ELSEVIER

DNA fragmentation based combinatorial approaches to soluble protein expression Part II: Library expression, screening and scale-up

Renos Savva[1,4], Chrisostomos Prodromou[2,4] and Paul C. Driscoll[3,4]

[1] Institute of Structural Molecular Biology, School of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX, United Kingdom
[2] Section of Structural Biology, Institute of Cancer Research, Chester Beatty Laboratories, 237 Fulham Road, London SW3 6JB, United Kingdom
[3] Institute of Structural Molecular Biology, Department of Biochemistry and Molecular Biology, University College London, Gower Street, London WC1E 6BT, United Kingdom
[4] Domainex Ltd., The London Bioscience Innovation Centre, 2 Royal College Street, Camden, London NW1 0NH, United Kingdom

In this second of a two-part review encompassing random, combinatorial methods for soluble protein 'domain hunting', we focus upon the expression screening from DNA fragment libraries. Given a library of domain length-encoding DNA fragments assembled in expression vectors, it is necessary to devise reliable means to screen the sample DNA fragment population to find those that express stable, soluble target protein fragments, suitable for the required downstream aims. This review summarizes a variety of alternative strategies that have been employed to identify such stable truncates of full-length proteins. In addition, we review measures that can determine the quality of the expressed protein, the likely reliability of these measures, and the apparent extent of their application within the featured studies.

Part I of this review [1] describes the general strategy of random, combinatorial approaches to the generation of DNA fragment libraries and the relationship of the protein expression screening paradigm to more conventional approaches to the expression of targeted protein domains. Part I also describes in detail practical aspects of the DNA fragmentation step that is intrinsic to a domain hunting project. In this second part, we focus upon the detection of soluble protein expression and provide brief descriptions of instructive examples of domain hunting-type exercises from the recent literature.

Choice of expression vector

In any screening procedure there is a premium on making the component procedures as efficient as possible: minimization of false positive and false negative hits is paramount in design of the ensuing steps in the domain-hunting pipeline. Selection of the expression vector can make a significant difference to the apparent solubility and experimental tractability of a protein target. A typical scenario is that a given target protein can appear to be expressed in bacteria in an insoluble form when in isolation, but for soluble expression to be recovered when the same polypeptide chain is expressed in tandem with common protein fusion partners (e.g. glutathione *S*-transferase, maltose-binding protein, thioredoxin, NusA, and so on); see, for example, references [2–5]. There is a strong consensus that the solubility enhancement characteristic of the fusion partners is often a result of a 'passenger' effect: the intrinsic high solubility of the fusion partner protein is sufficient to balance out the intrinsic low solubility of the appended target protein, dragging it into the soluble phase. Biophysical characterization of several of these examples has shown that this 'passenger solubilization' leads to fusion proteins that are ostensibly soluble, yet are, in practice, polydisperse soluble aggregates (e.g. the protein elutes in the void volume of a size exclusion chromatography column, has light-scattering activity consistent with a polymeric particle, and often can be depleted from solution by relatively low-speed centrifugation) [6]. Such species have been named 'soluble inclusion bodies', a name that immediately hints at their compromised experimental tractability.

Because of the intrinsic risk of 'passenger solubilization' by globular fusion partners, a preferable situation is to adopt a minimal translational protein tag that allows for both detection of expression and subsequent generic affinity-based purification.

Corresponding author: Driscoll, P.C. (p.driscoll@ucl.ac.uk)

1359-6446/06/$ - see front matter © 2007 Elsevier Ltd. All rights reserved. doi:10.1016/j.drudis.2007.08.016

In principle, any peptide tag that does not drastically alter the solubility of an attached polypeptide, and for which there is an applicable affinity purification matrix will suffice for this purpose. For example, the almost ubiquitous polyhistidine tag/immobilized metal-ion affinity chromatography (IMAC) meets these specifications, though many alternative peptide detection/purification systems also exist. For example, the streptavidin-binding (Strep-) peptide tag could be used to circumvent any concerns with the potential interference of the high concentration imidazole wash steps inherent in IMAC.

Detection of soluble protein expression

Expression of the DNA fragment library typically leads to tens of thousands of independent clones, only a subset of which express the fragment DNA in the correct reading frame and orientation leading to a protein product (Figure 1). Ideally one requires at this stage to combine the detection of such clones with an assessment of whether the protein produced is tractably soluble, a soluble aggregate, or expressed in inclusion bodies. Performing this operation as a single step is a significant challenge. Typical solutions break this problem down into multiple steps. First, colony lifts and anti-tag immunoblotting can be applied, using standard methodology, to identify all clones that express a tagged protein product in any form. Subsequent liquid-phase culture of a 'hit' clone can then be adopted to generate a sufficient amount of protein in order that it can be purified and assessed by various biochemical and biophysical parameters. This 'bulking up', albeit often performed on a relatively small scale (e.g. 5 ml culture volume) and in parallel fashion utilizing robotic liquid-handling equipment, can prove to be a limiting factor in the overall application of the expression screen. Nevertheless, the prescreen and postscreen paradigm is very commonly adopted, since in many respects it represents a standardization of familiar concepts that have been used in low-throughput protein expression applications for some time.

More sophisticated approaches to the combined detection of protein expression and assessment of solubility have been reported. Thus, in the colony-filtration immunoblot (COFI-blot) procedure of Nordlund and co-workers, mature colonies are lifted from the plate with a hydrophilic 0.45 μm membrane and transferred to an inducing environment for several hours [7,8]. The expressed colonies are then lysed by placing the PVDF membrane, colony side up, onto a nitrocellulose membrane that is, in turn, on top of a filter paper soaked in a lysis buffer. Soluble contents of the lysed cells will leach through the PVDF onto the nitrocellulose, whereupon binding of proteins will occur. The process appears to provide a credible, if somewhat manually intensive, route to the rapid detection of apparently soluble protein products.

Several researchers have utilized the intrinsic development of the fluorophore of green fluorescent protein (GFP) tagged to the expressed protein fragment as a marker of when the fusion protein is productively expressed in the cell [9–14]. As elegantly demonstrated in a series of examples by Waldo and co-workers, one finds that if GFP is tagged to the C-terminus of a polypeptide chain, then the fluorophore 'matures' only when the properties of that chain are such as not to interfere with the proper folding of GFP. Insoluble or aggregating polypeptides mean that the bacteria do not fluoresce; a stable 'soluble' (folded or not) polypeptide permits the generation of the fluorophore and imbues the bacterial cells with a dramatic fluorescent phenotype. In principle, extremely rapid and automatable optical detection of the clones that express such stable chains becomes possible and has clearly been demonstrated to be of practical value, particularly in the application to the random mutagenesis of the target polypeptide for improved folding kinetics (where folding kinetics can correlate strongly with apparent solubility). Recently, Waldo *et al.* have extended the GFP tag methodology to counter the prospect that the GFP tag itself can provide passenger solubilization to the appended polypeptide, by devising a two-plasmid split-GFP complementation system [9,15,16]. GFP is divided into an 11-residue fragment used to tag the target polypeptides expressed first, followed by the separate induction of the remaining 227 residues of GFP. Convincingly, only when the 11-residue tag of the expressed target is available for combination with the remainder of GFP does fluorophore-maturation take place. Provided that the 11-residue fragment of GFP is generically non-perturbing of the expressed protein's solubility properties, this and related methodologies [17] would appear to hold great promise for the true high-throughput assessment of protein expression in either colony or small-scale liquid culture formats. In a related development, Hart and co-workers have reported the use of a method that relies upon *in vivo* enzyme-dependent modification with biotin of an appended peptide tag [18–20] and is similarly thought to discriminate soluble expressed protein from aggregates or inclusions (see below), though full details have not yet been reported.

'Hit' scale-up

Once positive clones have been detected, by one means or other, essentially standard procedures can be employed for scaling up of protein expression and purification of the candidate proteins. One has to bear in mind that soluble protein detection is an inexact science. It is in the nature of the beast that different protein chains have individual physical and biochemical characteristics and, unlike their encoding DNAs, are likely to have idiosyncratic properties. The rate of false positive detection in domain hunting is likely to be higher than one would like and, therefore, the ability to make an early assessment of the experimental tractability of a given protein hit is paramount. Moreover, the protein that is eventually purified from the cell might be subject to post-translational proteolytic trimming or other modifications, so confirmatory identification of the protein size must bear this prospect in mind. Also, exactly as described for solubilizing protein tags above, it is entirely credible that a stable tractable protein domain can carry with it substantial additional disordered protein material—the domain itself imposes passenger solubilization on the neighbouring parts of the chain, as it was 'in *cis*'. Thus, it should be appreciated at this stage that assessment of foldedness of hit candidates by circular dichroism (CD) or NMR spectroscopy may reveal evidence of extended flexible tails. One needs also to bear in mind that because of their overall hydrophilic characteristics natively disordered regions of large proteins may emerge on their own as true positives in such domain hunting exercises, but these 'hits' will clearly lack the biophysical signatures of a globular protein (e.g. see reference [21]).

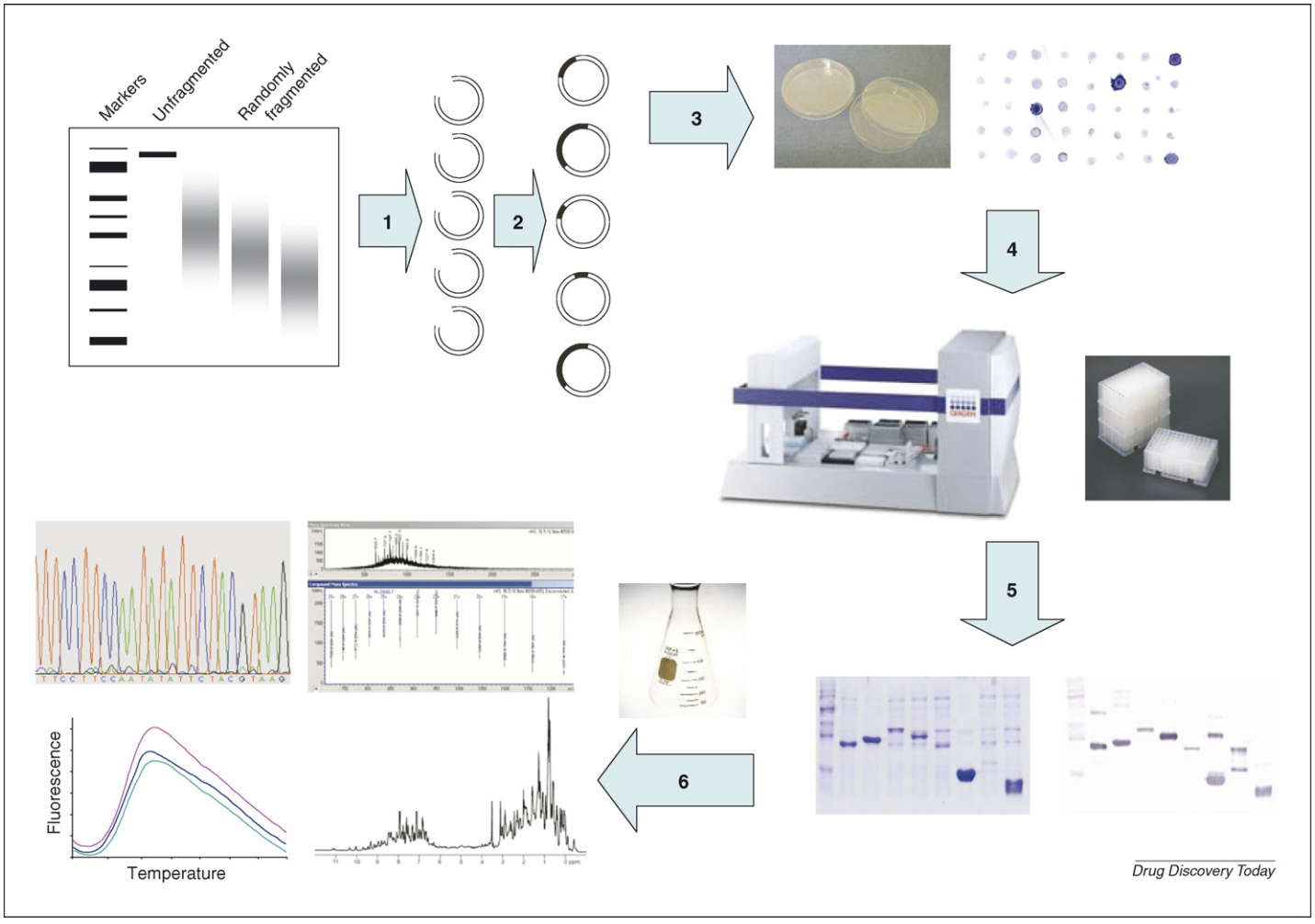

FIGURE 1

Schematic illustration of a generic domain-hunting workflow starting from **(1)** random fragmentation of a DNA template; **(2)** capture of the DNA fragments into an expression vector; **(3)** plating out of the expression library and assessment of protein expression per colony; **(4)** low volume scale-up and semi-automated lysis and purification; **(5)** assessment of protein yield and solubility by SDS-PAGE and western blot and **(6)** large volume scale-up of 'hit' expression and purification followed by biochemical and biophysical characterization, including (as depicted) identification at the nucleotide (DNA sequencing) and amino acid level (ESI mass spectrometry) of the expressed protein domain boundaries, detection of an (un)folding transition in a differential scanning fluorimetry experiment and assessment of foldedness by one-dimensional ^{1}H NMR spectroscopy.

Assessment of solubility/aggregation state and foldedness

Once sufficient quantities of expressed protein have been obtained, one wants to proceed to functional and biophysical characterization. In its purest form, the pursuit of domain hunting should not require any knowledge of the biological function, and so there is perhaps a bias towards a general biophysical assessment of the state of the protein domains. In this respect, there is a plethora of options, many of which are listed in brief in Table 1, together with short comments describing the relative advantages and disadvantages of each method. Because of the intrinsic, non-generic, behaviour of proteins, it would be unwise to set too much store by any one measure of foldedness, and combinations of such tests are generally the most useful. Of course, where a functional assay exists that probes an enzymatic function or binding to a cognate ligand this property can be used to triage the domain hits before further analysis.

Each method of analysis has a particular subset of outputs that are useful to a greater or lesser extent, sometimes depending on the specific characteristics of the particular protein, in determining the tractability of the protein hit for further structural or functional characterization. Suffice to say here that many of the methods listed can give an early indication that the protein is an adventitious 'molecular inclusion' (e.g. low speed centrifugation, SEC, AUC, DLS) and that if one has sufficient protein sample to perform the characterization of the protein by NMR spectroscopy, then this can be an extremely powerful way to assess the overall foldedness of the protein, particularly where the folded element of the domain hit is less than ca. 40 kDa in size [22,23]. The disadvantage of the NMR method is that it is relatively demanding on the quantity of protein required to perform the experiment. Once folded domain hits have been identified, then scale-up and optimization of protein expression and full-blown functional and structural characterization can be contemplated.

Examples of 'real world' domain hunting

The domain hunting concept addressed here, defined as an essentially random search for stable protein fragments expressed by

TABLE 1

Typical means to assess the folding properties of polypeptide chains derived from random screening approaches to expression

Biophysical technique	Indication	Caveats
Limited proteolysis by protease(s)	Stability of proteolytically cleaved protein fragments can be an indication of folding stability	An absolute requirement for target protein expression at the outset, from which to proceed. Even folded protein can be cleaved in exposed loops leading to ambiguities in the analysis. In some cases, proteolytically resolved termini are not related to the nascent folding, that is, cloning the corresponding encoding ORF will not always result in high yield, soluble or folded protein
N-terminal amino acid sequence determination (Edman degradation)	Results can aid unambiguous identification of the expressed protein fragment N-terminus	Sample purity is important, and not many laboratories maintain Edman sequencing as a routine service
Determination of intact protein mass, for example, electrospray ionisation-mass spectrometry (ESI-MS)	Techniques are useful for the corroboration of the polypeptide sequence, which can be different from that predicted by the encoding DNA fragment; tryptic digestion coupled and peptide mass mapping can aid the precise identification of large protein termini	The accuracy of intact masses may be diminished for high molecular weight polypeptides. Requires access to expensive equipment, often dedicated to other types of experiment. Analysis may require sophisticated software
Resistance to sedimentation by centrifugation	High molecular weight aggregates will be sedimented, 'soluble' proteins will remain in solution	Outcome can be a sensitive function of the centrifugation speed, temperature and buffer conditions
Measurement of regular secondary structure content by CD spectropolarimetry	Can give quantitative estimates of the relative secondary structure content	Aggregated or marginally folded proteins can give misleadingly high quality CD spectra
Measurement of apparent hydrodynamic parameters and mass dispersion of the protein sample by analytical size exclusion chromatography (SEC), dynamic light scattering (DLS) and/or analytical ultracentrifugation (AUC)	These methods give quantitative or semi-quantitative measures of the apparent protein molecular size and coupled with predicted or actual intact mass data can lead to the assessment of oligomeric state. SEC and DLS can give a relatively rapid indication of sample aggregation or monodispersity	SEC is a relatively low-resolution technique. DLS can be very sensitive to small amounts of contaminating high molecular weight aggregates. AUC experiments require expensive equipment and an experienced operator. Appropriate concentrations of proteins can be lower than that required for downstream applications (e.g. structural biology)
Assessment of thermal stability by temperature dependence of the intrinsic fluorescence, response of an extrinsic fluorophore ('Thermofluor'), CD signature, light scattering, or microcalorimetry response (e.g. DSC)	These techniques can reveal the presence of a folding transition, and the amplitude may give a semi-quantitative measure of the thermodynamic stability of the folded state	Optical measurements are relatively sensitive requiring small amounts of sample and can be engineered to be assessed in multi-well microplate format; calorimetry methods are less sensitive
Assessment of folding status by 1D ^{1}H or 2D heteronuclear nuclear magnetic resonance spectroscopy (NMR)	The extent of chemical shift dispersion of polypeptide NMR spectra is highly indicative of the folding status of the chain	NMR is an insensitive method requiring mg quantities of protein; the utility of NMR is generally limited to single chain proteins of limited overall molecular size (ca. $<$40 kDa); the presence of long flexible tails can obscure the presence of a smaller folded core; application of NMR requires high cost instrumentation and dedicated expertise; the more informative 2D heteronuclear NMR paradigm requires random isotope labelling

particular subregion of an encoding DNA molecule, has appeared in the literature only on limited occasions to date. Here we list some published examples that illustrate specific applications of some of the domain hunting principles outlined above. We suspect that aspects of domain hunting may be practised more widely, particularly in groups engaged in SP, and especially in the context of 'primer pair walking' [1], than we are presently aware, and we apologise here for any accidental omission.

T-PCR and GFP fusions applied to domain footprinting in the signalling protein Vav

Kawasaki and Inagaki's early work in this field adopted random-tagged PCR (T-PCR) [1] combined with capture of the DNA fragments into an expression vector encoding a 'standard' C-terminal GFP tag [10]. In a screen of 100 000 clones corresponding to the fragmented Vav protein (816 residues intact) they discovered multiple hits for the short acidic domain (2×), the DH domain (5×) and the short Cys-rich domain (3×). Whilst it is unclear whether the acidic and Cys-rich domain hits corresponded to folded globular proteins, the CD spectrum for the product of one of the subcloned DH constructs (i.e. without GFP) gave a signature consistent with a very high α-helical content, as expected from the known 3D structure. By elimination, fragmentation of a shorter template coding for the C-terminal region yielded two hits corresponding to the predicted SH3 domain within sampling of 30 000 clones. Additional notable aspects of this report include the early elucidation of the problem of false positives (e.g. GFP-encoded passenger solubilization), and the fact that the T-PCR approach appears to be accompanied by a significant degree of inadvertent mutagenesis: only one of 14 'soluble clones' had fewer than two mutations, and three had five amino acid substitutions.

Colony filtration methods coupled to 5′ deletion mutagenesis for protein solubilization

Nordlund and co-workers have been active in various aspects of high-throughput protein expression and have made some very interesting technical contributions relevant to domain hunting [7,8,24–26]. They introduced the colony-filtration (CoFi) concept [7,8] described above and demonstrated an application of the methodology to the screening of a library of N-terminal truncation mutants of the product of a human gene HP16/LCMT1 that was not expressed as a soluble full-length polypeptide in bacteria. The library of DNA fragments was generated using the commercially available Erase-A-Base system (Promega) that employs time-dependent exonuclease III (Exo III) digestion of the vector-ligated template that has been asymmetrically cut at the 5′ end to leave a 5′ overhang on the template side and a 3′ overhang on the vector side. ExoIII is a 3′–5′ exonuclease that is inactive against 3′ overhangs of more than three bases but will progressively digest 3′ recessed or blunt DNA ends. Thus, in this application the target DNA is progressively shortened on the insert side. The rate of digestion can be controlled by variation of the salt level and the reaction temperature. Timed aliquots of the reaction are removed for treatment with S1 nuclease to remove ssDNA and Klenow DNA polymerase to fill in the 3′ recessed end before re-ligation to create intact expression plasmids encoding N-terminally truncated polypeptides. This approach for incremental shortening of a DNA molecule is borrowed from the procedures that have previously been exploited in the methodology dubbed ITCHY (incremental truncation for the creation of hybrid enzymes) that is designed to create combinatorial libraries of two genes in a manner independent of DNA sequence homology [27–29].

Screening of the HP16/LCMT1 library using the CoFi blot method yielded 24 clones producing apparently soluble poly-His-tagged deletion constructs, nine of which could be purified and were shown to have translational starts that cluster close to the amino-terminal boundary of a predicted methyltransferase domain. The Nordlund group have described a survey of the application of their approach to a panel of mammalian protein targets [30]. They report that from a screen of ~2000 colonies per target, they were able to detect soluble protein fragment expression for 17 of 19 otherwise 'hard-to-express' proteins (between 20 and 100s of hits per target). In liquid-phase scale-up of a sample 24 colonies per target, 14 targets yielded sufficient quantities of N-deleted protein to be purified, 11 of which were detectable on a Coomassie-stained gel. At first glance the deduced translational starts of the identified fragments appeared to correlate reasonably with domain boundaries predicted by at least one computational prediction. However, without, as yet, further reports of the biophysical characterization of these protein products it is difficult to assess the folding status of these soluble proteins.

5′ incremental truncation applied to influenza virus RNA polymerase PB2 C-terminal domain

In the most recent report of its type, Hart and co-workers describe the application of a similar random 5′ deletion strategy to the high-throughput sampling of solubly expressed domains from the 759 residue PB2 subunit of the heterotrimeric RNA polymerase of influenza virus [20]. In their procedure (dubbed expression of soluble proteins by random incremental truncation, ESPRIT [18]) the target cDNA is directionally cloned into an expression vector, which is then opened with restriction endonucleases at the 5′ end of the insert and subjected to time-dependent 'incremental truncation' with Exo III. The clever aspect of the ESPRIT approach is that the expression vector employed adds a 15-residue C-terminal peptide (GLNDIFEAQKIEWHE) to the expressed protein fragment that is biotinylated in the bacterial cell by the endogenous BirA enzyme [19]. Biotinylation is thought to proceed efficiently only for soluble, non-aggregated polypeptides, and the attached biotin provides a facile means for detection of soluble expressed proteins and quantitation with a fluorescent streptavidin conjugate. In the application to flu PB2, colonies representing nearly 27 000 constructs were robotically arrayed, induced and assessed for expression of soluble biotinylated proteins by *in situ* lysis and fluorescent imaging. No statistics of the positive clone counts is reported, but by analysis of several positive clones it was possible to identify that the C-terminal 80–110 amino acid residues of PB2 were highly expressed. Following the subcloning of two inserts (residues 678–759 and 661–759) into standard expression vectors, sufficient protein for heteronuclear NMR spectroscopy-based structure determination was obtained, leading to novel insights in the biological role of this domain and its interactions with other proteins [20]. The authors of this study make the important point that in the examples described above that adopted the exonuclease III 'single-end'-based approach for the generation of random

protein fragments by specific 5′-directed (or 3′-directed) truncation of the encoding DNA template, one can realistically achieve oversampling of the potential 'construct space' [31]. Thus, in the case of flu, PB2 the ESPRIT screen of 27 000 colonies clearly represents multiple coverage of the potential $759 \times 3 = 2277$ incrementally truncated DNA variants. This situation contrasts strongly with that in which one is seeking domains that may occur in the middle of the target cDNA where the construct space is two to three orders of magnitude larger (see reference [1]).

Identification of a novel structured domain in telomerase by T-PCR/GFP-fusion protein screening

Using a strategy very similar to Kawasaki and Inagaki, Jacobs *et al.* screened DNA fragments of the *Tetrahymena thermophilia* telomerase reverse transcriptase (TERT) gene to identify soluble poplypeptide fragments [14]. Telomerase is a ribonucleoprotein complex that extends the ends of eukaryotic chromosomes, and the ca. 120 kDa catalytic protein subunit of the complex has properties that are distinct from those of other reverse transcriptases. Although some elements of the TERT protein sequences could be related to aspects of known biological properties, little was known about the boundaries of the structured domains. The Cech group generated T-PCR products between 600 and 3400 bp and cloned them into a vector for expression as C-terminally GFP-tagged proteins. Optical UV screening ($\lambda = 366$ nm) of 10 000 clones yielded ~1% with a strong fluorescence signal, and several were selected for overexpression of the fusion proteins in liquid culture. The authors report that 27 colonies expressed fusion proteins larger than 35 kDa in size (GFP alone is ca. 27 kDa). Sequencing results indicated mutations in the N-termini and C-termini of several of the expressed protein fragments, a consistent risk of the T-PCR fragmentation approach. Five of the most highly expressed fragments corresponded closely to the conserved N-terminal region of TERT spanning the so-called 'GQ motif'. Subcloning of the DNA fragment coding for TERT(2-191) to a standard His_6-expression vector led to 30 mg/l pure protein when expressed at 18 °C. This protein product was subjected to limited proteolysis, yielding stable subfragments identified as encompassing constructs between residues 4 and 13 at the N-terminus and 184 or 185 at the C-terminus. Separate expression of TERT(13–184) gave a clearly stable globular protein, as assessed by 2D heteronuclear NMR. The 3D structure of this so-called telomerase essential (TEN) domain has been solved by using crystals of the TERT(2–191) fragment, revealing a novel polypeptide fold [32]. Interestingly, it was the longer construct that includes flexible tail regions that was crystallized; the shorter TERT(13–184) polypeptide lacks the ability to bind strongly to RNA shown by the longer version. This result perhaps serves to illustrate that the optimal construct for structural analysis is not necessarily the one that corresponds precisely with boundaries between chain order and disorder.

Massively parallel shotgun proteolysis for stable domain identification

A different approach to identifying stable protein domains was described recently by Christ and Winter in which they combined fragmented DNA and phage display technology to screen very high numbers of affinity protein-tagged clones for resistance to protease digestion [33]. Thus, shear-fragmented DNA was captured, via oligonucleotide adaptors, in a phagemid expression cassette bracketed by DNA for the pelB leader sequence connected to N-terminal protein barnase and the phage pIII protein. After expansion of the phage library to 10^{10}–10^{11} clones, the phages were incubated with trypsin at low temperature. Clones that remain intact under these conditions are retained on immobilized barstar protein because of the high affinity of the barnase/barstar complex. After washing, the surviving phages are recovered and amplified again. PCR screening of these phage clones is then used to isolate the DNA fragments corresponding to protease-resistant domains, which can be subcloned into standard expression vectors and tested for protein expression. Whilst this method does not ostensibly screen for mid-to-high protein solubility, the demonstrations included in the report suggest that one can approach very significant coverage of the potential protein coding space because of the very high numbers of independent clones in the phage library. To this extent, concerns about the 1-in-18 in-frame cloning efficiency of blunt-ended DNA [1] can be safely circumvented. The authors make the claim that this 'shotgun proteolysis' strategy, which represents a kind of massively parallel approach to performing limited proteolysis trials, is able to identify whole structural domains as well as 'variants lacking peripheral elements'. There is also a degree of correlation between the apparent domain boundaries identified by this method with bioinformatic predictions, though there are some significant differences. However, whether the 'peripherally trimmed' protein domain 'hits' reflect tractably stable isolated polypeptides is not immediately clear. Moreover, the examples given are exclusively of bacterial proteins, and it will be interesting to discover whether this approach works well for eukaryotic targets.

Nested deletion libraries applied to domain mapping of the silent information regulator Sir3

Ellenberger and co-workers reported the application of mung bean nuclease to generate a large number of overlapping fragments with random 5′ and 3′ termini of the gene encoding the 978-residue *Saccharomyces cerevisiae* Sir3 protein [34]. Sir3 is part of the silent information regulator machinery that makes heterochromatin transcriptionally silent. Mung bean nuclease is most commonly used for removing single-stranded 3′ overhangs but will slowly cleave double-stranded DNA at high concentrations of enzyme. In the application to Sir3, the nested gene deletions were cloned into a vector that leads to fusion of the expressed protein products to the N-terminus of *E. coli* type I chloramphenicol acetyltransferase (CAT), pre-pended with a six-His affinity tag. The use of CAT provided a means for a first round of selection of in-frame coding sequences by growth on chloramphenicol, though the researchers also reported that only about one-third of the clones so obtained expressed Coomassie-detectable amounts of soluble protein, necessitating further rounds of screening; apparently high level expression of CAT and CAT-fusion proteins is itself toxic to *E. coli* cells. Nevertheless, the strategy adopted here allowed for substantial enrichment of in-frame gene fragments, and the subsequent detection by a functional assay of eight overlapping segments of the Sir3 protein that bind the coiled-coil domain of the cognate partner Sir4, from which it was deduced that residues 464–728 represent the minimal domain that gives a stable interaction. In a

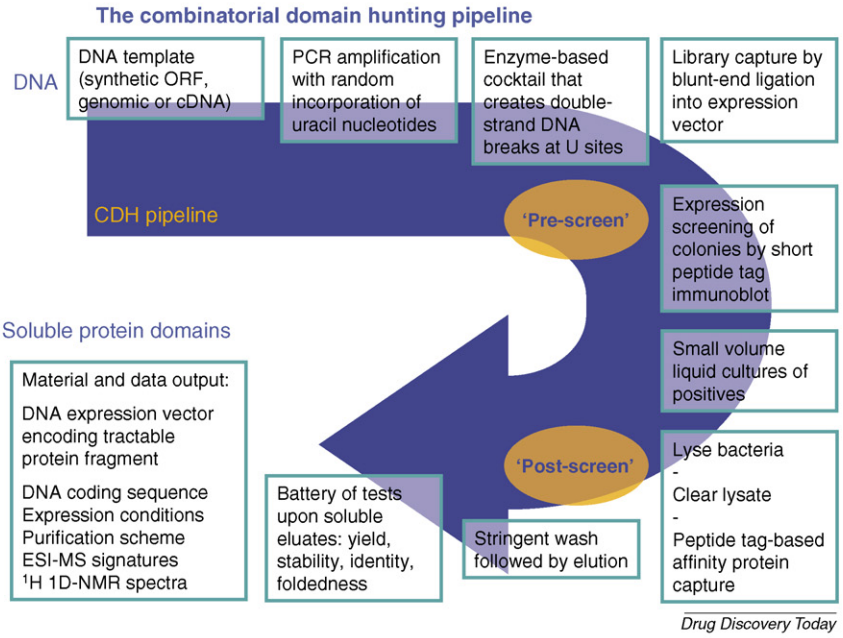

FIGURE 2

A schematic representation of the Combinatorial Domain Hunting pipeline as described by Reich *et al.* [21]. Here 'Prescreen' refers to the immunodetection of total tagged protein fragment expression, whereas 'Postscreen' describes immunodetection of soluble tagged protein domains. A description of the CDH DNA fragmentation methodology is also given in Part 1 of this review [1].

separate screen, Sir3 residues 832–978 were found to operationally define the minimal region of the protein's C-terminus that forms homodimers. The identification of these stable, functional domains of Sir3 permitted various additional biophysical analyses of their interaction with Sir4, leading to insightful models of the architecture of their mode of interaction [34].

'Combinatorial Domain Hunting' applied to the p85 subunit of phosphoinositide 3-kinase

Together with Domainex Ltd., we have described a set of procedures that combines a uracil-doped PCR/DNA mismatch base excision repair based approach to fine-grained DNA fragmentation (see reference [1,21]), coupled to automated filter-capture screening of the resulting expression libraries. The overall process is named 'Combinatorial Domain Hunting' (CDH) (see Figure 2) and has been demonstrated in a proof-of-principle application to the multimodular 85 kDa regulatory subunit of class IA phosphoinositide 3-kinase. The design of the CDH process focused upon: optimization of the sampling fidelity of the DNA fragmentation step; minimization of the influence of passenger solubilization effects arising from the effectively obligatory tandemly expressed polypeptide affinity tag; and consideration of the optimal filter/capture procedure for isolation and detection of soluble protein fragments for discrimination of soluble aggregates. Thus, CDH incorporates the novel DNA fragmentation method that allows for unbiased DNA cleavage on either side of any A:T base pair in the template, utilises efficient capture of the DNA fragment library into a series of topoisomerase-modified vectors designed for the shortest possible C-terminal poly-histidine affinity tag, and adopts sampling, rather than solubilization, of the cell contents coupled with 96-well-plate-based tandem filtration/immobilized Ni(II) or Strep-tag capture and high-speed centrifugation for discrimination of soluble polypeptides. The p85 protein contains four globular domains with well-characterized 3D structures. Although the overall 3D structure of p85 is not known and, in fact, no structure of tandemly linked p85 domains has been reported to date, it is expected that the domains are rather loosely tethered in the intact protein. The application of CDH to p85 recovered, from a screen of ca. 1400 clones containing inserts, 16 apparently soluble protein constructs that could be purified on a small scale. Fourteen of these clones provided sufficient quantities for further characterization when expressed at the 1 L scale, and eight of these were examined for foldedness by 1D ^{1}H NMR spectroscopy. Seven of the eight proteins examined this way gave chemical shift dispersion consistent with a substantially folded globular structure; the remaining construct was a false positive soluble and non-aggregating polypeptide with no ordered conformation that mapped to one of the interdomain linker regions of p85. In fact, CDH recovered all of the known globular folded domains of p85 either in isolation or in tandem. The two Src 2 homology (SH2) domains were recovered multiple times (3×), and the N-terminal SH3 domain was obtained with a substantial portion of extra sequence at its C-terminus, presumably 'passenger solubilized' in *cis* by the SH3 domain itself.

In applications to a variety of targets that have been explored since the p85 study, CDH has yielded stable protein fragments, usually within a screen of fewer than 20 000 colonies. These targets range from bacterial proteins through to giant human kinases and viral polyproteins. Clearly, the combinatorial possibilities within a single protein cannot be fully sampled in 20 000 or fewer clones (see reference [1]); however, the success rate is such that alternative phenomena in cellular selection are surely at work. It is possible

that an unknown proportion of constructs do not give rise to colonies due to toxic effects of the random fragment/reading frame arrangement in the expression vector. If so, the viable colonies may appear to be enriched for random fragment sequences with a greater potential to generate stable recombinant protein than might otherwise be expected. In some cases, CDH targets have been DNA-recoded to potentially enhance expression in *E. coli*, which could have had a small positive effect. This might then be bolstered by expression from alternative initiation codons, perhaps even by translational frame-shift tolerance, through to intracellular proteolysis leading to more stable polypeptide sub-fragments.

Summary

In combination, the various descriptions of domain footprinting reveal that the randomized approach can massively increase the scope for grazing the landscape of polypeptide solubility for a given, possibly recalcitrant, protein target. Even though the numbers of candidate clones screened largely remain well below what might be considered necessary to achieve success, the early signs are that such approaches demonstrate an unexpectedly high hit rate for the discovery of soluble proteins. Effects of frame-shift tolerance in the cellular translation machinery, as well as some other selection phenomena that are, as yet, unclear may be playing a part. Undoubtedly bottlenecks remain in the pipelines that have been devised for such procedures, perhaps most notably in terms of the scale-up and structural validation of the identified 'hits'. Thus, coupled with anticipated advances in the parallelization of high level protein expression, for example, in small-scale bioreactors, and the miniaturization of assays of biophysical properties of expressed proteins (e.g. Thermofluor [35,36], Stargazer [37], SUPREX mass spectrometry [38,39]), and very low volume heteronuclear NMR spectroscopy [40], one can hope that the scope of such screens can be increased, and the unit costs (both in money and time) of each screening project be reduced. More challenging will be the extension of the random DNA fragment approach to target expression in heterologous expression hosts other than *E. coli* that would open up the potential for a wide range of post-translational modifications that might prove to be crucially determinant of the final outcome for many important therapeutic targets. In the meantime one can anticipate extending the applications of domain hunting in the not too distant future to examples of multidomain-secreted protein targets or the ectodomains of transmembrane proteins and to systems of searching for the functionally and structurally tractable parts of obligate heterodimers.

Acknowledgements

We thank our colleagues Professor Laurence Pearl and Dr Keith Powell for comments on the manuscript and their important contributions to the realization of CDH concept in our laboratory.

References

1 Prodromou, C. *et al*. (2007) DNA Fragmentation-based combinatorial approaches to soluble protein expression. Part I: Generating DNA fragment libraries. *Drug Discov. Today*, in press, doi:10.1016/j.drudis.2007.08.012.
2 Hockney, R.C. (1994) Recent developments in heterologous protein production in *Escherichia coli*. *Trends Biotechnol.* 12, 456–463
3 Uhlen, M. *et al*. (1992) Fusion proteins in biotechnology. *Curr. Opin. Biotechnol.* 3, 363–369
4 Kapust, R.B. and Waugh, D.S. (1999) *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Sci.* 8, 1668–1674
5 LaVallie, E.R. and McCoy, J.M. (1995) Gene fusion expression systems in *Escherichia coli*. *Curr. Opin. Biotechnol.* 6, 501–506
6 Nomine, Y. *et al.* (2001) Formation of soluble inclusion bodies by hpv e6 oncoprotein fused to maltose-binding protein. *Protein Exp. Purif.* 23, 22–32
7 Cornvik, T. *et al.* (2005) Colony filtration blot: a new screening method for soluble protein expression in *Escherichia coli*. *Nat. Methods* 2, 507–509
8 Dahlroth, S.L. *et al.* (2006) Colony filtration blotting for screening soluble expression in *Escherichia coli*. *Nat. Protoc.* 1, 253–258
9 Cabantous, S. and Waldo, G.S. (2006) *In vivo* and *in vitro* protein solubility assays using split GFP. *Nat. Methods* 3, 845–854
10 Kawasaki, M. and Inagaki, F. (2001) Random PCR-based screening for soluble domains using green fluorescent protein. *Biochem. Biophys. Res. Commun.* 280, 842–844
11 Pedelacq, J.D. *et al*. (2002) Engineering soluble proteins for structural genomics. *Nat. Biotechnol.* 20, 927–932
12 Waldo, G.S. (2003) Improving protein folding efficiency by directed evolution using the GFP folding reporter. *Methods Mol. Biol.* 230, 343–359
13 Waldo, G.S. *et al.* (1999) Rapid protein-folding assay using green fluorescent protein. *Nat. Biotechnol.* 17, 691–695
14 Jacobs, S.A. *et al.* (2005) Soluble domains of telomerase reverse transcriptase identified by high-throughput screening. *Protein Sci.* 14, 2051–2058
15 Cabantous, S. *et al.* (2005) Protein tagging and detection with engineered self-assembling fragments of green fluorescent protein. *Nat. Biotechnol.* 23, 102–107
16 Cabantous, S. *et al.* (2005) 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
17 Waldo, G.S. (2003) Genetic screens and directed evolution for protein solubility. *Curr. Opin. Chem. Biol.* 7, 33–38
18 Alzari, P.M. *et al.* (2006) Implementation of semi-automated cloning and prokaryotic expression screening: the impact of SPINE. *Acta Crystallogr. D. Biol. Crystallogr.* 62, 1103–1113
19 Beckett, D. *et al.* (1999) A minimal peptide substrate in biotin holoenzyme synthetase-catalyzed biotinylation. *Protein Sci.* 8, 921–929
20 Tarendeau, F. *et al.* (2007) Structure and nuclear import function of the C-terminal domain of influenza virus polymerase PB2 subunit. *Nat. Struct. Mol. Biol.* 14, 229–233
21 Reich, S. *et al.* (2006) Combinatorial domain hunting: an effective approach for the identification of soluble protein domains adaptable to high-throughput applications. *Protein Sci.* 15, 2356–2365
22 Page, R. *et al.* (2005) NMR screening and crystal quality of bacterially expressed prokaryotic and eukaryotic proteins in a structural genomics pipeline. *Proc. Natl. Acad. Sci. U.S.A.* 102, 1901–1905
23 Rehm, T. *et al.* (2002) Application of NMR in structural proteomics: screening for proteins amenable to structural analysis. *Structure* 10, 1613–1618
24 Knaust, R.K. and Nordlund, P. (2001) Screening for soluble expression of recombinant proteins in a 96-well format. *Anal. Biochem.* 297, 79–85
25 Vedadi, M. *et al.* (2006) Chemical screening methods to identify ligands that promote protein stability, protein crystallization, and structure determination. *Proc. Natl. Acad. Sci. U.S.A.* 103, 15835–15840
26 Ericsson, U.B. *et al.* (2006) Thermofluor-based high-throughput stability optimization of proteins for structural studies. *Anal. Biochem.* 357, 289–298
27 Ostermeier, M. *et al.* (1999) A combinatorial approach to hybrid enzymes independent of DNA homology. *Nat. Biotechnol.* 17, 1205–1209
28 Ostermeier, M. *et al.* (1999) Combinatorial protein engineering by incremental truncation. *Proc. Natl. Acad. Sci. U.S.A.* 96, 3562–3567
29 Ostermeier, M. and Lutz, S. (2003) The creation of ITCHY hybrid protein libraries. *Methods Mol. Biol.* 231, 129–141
30 Cornvik, T. *et al.* (2006) An efficient and generic strategy for producing soluble human proteins and domains in *E. coli* by screening construct libraries. *Proteins* 65, 266–273
31 Hart, D.J. and Tarendeau, F. (2006) Combinatorial library approaches for improving soluble protein expression in *Escherichia coli*. *Acta Crystallogr. D. Biol. Crystallogr.* 62, 19–26

32 Jacobs, S.A. *et al.* (2006) Crystal structure of the essential N-terminal domain of telomerase reverse transcriptase. *Nat. Struct. Mol. Biol.* 13, 218–225
33 Christ, D. and Winter, G. (2006) Identification of protein domains by shotgun proteolysis. *J. Mol. Biol.* 358, 364–371
34 King, D.A. *et al.* (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
35 Pantoliano, M.W. *et al.* (2001) High-density miniaturized thermal shift assays as a general strategy for drug discovery. *J. Biomol. Screen.* 6, 429–440
36 Lo, M.C. *et al.* (2004) Evaluation of fluorescence-based thermal shift assays for hit identification in drug discovery. *Anal. Biochem.* 332, 153–159
37 Senisterra, G.A. *et al.* (2006) Screening for ligands using a generic and high-throughput light-scattering-based assay. *J. Biomol. Screen.* 11, 940–948
38 Ghaemmaghami, S. *et al.* (2000) A quantitative, high-throughput screen for protein stability. *Proc. Natl. Acad. Sci. U.S.A.* 97, 8296–8301
39 Ghaemmaghami, S. and Oas, T.G. (2001) Quantitative protein stability measurement *in vivo*. *Nat. Struct. Biol.* 8, 879–882
40 Peti, W. *et al.* (2005) Towards miniaturization of a structural genomics pipeline using micro-expression and microcoil NMR. *J. Struct. Funct. Genomics* 6, 259–267

Five things you might not know about Elsevier

1.
Elsevier is a founder member of the WHO's HINARI and AGORA initiatives, which enable the world's poorest countries to gain free access to scientific literature. More than 1000 journals, including the *Trends* and *Current Opinion* collections and *Drug Discovery Today*, are now available free of charge or at significantly reduced prices.

2.
The online archive of Elsevier's premier Cell Press journal collection became freely available in January 2005. Free access to the recent archive, including *Cell, Neuron, Immunity* and *Current Biology*, is available on ScienceDirect and the Cell Press journal sites 12 months after articles are first published.

3.
Have you contributed to an Elsevier journal, book or series? Did you know that all our authors are entitled to a 30% discount on books and stand-alone CDs when ordered directly from us? For more information, call our sales offices:

+1 800 782 4927 (USA) or +1 800 460 3110 (Canada, South and Central America)
or +44 (0)1865 474 010 (all other countries)

4.
Elsevier has a long tradition of liberal copyright policies and for many years has permitted both the posting of preprints on public servers and the posting of final articles on internal servers. Now, Elsevier has extended its author posting policy to allow authors to post the final text version of their articles free of charge on their personal websites and institutional repositories or websites.

5.
The Elsevier Foundation is a knowledge-centered foundation that makes grants and contributions throughout the world. A reflection of our culturally rich global organization, the Foundation has, for example, funded the setting up of a video library to educate for children in Philadelphia, provided storybooks to children in Cape Town, sponsored the creation of the Stanley L. Robbins Visiting Professorship at Brigham and Women's Hospital, and given funding to the 3rd International Conference on Children's Health and the Environment.