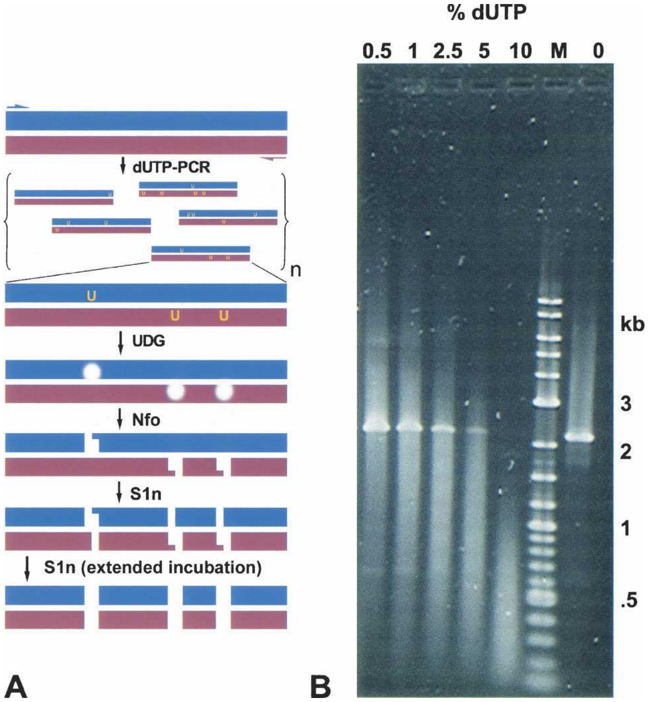

Figure 1. Gene fragmentation. (*A*) Schematic of the CDH gene fragmentation process. PCR with TTP/dUTP mixtures is used to generate copies of the target gene in which uracil is randomly incorporated in place of thymine. The uracil-doped amplified DNA is subjected to a modified base-excision cascade in which uracil-DNA glycosylase excises the uracil bases generating abasic sites, which are cleaved by endonuclease IV, giving a single-strand nick that is converted to a double-strand break and blunt-ended by S1 nuclease. As the reaction cascade is initiated only at uracils, whose distribution along the sequence and among the PCR reaction products is random, the cascade generates a random and unbiased library of gene fragments, whose size distribution is solely dictated by the TTP/dUTP ratio. (*B*) dUTP-dose dependent fragmentation. SYBR-Safe stained 1% agarose gel of an ~2.2-kb human p85α PCR-amplified cDNA (*right-hand* lane), alongside the products of CDH fragmentation reactions using increasing amounts of dUTP (as percent of total TTP+dUTP concentration). The progressive decrease in modal size of the DNA distribution with increasing dUTP concentration is clearly seen.

Although in this study wild-type human p85α DNA was used, resynthesis of the gene is desirable as this has several advantages. Firstly, it optimizes the DNA sequence for expression in the target host, and secondly, it can be used to disrupt G:C islands to ensure that a more even fragmentation of the DNA is observed, thus ensuring that all domains can be captured. To achieve fragmentation, we first amplified the target gene by PCR in which dUTP was included at 1% of the TTP concentration. The dUTP/TTP ratio determines the size distribution of the fragments generated in the subsequent reactions, and an optimal range for a desired modal size can be reliably estimated on the basis of the length of the gene. The purified PCR product is then exposed to a modified base excision pathway consisting of uracil-DNA glycosylase (UDG), endonuclease IV (Nfo), and S1 nuclease (S1n). The consecutive action of these three enzymes generates a double-strand break at each point where a uracil was present on either strand. The probability of uracil incorporation at any site in any cycle is entirely a function of the dUTP/TTP ratio used in the PCR reaction, and the initiation of the reaction cascade by UDG proceeds with very high efficiency wherever a uracil is present, regardless of local sequence. Furthermore, uracil, unlike other noncanonical bases such as oxyanine (Hitchcock et al. 2004), maintains authentic Watson-Crick base-pairing so that its incorporation is nonmutagenic. Given pure enzymes free of nonspecific nuclease activity, the reactions can be run to completion without need for time courses or titrations, and with the outcome entirely dictated by the dUTP/TTP ratio (Fig. 1B).

The products of the target gene fragmentation reaction are directly ‘‘captured’’ using the pCR-Blunt-TOPO ligase-free cloning system (Invitrogen). Although the fragments produced by the UDG/Nfo/S1n cascade can be ligated into general blunt-cut vectors by conventional ligase reactions, the very high efficiency and very low background of the topoisomerase-modified vectors is a facile way of capturing the fragment library generated. For the proof-of-principle study reported here, the standard nonexpressing version of the pCR-Blunt-II-TOPO vector was used, and inserts transferred to the pDXV3 (see Materials and Methods) series of expression vectors (Domainex Ltd, UK) as EcoRI fragments. The pDXV vector series provides three translation starts in three different reading frames, each with C-terminal ‘‘tags’’ and stop codons in three different reading frames. Although the use of multiple vectors does not increase the probability of inserts in the correct orientation and in frame per se (1/18), it guarantees that every generated fragment can be captured in frame, no matter where the initial point of DNA fragmentation. So even DNA fragments generated from a double-strand break of a rare A:T base pair inside an G:C island can be captured, whereas, with only one cloning vector it could be lost if it happens to be out of frame.

During development of the method and in this proof-of-principle study, we have sequenced selections of clones from libraries generated for several genes. While sequencing of a sufficient number of inserts to achieve formal statistical significance would be prohibitively expensive, the data we have obtained suggest that the fragmentation process is, indeed, generating the size distribution and sequence ‘‘cover’’ consistent with unbiased uracil incorporation and consequent fragmentation (Fig. 2).

Development of a holistic solubility screen

The yield, stability, and solubility of a given protein construct expressed in a bacterial cell depend on several interacting factors, including the stability of the mRNA, the processivity with which it is translated, the susceptibility of the nascent polypeptide product to aggregation, and the stability of the folded product. Some of these factors can be improved by, for example, recoding the target gene to give optimal codon usage and minimal mRNA secondary structure (Prodromou and Pearl 1992; Wheeler et al. 1996; Jaffe et al. 2000; Hamdan et al. 2002). Alternatively, the protein can be expressed intentionally in an insoluble state and then attempts made to refold it in vitro (Cabrita and Bottomley 2004). In all cases, success depends on the actual stability of the protein segment being expressed. If it cannot adopt a folded globular state in which uncompensated hydrophobic exposure and polar burial are minimized, then it will not be soluble in vivo, regardless of the expression system, nor will it be amenable to refolding in vitro. In designing our screening protocol, we have sought to identify those constructs within a gene fragment library that expresses protein segments, simultaneously satisfying the requirements of yield, stability, and solubility, sufficiently well to allow scale-up to the levels required for structural studies.

Our screening protocol for this proof-of-principle study operates in several stages (Fig. 3), and we were concerned with following the distribution of fragments and their behavior at all stages. Consequently, we acquired data at many levels to allow us to determine the possible origins of any false positives and negatives that might arise. We have therefore used a more laborious version than would be adopted by an optimized high-throughput protocol, as follows. Firstly, clones from the fragment library were analyzed at the DNA level by restriction digestion to determine whether or not insertion of a fragment had taken place. Vectors containing inserts were then transformed into an expression strain and a dot-blot analysis with an anti-‘‘tag’’ antibody used to detect clones expressing tagged protein.

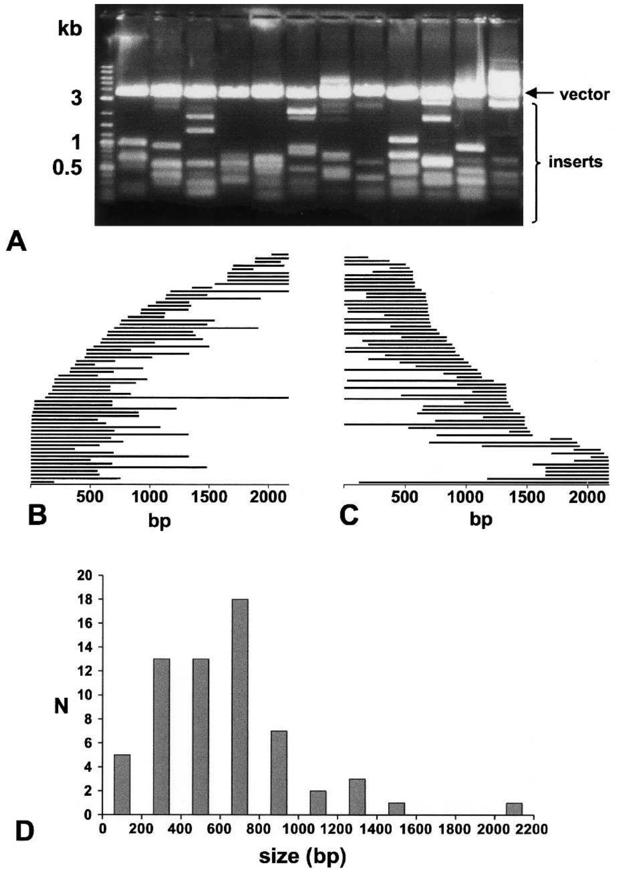

Figure 2. Fragment library distribution. (*A*) Fragment size distribution is unbiased. SYBR-Safe stained 1% agarose gel of 144 individual clones, generated by shotgun capture of the fragmentation reaction in the ligase-free cloning vector pCR-Blunt-II TOPO (Invitrogen). Clones were pooled in lots of 12 and miniprepped, and captured DNA inserts were released as EcoRI fragments, with 12 vector-derived bases still attached to each end. The distribution of fragment sizes populates the desired range 0.1–1.0 kb. (*B*) The fragment position is random. Coverage plot of 63 randomly selected and sequenced clones (black lines) from the p85α fragment library, ordered according to their 5′-end (*bottom* to *top*), arrayed against the 2175-bp sequence of human p85α. Apart from clones beginning at the actual 5′-end of the target gene, the start positions of the fragments are evenly distributed across the target gene, which is fully sampled. Although the sample size is far too small for statistical significance, it is fully consistent with random and unbiased fragmentation. (*C*) As *B*, but with the data sorted by 3′-end position. (*D*) Histogram of fragment size frequency (N). Fragment sizes are binned in intervals of 200 bp. Although the sample size is too small for statistical significance, the distribution is consistent with the expected Poisson distribution for a random fragmentation process.

The tag we have chosen consists of a minimally short peptide fused in-frame to the C-terminal end of the screened fragment, which is used for detection by antibodies and for reporting the foldedness of the construct. In our screen, the tag is kept as small as possible so that its presence has a minimal effect on the properties of the protein segment to which it is attached, thereby minimizing the potential for passenger solubilization effects that accompany the use of fusion proteins. Previous approaches to high-throughput solubility screens (Maxwell et al. 1999; Waldo et al. 1999; Pedelacq et al. 2002; Nakayama and Ohara 2003) have generally used expression systems in which the screened fragment is fused to a folded "carrier" protein, variously employed as a reporter (green fluorescent protein), a selective marker (chloramphenicol acyltransferase, kanamycin phosphotransferase), and/or affinity ligand (maltose binding protein, glutathione-*S*-transferase). While these systems have various strengths and have had some success, they can on occasion suffer in that the solubility, stability, and yield

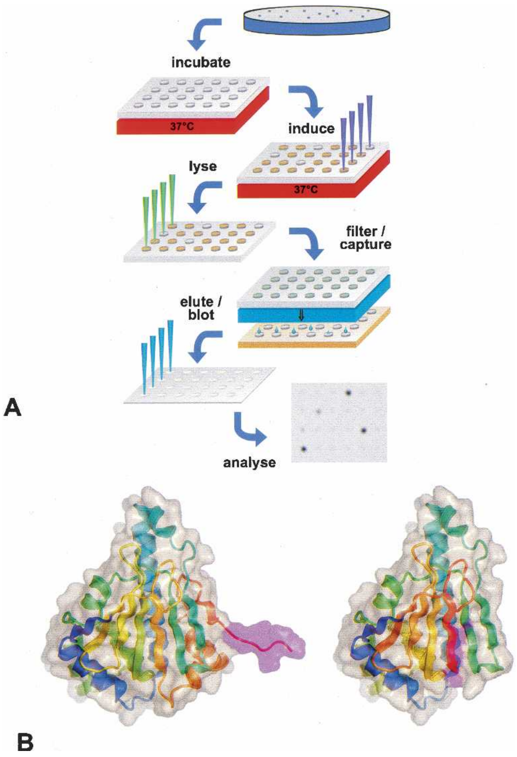

Figure 3. Solubility screening. (*A*) Schematic of CDH screening process. Colonies arrayed on membranes that react with an anti-tag antibody are picked and individually inoculated into small-scale liquid cultures in multiwell dishes, incubated under standard conditions and expression-induced. Gentle nondenaturing lysis releases cytoplasmic proteins that pass through a hydrophobic filter, and over an affinity resin for the attached tag. Eluates from the affinity resin are blotted onto membranes and detected using anti-tag antibody. Tagged protein that is abundant in the cytoplasm, soluble and nonaggregated, and properly folded is substantially enriched by this process and gives rise to strong signals in the dot blot. (*B*) Principle of "tag-availability." When a peptide tag is appended to the C terminus of a hypothetical target protein construct that encapsulates a folded globular region (*left*), the tag (magenta surface) is fully exposed and available for interaction with affinity resins. When the construct is too short (*right*), the tag becomes embroiled in the core of the protein and is unavailable to affinity resins. Even where a tag is appropriately positioned relative to the domain termini, aggregation and misfolding decrease the availability of the tag favoring retention of "good" constructs over "bad."

that are detected are properties not of the target protein fragment in isolation, but of the fusion with the carrier protein (Nakayama and Ohara 2003). One practical outcome of this is that inherently unstable or unfolded protein segments can become significantly stabilized and/or solubilized as ‘‘passengers’’ of the fusion protein, and give rise to false positives (Nakayama and Ohara 2003). Once separated from their fusion partner, however, and expressed as the isolated segment, they display their inherent properties and are revealed as negatives. In another study using green fluorescent protein as a fusion partner (Kawasaki and Inagaki 2001), out-of-frame spurious translation products up to 90 residues long were identified as strong positives. However, recently this problem has been addressed (Cabantous et al. 2005a). In contrast, in our screen such small fragments are not stabilized by a large fusion and do not appear as false positives.

At the first level of our screen, clones expressing a fragment in-frame from the start codon through to the C-terminal tag, regardless of solubility or stability, are readily detected by anti-tag antibody in a colony blot or dot blot. Inclusion of defined negative and positive control samples allows the significance of experimental antibody reactivity to be determined, allowing downstream processing to be restricted only to those with ‘‘strong’’ signals if desired. Central to the efficacy of the screen is the ability to discriminate between proteins that possess the desired properties of solubility, stability, and yield, and those that don’t. The mere presence of a detectable ‘‘tag’’ in the first-level screen gives an indication of yield and in vivo stability of a particular construct within the expressing bacterium, but gives no indication of its solubility. To determine this, we use a second-stage screen in which cultures of positive colonies from the first stage are lysed, and the supernatant passed through a hydrophobic filter to remove cell debris and aggregated material, and then over an affinity resin specific to the peptide tag appended to all constructs. Although the tag is present on all constructs at this stage, we have observed that the ability of a tagged construct to be retained on an affinity matrix is strongly influenced by the suitability of that construct. Thus, where a protein construct is highly aggregated or misfolded, the tag becomes buried and unavailable for interaction with the affinity matrix. A similar concept underlies structural complementation methods in which the ability of a peptide tag to bind to and functionally reconstitute a coexpressed reported protein is used to indicate the foldedness of the tagged protein construct (Wigley et al. 2001; Cabantous et al. 2005b). However, in our approach the availability of the tag is determined after release from the supportive cellular milieu, so that the protein is assessed on its own merits and the possibility of passenger solubilization by complexation with a reporter protein is eliminated. We find that this property of ‘‘tag-availability,’’ in combination with filtration, provides an effective holistic discriminator in favor of soluble, stable, and nonaggregated protein constructs amenable to at least affinity purification. Key to the efficacy of this second screen step is the use of a gentle enzymatic lysis process, whereby the cytoplasm of the bacterial cells is sampled, rather than solubilized. Aggressive lysis procedures involving sonication, mechanical disruption, or detergent resuspend a great deal of otherwise insoluble tagged material, which then binds to the affinity matrix regardless.

Application to p85α and results

Although CDH was developed for application to targets that lack structural data, for our proof-of-principle study, we applied it to a very well-studied target protein, the p85α regulatory subunit of class 1_A phosphoinositide 3-kinase (Otsu et al. 1991; Skolnik et al. 1991). Work over many years by ourselves and others (Booker et al. 1992, 1993; Liang et al. 1996; Musacchio et al. 1996; Nolte et al. 1996; Siegal et al. 1998; Hoedemaeker et al. 1999) has defined the domain architecture of this protein empirically, and elucidated three-dimensional structures for most of its folded regions, making p85α an ideal benchmark for testing CDH.

A cDNA for human p85α was PCR-amplified and fragmented using the UDG/Nfo/S1n system with a 100:1 TTP:dUTP ratio, and the resulting fragment library captured in pCR-Blunt-II TOPO (see Materials and Methods). For expression screening, the library was transferred to the pDXV3 vector series as EcoRI fragments (see Materials and Methods); 1404 clones with inserts were picked, grown in liquid culture, and lysed using a gentle enzymatic protocol (see Materials and Methods). In-frame protein expression was determined in ‘‘dot blots’’ with an antibody to the C-terminal His_5 tag appended onto expressed p85α fragments by the pDXV3 vectors (see Materials and Methods), and 191 clones gave signals sufficiently above background to warrant further analysis. Of these, 109 showed high or medium strength signals in a ‘‘dot blot’’ after the second stage, which selects against aggregated, insoluble, or misfolded protein (Fig. 4A). Ni-IMAC-eluates from these were subjected to SDS-PAGE and analyzed by immunoblot directed against the C-terminal tag (Fig. 4B), with clear, strong bands observed for 41 clones. Inserts from all clones yielding a high-level immunoblot were sequenced to allow determination of their location within the overall p85α cDNA. Sixteen of these clones also gave strong bands of corresponding molecular weight on Coomassie-stained gels, suggesting that they were producing protein at the levels required for structural studies, and were designated ‘‘hits’’ (Fig. 4C). These Coomassie-positive

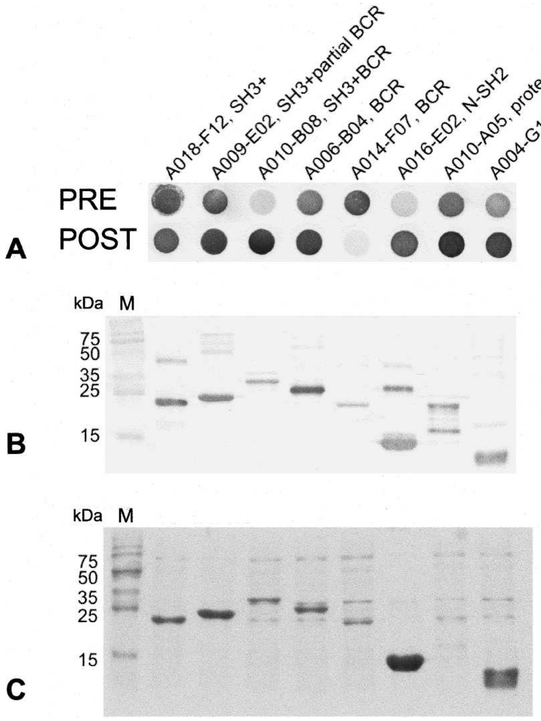

Figure 4. p85α “hits.” (*A*) Dot blots for eight clones that were taken through to preliminary structural assessment by ^{1}H-NMR. Pre-screen blots indicate reactive protein levels prior to any filtration or affinity enrichment that is sensitive to tag-availability. Post-screen blots indicate levels of folded, soluble, nonaggregated protein. A decrease in signal (as in A014-F07) suggests that this construct expresses at high levels, but is not as efficiently released from the cytoplasm as other constructs. Nonetheless, it produces sufficient protein for structural studies. (*B*) Western blots of SDS-PAGE gel of protein eluted from the final stage of the screen for eight clones taken through to preliminary structural assessment by ^{1}H-NMR. Consistent with the dot blots, all samples show bands that are immunoreactive to anti-tag antibody, and indicate that “hits” are in the expected size range for the experiment. One clone (A010-A05) shows clear evidence of proteolytic breakdown from the genetically predicted protein size. (*C*) As *B* but Coomassie-stained to indicate total protein. All clones show good correlation between the immunoreactive bands in the Western blots (*B*) and the major protein bands in this gel. In most cases, the level of the target band is substantially higher than other protein bands and should readily purify with one or two more steps to a suitable degree for detailed structural analysis by NMR or X-ray crystallography.

clones were grown up on a larger scale (1 L), lysed by sonication after an enzymatic incubation, clarified by centrifugation, and subjected to a single step of purification using a *Proteus* IMAC Mini spin-column. Fourteen clones produced sufficient semipure protein on scale-up to allow further analysis, and the eight purest clones were taken through to 1D ^{1}H-NMR spectroscopy (Fig. 5). Chemical shift dispersion in 1D ^{1}H-NMR spectra is a strong indicator of the presence of structured globular protein, and is an effective method we and others (Rehm et al. 2002; Page et al. 2005) have used to determine the foldedness or otherwise of protein constructs. Out of eight clones analyzed, seven gave spectra consistent with a substantially folded structure, and one gave a spectrum indicating an absence of ordered globular structure.

Gratifyingly, the distribution of the soluble folded hits across the p85α sequence corresponded very well with the known positions of domains (Fig. 6), and provided constructs suitable for determination of the individual domain structures were they not already known. For

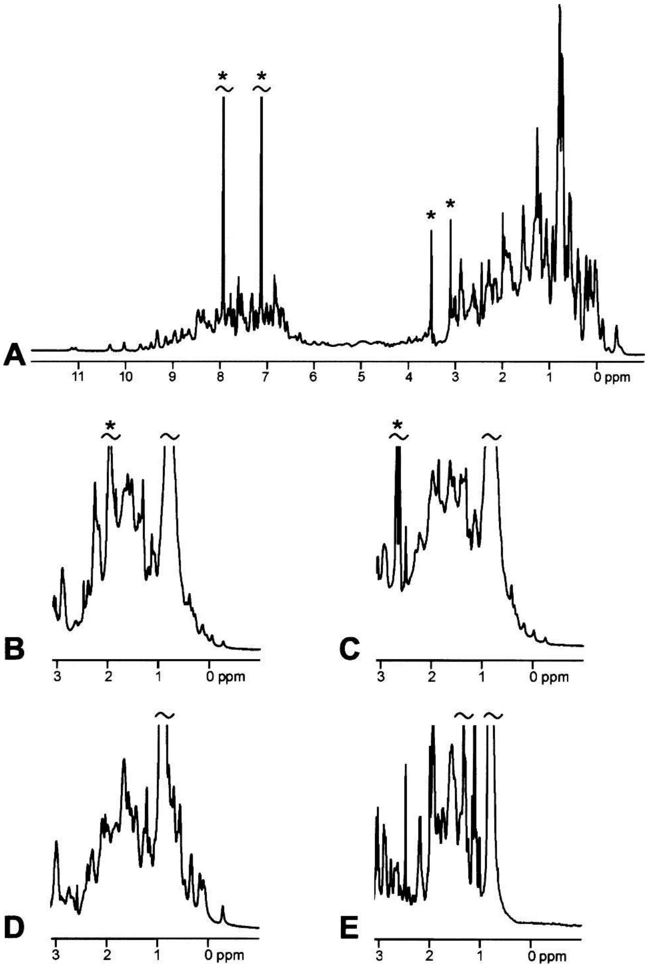

Figure 5. ^{1}H-NMR spectra of p85α “hits.” (*A*) ^{1}H-NMR spectrum of protein expressed by clone A016-E02 (corresponding to the N-SH2 domain). Clear resonances below 0 ppm arise from upfield-shifted methyl groups, which are strongly indicative of globular structure. Tildes (~) represent signals truncated for clarity, and asterisks (*) indicate sharp signals from buffer components. (*B*) Upfield methyl group region from clone A010-B08, corresponding to the tandem SH3-BCR domain pair. (*C*) As *B*, but for clone A014-F07 corresponding to the BCR domain. (*D*) As *B*, but for clone A010-A05 corresponding to the C-SH2 domain. (*E*) As *B*, but for clone A004-G10 corresponding to a short segment of polypeptide containing the low-sequence complexity linker between the BCR and N-SH2 domains and a fragment of the N-SH2 domain. The absence of upfield-shifted signals with chemical shifts <0.8 ppm indicates a non-globular piece of protein representing a rare false positive from CDH.

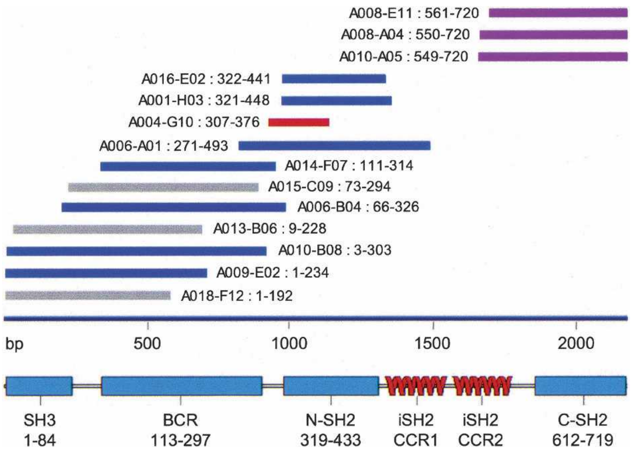

Figure 6. Coverage of p85α CDH "hits." Bars indicate the positions of the 14 final "hits" relative to the p85α protein sequence and known domain structure. These 14 clones gave pre- and post-screen dot blots significantly above background, gave immunoreactive bands in Western blots that correlated with strong protein bands in Coomassie-stained gels, and produced good levels of protein in one simple scale-up from the small-scale parallel growth conditions used in the screens. Constructs in blue have been shown by NMR to encode folded globular protein; those in magenta also give NMR spectra consistent with folded globular structures, but display some proteolysis in gels, suggesting that they contain poorly ordered but nonaggregating termini attached to a folded core. Constructs in gray have not been further characterized. Nearly all of the constructs shown (except the one unfolded construct, red) would be suitable for structural studies of p85α component domains and/or screening assays for small-molecule ligands.

example, two hits, A016-E02 and A001-H03 (amino acid residues 322–441 and 321–448, respectively), were obtained for the central N-SH2 domain, which closely encapsulate the domain boundaries used in NMR and crystal structure analyses of that domain (NMR, residues 314–431; crystallography, 321–440) (Booker et al. 1992; Nolte et al. 1996). Similarly, two hits, A006-B04 and A014-F07 (residues 66–326 and 111–314 respectively), encapsulate the construct used in crystal structure determination of the BCR domain (residues 105–319) (Musacchio et al. 1996). Although no hit was obtained in this sampling of the fragment library for the small N-terminal SH3 domain in isolation, a highly expressed soluble hit was obtained for a larger construct, A010-B08 (3–303), encapsulating the tandem SH3 and BCR domains, whose structure in combination has not yet been described. Three hits were obtained—A008-E11, A008-H04, and A010-A05 (residues 561–720, 550–720, and 549–720, respectively)—that extend nearly to the C terminus (residue 724), encapsulating the C-SH2 domain with varying amounts of the predicted inter-SH2 coiled-coil region. These constructs, whose N termini are longer than that used in previous structural studies (residues 614–720) (Hoedemaeker et al. 1999), show a degree of N-terminal proteolysis, but give excellent ^{1}H-NMR spectra and display better behavior in terms of solubility and aggregation than the original constructs used in structural studies, suggesting that those may have been suboptimal.

One highly expressed hit (A004-G10) nonetheless gave an NMR spectrum indicating a substantial lack of folded structure. This construct corresponds to a short segment of polypeptide (residues 307–376) running from the end of the BCR domain into the first third of the N-SH2 domain and incorporating the BCR – N-SH2 "linker" region. Interdomain "linkers" are commonly natively unfolded and flexible, and are by their nature polar and hence soluble. In a screening process intended to find folded globular segments, this hit must be formally considered a false positive. While the short length of this type of linker segment allows their representation in the fragment library to be substantially reduced by applying a size cut-off during gel-purification of the products of the DNA fragmentation reaction, fragments of this and much shorter size have been a source of false positives, resulting from out-of-frame translation, in fusion systems where they are stabilized by association with the larger globular protein, and not subject to the degradation of short polypeptides that occurs in the *E. coli* cytoplasm. However, this has recently been addressed (Cabantous et al. 2005a,b). The use of a minimal tag in our system provides no such stabilization and serves to minimize the occurrence of such false positives. Interestingly, the number of clones expressing in-frame DNA fragments that we obtained was higher (191 clones) than the theoretically expected number (78 clones; 1404 clones with a 1/18 chance of being in-frame). It is difficult to rationalize why this is the case, and we can only speculate as to the reasons. Often we have observed that DNA fragments prefer a particular orientation during nondirected cloning experiments. This bias may result because the growth of *E. coli* harboring expression plasmids with out-of-frame and/or inserts in a reverse orientation may in some way be suppressed. This suppression may result from deleterious effects of the translated products or their unusual demand for rare codons. Ongoing work will show how other proteins behave in this respect. However, the number of positive clones decreases substantially in subsequent stages of our screen such that we obtain a pool of clones that represent in-frame translated p85α DNA fragments expressing protein in multimilligram amounts.

Modifications toward a high-throughput protocol

The proof-of-principle study was concerned with closely following the distribution and behavior of the gene fragments at all stages. However, a few modifications would adapt the screen to a high-throughput procedure. We propose that the fragments be cloned into proprietary TOPO-charged expression vectors (pDXV4) that directly express the captured fragment. Transformants are then arrayed, using a colony-picking robot, onto nitrocellulose membranes, and clones expressing ‘‘tagged’’ protein are identified by standard colony-blotting using an anti-‘‘tag’’ antibody. This would complete the high-throughput screen, and only positive clones are then analyzed further to identify those expressing at multimilligram quantities in a soluble form. Characterization could include DNA sequencing, protein purification, mass spectrometry, and 1D-NMR to determine their folded state. A recent in-house project with human Hsp90-β using such a high-throughput screen showed that the expected domains are identified and that false positives are, indeed, very rare (data not shown), indicating that the screen effectively works as we propose.

Conclusions

We have developed an effective means for the production of stable, soluble, and highly expressed segments of protein encoded by a target gene, in a high-throughput and semiautomated process. The modified base-excision cascade delivers predictable and positionally unbiased fragmentation of target genes without the need for case-by-case titration and/or time-course experiment. The implementation of the minimal tag-availability screen provides effective detection and selection of soluble constructs, free from the high levels of false positives generated by passenger solubilization observed occasionally in some fusion-protein systems. We have clearly demonstrated the efficacy of the Combinatorial Domain Hunting approach on the p85α subunit of phosphatidylinositol-3-kinase, rapidly generating ‘‘structure-friendly’’ expression constructs that encapsulate the known globular domain structure of the protein within a time frame of a few months rather than years as was achieved by more conventional means.

While this manuscript was in preparation, recent advances in colony screening and reduction of false positives have been published (Cabantous et al. 2005a; Cornvik et al. 2005) that could also be integrated within our screen to make it more effective.

Materials and methods

Construction of pDXV3 vectors

The pDXV3 vector series is designed to allow expression of all DNA fragments generated from the target gene, with a short peptide tag (e.g., His$_5$) added to the C terminus of the encoded protein segment. As fragmentation does not preserve the reading frame, variants are required both upstream and downstream of the plasmid-encoded start codon in order to restore all possible forward reading frames. pDXV3 vectors were constructed by replacing the NdeI–HindIII segment of the multiple cloning site of pRSET T7 (Invitrogen), with overlapping oligonucleotides based on the following core sequence:

CATATG**X**CAATTGCAGCTG**X**CACCATCACCATCACTGATTGAATAAGCTT

where **X** represents frameshift positions. At these positions, all combinations of (1) no additional nucleotide, (2) cytosine mononucleotide, or (3) cytosine-guanine dinucleotide were represented, resulting in nine variants. The pDXV3 cloning site thus provides (1) a start codon embedded in an NdeI restriction site, (2) a 5′ frame-correction sequence, immediately upstream of (3) restriction sites for cohesive-ended (MfeI) or blunt-ended (PvuII) cloning of fragments, (4) a 3′ frame-correction sequence, followed by (5) a peptide tag-encoding sequence, followed by (6) stop codons in all forward reading frames, the last of which is part of a HindIII restriction site.

Human p85α gene cloning and fragmentation

A full-length native coding sequence for human p85α, fully consistent with the GenBank deposition NM_181523, was assembled from DNA sequences generated by standard PCR with Taq polymerase from a human brain cDNA pool (Invitrogen). This sequence-verified cDNA provided the template for a further PCR reaction with a modified dNTP pool containing dATP, dCTP, dGTP as normal, and a mixture of TTP and dUTP in a ratio of 100/1. Amplified DNA was agarose gel-purified, spectrophotometrically quantitated, and incubated in restriction buffer 3 (NEB) with a cocktail of *E. coli* uracil-DNA glycosylase (UDG–NEB), *E. coli* endonuclease IV, S1-nuclease (Invitrogen), and calf intestinal phosphatase (NEB) at 37°C for 16 h. The resultant DNA fragment pool was purified using the Min-Elute kit (QIAGEN), size-selected by excision from agarose gels, and requantitated prior to capture in pCR-Blunt-II TOPO (Invitrogen). Colonies were picked and transferred to 96-well blocks for growth, subsequently miniprepped in pools of 12, and digested with EcoRI (NEB). Excised fragments were agarose gel-purified and pooled and ligated with a pool of the nine pDXV3 vectors digested with MfeI (NEB) using T4-DNA ligase (Promega), and the mixture was used to transform TOP10 Chemically Competent *E. coli* (Invitrogen). Resultant clones were miniprepped, and plasmid DNA was cut with NdeI/HindIII (NEB) and run on agarose gels to verify successful insertion.

Fragment library screening

For screening, plasmids with inserts were used to transform *E. coli* BLR(DE3) cells, and picked colonies were grown in 24-well blocks containing LB media supplemented with Overnight Express Autoinduction System 1 (Novagen) at 37°C for 12 h. Aliquots from each well were aggregated into 96-well blocks and lysed with 2 μg/mL RNaseA (AbGene), 0.6 μg/mL DNase I (Roche), and 2.5 μg/mL lysozyme (Sigma) at 30°C for 1 h.

To determine ‘‘in-frame’’ expression, lysates were ‘‘dotted’’ onto a Protran nitrocellulose membrane (Schleicher & Schuell), which was then probed with Anti-His$_6$ mAb (BD Biosciences) and developed with anti-mouse IgG-AP Conjugate (Promega) and BCIP/NBT (Sigma). Dark spots on the membrane were registered as positive expression hits.

A second aliquot of each culture was transferred to a new 96-well block, which was centrifuged, and the pellets were stored at −20°C for later DNA analysis. The remaining cultures were spun down, and the supernatants were discarded.

To determine soluble expression, lysates were subjected to filtration and affinity purification using the Ni-NTA Superflow 96 BioRobot Kit (QIAGEN) on a BioRobot 8000 (QIAGEN), and the eluate was dotted onto nitrocellulose membrane for immunodetection, as above. Dark spots on the membrane were registered as potential soluble hits. Aliquots of these samples were run on each of two SDS-PAGE gels—one stained with Coomassie Brilliant Blue R (Sigma) and the other blotted onto nitrocellulose membrane for immunodetection as described above. Clones presenting visible and correlated bands in both detection modes were registered as positive soluble hits. The gene fragmentation and screening process that comprises CDH is the subject of the published patent application WO 03/040391.

Verification of soluble fragment identities and protein quality assessment

Plasmid DNA was prepared from the stored pellets for clones identified as soluble hits, and the DNA sequence of the insert was determined. The protein samples from the soluble hits were analyzed by peptide mass spectrometry to verify the size and composition predicted from the fragment DNA sequence. To determine the foldedness of the expressed protein segments, *E. coli* BLR(DE3) cells were transformed with plasmid from soluble hits and grown in 1 L of LB media to an OD of ~0.8. The autoinduced cells were lysed by enzymatic incubation followed by sonication, the lysate was partitioned at 45,000*g*, and the soluble fraction was purified on a *Proteus* Ni-NTA Mini spin-column (Generon). Eluted protein with no further purification was buffer-exchanged into a low salt buffer (50 mM potassium phosphate at pH 8, 50 mM sodium chloride, 1 mM dithiothreitol, 1 mM ethylenediaminetetraacetic acid) for ^{1}H-NMR spectroscopy. One-dimensional ^{1}H-NMR spectra were obtained at 25°C using either a 500 MHz or 600 MHz Varian NMR spectrometer equipped with a 5-mm room temperature triple resonance probehead with *Z*-axis pulse field gradient capability. Typical acquisition parameters were: 1.5 sec relaxation delay; 128–256 transients of 4 K complex points; 0.4 sec acquisition time. Solvent suppression was achieved with the WATERGATE pulse sequence element (Piotto et al. 1992). Foldedness was determined by qualitative comparison of spectra with those previously obtained for known folded globular proteins and for natively unfolded proteins (Rehm et al. 2002; Page et al. 2005).

Acknowledgments

We thank Stephane Mery and The Bloomsbury Bioseed Fund for their faith and financial support, and Adeel Mustaq and David Knight for advice and useful discussion. We acknowledge early contributions to the development of CDH from Mark McAlister, and the BBSRC-supported Bloomsbury Centre for Structural Biology. This work was supported by a grant from the BBSRC in the Exploiting Genomics Initiative.

References

Blundell, T.L., Jhoti, H., and Abell, C. 2002. High-throughput crystallography for lead discovery in drug design. *Nat. Rev. Drug Discov.* **1:** 45–54.

Booker, G.W., Breeze, A.L., Downing, A.K., Panayotou, G., Gout, I., Waterfield, M.D., and Campbell, I.D. 1992. Structure of an SH2 domain of the p85 α subunit of phosphatidylinositol-3-OH kinase. *Nature* **358:** 684–687.

Booker, G.W., Gout, I., Downing, A.K., Driscoll, P.C., Boyd, J., Waterfield, M.D., and Campbell, I.D. 1993. Solution structure and ligand-binding site of the SH3 domain of the p85 α subunit of phosphatidylinositol 3-kinase. *Cell* **73:** 813–822.

Cabantous, S., Pedelacq, J.D., Mark, B.L., Naranjo, C., Terwilliger, T.C., and Waldo, G.S. 2005a. Recent advances in GFP folding reporter and split-GFP solubility reporter technologies. Application to improving the folding and solubility of recalcitrant proteins from *Mycobacterium tuberculosis*. *J. Struct. Funct. Genomics* **6:** 113–119.

Cabantous, S., Terwilliger, T.C., and Waldo, G.S. 2005b. Protein tagging and detection with engineered self-assembling fragments of green fluorescent protein. *Nat. Biotechnol.* **23:** 102–107.

Cabrita, L.D. and Bottomley, S.P. 2004. Protein expression and refolding—A practical guide to getting the most out of inclusion bodies. *Biotechnol. Annu. Rev.* **10:** 31–50.

Cornvik, T., Dahlroth, S.L., Magnusdottir, A., Herman, M.D., Knaust, R., Ekberg, M., and Nordlund, P. 2005. Colony filtration blot: A new screening method for soluble protein expression in *Escherichia coli*. *Nat. Methods* **2:** 507–509.

Hamdan, F.F., Mousa, A., and Ribeiro, P. 2002. Codon optimization improves heterologous expression of a *Schistosoma mansoni* cDNA in HEK293 cells. *Parasitol. Res.* **88:** 583–586.

Hitchcock, T.M., Gao, H., and Cao, W. 2004. Cleavage of deoxyoxanosine-containing oligodeoxyribonucleotides by bacterial endonuclease V. *Nucleic Acids Res.* **32:** 4071–4080.

Hoedemaeker, F.J., Siegal, G., Roe, S.M., Driscoll, P.C., and Abrahams, J.P. 1999. Crystal structure of the C-terminal SH2 domain of the p85α regulatory subunit of phosphoinositide 3-kinase: An SH2 domain mimicking its own substrate. *J. Mol. Biol.* **292:** 763–770.

Jacobs, S.A., Podell, E.R., and Cech, T.R. 2006. Crystal structure of the essential N-terminal domain of telomerase reverse transcriptase. *Nat. Struct. Mol. Biol.* **13:** 218–225.

Jaffe, E.K., Volin, M., Bronson-Mullins, C.R., Dunbrack Jr., R.L., Kervinen, J., Martins, J., Quinlan Jr., J.F., Sazinsky, M.H., Steinhouse, E.M., and Yeung, A.T. 2000. An artificial gene for human porphobilinogen synthase allows comparison of an allelic variation implicated in susceptibility to lead poisoning. *J. Biol. Chem.* **275:** 2619–2626.

Kawasaki, M. and Inagaki, F. 2001. Random PCR-based screening for soluble domains using green fluorescent protein. *Biochem. Biophys. Res. Commun.* **280:** 842–844.

King, D.A., Hall, B.E., Iwamoto, M.A., Win, K.Z., Chang, J.F., and Ellenberger, T. 2006. Domain structure and protein interactions of the silent information regulator SIR3 revealed by screening a nested deletion library of protein fragments. *J. Biol. Chem.* **281:** 20107–20119.

Liang, J., Chen, J.K., Schreiber, S.T., and Clardy, J. 1996. Crystal structure of P13K SH3 domain at 20 angstroms resolution. *J. Mol. Biol.* **257:** 632–643.

Maxwell, K.L., Mittermaier, A.K., Forman-Kay, J.D., and Davidson, A.R. 1999. A simple in vivo assay for increased protein solubility. *Protein Sci.* **8:** 1908–1911.

Musacchio, A., Cantley, L.C., and Harrison, S.C. 1996. Crystal structure of the breakpoint cluster region-homology domain from phosphoinositide 3-kinase p85 α subunit. *Proc. Natl. Acad. Sci.* **93:** 14373–14378.

Nakayama, M. and Ohara, O. 2003. A system using convertible vectors for screening soluble recombinant proteins produced in *Escherichia coli* from randomly fragmented cDNAs. *Biochem. Biophys. Res. Commun.* **312:** 825–830.

Nolte, R.T., Eck, M.J., Schlessinger, J., Shoelson, S.E., and Harrison, S.C. 1996. Crystal structure of the PI 3-kinase p85 amino-terminal SH2 domain and its phosphopeptide complexes. *Nat. Struct. Biol.* **3:** 364–374.

Otsu, M., Hiles, I., Gout, I., Fry, M.J., Ruiz-Larrea, F., Panayotou, G., Thompson, A., Dhand, R., Hsuan, J., Totty, N., et al. 1991. Characterization of two 85 kD proteins that associate with receptor tyrosine kinases, middle-T/pp60c-src complexes, and PI3-kinase. *Cell* **65:** 91–104.

Page, R., Peti, W., Wilson, I.A., Stevens, R.C., and Wuthrich, K. 2005. NMR screening and crystal quality of bacterially expressed prokaryotic and eukaryotic proteins in a structural genomics pipeline. *Proc. Natl. Acad. Sci.* **102:** 1901–1905.

Pedelacq, J.D., Piltch, E., Liong, E.C., Berendzen, J., Kim, C.Y., Rho, B.S., Park, M.S., Terwilliger, T.C., and Waldo, G.S. 2002. Engineering soluble proteins for structural genomics. *Nat. Biotechnol.* **20:** 927–932.

Piotto, M., Saudek, V., and Sklenar, V. 1992. Gradient-tailored excitation for single-quantum NMR spectroscopy of aqueous solutions. *J. Biomol. NMR* **2:** 661–665.

Prodromou, C. and Pearl, L.H. 1992. Recursive PCR: A novel technique for total gene synthesis. *Protein Eng.* **5:** 827–829.

Rehm, T., Huber, R., and Holak, T.A. 2002. Application of NMR in structural proteomics: Screening for proteins amenable to structural analysis. *Structure* **10:** 1613–1618.

Rowlands, M.G., Newbatt, Y.M., Prodromou, C., Pearl, L.H., Workman, P., and Aherne, W. 2004. High-throughput screening assay for inhibitors of heat-shock protein 90 ATPase activity. *Anal. Biochem.* **327:** 176–183.

Siegal, G., Davis, B., Kristensen, S.M., Sankar, A., Linacre, J., Stein, R.C., Panayotou, G., Waterfield, M.D., and Driscoll, P.C. 1998. Solution structure of the C-terminal SH2 domain of the p85 α regulatory subunit of phosphoinositide 3-kinase. *J. Mol. Biol.* **276:** 461–478.

Skolnik, E.Y., Margolis, B., Mohammadi, M., Lowenstein, E., Fischer, R., Drepps, A., Ullrich, A., and Schlessinger, J. 1991. Cloning of PI3 kinase-associated p85 utilizing a novel method for expression/cloning of target proteins for receptor tyrosine kinases. *Cell* **65:** 83–90.

Waldo, G.S., Standish, B.M., Berendzen, J., and Terwilliger, T.C. 1999. Rapid protein-folding assay using green fluorescent protein. *Nat. Biotechnol.* **17:** 691–695.

Wheeler, V.C., Prodromou, C., Pearl, L.H., Williamson, R., and Coutelle, C. 1996. Synthesis of a modified gene encoding human ornithine transcarbamylase for expression in mammalian mitochondrial and universal translation systems: A novel approach towards correction of a genetic defect. *Gene* **169:** 251–255.

Wigley, W.C., Stidham, R.D., Smith, N.M., Hunt, J.F., and Thomas, P.J. 2001. Protein solubility and folding monitored in vivo by structural complementation of a genetic marker protein. *Nat. Biotechnol.* **19:** 131–136.