Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Combinatorial library approaches for improving soluble protein expression in *Escherichia coli*

Darren J. Hart* and Franck Tarendeau

EMBL Grenoble Outstation, 6 Rue Jules Horowitz, 38042 Grenoble, France

Correspondence e-mail: hart@embl.fr

Received 27 June 2005
Accepted 2 November 2005

High-throughput screening methodologies are already used in structural biology to define efficient protein crystallization and expression conditions. Recently, screening approaches have been extended to the optimization of genetic constructs for improved soluble protein expression. With similarities to the directed evolution strategies used in protein engineering, a target gene encoding a poorly expressed protein is mutated by truncation, fragmentation or point mutation. Rare clones with improved protein expression characteristics are then isolated from the random library using a phenotypic screen or selection. This article reviews the progress in this field and provides a general overview of relevant mutation methods, screens and selections.

1. Introduction

In structural biology, both protein expression and crystallization processes contain large numbers of experimental variables that interact in complex ways and outcomes are often difficult or impossible to predict. In these situations, the desired result can be achieved by searching through a high diversity of possible solutions rather than by rational design. Screening for optimal protein-expression conditions for a specific open reading frame (ORF) is a well established approach. In contrast, the design of constructs for expression testing is usually performed in a rational manner based upon careful analysis of the primary sequence of the protein. Generation of combinatorial libraries of randomly mutated genetic constructs coupled to a phenotypic screen or selection, a process often termed 'directed evolution' (Fig. 1), has long been used in protein engineering for improving characteristics such as thermostability, enzyme specificity and binding-protein affinity. Attempts to apply directed evolution to the problem of heterologous protein expression through random mutation of genetic constructs have been described and will be reviewed here (see also Magliery & Regan, 2004; Roodveldt *et al.*, 2005) together with a general overview of library construction and screening methods (see also Aharoni *et al.*, 2005; Neylon, 2004).

2. Problems associated with protein production

Soluble proteins are required for most functional investigations including structural characterization by crystallography and NMR, high-throughput screening and protein biochemistry. Isolation of sufficient amounts of protein from native sources is often impossible owing to low yields, culturing difficulties of the organism and instability during purification.

Therefore, protein production from cloned recombinant DNA is normally the method of choice and *Escherichia coli* the preferred organism owing to ease of cloning, potentially high expression levels and simplicity of downstream processing (Baneyx, 1999). However, despite the attractiveness of this route, problems are commonly encountered in the expression of recombinant proteins, especially low expression levels and production of misfolded material (Baneyx & Mujacic, 2004). The latter may be proteolytically degraded or form insoluble inclusion bodies that can be easily isolated and sometimes (although unpredictably) refolded into active soluble material. Aggregation is also observed during purification, either as precipitation of solid material or formation of 'soluble aggregates' of high molecular weight that are visible during gel-filtration steps. Problems such as these create frequent insurmountable bottlenecks in projects and result in a high attrition rate of targets.

The advent of structural genomics, in which large sets of targets are cloned and expressed as part of a structure-determination pipeline (Burley, 2000), provides an apparent snapshot of the efficiencies of individual steps including cloning, expression, soluble expression, purification, crystallization, diffraction and structure solution. However, the figures extracted from target progression tables should be treated with caution since they are easily biased. One group analysed target progression data from all current structural genomics projects, identified biases and then quantified the target progression rates through the sequential steps of the structure pipeline using statistical techniques commonly applied in epidemiology (O'Toole *et al.*, 2004). In this work, cumulative occupancies of target states (*e.g.* cloned, expressed or purified) were used to calculate inter-step progression rates together with a 'survival analysis' that tracked the progress of targets over time. An example of a bias that affects the value for, say, cloning-to-expression or expression-to-purification, is the selection of easier targets (*e.g.* of bacterial origin) and exclusion of 'difficult' proteins (*e.g.* membrane proteins or those from complexes). This process of target selection (Brenner, 2000) is important to maximize the chances of success when working in a structural genomics context, but at the same time may result in some interesting proteins being ignored. Another bias is the abandonment of targets owing to successful structure determination of a homologue (and not owing to technical problems). From their statistical analyses, the authors conclude that there is an approximate 45% rate of progression through each step (cloning, expression, purification and crystallization). Therefore, averaged over the entire structural genomics data set, the success rate for obtaining soluble purifiable material from a clone is 0.45^2 or about 20%. This analysis demonstrates in a quantitative manner the experience of many individual researchers: that production of soluble purifiable protein remains a challenge for many targets.

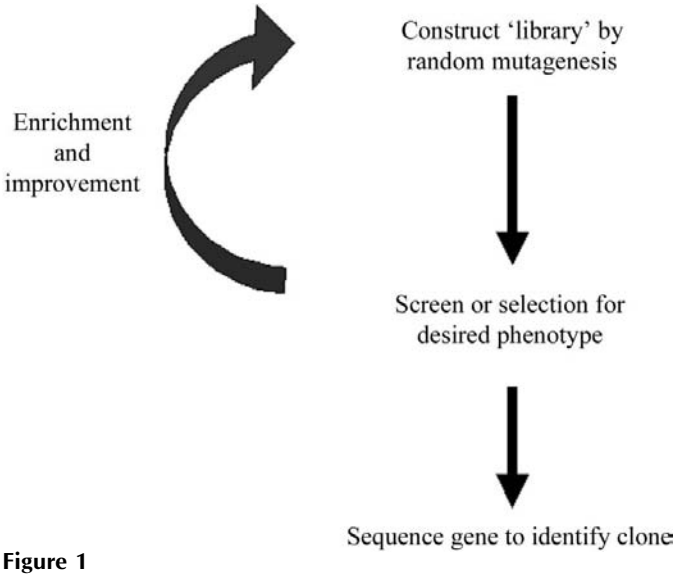

Figure 1
Schematic of the directed-evolution approach. A combinatorial library of randomly mutated genes is synthesized. Improved variants are identified *via* a high-throughput screen or selective process. Further improvements may be gained from iterative cycles of mutation and selection. Finally, clones are characterized by DNA sequencing to identify beneficial mutations.

3. Screening strategies for protein expression

The variables in a protein expression experiment can be divided into two types: genetic and environmental. Genetically encoded variables that affect protein expression include the sequence of the open reading frame, the choice of promoter, codon usage, mRNA secondary structure and addition of fused tags (Sorenson & Mortensen, 2004). These are typically manipulated by the researcher in a cloning step and are time-consuming. Environmental variables include host strain, growth medium and induction parameters, *e.g.* temperature, IPTG concentration and duration of induction step. In contrast to the genetic variables, they can be screened relatively quickly. When attempting to express a particular target, permutations of genetic and environmental conditions can be tested (Baneyx, 1999; Makrides, 1996) and dramatic and unpredictable differences are often observed. Whilst there are no absolute rules, it is often observed that good genetic constructs can express well after optimization of environmental conditions, whilst poor genetic constructs (*e.g.* with incorrectly defined domain boundaries) rarely express as desired, even after extensive screening of strains and conditions.

Many structural genomic projects are now in progress (see http://www.isgo.org for a list of current projects) and these have resulted in the tooling up of laboratories for automated work to increase the throughput of cloning and expression experiments. Robots and protocols for rapid synthesis of genetic constructs and screening of environmental conditions during protein-expression trials are now well established and accessible to many previously low-throughput structural biology laboratories. The usual application of this high experimental capacity has been to process many targets in parallel through a 'structure pipeline', either to obtain broad coverage of representatives of each protein fold or to understand structurally a themed set of proteins, *e.g.* from an

organism, an enzyme class or a biochemical pathway (Mittle & Gruetter, 2001). Generally, the genetic diversity is spread across many different targets, perhaps with only the full-length ORF and a few sub-constructs being synthesized and tested for each. Indeed, the actual set of genetic variants (truncated constructs or mutants) studied for any single target in a structural genomics project is relatively low and usually less than if the protein were being worked upon intensively by an individual scientist.

An obvious adaptation of the structural genomics approach is to focus the high experimental capacity afforded by its automated methods on individual targets of particular interest. Multiple constructs of an individual ORF are generated by PCR-based cloning and screened for improved soluble expression. These sub-constructs are often designed with the aim of expressing putative domains or multidomains and the challenge is to define their edges from the available information. Sequence-homology alignments, secondary-structure predictions and other computational tools are used for this task. Interesting examples of the latter include *DisEMBL* (Linding, Jensen *et al.*, 2003), *RONN* (Yang *et al.*, 2005), *GLOBPLOT* (Linding, Russell *et al.*, 2003) and *FOLDINDEX* (Jaakola *et al.*, 2005), which are predictors of disorder and/or globularity. Dozens of constructs might be designed based upon hypothesized domain-boundary positions, synthesized and screened for protein expression.

4. Combinatorial library-based strategies to optimize proteins

The strategy of screening randomly generated genetic diversity to obtain new or better phenotypes has ancient origins in plant or animal breeding. At the molecular level, random mutagenesis coupled with screening and the related field of directed evolution have been used successfully in many areas such as the improvement of industrial enzymes, *e.g.* enhancement of kinetics, thermostability or substrate specificity (Bornscheuer & Pohl, 2001; Chartrain *et al.*, 2000; Lehmann & Wyss, 2001; Tobin *et al.*, 2000), and in the generation of high-affinity binding proteins using technologies such as phage and ribosome display (Amstutz *et al.*, 2001; Kretzschmar & von Rueden, 2002) for use as biopharmaceuticals or research tools. One recurring conclusion when analysing the output of directed-evolution projects is that the solutions identified from random library approaches would have been unpredictable at the start or are different from (often better than) those that would have been rationally designed (Tobin *et al.*, 2000). Typically, the DNA encoding the target protein is mutated in some way such that new forms of protein are produced in the expression system being employed (Neylon, 2004). The improved phenotype is detected either by direct measurement (*e.g.* high-throughput enzyme-activity screens) or by linking the desired phenotype to observable traits such as survival in the presence of toxic chemicals (*e.g.* antibiotics), change of host cell colour or separation of positives by selective binding (*e.g.* phage, plasmid and ribosome display technologies). The mutation strategies that are available are fairly generic in that they may be applied to any gene irrespective of its function. In contrast, the screening or selection process is usually tightly linked to the activity of the encoded protein and it is often a challenge to devise an efficient approach (Olsen *et al.*, 2000).

Combinatorial library methods are starting to be applied to problems in structural biology and are ideally suited to the area of protein production. Indeed, there is a clear overlap between protein expression and protein engineering as areas for study by directed evolution since expression of correctly folded molecules is implicitly required prior to identification of a new activity. In fact, since solutions to an expression problem are often more abundant than those required for specifically altering binding or catalysis, improving protein expression is likely to emerge as a simpler problem than, say, remodelling an enzyme active site or generating high-affinity antibodies, both of which are commonly achieved.

Methods to generate diversity in a target protein are well established and are reviewed below (see also Neylon, 2004). Before starting a mutagenesis program for improved protein expression, it is prudent to establish an effective screen or selection for the desired phenotype. Possible approaches to detecting improvements in soluble expression are described

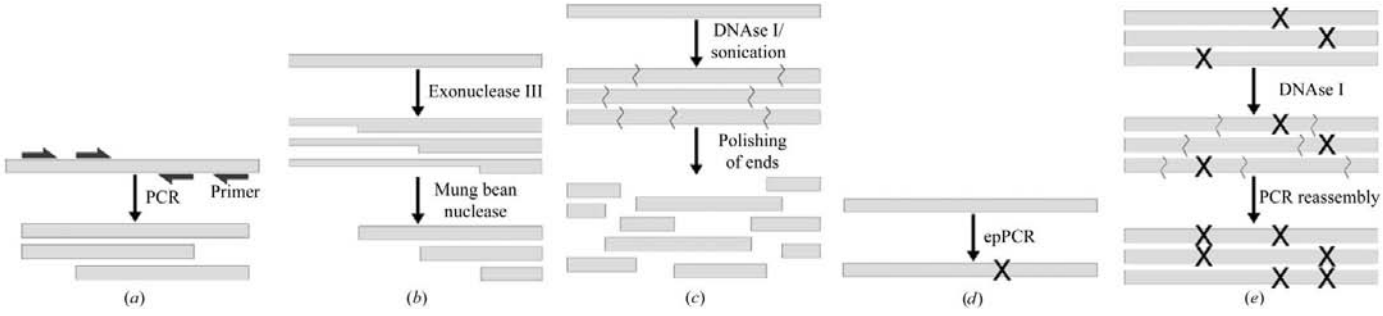

Figure 2
DNA truncation and mutagenesis methods. (*a*) Generation of specific constructs by PCR, (*b*) unidirectional DNA truncation using exonuclease III and nuclease (mung bean or S1), (*c*) fragmentation using enzymatic or physical breakage of the DNA, (*d*) random point mutation by error-prone PCR (epPCR) or bacterial mutator strain and (*e*) DNA shuffling for combining different lineages of point mutations on the same DNA strand.

after the following section on mutagenesis, but the important point is that the mutation and screening strategies need to be considered together.

5. Mutation strategy

The mutation methods available to the researcher mirror those that underlie natural evolution. Either the gene can be mutated at individual positions, resulting in a full-length protein containing amino-acid substitutions, or regions of the reading frame can be deleted to generate shorter constructs encoding single or multidomain fragments. Since the underlying cause of a protein-expression problem is often unclear at the outset, it may be difficult to predict the most appropriate strategy; however, the following suggestions may provide a starting point in selecting a mutation strategy.

(i) Single-domain proteins may be better suited to point mutagenesis than truncation.

(ii) Multidomain proteins are perhaps most likely to yield soluble material by truncation.

(iii) Expression problems arising from unfolded or disordered regions of a multidomain protein are probably best addressed by truncation strategies than point mutagenesis.

(iv) Components of protein complexes expressed in isolation may benefit from truncation to remove interaction regions that are unstable or disordered in the absence of their partner.

The issue of library size and diversity is also worth considering. Truncation libraries are relatively small and much, if not all, of the possible diversity can be synthesized and tested. In contrast, point-mutation libraries are huge and only a small proportion of the possible diversity can be sampled.

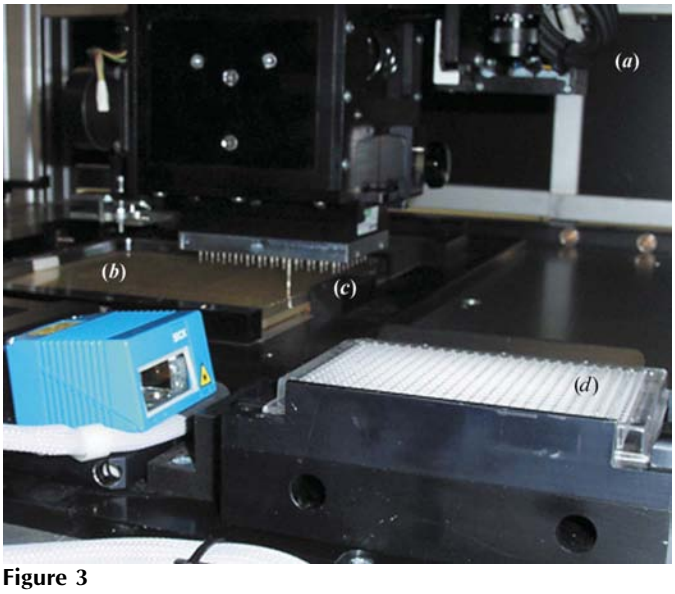

Figure 3
A colony-picking robot uses a video camera (*a*) to map the positions of colonies plated out on agar trays (*b*). 2500 single colonies are picked per hour with a pneumatic pin tool (*c*) into the wells of microtitre plates filled with growth medium (*d*). In this way, members of a random library can be spatially separated and each clone provided with a unique identification number.

5.1. Truncated constructs

Expression of domains or multidomains using PCR to generate truncated forms of genes is a well established method for obtaining soluble protein (Fig. 2*a*). The success of this approach is highly dependent on defining domain boundaries for design of expression constructs and if the positions of these boundaries cannot be predicted successfully, *e.g.* using sequence alignments or bioinformatic tools, expression of a target protein may fail. In these cases, the systematic truncation of the gene by positioning of PCR primers every few residues, coupled with protein-expression testing, can provide a way forward. This has its limitations since large numbers of PCR cloning experiments are expensive owing to the cost of reagents (*e.g.* primer synthesis, proof-reading polymerases, cloning reagents such as Gateway enzymes and DNA sequencing). The handling and tracking of multiple PCR cloning experiments can also become logistically complex. Some of these technical limitations have been overcome by applying cloning and expression platforms developed for structural genomics, although cost and limited throughput means that this approach is not suitable for large numbers of targets.

5.2. Unidirectional truncations

One highly efficient and economical solution to generating large numbers of gene truncations is to replace primer-dependent PCR strategies with the generation of gene fragments by enzymatic digest of the full-length DNA (Fig. 2*b*). Unidirectional truncations can be generated effectively using exonuclease III, with subsequent removal of the remaining single strand using mung bean or S1 nucleases (Henikoff, 1984). Commercial kits for this procedure are available including the Erase-a-Base Kit from Promega and the ExoIII/S1 Deletion Kit from Fermentas. In our hands, protocols published for directed-evolution experiments also work well (Ostermeier & Lutz, 2003). Unidirectional deletions are achieved either by blocking one end with a 3′ overhang or by filling in 5′ extensions with α-S-dNTPs which generate ends resistant to exonuclease-catalysed hydrolysis (Putney *et al.*, 1981). These fragments can be size fractionated by agarose electrophoresis prior to cloning and transformation. Recently, bacterial expression of a potato mop-top virus protein was improved by progressive deletion of its N-terminus (Pecenkova *et al.*, 2005) using a simple screen employing blotting of colonies of transformants followed by detection of protein expression using an antibody against the fused hexahistidine tag. A few clones showed improved levels of expression, although they were mostly insoluble.

While such sets of nested deletions are simple to produce and far cheaper than primer-dependent PCR for the equivalent level of sequence coverage, there are several significant differences. The first is that, unlike PCR-cloned constructs, two-thirds are out of frame with the terminus that must be added during cloning. In the case of 5′ or 3′ deletions where a peptide tag is fused, this is unavoidable. A workaround can be found in the case of 3′ deleted genes where no fusion tag is

required by ligation into a vector that fuses three stop codons in different frames, although this may add one or two extra amino acids to the C-terminus of the protein. A second difference is that the reaction is performed in a single tube that then contains all the truncated genes. Upon cloning, it is necessary to clonally separate these, usually by colony picking. For small libraries of up to 1000 clones this can be performed by hand, but for large libraries a colony-picking robot is required (Fig. 3). In our laboratory, we routinely pick 30 000 colonies into microtitre plates for further analysis; such a number would be impossible by hand and even experiments in the low thousands risk contamination. Thirdly, during PCR cloning, the identity of the construct is known from the outset of the experiment from the primers used. Thus, sample tracking is logical since the identity and position of this construct is known from start to finish. In contrast, random library experiments generate an output that, even after clonal separation into plates, is uncharacterized until the DNA sequence is obtained. Since sequencing is expensive, constructs are analysed with little knowledge of their identity until the end of the experiment when positive clones are sequenced. Fourthly, if the aim is to comprehensively sample the diversity in a random library, it becomes necessary to oversample in order to obtain a reasonable confidence of having achieved coverage. For example, a 3000-base-pair gene can give rise to 3000 unidirectional nucleotide truncations (either 5′ or 3′) and, when expressed, 1000 will be in frame with start or stop signals or vector-borne tags. However, more than 3000 constructs must be tested to observe all expression constructs within certain confidence limits. Since experiments contain many small biases that are difficult or impossible to measure, we aim for a fivefold to tenfold oversample (of each nucleotide position) and usually achieve sixfold to sevenfold by the end of the expression-testing procedure. This level of oversampling should compensate for effects such as an imperfect distribution of truncation lengths, any sequence biases of exonucleases during the truncation digest and any inefficiencies encountered during automated steps. In downstream processes, the oversample can be reduced or eliminated as required by proceeding with only a fraction of the clones.

5.3. Gene fragmentation

The next level of complexity from the point of view of library size is the generation of gene fragments (Fig. 2*c*). To calculate the library size required when both ends of a reading frame are varied, the number of polypeptide fragments of a protein of N amino acids is approximately $N^2/2$ and, since the DNA insert must be in frame at both ends and in the correct orientation, it is necessary to multiply this by 18 to obtain the number of ligation products required to encode this diversity (Kawasaki & Inagaki, 2001). Thus, a library encoding all fragments of a 500-amino-acid protein will comprise 2.25 million clones. It is not a trivial issue to clone this number of DNA inserts into a plasmid backbone and a very high throughput screen or life-or-death selection strategy is then required to analyse this number of clones.

Several methods have been developed for generating fragmentation libraries, usually for large-scale DNA sequencing purposes, although there is no reason why the fragments could not be tested for expression. These include digestion with DNAseI (Anderson, 1981), tagged PCR (T-PCR; Grothues *et al.*, 1993), physical fragmentation by sonication (Deininger, 1983) and point-sink methods whereby DNA is sheared by forcing it through small apertures (Oefner *et al.*, 1996; Thorstenson *et al.*, 1998). DNAseI normally cuts single strands of a DNA template in the presence of Mg^{2+}, but substitution to Mn^{2+} results in double-stranded breaks leaving ends that, after polishing with a proofreading polymerase, can be cloned. A recent patent application describes a method in which uracil is incorporated into the gene by PCR in a highly controllable manner and uracil-DNA glycosylase and apurinic/apyrimidinic endonuclease subsequently used to nick the DNA at that position (McAlister *et al.*, 2003). It is claimed that this results in an even and tunable distribution of cut positions and therefore a high-quality unbiased library of DNA fragments that the authors use for domain screening.

A successful example of screening a fragmentation library for soluble protein has been described (Kawasaki & Inagaki, 2001) in which T-PCR generated fragments of the gene *vav* were cloned as fusions with GFP. Analysis of green fluorescing clones yielded information on domain architecture of the Vav protein and resulted in soluble protein domains for further structural characterization. This method was also used to identify soluble domains of telomerase reverse transcriptase (Jacobs *et al.*, 2005).

5.4. Point-mutation libraries and DNA shuffling

A number of methods can be used to introduce random point mutations into reading frames (reviewed in Neylon, 2004). Most methods use PCR in some way, but a notable exception is the use of mutator strains of *E. coli* (Nguyen & Daugherty, 2003) for low-fidelity propagation of plasmids, *e.g.* XL1-red (Stratagene), that are deficient in enzymes of the DNA-repair pathways (*mutS*, *mutD* and *mutT*). Error-prone PCR (epPCR) methods are designed to introduce errors by misincorporation of bases during amplification of a target sequence (Fig. 2*d*). This is achieved either by manipulating the salt conditions, *e.g.* use of elevated concentrations of $MgCl_2$ (Cadwell & Joyce, 1994) or $MnCl_2$ (Cirino *et al.*, 2003), use of mutant polymerases with lower fidelity, *e.g.* Mutazyme from Stratagene, or nucleotide analogues that promote transitions after incorporation (Zaccolo *et al.*, 1996). It is important to realise that most mutagenic methods are usually biased in one or more ways. Firstly, different methods favour different types of base substitution that will clearly affect the frequency of some codons over others. Secondly, mutations will be propagated unevenly within the collection of PCR products depending on whether they occur in early or late cycles of the amplification reaction. Thirdly, there is also a significant bias in the amino-acid sequence space that can be explored by point

mutation since a codon can only be mutated into a subset of others by a single-nucleotide change. For example, a valine codon can only be converted into a phenylalanine, leucine, isoleucine, alanine, aspartate or glycine by a single-point mutation. Therefore, a library generated by these methods will not be truly random regarding the amino-acid distribution in the resulting proteins.

A significant advance over simple point-mutagenesis techniques is DNA shuffling (Fig. 2*e*) that introduces an *in vitro* recombination-like step to mix PCR-generated mutations between lineages of mutated DNA strands (Stemmer, 1994). DNA shuffling also permits a process analogous to 'backcrossing', as used in plant or animal breeding, in which a mutated ORF is recombined with the original gene, but still under conditions of selective pressure, to dilute out silent or deleterious mutations that accompany those that are beneficial.

An early example of the power of this method was provided by an experiment to improve the fluorescent properties of GFP in *E. coli* by DNA shuffling (Crameri *et al.*, 1996). Clones with improved fluorescence were identified by transillumination of colonies on agar over a standard 360 nm light box and the basis of this improvement was revealed by sequencing of the *gfp* gene. The level of protein expression was unchanged from the wild type; however, the protein was highly soluble compared with the original. It was hypothesized that the improvement observed was largely a consequence of the substitution of three surface-borne solvent-exposed hydrophobic amino acids for more hydrophilic residues. This region was thought to form the protein-interaction interface between GFP and a second protein, aequorin, in the native jellyfish host. This example might highlight a more general route to the study of single components of some protein complexes in which stable expression could be achieved through mutation of interfacial regions.

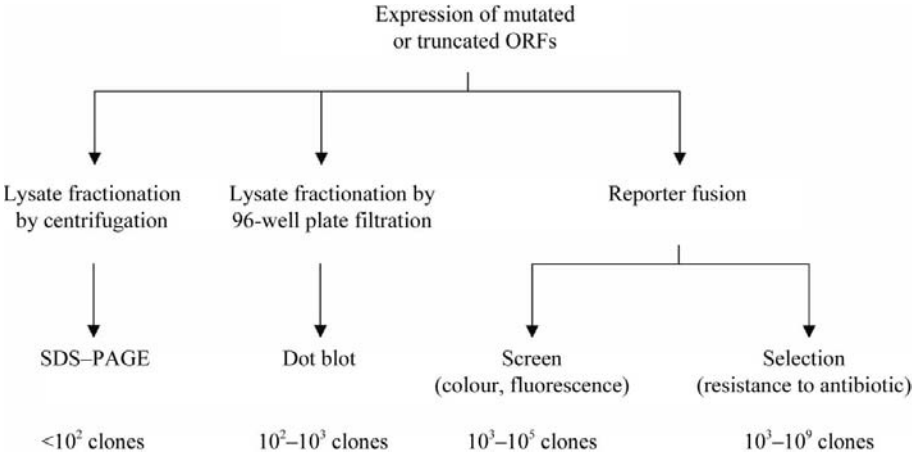

Figure 4
Overview of different methods of assaying for protein solubility. Low-throughput studies typically involve fractionation of expression lysates by centrifugation and SDS–PAGE. Increased throughput can be achieved by fractionating lysates using a filter plate and analysis of fractions by dot blot using an antibody against a fused epitope. For larger libraries, cell-based screens or selections employing reporter fusions can be used in which the solubility of the protein is coupled to the phenotype of the fusion partner. Approximate numbers of clones that can be measured are indicated.

Family shuffling is a variant of DNA shuffling that allows genetic homologues of a target to be shuffled together to form folded chimeric proteins encoded by many short regions of the different parental ORFs spliced together (Crameri *et al.*, 1998). The advantage of shuffling homologues over point mutants of the same gene (DNA shuffling) is that regions of a gene are exchanged for those encoding chemically similar polypeptides from a second protein. The probability of avoiding deleterious mutations is therefore higher and so folded proteins may be obtained by screening smaller libraries. Such an approach was used to achieve soluble expression in *E. coli* of paraoxanase enzymes by shuffling together homologues of human, mouse, rat and rabbit *PON* genes and screening of *E. coli* colonies that exhibited improved esterolytic activity (Aharoni *et al.*, 2004). The structure of a protein derived from a mutant clone was then solved, revealing a six-bladed β-propeller (Harel *et al.*, 2004).

6. Screens and selections for improved protein expression

Described above are various approaches for generating genetic diversity in a target ORF. A high-throughput compatible strategy for identifying rare variants with improved soluble yield is then required to find the 'needle in the haystack'. These can be loosely classified into true screens, in which all members are observed and scored, and selections, in which only those that resist a selective pressure (such as an antibiotic or affinity purification) survive to be measured. An important consideration when designing a screen or selection is that its throughput must be high enough to analyse the diversity required to yield a solution. However, since it is usually impossible to estimate the number of possible solutions and since the sizes of random libraries can be huge, it is generally the case that the more clones that can be screened, the better the chance of finding an answer. In many directed-evolution experiments it is only ever possible to undersample the diversity of mutants owing to the vast library size, although in the case of unidirectional truncations the smaller libraries may permit screening of all clones or even a several-fold oversample.

There are a number of methods for assaying soluble expression of proteins with different levels of throughput (Fig. 4). When working with low numbers of clones by hand, the method of choice is physical fractionation of lysates by centrifugation followed by visualization of soluble and total fractions by SDS–PAGE. The use of filter plates and a vacuum manifold to fractionate lysates provides an alternative to centrifugation and permits the assay to be automated on a liquid-handling

robot (Knaust & Nordlund, 2001). The throughput of such a method is limited only by the expense of the filter plates and the preparation of bacterial cultures, but it should be capable of handling thousands of clones. Since analysis by SDS–PAGE of this number of fractionated samples is not feasible, the authors used a dot blot format with antibody detection of histidine-tagged proteins. Whilst no information is provided about the size or quality of the proteins at this step, SDS–PAGE and Western blotting can be used downstream on the reduced number of positive clones identified from the dot blot. The same authors have recently reported a 'colony filtration' version of this approach that significantly increases the throughput (Cornvik *et al.*, 2005).

In some cases, usually an enzyme or ligand-binding protein, it may be possible to assay directly the function of the protein being studied. Direct activity assays provide the most rigorous route to identifying correctly folded soluble material since a wild-type protein conformation is normally required for function. The exact nature of the screen and therefore its throughput will be on an individual basis and a major disadvantage of assaying by activity is that many proteins or their domains do not have a measurable function. Such a strategy was used successfully in the directed evolution of soluble variants of paraoxonases by DNA shuffling (described above) that interestingly used not the primary activity of the protein which is unknown, but promiscuous esterase and phosphotriesterase activities (Aharoni *et al.*, 2004). A colour-based colony screen of 10^3–10^4 clones per round of shuffling identified active folded proteins. Colony screens, where possible, provide a high-density format capable of screening up to 10^6 constructs individually since it avoids the logistical problems associated with growing and inducing clones in individual containers, *e.g.* tubes or multiwell plates. Sometimes the activity of the protein being studied can be linked to the survival of the host organism, for example, by complementation of a mutation an auxotrophic strain of *E. coli* or when studying specific problems such as proteins responsible for antibiotic resistance mechanisms (Crameri *et al.*, 1998; Stemmer, 1994). In these cases, much larger libraries can be assayed since plating density is not limiting in contrast to colony screens.

Where libraries are too large to screen clones by lysate fractionation, or where the activity of the target protein is not easily assayable, an attractive approach is that of fusion-protein reporters in which the solubility of the mutated construct is assayed *via* the observable phenotype of a fused enzyme or fluorescent protein. Probably the best known of these strategies uses GFP (Waldo *et al.*, 1999). Here, insoluble fusion constructs result in aggregation and inhibition of GFP fluorophore maturation that can be observed at the cellular level through lack of cell or colony fluorescence, whilst mutated ORFs that express soluble GFP-fusion proteins result in green cells. This was used to improve the soluble expression of methyl transferase, tartrate dehydratase β-subunit and nucleoside diphosphate kinase from *Pyrobaculum aerophilum* to 50, 95 and 90%, respectively (Pedelacq *et al.*, 2002) from variants generated by DNA shuffling. The soluble kinase variant was then crystallized and its structure solved. In another study, the GFP-fusion method was used to screen a *vav* gene-fragmentation library generated by T-PCR (Kawasaki & Inagaki, 2001). Another interesting application of this system was demonstrated in which the GFP-folding reporter was used to monitor for solubilization of insoluble integration host factor β by coexpression of integration host factor α (Wang & Chong, 2003). Complex formation *in vivo* stabilized the integration host factor β fusion protein and produced detectable fluorescence. An approach such as this could be used to screen a combinatorial library of partner proteins to identify stabilizing interactors *via* formation of complexes. Life-or-death selections have been configured in a similar way to the GFP screen using fusions to chloramphenicol acetyl transferase (CAT; Maxwell *et al.*, 1999). Unlike monomeric GFP, CAT is an obligate trimer that may lead to unpredictable higher order complexes with unpredictable effects on solubility. However, the powerful selection offered by CAT fusion permits larger libraries to be analysed than by GFP fusion.

A disadvantage of fusing targets onto large reporter proteins is that the solubility may be perturbed in a similar way to that observed when fusing to maltose-binding protein or other large affinity tags. This may result in false positive clones that are soluble only in the fused form or false negatives that are insoluble as fusions. One method to circumvent this is the fusion of small peptides that complement a folding mutant of a coexpressed partner protein. In this way, protein folding and solubility of the target protein–peptide fusion is monitored by activation of a misfolded signalling protein. The well known LacZα complementation system of β-galactosidase (Ullmann *et al.*, 1967) has been modified for this purpose using a C-terminal LacZα peptide tag (Wigley *et al.*, 2001). Soluble expression of a fusion protein was shown to complement the inactive cytoplasmically expressed ω-fragment of β-galactosidase and permit blue–white screening. A recent development of the GFP-folding screen has been described (Cabantous *et al.*, 2004) with similarities to the LacZα system. In this work, GFP, an 11-stranded β-barrel, was divided to form split protein comprising a non-fluorescent ten-β-strand GFP fragment and a peptide corresponding to the eleventh β-strand. The solubility of both components was optimized using directed evolution to generate the final working system. When fused to a soluble protein, the peptide complemented the non-fluorescent GFP fragment, resulting in a protein-folding event and cyclization of the GFP fluorophore.

7. Conclusions

Combinatorial library strategies for improving protein solubility are being developed as tools for the structural biologist and these offer promising solutions to the production of difficult targets. These methods are well established in other fields, such as the improvement of industrial enzymes, and should transfer readily to problems in protein expression for structural biology. A mutagenic method is first used to

generate diversity in the target gene and then a screen or selective process identifies the rare variants with improved protein-expression characteristics. Random mutation of the full-length gene and generation of DNA fragments are both strategies that can provide solutions to protein insolubility. An effective screen or selection process for identification of soluble constructs is required and this can either exploit an assayable activity of the target, as with an enzymatic activity screen, or may use a more generic readout such as a reporter fusion. The aim of the screen or selection is to rapidly reduce the large numbers of clones from a random library to a level where constructs can be assayed directly for solubility, usually by hand, using lysate fractionation and analysis by SDS–PAGE.

DJH wishes to thank the organisers of the CCP4 Study Weekend 2005 for the invitation to present this subject and the EC-funded Fifth Framework Program, SPINE for funding FT.

References

Aharoni, A., Gaidukov, L., Yagur, S., Toker, L., Silman, I. & Tawfik, D. S. (2004). *Proc. Natl Acad. Sci. USA*, **101**, 482–487.

Aharoni, A., Griffiths, A. D. & Tawfik, D. S. (2005). *Curr. Opin. Chem. Biol.* **9**, 210–216.

Amstutz, P., Forrer, P., Zahnd, C. & Plueckthun, A. (2001). *Curr. Opin. Chem. Biol.* **12**, 400–405.

Anderson, S. (1981). *Nucleic Acids Res.* **9**, 3015–3027.

Baneyx, F. (1999). *Curr. Opin. Biotechnol.* **10**, 411–421.

Baneyx, F. & Mujacic, M. (2004). *Nature Biotechnol.* **22**, 1399–1408.

Bornscheuer, U. T. & Pohl, M. (2001). *Curr. Opin. Chem. Biol.* **5**, 137–143.

Brenner, S. E. (2000). *Nature Struct. Biol.* **7**, *Suppl.*, 967–969.

Burley, S. K. (2000). *Nature Struct. Biol.* **7**, *Suppl.*, 932–934.

Cabantous, S., Terwilliger, T. C. & Waldo, G. S. (2004). *Nature Biotechnol.* **23**, 102–107.

Cadwell, R. C. & Joyce, G. F. (1994). *PCR Methods Appl.* **3**, S136–S140.

Chartrain, M., Salmon, P. M., Robinson, D. K. & Buckland, B. C. (2000). *Curr. Opin. Biotechnol.* **11**, 209–214.

Cirino, P. C., Mayer, K. M. & Umeno, D. (2003). *Methods Mol. Biol.* **231**, 3–9.

Cornvik, T., Dahlroth, S., Magnusdottir, A., Herman, M. D., Knaust, R., Ekberg, M. & Nordlund, P. (2005). *Nature Methods*, **2**, 507–509.

Crameri, A., Raillard, S. A., Bermudez, E. & Stemmer, W. P. (1998). *Nature (London)*, **391**, 288–291.

Crameri, A., Whitehorn, E. A. & Stemmer, W. P. (1996). *Nature Biotechnol.* **14**, 315–319.

Deininger, P. L. (1983). *Anal. Biochem.* **129**, 216–223.

Grothues, D., Cantor, C. R. & Smith, C. L. (1993). *Nucleic Acids Res.* **21**, 1321–1322.

Harel, M., Aharoni, A., Gaidukov, L., Brumshtein, B., Khersonsky, O., Meged, R., Dvir, H., Ravelli, R. B., McCarthy, A., Toker, L., Silman, I., Sussman, J. L. & Tawfik, D. S. (2004). *Nature Struct. Mol. Biol.* **11**, 412–419.

Henikoff, S. (1984). *Gene*, **28**, 351–359.

Jaakola, V. P., Prilusky, J., Sussman, J. L. & Goldman, A. (2005). *Protein Eng. Des. Sel.* **18**, 103–110.

Jacobs, S. A., Podell, E. R., Wuttke, D. S. & Cech, T. R. (2005). *Protein Sci.* **14**, 2051–2058.

Kawasaki, M. & Inagaki, F. (2001). *Biochem. Biophys. Res. Commun.* **280**, 842–844.

Knaust, R. K. C. & Nordlund, P. (2001). *Anal. Biochem.* **297**, 79–85.

Kretzschmar, T. & von Rueden, T. (2002). *Curr. Opin. Biotechnol.* **13**, 598–602.

Lehmann, M. & Wyss, M. (2001). *Curr. Opin. Biotechnol.* **12**, 371–375.

Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J. & Russell, R. B. (2003). *Structure*, **11**, 1453–1459.

Linding, R., Russell, R. B., Neduva, V. & Gibson, T. J. (2003). *Nucleic Acids Res.* **31**, 3701–3708.

McAlister, M., Savva, R., Pearl, L. H., Chrisostomos, P. & Driscoll, P. (2003). Patent WO03 040 391.

Magliery, T. J. & Regan, L. (2004). *Eur. J. Biochem.* **271**, 1595–1608.

Makrides, S. C. (1996). *Microbiol. Rev.* **60**, 512–538.

Maxwell, K. L., Mittermaier, A. K., Forman-Kay, J. D. & Davidson, A. R. (1999). *Protein Sci.* **8**, 1908–1911.

Mittle, P. R. E. & Gruetter, M. G. (2001). *Curr. Opin. Chem. Biol.* **5**, 402–408.

Neylon, C. (2004). *Nucleic Acids Res.* **32**, 1448–1459.

Nguyen, A. W. & Daugherty, P. S. (2003). *Methods Mol. Biol.* **231**, 39–44.

Oefner, P. J., Hunicke-Smith, S. P., Chiang, L., Dietrich, F., Mulligan, J. & Davis, R. W. (1996). *Nucleic Acids Res.* **24**, 3879–3886.

Olsen, M., Iverson, B. & Georgiou, G. (2000). *Curr. Opin. Biotechnol.* **11**, 331–337.

Ostermeier, M. & Lutz, S. (2003). *Methods Mol. Biol.* **231**, 129–141.

O'Toole, N., Grabowski, M., Otwinowski, Z., Minor, W. & Cygler, M. (2004). *Proteins*, **56**, 201–210.

Pecenkova, T., Filigarova, M. & Cerovska, N. (2005). *Protein Expr. Purif.* **41**, 128–135.

Pedelacq, J. D., Piltch, E., Liong, E. C., Berendzen, J., Kim, C. Y., Rho, B. S., Park, M. S., Terwilliger, T. C. & Waldo, G. S. (2002). *Nature Biotechnol.* **20**, 927–932.

Putney, S. D., Benkovic, S. J. & Schimmel, P. R. (1981). *Proc. Natl Acad. Sci. USA*, **78**, 7350–7354.

Roodveldt, C., Aharoni, A. & Tawfik, D. S. (2005). *Curr. Opin. Struct. Biol.* **15**, 50–56.

Sorenson, H. P. & Mortensen, K. K. (2004). *J. Biotechnol.* **115**, 113–128.

Stemmer, W. P. (1994). *Nature (London)*, **370**, 389–391.

Thorstenson, Y. R., Hunicke-Smith, S. P., Oefner, P. J. & Davis, R. W. (1998). *Genome Methods*, **8**, 848–855.

Tobin, M. B., Gustafsson, C. & Huisman, G. W. (2000). *Curr. Opin. Struct. Biol.* **10**, 421–427.

Ullmann, A., Jacob, F. & Monod, J. (1967). *J. Mol. Biol.* **24**, 339–343.

Waldo, G. S., Standish, B. M., Berendzen, J. & Terwilliger, T. C. (1999). *Nature Biotechnol.* **17**, 691–695.

Wang, H. & Chong, S. (2003). *Proc. Natl Acad. Sci. USA*, **100**, 478–483.

Wigley, W. C., Stidham, R. D., Smith, N. M., Hunt, J. F. & Thomas, P. J. (2001). *Nature Biotechnol.* **19**, 131–135.

Yang, Z. R., Thomson, R., McNeil, P. & Esnouf, R. (2005). *Bioinformatics*, **21**, 3369–3376.

Zaccolo, M., Williams, D. M., Brown, D. M. & Gheradi, E. (1996). *J. Mol. Biol.* **255**, 589–603.