ELSEVIER

DNA fragmentation-based combinatorial approaches to soluble protein expression Part I. Generating DNA fragment libraries

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

[1] Section of Structural Biology, Institute of Cancer Research, Chester Beatty Laboratories, 237 Fulham Road, London SW3 6JB, United Kingdom
[2] Institute of Structural Molecular Biology, School of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX, 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 addressing a new drug discovery target, the generation of tractable protein substrates for functional and structural analyses can represent a significant hurdle. Traditional approaches rely on protein expression trials of multiple variants in various systems, frequently with limited success. The increasing knowledge base derived from genomics and structural proteomics initiatives assists the bioinformatics-led design of these experiments. Nevertheless, for many eukaryotic polypeptides, particularly those with relatively few homologues, the generation of useful protein products can still be a major challenge. This review describes the basis of efforts to forge an alternative 'domain-hunting' paradigm, based upon combinatorial sampling of expression construct libraries derived by fragmentation of the encoding DNA template, namely the methods and considerations in generating fragment length DNA from target genes. An accompanying review focuses upon the expression screening of such combinatorial DNA libraries for the sampling of the corresponding set of protein fragments.

It has been an implicit, if not explicit, assumption that the undertaking of the Human Genome Project and other efforts to catalogue the genes of pathogenic organisms will drive a major expansion of the prescription pharmaceutical market [1–5]. One basis for this prediction is that recent analyses show that the majority of present-day drugs act upon a rather small number (perhaps <1000) of distinct macromolecular targets (e.g. GPCRs, nuclear receptors, ion channels, proteases, kinases, phosphodiesterases and so on) [6]. Knowledge of the human genome sequence opens up the potential to explore many novel, perhaps rare, target types that have previously evaded identification by classical methods. Exploitation of these new targets will depend upon our ability to translate the emerging genomic information into tractable *in vitro* and cell-based assays of macromolecular function. In this context, within many areas of biomedicine, there is an increasing need to understand protein function at the atomic level, which implies having 3D structural information derived from X-ray crystallography and, where applicable, multi-dimensional solution NMR spectroscopy [7,8]. Both methods of analysis place a relatively high burden on the quantity, solubility, stability and 'foldedness' of the macromolecular analyte. An intrinsic bottleneck in such efforts is often the generation of recombinant, soluble, tractable protein materials that can be used for both inhibitor screening and structure-based drug discovery approaches. In general, one finds that proteins corresponding to whole open reading frames (ORFs) of cloned cDNAs can turn out to be difficult or impossible to produce in a facile manner. Therefore, a great deal of resource, both in academia and the commercial biotechnology and pharmaceutical sectors, is expended on efforts to obtain tractable fragments of such proteins in a paradigm that is led by bioinformatics-driven prediction of the likely stable globular domains, or by limited proteolysis of isolated full length proteins. In this review, we discuss an emerging principle that attempts to bypass these 'traditional' approaches, by appealing to high-throughput screening of DNA fragment libraries to identify stable, functionally and structurally tractable fragments of poly-

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

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2007.08.012

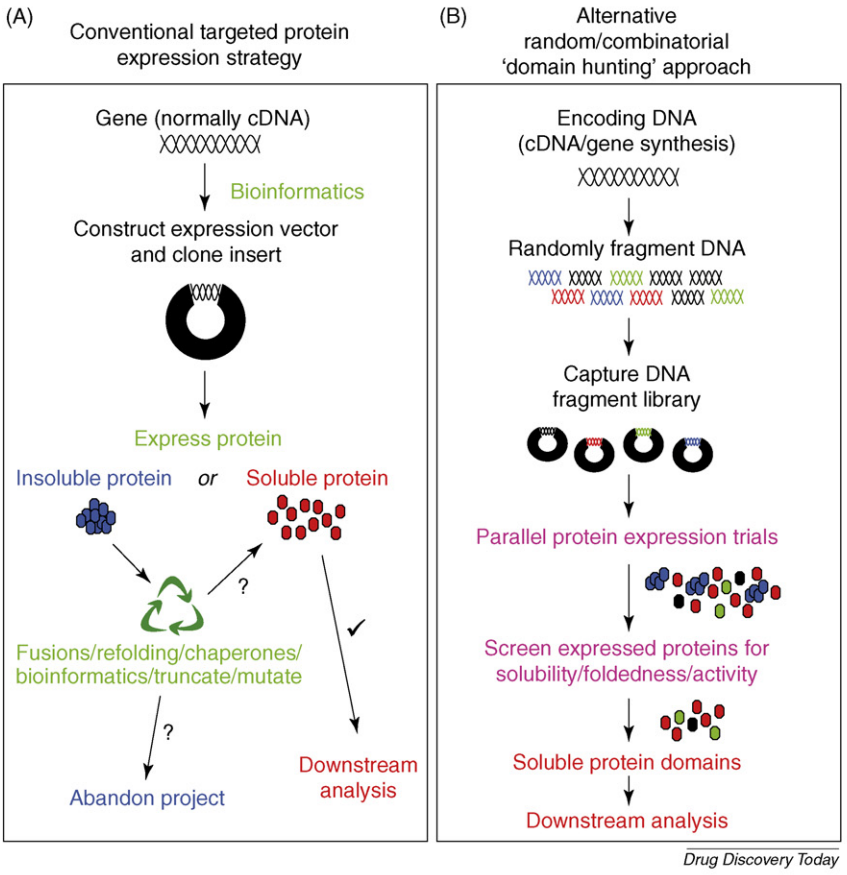

FIGURE 1

(A) A schematic capturing the conventional approach to targeted expression of protein products for structural and functional investigations. (B) Describes in outline the basic strategy adopted in protein expression screening approaches based upon random or combinatorially generated DNA fragments.

peptides, a process that has been variously known as 'domain footprinting' and 'combinatorial domain hunting' [9–14].

In the realm of recombinant protein production, particularly in a heterologous organism, the importance of selecting the optimal piece of the target polypeptide chain barely needs any exposition. It is the universal experience of protein expression scientists that:

(a) full length proteins, particularly from more complex eukaryotic genes, are difficult to express, unstable, or structurally intractable because of the overall size, intrinsic segmental flexibility, susceptibility to proteases, or requirements for obligate binding partners or post-translational modifications;
(b) despite the undoubted power of bioinformatic analyses to predict functional and often structural classification of the gene and encoded polypeptide, accurate prediction of structured domain boundaries is compromised exactly because the sequence identity in these boundaries is often lower than in the core regions; and
(c) even where domain boundaries can be estimated with some confidence, this does not always translate into successful generation of the corresponding protein product; it appears that the level of expression and stability of the final purified protein 'fragment' is a complex function of the choice of chain termini (amongst other factors), with possibly just a couple of amino acid residues either way making for success or failure.

Figure 1A describes a simplistic view of the traditionally employed approach to soluble protein expression. The diagram indicates that perhaps the majority of protein expression trials meet with some degree of failure, in that the target polypeptide is found to be insoluble or unstable. The protein chemist then resorts to many different alternative (non-linear, and frustratingly, often pseudo-circular) strategies to bypass the 'roadblock' to further investigations. This paradigm is contrasted with the 'upside-down logic' of a typical random domain sampling approach. Here, the aim is to isolate soluble sub-fragments of a given protein target, and the balance of effort is transferred from bespoke and a potentially futile 'local' search for stable protein fragments to a generic, potentially holistic screening process that in many aspects might be automated, or at minimum requires a lower level of experimental expertise for its execution.

The standard approach to maximise the chance of successfully procuring suitable quantities of the target protein is to alter the design of the target protein product with various combinations of silent and strategic mutagenesis including *de novo* gene design,

translational fusion strategies, alternative heterologous expression hosts, variation of the expression conditions, cell lysis protocols and protein solubilisation and purification routes [15–33]. Results from both our own and many other laboratories find that this naturally expansionist 'trial and error' approach quickly becomes a challenge to the available resources (including the scientist's morale) and can generate apparently stochastic outcomes. As a result of this frustration, we and others have considered whether there might be a more standardised, holistic and perhaps exhaustive solution to this general problem [9,10,12–14,34–36].

In some respects, this thinking is along similar lines to those engaged in structural proteomics (SP) endeavours. In SP, protein chemists set up a generic linear pipeline for construct design, cDNA cloning, protein expression and purification. At least in the early phase of SP programmes, the object was to cycle through input intact cDNA molecules and simply take the tractable protein products through to structural analysis. To first approximation the great hope was that SP practitioners would be kept busy with the proteins that emerged, harvesting the 'low-hanging fruit' of the accessible proteome, before moving onto the more difficult cases that require some sort of variation in the process; in an oversimplified nutshell: many cDNAs in → some 3D protein structures out (and effectively ignore the unsuccessful targets, at least for now).

For other structural biologists and their colleagues, the emphasis is somewhat different: here the target cDNA (or gene) is the be-all and end-all of the process. We want either one or more of the stable, soluble structured domains encoded by that stretch of DNA, and either we do not know where to start (because of the singleton nature of the corresponding amino acid sequence) or we want to find the predicted domains by another route than the standard (frustratingly challenging) one. Here, the aim is to vary the expressed DNA constructs to generate a library of expression constructs of arbitrary size, coupled with a process to sift out those stretches of DNA that produce the stably folded, globular fragments that we desire. This straightforward linear approach is illustrated schematically in Figure 1B.

The typical domain footprinting pipeline

The complete generalisation of the bioinformatics-blind expression screening concept that can be described in its most general sense as 'domain footprinting' is illustrated in more detail as a 10-step work-flow in Figure 2. Almost all steps of this pipeline are easily recognisable to any practising protein expression specialist; the focus of our attention here is that this procedure is not iterated (as in SP) over the template cDNA in Step 1, but rather over the systematically or randomly generated sub-fragments of a single DNA source that are obtained in Step 2. The implementation of Step 2 can, in principle, be achieved in many different ways, and different examples of the potential solutions are described briefly here. We particularly describe methods that lead to fragmentation of the template cDNA, as it were, at both ends; alternative strategies that target just one end of a target cDNA to generate deletion constructs with a constant 5′ or 3′ terminus represent a slightly different approach and, together with later stages of the work-flow implied in Figure 2, are considered in the second part of this review [53].

'Ordered' (systematic) DNA fragmentation

Here the DNA fragments of the library are designed *a priori* by the selection of construct ends:

'Primer pair walking'

A simple approach to Step 2 of Figure 2 would be multiple, presumably parallel PCR reactions using PCR oligonucleotide primers that are designed to generate specific PCR products that have sequence-defined 5′ and 3′ ends, and that can incorporate specific restriction sites for directional and in-frame capture into the expression vector of choice. Sequential variation of the priming co-ordinates of the individual PCR primers and combinatorial assortment of the sets of primer pairs can then be used to generate multiple candidate constructs, guaranteed to be in frame with any vector-encoded translational tag. These last features make such an approach a highly appealing strategy, but ultimately the sampling efficiency, defined as the degree to which the process can lead to protein products that differ by as little as a single residue, will be limited by the cost of oligonucleotide primers and the capacity to perform individual combinatorial PCR reactions in parallel. We are certain that such 'primer pair walking PCR' searches are often performed, but are not aware of any examples that report truly high-throughput screening applications.

'Random' DNA fragmentation

Under this heading we briefly describe a number of methods that can be used to generate DNA fragment libraries, each of which has little or no control over the nature of the DNA ends that are produced. These methods, therefore, have the potential advantage of fine-sampling the fragment space, but often at the expense of the fragments not being clonable in-frame and in a unique 5′–3′ orientation. The random fragmentation of DNA can be achieved either by physical, enzymatic or PCR-based methods. Although all methodologies described here are suitable for generating DNA fragments of size appropriate for the expression of typically sized protein products and of sufficiently high quality for expression library construction, each procedure differs in the particular set of advantages and limitations (see Table 1).

Physical methods

Physical methods that lead to DNA fragmentation (sonication, nebulisation and hydrodynamic shearing) are sequence-independent and, therefore, theoretically produce a more uniform and random distribution of DNA pieces than enzymatic methods, though end-repair to blunt the duplexes is usually required before cloning. Sonication generally requires large amounts of template DNA (10–100 μg). The fragments that are produced are distributed over a broad size range [37] and, consequently, only limited amounts of the fragmented DNA are suitable for cloning purposes. The DNA fragments must, therefore, be fractionated to obtain the desired type of protein product. A disadvantage of sonication is that the DNA may be damaged by hydroxyl radicals that are produced during cavitation [38]. Nebulisation generally requires smaller amounts of DNA (0.5–5 μg) [39,40]. The DNA concentration should be relatively low, as the technique requires relatively large solution volumes. The major advantage of nebulisation over sonication is that altering the pressure of the gas blowing through the nebuliser allows some control of the size distribution of frag-

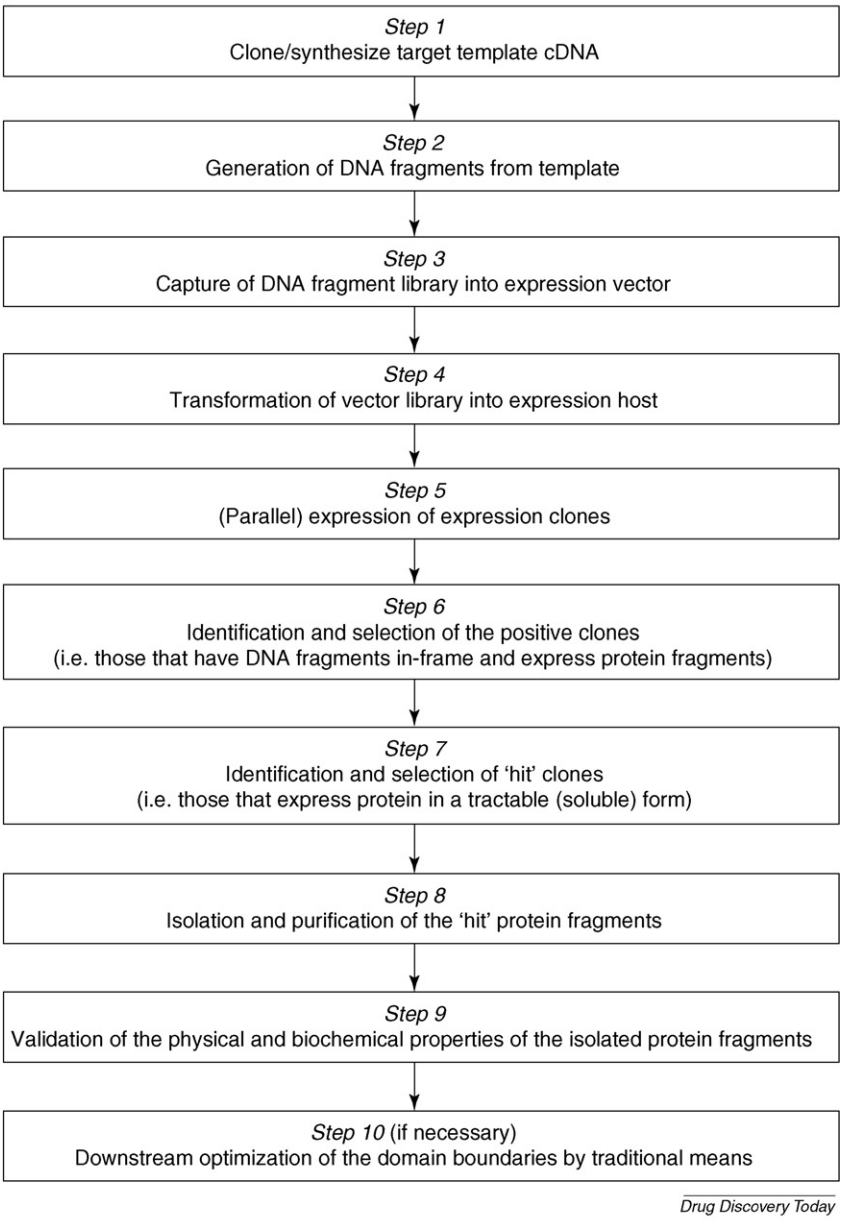

FIGURE 2

Conceptual representation of the 10-step generic work-flow for soluble protein domain 'footprinting', or domain hunting using a single template DNA as a starting point. Shotgun domain hunting is in principle possible using a collection of cDNA molecules (e.g. from a viral genome, or a cDNA library) as input. This review focuses particularly on Step 2 corresponding to the random fragmentation of DNA. The second part of this review [53] describes other aspects of the work-flow and describes specific examples of the application of domain hunting from the literature.

ments. Fragment sizes of 0.7–1.3 kbp can be readily obtained. Hydrodynamic shearing relies on the passage of a DNA solution, driven by a HPLC pump, through a hole of very small diameter. The process has recently been automated [41,42] and is often called the 'point-sink' flow system. Repeated passage through the small orifice results in DNA fragments of predictable size that is tailored by the flow rate and absolute size of the orifice. In fact, 90% of the fragmented DNA can fall in the required size range.

Enzymatic methods

A variety of enzymes can fragment or degrade DNA to smaller sizes than typically achieved with physical methods, either by internal cleavage of the dsDNA or by selective removal of bases at the 3′ or 5′ ends of DNA strands. Others can nick strands of dsDNA, which can then be followed by reactions that produce dsDNA breakage at the nicked sites. The main disadvantage of all enzymatic methods is that they are to some degree-dependent on the DNA sequence. Typically the DNA fragments that result are therefore less random than those resulting from physical methods. Size fractionation of the DNA following fragmentation is generally required, but the main advantage is that the DNA is generally of high quality and can often be used without further treatment such as DNA end-repair.

Restriction endonucleases have DNA sequence-specific recognition sites and, consequently, the fragments they produce are non-

TABLE 1

DNA fragmentation methods useful for the generation of DNA fragment libraries, detailing specific advantages and disadvantages

DNA fragmentation method	Advantages	Disadvantages
Primer pair walking	• Small amount of template required	• Costly oligonucleotide synthesis
	• Designed DNA fragmentation means that the ends of the DNA are precisely known and of high quality	• Time-consuming multiple PCR reactions
	• Directional cloning is possible	
	• DNA fragments can be systematically generated and are clonable in-frame with vector-encoded elements of the construct	
Physical methods (sonication, nebulisation, hydrodynamic shearing)	• Sequence-independent random method	• Large amounts of DNA may be required particularly for sonication
	• Straightforward implementation	• DNA fragments need end-repair
	• High yields from nebulisation and hydrodynamic shearing	• DNA damage can result with sonication
		• Size fractionation is required although nebulisation gives a narrow size distribution
		• Potential low efficiency cloning of blunt-ended DNA fragments
Restriction enzyme digestion	• Straightforward implementation	• Non-random and sequence-dependent DNA fragment ends
	• DNA fragments easily cloned and are of high quality	• Very few cut sites result from use of a single restriction enzyme
	• Directional cloning is possible	
DNase I digestion	• Straightforward implementation	• DNA fragments require end-repair
	• DNA cut sites are randomly distributed, though some sequence bias can occur	• DNA fragment size distribution can be broad
		• Size fractionation is required
		• Potential low efficiency cloning of blunt-ended DNA fragments
Tagged-PCR (T-PCR)	• Small amounts of template required	• Intermediate DNA purification step required
	• DNA fragments easily cloned and of high quality	• Sampling bias arises because of different efficiencies in annealing of the random PCR primers
	• Directional cloning is possible	
Combinatorial domain hunting (CDH)	• Simple one-step DNA fragmentation	• Cleavage sites biased to the position of A:T base pairs; synthetic gene synthesis can limit this bias by optimisation of the template A:T base pair composition
	• Fragmentation reaction goes to completion: no time-course dependence	• Size fractionation is required
	• Tunable end-point: DNA fragment size distribution varies only with the PCR step thymine:uracil ratio	• Use of Taq polymerase leads to risk of adventitious mutations
	• DNA ends of high quality	• Potential low efficiency cloning of blunt-ended DNA fragments

randomly distributed throughout the template. However, the DNA fragments generated are of high quality and can be used directly for cloning following a fractionation step that selects the desired fragment size [43]. Depending upon the choice of enzymes and the design of the cloning vector, one can arrange for the fragment cloning to yield directional insertion of the DNA fragments. The technique is quick and simple to use but its non-random nature and the relatively low frequency of cut sites limits the scope of their application.

Deoxyribonuclease I (DNase I) is the classical non-specific DNA endonuclease that can cleave a given DNA substrate in a manner that is often presumed to be without sequence bias. The precise output of DNase I cleavage depends on the specific conditions of the reaction. In the presence of Mn^{2+} ions DNase I cleaves both strands of the DNA duplex at approximately the same site [44–46]. However, in the presence of Mg^{2+} ions DNase I introduces nicks into either strand of the DNA in an independent fashion [46]. The former reaction conditions are more suitable in the context of the present applications in that the nicks on either side of the DNA are close together. However, the resulting DNA fragments require end-repair before cloning and subsequent fractionation of the library may be required since the resulting DNA fragment size distribution can be rather broad [47]. Furthermore, in spite of the statement above there is evidence that the local structure of the DNA can influence the efficiency at which DNase I can cleave the DNA.

It is well known that S1 nuclease will degrade both single-stranded RNA and DNA [48], but at high concentrations it will continuously decompose double-stranded RNA and DNA. Empirically, we and others find that moderate concentrations of the enzyme can usefully lead to complete cleavage at nicks or gaps in a dsDNA if these can be introduced by artificial means [49].

PCR-based DNA fragment generation

Sections of a DNA template can be efficiently generated using PCR amplification with tagged random sequence DNA primers ('tagged' PCR or T-PCR) [50]. In this procedure, the 'tagged'

primers are designed with a fixed region of *ca.* 15–20 bp of 5′ sequence non-complementary to any sequence in the template DNA appended with 9–15 bp of randomly encoded 3′ sequence. The method requires two or more PCR cycles. In the first PCR cycle, the random section of a given tagged primer has the opportunity to anneal to any cognate complementary sequence in the sequence in the template and the polymerase then extends the DNA as usual. In the second cycle, productive annealing of a second random tagged primer to the reaction products of the first PCR cycle will lead to the generation of DNA fragments terminated with two tag sequences. An intermediate purification step to remove unreacted tagged primers and primer dimers is then required before the amplification of the DNA fragments by secondary primers specifically complementary to the constant region of the tag primers. An advantage of T-PCR is that it requires only very small amounts of DNA template. The generated DNA fragments can then be cloned, either blunt-ended, or by the inclusion of restriction sites in the tagged section of the primer. A practical aspect of the T-PCR procedure is that the sampling profile of this method is presumably subject to potential bias, arising from differential annealing efficiency of G:C-rich versus A:T-rich random primer sequences and effects of differential secondary structure propensity within individual primers. In addition because two rounds of PCRs are required, variation in the efficiency of PCR amplification with template length is encountered twice, tending to lead to a bias of shorter DNA products compared to methods that require only a single PCR step (see below).

Based upon experience of researching base excision DNA repair, we have devised a method that we believe comes close to providing a means to fragment randomly a DNA template in a facile manner with very close to the theoretical limit of an absence of any sequence bias of the fragment DNA ends [13]. This is a PCR-based procedure that, combined with the more standard steps of clone selection, identification and validation, we have named Combinatorial Domain hunting (CDH). In CDH we perform an essentially standard PCR amplification of the target template using a non-proofreading DNA polymerase (e.g. Taq polymerase) in which the TTP component of the reaction mix is substituted with a TTP/dUTP mixture. Taq polymerase has the ability to incorporate a uracil in place of a thymine in the PCR product leading to a mixture of PCR product molecules which include essentially randomly distributed uracil bases at the frequency related to the input TTP/dUTP ratio (typically in the range 100:1 to 20:1). The uracil-doped amplified DNA is then subjected to uracil-DNA glycosylase, which cleanly excises the uracil bases, generating abasic sites that are then cleaved by endonuclease IV to generate single-strand nicks in the DNA. Application of S1 nuclease converts these nicks to double-strand breaks, yielding a library of blunt-ended PCR products that can be captured directly into an expression vector ready for screening. We have developed this procedure so that the enzymatic generation of the DNA fragments following the initial doped PCR can be conducted in a single-tube reaction that is simply allowed to reach an equilibrium end point. Compared to the T-PCR method one requires greater quantities of the template DNA, though this is not an economically limiting factor, since one can expand the bulk of the initial PCR reaction at will. Also the method does not suffer from a double dose of template length dependence of the two PCR amplification rounds inherent in T-PCR. Moreover, a very significant advantage of the CDH DNA fragmentation method, over most of the other procedures that rely on degradation of the DNA template, is that there is no time-dependent element to the process. The use of high-quality enzyme preparations means that the final distribution of the DNA fragments is directly determined only by the starting TTP/dUTP ratio. The sampling efficiency of the CDH process is determined only by the distribution of A:T base pairs in the DNA template, though we have noted that the intrinsic less-than-100% fidelity of the combined enzyme reactions means that the process can lead to DNA fragment generation in regions of the template corresponding to contiguous G:C base pairs. Nevertheless, a further degree of experimental control of the CDH process can be exercised by prior synthetic gene assembly of the template to approach the desired A:T base pair composition, a step that can also permit accounting for the preferred codon usage of the expression host (*E. coli* in all applications reported till date). The relatively low fidelity of Taq polymerase means that there is a risk of unexpected mutations in the generated DNA fragments, though experience suggests that this is not a major issue. Although one might expect mutations to be deleterious to protein folding, when mutation in a protein expression screening 'hit' is found it possible that the substitution has a positive influence on protein solubility. Indeed one can imagine wanting to reduce the fidelity of Taq to increase the mutation rate when dealing with very problematic protein targets. By contrast, it remains to be seen if application of recently commercialised high-fidelity thermostable DNA polymerases that can also incorporate uracil [51] will permit CDH without risk of unwanted mutations.

Fragment library cloning

The precise nature of the fragment library cloning step will depend upon the nature of the DNA ends. Both primer pair walking, T-PCR and restriction digest library generation procedures allow the potential for sticky-end ligase-dependent capture into the expression vector. The other physical and enzyme-based DNA fragmentation methods described tend to generate flush-ended DNA products that can be cloned either by blunt-ended ligase-dependent cloning, or by ligase-independent capture methods such as is common with commercially available topoisomerise-modified cut-vector products. Typically one requires the expression vector to provide translational start and stop codons, and coding sequence for a translational fusion peptide or protein that will be required for detection, selection and isolation of the expressed protein products.

Considerations of library size

Kawasaki and Inagaki [14] were the first to report a random approach to finding the soluble domains of a large protein (see [53]) in the process highlighting many of the salient issues that arise in approaching domain footprinting by random DNA fragmentation methods. Thus, they correctly state that for an intact protein of N amino acids, there are approximately $N^2/2$ (actually $N(N+1)/2$) possible sub-polypeptide products of arbitrary length P, $0 < P < N$. Nevertheless, at the nucleotide level the number of conceivable unique DNA fragments is much larger because of the tripartite nature of the codons, $\# = 3N(3N+1)/2$. However, this number significantly overestimates the number of experimentally sensible sub-constructs since one is not likely to be interested in any polypeptides shorter than a given length (say $P < 50$–100 residues), or longer than a particular defined or arbitrary limit.

By selecting DNA fragments from a range of DNA molecules with sizes between 3L (long) and 3S (short) base pairs (e.g. by excision from an agarose sizing gel), the corresponding number of unique DNA constructs is reduced to

$$= \frac{[(3N-3S)(3N-3S+1)-(3N-3L-1)(3N-3L)]}{2}.$$

For $3N = 2.1$ kbp (corresponding to a protein 'template' of $N = 700$ amino acid residues), and 3S = 300 and 3L = 750, appropriate for targeting polypeptide fragments between 100 and 250 residues in length, the potential unique DNA fragment space is ~710,000. Unfortunately the prospect that a given blunt-ended DNA fragment will be captured both in the correct orientation ($p = 0.5$) and in-frame with the 5′ start codon ($p = 0.333$) and any 3′-coded affinity appendage ($p = 0.333$) is 1-in-18. Thus, to sample exhaustively all possible truncated polypeptides one requires aiming for library sizes running to 10^6–10^7 independent clones. However, experience shows that exhaustive screening appears not to be required, since with much smaller libraries it has been possible to discover soluble protein fragments from large genes [9,13,52].

The accompanying review considers the application of protein expression screening procedures to libraries generated by methods such as those described above [53].

Summary

Alongside traditional, more systematic, approaches, random truncation of protein coding nucleotide sequences could well provide an alternative source of material to be used for domain footprinting. Rather than the necessity for high levels of target holo-protein to be assessed for its proteolytic lability and the stability of the resulting fragments, there is now the potential to attempt isolation of stable sub-domains from proteins which have not previously been expressed in a soluble form, if at all. The emphasis is on protein discovery, and truncated proteins have been shown to be tractable in overcoming bottlenecks in structural biology. Thus, methods to generate fragments of longer nucleic acids suddenly become tools for protein discovery and expression optimisation. The array of methods presented here may differ in their ultimate resolution and degree of bias in sampling, thus attention needs to be paid to the ultimate success of each method in providing stable, soluble protein material. Given that the difference between a soluble and insoluble protein, and between a protein that crystallises and one that does not, may be as little as one or a few amino acid residues at one or other terminus raises the bar for what will constitute the most reliable methods in endeavours in empirical domain identification. Part II of this review [53] examines recently applied methods in combinatorial protein expression screening that variously rely on the techniques for DNA fragmentation described herein.

Acknowledgements

We thank our colleagues Professor Laurence Pearl and Dr Keith Powell for comments on the manuscript and their input and critical thinking in all aspects of domain hunting.

References

1 Little, P.F. (2005) Structure and function of the human genome. *Genome Res.* 15, 1759–1766
2 Bentley, D.R. (2000) Decoding the human genome sequence. *Hum. Mol. Genet.* 9, 2353–2358
3 Bentley, D.R. (2004) Genomes for medicine. *Nature* 429, 440–445
4 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
5 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
6 Blundell, T.L. *et al.* (2006) Structural biology and bioinformatics in drug design: opportunities and challenges for target identification and lead discovery. *Philos. Trans. R. Soc. London B: Biol. Sci.* 361, 413–423
7 Blundell, T.L. *et al.* (2002) High-throughput crystallography for lead discovery in drug design. *Nat. Rev. Drug Discov.* 1, 45–54
8 Congreve, M. *et al.* (2005) Structural biology and drug discovery. *Drug Discov. Today* 10, 895–907
9 Jacobs, S.A. *et al.* (2005) Soluble domains of telomerase reverse transcriptase identified by high-throughput screening. *Protein Sci.* 14, 2051–2058
10 Christ, D. and Winter, G. (2006) Identification of protein domains by shotgun proteolysis. *J. Mol. Biol.* 358, 364–371
11 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
12 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
13 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
14 Kawasaki, M. and Inagaki, F. (2001) Random PCR-based screening for soluble domains using green fluorescent protein. *Biochem. Biophys. Res. Commun.* 280, 842–844
15 Knaust, R.K. and Nordlund, P. (2001) Screening for soluble expression of recombinant proteins in a 96-well format. *Anal. Biochem.* 297, 79–85
16 Stevens, R.C. (2000) Design of high-throughput methods of protein production for structural biology. *Structure* 8, R177–R185
17 Berrow, N.S. *et al.* (2006) Recombinant protein expression and solubility screening in *Escherichia coli*: a comparative study. *Acta Crystallogr. D: Biol. Crystallogr.* 62, 1218–1226
18 Scheich, C. *et al.* (2004) Fast identification of folded human protein domains expressed in *E. coli* suitable for structural analysis. *BMC Struct. Biol.* 4, 4
19 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
20 Busso, D. *et al.* (2003) Expression of soluble recombinant proteins in a cell-free system using a 96-well format. *J. Biochem. Biophys. Methods* 55, 233–240
21 Busso, D. *et al.* (2005) Structural genomics of eukaryotic targets at a laboratory scale. *J. Struct. Funct. Genomics* 6, 81–88
22 Vincentelli, R. *et al.* (2003) Medium-scale structural genomics: strategies for protein expression and crystallization. *Acc. Chem. Res.* 36, 165–172
23 Vincentelli, R. *et al.* (2004) High-throughput automated refolding screening of inclusion bodies. *Protein Sci.* 13, 2782–2792
24 Vincentelli, R. *et al.* (2005) Automated expression and solubility screening of His-tagged proteins in 96-well format. *Anal. Biochem.* 346, 77–84
25 Esposito, D. and Chatterjee, D.K. (2006) Enhancement of soluble protein expression through the use of fusion tags. *Curr. Opin. Biotechnol.* 17, 353–358
26 Hammarstrom, M. *et al.* (2002) Rapid screening for improved solubility of small human proteins produced as fusion proteins in *Escherichia coli*. *Protein Sci.* 11, 313–321
27 Hammarstrom, M. *et al.* (2006) Effect of N-terminal solubility enhancing fusion proteins on yield of purified target protein. *J. Struct. Funct. Genomics* 7, 1–14
28 Sorensen, H.P. and Mortensen, K.K. (2005) Soluble expression of recombinant proteins in the cytoplasm of *Escherichia coli*. *Microb. Cell Fact.* 4, 1
29 Sorensen, H.P. and Mortensen, K.K. (2005) Advanced genetic strategies for recombinant protein expression in *Escherichia coli*. *J. Biotechnol.* 115, 113–128
30 Doyle, S.A. *et al.* (2002) High-throughput proteomics: a flexible and efficient pipeline for protein production. *J. Proteome. Res.* 1, 531–536

31 Doyle, S.A. (2005) Screening for the expression of soluble recombinant protein in *Escherichia coli*. *Methods Mol. Biol.* 310, 115–121

32 Shih, Y.P. *et al.* (2002) High-throughput screening of soluble recombinant proteins. *Protein Sci.* 11, 1714–1719

33 Dyson, M.R. *et al.* (2004) Production of soluble mammalian proteins in *Escherichia coli*: identification of protein features that correlate with successful expression. *BMC Biotechnol.* 4, 32

34 Cornvik, T. *et al.* (2005) Colony filtration blot: a new screening method for soluble protein expression in *Escherichia coli*. *Nat. Methods* 2, 507–509

35 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

36 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

37 Deininger, P.L. (1983) Approaches to rapid DNA sequence analysis. *Anal. Biochem.* 135, 247–263

38 McKee, J.R. *et al.* (1977) Effects of ultrasound on nucleic acid bases. *Biochemistry* 16, 4651–4654

39 Bodenteich, A. *et al.* (1994) Shotgun cloning as the strategy of choice to generate templates for high-throughput dideoxynucleotide sequencing. In *Automated DNA Sequencing and Analysis* (Adams, M.D. *et al.* eds), Academic Press

40 Hengen, P.N. (1997) Shearing DNA for genomic library construction. *Trends Biochem. Sci.* 22, 273–274

41 Oefner, P.J. *et al.* (1996) Efficient random subcloning of DNA sheared in a recirculating point-sink flow system. *Nucleic Acids Res.* 24, 3879–3886

42 Thorstenson, Y.R. *et al.* (1998) An automated hydrodynamic process for controlled, unbiased DNA shearing. *Genome Res.* 8, 848–855

43 Maniatis, T. *et al.* (1982) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press

44 Campbell, V.W. and Jackson, D.A. (1980) The effect of divalent cations on the mode of action of DNase I. The initial reaction products produced from covalently closed circular DNA. *J. Biol. Chem.* 255, 3726–3735

45 Melgar, E. and Goldthwait, D.A. (1968) Deoxyribonucleic acid nucleases. II. The effects of metals on the mechanism of action of deoxyribonuclease I. *J. Biol. Chem.* 243, 4409–4416

46 Price, P.A. (1972) Characterization of Ca++ and Mg++ binding to bovine pancreatic deoxyribonuclease A. *J. Biol. Chem.* 247, 2895–2899

47 Anderson, S. (1981) Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Res.* 9, 3015–3027

48 Vogt, V.M. (1973) Purification and further properties of single-strand-specific nuclease from Aspergillus oryzae. *Eur. J. Biochem.* 33, 192–200

49 Kroeker, W.D. and Kowalski, D. (1978) Gene-sized pieces produced by digestion of linear duplex DNA with mung bean nuclease. *Biochemistry* 17, 3236–3243

50 Grothues, D. *et al.* (1993) PCR amplification of megabase DNA with tagged random primers (T-PCR). *Nucleic Acids Res.* 21, 1321–1322

51 Fogg, M.J. *et al.* (2002) Structural basis for uracil recognition by archaeal family B DNA polymerases. *Nat. Struct. Biol.* 9, 922–927

52 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

53 Savva, R., Prodromou, C. and Driscoll, P.C. (2007) DNA fragmentation-based combinatorial approaches to soluble protein expression. Part II. Library expression, screening, and scale-up. *Drug Discov. Today*, in press