Contents lists available at ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

ESPRIT: An automated, library-based method for mapping and soluble expression of protein domains from challenging targets

Hayretin Yumerefendi [1], Franck Tarendeau [1], Philippe J. Mas, Darren J. Hart *

Unit of Virus Host–Cell Interactions, UJF-EMBL-CNRS, UMI 3265, 6 rue Jules Horowitz, BP181, 38042 Grenoble Cedex 9, France
Grenoble Outstation, European Molecular Biology Laboratory, 6 rue Jules Horowitz, BP181, 38042 Grenoble Cedex 9, France

ARTICLE INFO

Article history:
Received 12 January 2010
Received in revised form 24 February 2010
Accepted 28 February 2010
Available online 4 March 2010

Keywords:
ESPRIT
NF-κB
PB2
TBK1
High-throughput
Directed evolution
Protein structure
Protein expression

ABSTRACT

Expression of sufficient quantities of soluble protein for structural biology and other applications is often a very difficult task, especially when multimilligram quantities are required. In order to improve yield, solubility or crystallisability of a protein, it is common to subclone shorter genetic constructs corresponding to single- or multi-domain fragments. However, it is not always clear where domain boundaries are located, especially when working on novel targets with little or no sequence similarity to other proteins. Several methods have been described employing aspects of directed evolution to the recombinant expression of challenging proteins. These combine the construction of a random library of genetic constructs of a target with a screening or selection process to identify solubly expressing protein fragments. Here we review several datasets from the ESPRIT (Expression of Soluble Proteins by Random Incremental Truncation) technology to provide a view on its capabilities. Firstly, we demonstrate how it functions using the well-characterised NF-κB p50 transcription factor as a model system. Secondly, application of ESPRIT to the challenging PB2 subunit of influenza polymerase has led to several novel atomic resolution structures; here we present an overview of the screening phase of that project. Thirdly, analysis of the human kinase TBK1 is presented to show how the ESPRIT technology rapidly addresses the compatibility of challenging targets with the *Escherichia coli* expression system.

1. Introduction

The success of structural characterisation of proteins largely depends on the ability to produce sufficient quantities, generally tens of milligrams, of soluble purified protein (Blundell et al., 2002). Similar amounts of protein may be required for other non-structural applications such as inhibitor screening or biophysical analyses. The challenges associated with overexpression and purification of monodisperse, soluble, purifiable protein are well-appreciated by any structural biology laboratory. Over the last decade, structural genomics projects have permitted a more quantitative measure of the efficiency of different steps in the structure solution process under a relatively standardised set of conditions (Burley, 2000; O'Toole et al., 2004). Clearly, proteins do not behave uniformly during recombinant expression steps, notably the success rate of a single-domain protein may be very different to a large multi-domain protein when *Escherichia coli* is used as the preferred production system. Success rates for bacterial expression are generally higher with prokaryotic proteins than those from human or viral origin, perhaps because the bacterial proteins are frequently smaller with a simpler domain arrangement. Human or viral proteins are often larger and comprise multiple domains connected by longer linkers or low-complexity regions (Ward et al., 2004). They may be subunits of multi-component complexes, or at least require interaction of partners for stability, either via binding or post-translational modification. Expression of such proteins in full-length form in *E. coli* frequently results in aggregation or degradation (Dobson, 2004). Sometimes targets can be expressed more successfully in eukaryotic expression hosts, for example in insect cells using the baculovirus expression system (Hofinger et al., 2007), but these are far from universal solutions and present their own set of obstacles such as cost, duration, variable post-translational modification and incompatibility with isotopic or heavy atom labelling.

When a full-length protein fails to express in soluble form, or a purified protein does not crystallise, isolation of shorter constructs encompassing single- or multi-domain protein fragments is a common approach. Obtaining well-expressing protein domains is, traditionally, a time-consuming process involving repeated

Abbreviations: ESPRIT, expression of soluble proteins by random incremental truncation; GFP, green fluorescent protein; TBK1, TANK-binding kinase 1; IPTG, isopropyl-beta-D-thiogalactopyranoside; TEV, tobacco etch virus; BCCP, biotin carboxyl carrier protein..

* Corresponding author. Address: Unit of Virus Host–Cell Interactions, UJF-EMBL-CNRS, UMI 3265, 6 rue Jules Horowitz, BP181, 38042 Grenoble Cedex 9, France.
E-mail address: hart@embl.fr (D.J. Hart).

[1] These authors contributed equally to this work.

1047-8477/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jsb.2010.02.021

iterations of cloning and expression testing, and is most successful when a good level of prior knowledge of the target is available to guide construct design. If small amounts of material can be purified, limited proteolysis can lead to identification of stable fragments for further characterisation by mass spectrometry. Multiple sequence alignments can suggest domain location via evolutionarily conserved regions given the availability of related sequences, although protein expression-compatible boundaries do not always coincide with those predicted from sequence conservation in alignments (Dyson et al., 2008). The situation becomes even more difficult with targets that have no sequence or structural homologues. Order and disorder prediction algorithms, e.g. DisEMBL (Linding et al., 2003), RONN (Yang et al., 2005), IUPRED (Dosztanyi et al., 2005), DISOPRED2 (Ward et al., 2004) may suggest which regions of a protein are folded or unfolded, but the predictors are only of limited accuracy and are unable to define reliably the positions of primers for subcloning experiments. Even when domain locations can be predicted with accuracy, various underlying factors beyond those identifiable from sequence analyses can inhibit soluble expression; these include folding efficiency, requirement for intermolecular interactions by chaperones, ligands and protein partners, toxicity, inhibitory mRNA secondary structure, rare codon usage and intramolecular stabilisation from contacts by other parts of the protein.

One way to shortcut this lengthy route from DNA to protein is to make a large library of random genetic constructs and then screen them for soluble expression in a single, linear (non-iterative) experiment (Hart and Tarendeau, 2006). Such approaches have their origins in the directed evolution strategies used in protein engineering (Waldo, 2003). Several library-based construct screening strategies have been described (Cornvik et al., 2006; Kawasaki and Inagaki, 2001; Reich et al., 2006) and they all comprise a method of generating random libraries of sub-full-length DNA molecules (Prodromou et al., 2007) coupled to a high-throughput strategy to screen for rare clones expressing soluble proteins amongst the high background of non-productive constructs (Savva et al., 2007). Notable advantages of random library approaches are that, firstly, in contrast to classical construct design, they address empirically the confounding, unpredictable factors described above that complicate expression in a manner that has been described as ''holistic'' (Reich et al., 2006). Secondly, they can identify suspected or unsuspected domains in novel targets since they circumvent the need for multiple sequence alignments or domain prediction algorithms. Thirdly, when purifiable domains are identified by screening, several similar variants may be obtained and this can increase the chances of successful crystallisation.

Amongst the various random mutation strategies, those that best imitate the conventional approach of generating genetic constructs are exonuclease III truncation (Henikoff, 1984; Ostermeier and Lutz, 2003), or DNA fragmentation by endonucleases or physical methods (Anderson, 1981; Oefner et al., 1996). Point mutagenesis protocols employing DNA shuffling technologies have also been used to improve poorly soluble single-domain proteins (Aharoni et al., 2004; Pedelacq et al., 2006, 2002), but do not address the definition of sub-full-length constructs.

Critical to the success of these experiments is the selection or screening process for identification of clones exhibiting the desired phenotype: sufficient yields of soluble and stable protein. There are two major groups of screens (1) *direct* solubility assay by physical separation of the protein and (2) *indirect* assay such as solubility reporters and small peptides. In all approaches where activity or function of the target is not the readout in the screen, an initial assumption is made that a soluble protein is folded. Downstream biophysical analyses including circular dichroism spectroscopy, NMR and Thermofluor (Ericsson et al., 2006) can then be used to confirm foldedness.

Although the classic, direct method of solubility determination by lysate centrifugation is not readily adaptable for high-throughput screening, lysate filtration coupled with protein tag detection has been used successfully for assessing solubility of collections of proteins. Initially demonstrated in a plate format (Knaust and Nordlund, 2001; Vincentelli et al., 2005) it exhibited a limited throughput incompatible with screening random libraries since clones were tested in 96-well format. More recently, a colony-based filtration approach has been developed with the capacity to screen libraries of thousands of constructs (Cornvik et al., 2006; Dahlroth et al., 2006).

The majority of library-compatible solubility screens are indirect in that they report solubility via C-terminally fused partners such as fluorescent proteins (Heddle and Mazaleyrat, 2007; Waldo, 2003) or antibiotic resistance proteins (Dyson et al., 2008; Maxwell et al., 1999). These solubility assays permit enrichment of positive hits from the library by screening for colony fluorescence or survival on selective media respectively, but require further experiments to confirm the true solubility phenotype. False positives are frequent in these assays, often due to a ''passenger solubilisation'' effect resulting from the fused reporter itself (Reich et al., 2006); constructs identified are often soluble only in the context of the fusion, but not when the reporter protein is removed upon recloning or proteolytic cleavage. This may not be a problem when active protein preparations are required for assays or biochemical studies, but it is essential to remove larger tags prior to many applications including crystallisation trials. A further form of this passenger solubilisation effect is that small proteolytic or mistranslated fragments of the target can be observed fused to an active, well-folded reporter that signals as a (false) positive. This becomes apparent when the protein observed by SDS–PAGE or western blot is significantly smaller than that predicted from the encoding genetic insert.

One way to minimise these effects of passenger solubilisation by C-ter folded reporter proteins is to design a screen employing small linear peptide tags instead. Short peptides do not appear to perturb solubility significantly (Waugh, 2005) and are undetectable, presumably degraded, if not fused to a stabilising domain. In this manner, the passenger solubilisation problems observed when using GFP as a reporter have been elegantly addressed using a split-GFP strategy (Cabantous and Waldo, 2006; Cabantous et al., 2005). Expression and solubility are monitored by *in vivo* complementation between the target protein fused to beta strand 11 (only) of GFP and a co-expressed, truncated GFP (beta strands 1–10): if the target is expressed solubly, α-complementation results in detectable cellular fluorescence. The ESPRIT method reviewed here has similarities in that it uses only a short C-ter extension, the biotin acceptor peptide which is a well-established tool for protein immobilisation on streptavidin-coated surfaces (Beckett et al., 1999; Schatz, 1993). However, the mechanism of solubility sensing is quite different to the split GFP system; the peptide is post-translationally biotinylated *in vivo* by a transient interaction with the cytoplasmic BirA enzyme. The efficiency of this reaction can be measured quantitatively on many thousands of constructs simultaneously by streptavidin binding to robotically printed colony blots. Ranking of signal intensity permits enrichment of putatively soluble constructs from the starting library that are further analysed in a second screening phase of small-scale liquid expression and nickel affinity purification to identify well-expressing, purifiable proteins (Fig. 1).

2. The ESPRIT technology

2.1. Random construct libraries

ESPRIT uses exonuclease III degradation protocols for the generation of nested deletions of genes (Henikoff, 1984; Ostermeier and

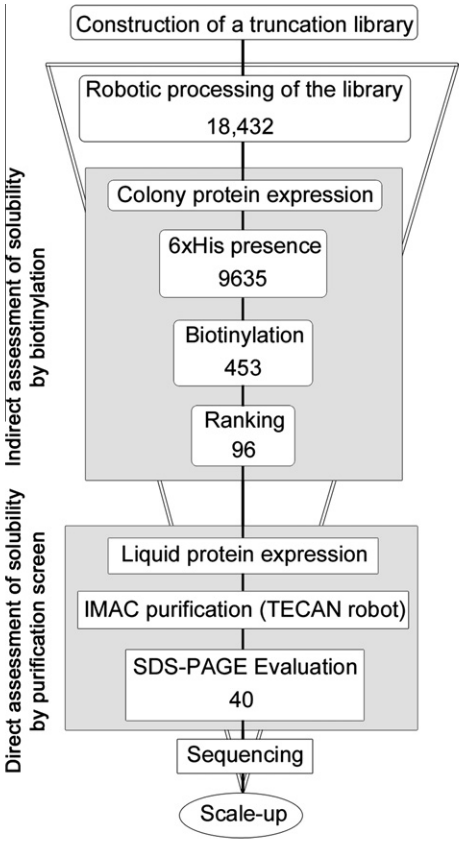
Construction of a truncation library
Robotic processing of the library
18,432
Colony protein expression
6xHis presence
9635
Biotinylation
453
Ranking
96
Liquid protein expression
IMAC purification (TECAN robot)
SDS-PAGE Evaluation
40
Sequencing
Scale-up
Indirect assessment of solubility by biotinylation
Direct assessment of solubility by purification screen

Fig. 1. ESPRIT construct screening workflow. Random DNA truncation libraries of the target gene are processed with colony picking robots to isolate individual clones. In a first screening step, putative soluble constructs are enriched using measurement of *in vivo* biotinylation as an indirect, but high-throughput, solubility assay. In a second step, constructs are directly assessed for yield and solubility by purification from small-scale liquid cultures. Positive clones are sequenced to characterise expression-compatible construct boundaries. Numbers shown are representative for a C-ter deletion library screen.

Lutz, 2003). The target genes are cloned into a vector encoding a C-ter biotin acceptor peptide and, at the end of the insert to be truncated, a pair of restriction sites that leave exonuclease III sensitive 5′ and resistant 3′ overhangs (Fig. 2a). In our optimised vectors (used for PB2 and TBK1 analyses below), an N-ter hexahistidine tag has also been included to facilitate direct purification testing of constructs. Once the exonuclease III reaction is initiated, small aliquots are removed at 1 min intervals for an hour and pooled in a quenching solution of 2 M NaCl (Tarendeau et al., 2007). Mung bean nuclease is then used to remove the single stranded overhang and the vector recircularised with T4 DNA ligase. Recovery of tens of thousands of colonies after transformation of *E. coli* is easy due to the highly efficient intramolecular religation reaction.

Vectors have been designed to enable three possible truncation strategies: unidirectional gene truncations from either the 5′ (Fig 2a) or 3′ end, or a combined protocol resulting in a bidirectional truncation library (Fig. 2b). In unidirectional truncation experiments, one end of the construct is fixed by in-frame fusion with hexahistidine tag or biotin acceptor peptide, the other end being truncated. All possible variants of the other end are generated with one-third being in-frame with the tag following ligation. The library diversity is *N* (the gene length in base pairs) and relatively small compared to classic directed evolution approaches. With a screening capacity of approximately 28,000 clones (10 h of colony picking), several-fold oversampling of all constructs can be achieved fairly easily.

The bidirectional truncation library method was designed to avoid the ‘‘fixed-ends” approach of the unidirectional method and to search internal regions of a gene for solubly expressing fragments. In a two step process, one end is truncated randomly and a plasmid library recovered. A second truncation reaction is then performed on a pool of these plasmids which are then religated again. An advantage of this two-step approach is that the true orientation of the insert is maintained throughout in contrast to DNA fragmentation protocols where only 50% of inserts are present in the correct sense after cloning. The diversity of bidirectional truncation libraries is much higher than the unidirectional libraries approximating to $N(N+1)/2$ (Prodromou et al., 2007). In order to reduce the number of clones to screen, and thereby improve the sampling efficiency, a subset of inserts is isolated by size fractionation of the library on agarose gels (Fig. 2c). For example, if a domain of 300 aa were suspected to exist, constructs in the range 200–400 aa can be isolated for expression testing. For a typical library screened, for example when the influenza *pb2* gene was screened (Guilligay et al., 2008), only a few percent of the total theoretical diversity of constructs was sampled in a 28,000 clone screen. However, since it is commonly observed that multiple, similar soluble fragments may exist for a given protein domain, the chance of identifying domains with this strategy is still high should such a domain exist.

2.2. Colony blot analysis of protein expression

Protein expression is induced in a high-density colony array format of up to 56,000 colonies, usually 28,000 in duplicate, each colony expressing a unique target gene insert. Colony blots are prepared by *in situ* lysis on nitrocellulose membranes (Bussow et al., 1998) and blots are probed with fluorescent streptavidin to detect the biotinylation status of the C-ter biotin acceptor peptide as an indicator for solubility (Fig. 3a and b). An additional quality filter was introduced by simultaneously probing the array with a monoclonal antibody against the N-ter hexahistidine tag; only clones exhibiting both N- and C-ter tags by colony blot analysis were analysed further since these comprised undegraded, intact protein constructs (Fig. 3c).

2.3. Protein expression and purification from liquid cultures

Isolation of the most efficiently biotinylated constructs leads to an enrichment of putatively soluble clones that, in the second level of screening, are expressed in small-scale liquid format, lysed and purified on Ni^{2+} NTA affinity resin. Soluble, purifiable constructs are visualised by SDS–PAGE and their sequence boundaries determined by DNA sequencing. A fairly common result is that the best constructs identified are only partially soluble with the purifiable material constituting a minor, but sometimes useable proportion of the total material. It is also clear that inclusion bodies do exhibit some level of biotinylation, perhaps on surface-exposed biotin acceptor peptides, or through BirA-mediated labelling of slowly aggregating material. Thus, a second step of purification of the primary hits functions to distinguish between fully or partially soluble and insoluble constructs.

2.4. The advantages of automation

In principle, the ESPRIT method could be performed manually by titrating and replicating the library directly on a membrane. However, a standard colony picking and gridding robot provides an efficient solution to the handling, analysis and sample-tracking of 10^4–10^5 individual clones. The library, once picked, can be replicated and frozen so that it is available for experimental repeats (e.g. under different expression conditions), providing significant

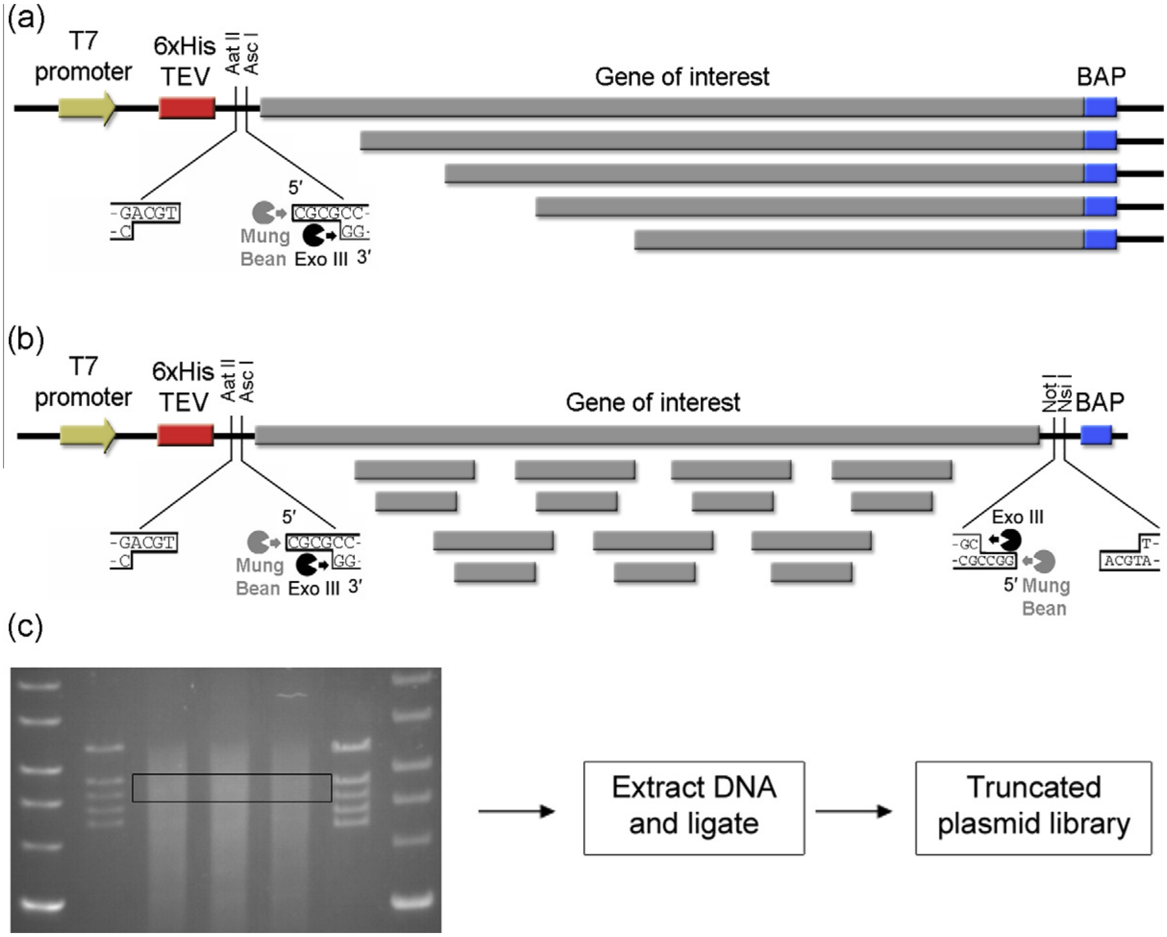

Fig. 2. ESPRIT truncation molecular biology. (a) Expression cassette region of unidirectional ESPRIT truncation vector (5′ truncation shown). Restriction digest with AatII generates exonuclease III resistant 3′ overhang, and with AscI generates substrate 5′ overhang. After exonuclease III digest, mung bean nuclease removes the remaining single strand prior to vector recircularisation by ligation. (b) Bidirectional ESPRIT truncation vector combines two pairs of restriction enzyme sites (AatII and AscI, NotI and NsiI) to permit two sequential unidirectional reactions that generate internal fragments. (c) Size fractionation of vector plus insert DNA from a bidirectional truncation library by excision of the desired DNA size range from agarose gel.

practical advantages over manual plating where the library must be regenerated each time. The geometrically arrayed colony blots on a 22 cm square membrane can be analysed with only small quantities of antibody/streptavidin fluorescent conjugates, and the digitised array images analysed with software already developed for DNA arrays. Later in the workflow, putative positive clones are rearrayed robotically into a 96-well plate for DNA insert analysis and protein expression trials.

3. Overview of ESPRIT screening data

The application of ESPRIT to expression of several proteins for structural studies has been reported including *Helicobacter pylori* CagA (Angelini et al., 2009), influenza polymerase PB2 subunit (Guilligay et al., 2008; Tarendeau et al., 2007, 2008) and the neurofibromatosis type 1 protein neurofibromin (NF1) (Bonneau et al., 2009). Here, we compare data from several ESPRIT screening experiments. We show a proof-of-concept study on the two-domain human transcription factor NF-κB demonstrating single amino acid resolution boundary mapping. We then review construct screening data from a challenging structural project on the influenza PB2 polymerase subunit that resulted in determination of three crystal structures described elsewhere (Guilligay et al., 2008; Tarendeau et al., 2007, 2008). Finally, we discuss results obtained on the screening of the recalcitrant human kinase TBK1 demonstrating how screening of all potential constructs around a targeted domain can provide a rapid and definitive result valuable for reorienting the experimental strategy.

3.1. Fine resolution domain mapping of NF-κB p50

As a proof-of-concept we applied our method to nuclear factor κB (NF-κB) p50, a human transcription factor of known structure (Muller et al., 1995). It has high sequence similarity with other Rel transcription factors and comprises separate dimerisation and DNA binding domains. The domains have been studied independently (de Lumley et al., 2004; Huang et al., 1997), so can be considered as autonomously folded units joined by a flexible linker region.

Both 5′ and 3′ unidirectional truncation libraries were synthesised and their quality verified by colony PCR showing an even distribution of insert sizes in each library. Approximately 27,000 colonies from each library were isolated with a colony picking robot. Constructs of the 5′ truncation library were 8-fold oversampled and the 3′ library 14-fold. Colonies were gridded in duplicates onto nitrocellulose membranes and protein expression induced on LB-agar supplemented with IPTG and biotin. After *in situ* lysis and hybridisation of the colony blots with fluorescent streptavidin, the blots were visualised with a fluorescent scanner and clones ranked according to their biotinylation levels. Colony

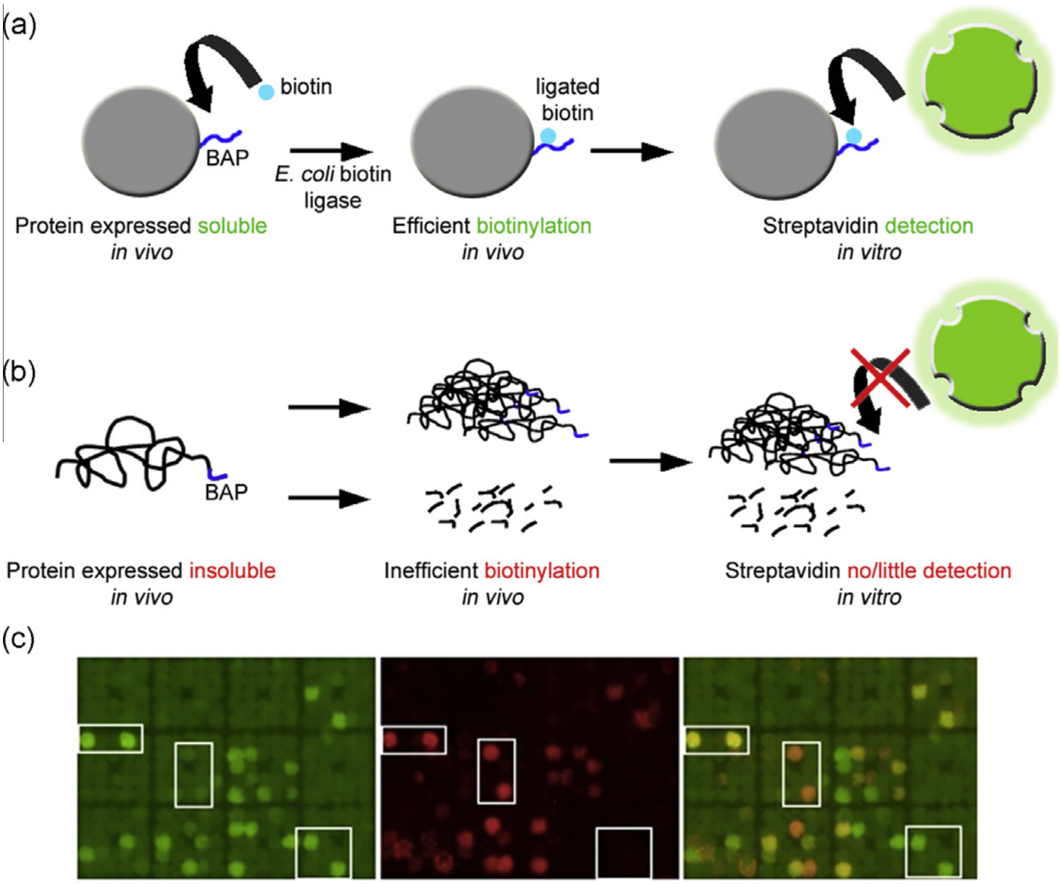
(a)
biotin
BAP
E. coli biotin ligase
ligated biotin
Protein expressed soluble in vivo
Efficient biotinylation in vivo
Streptavidin detection in vitro
(b)
BAP
Protein expressed insoluble in vivo
Inefficient biotinylation in vivo
Streptavidin no/little detection in vitro
(c)

Fig. 3. Colony array screen for solubility using *in vivo* biotinylation levels. Principle of screen. The target of interest is fused to the biotin acceptor peptide via a short linker. Soluble proteins (a) are better biotinylated by *E. coli* than those that are degraded or insoluble (b). (c) Measurement of putative solubility and intactness of construct through simultaneous hybridisation of fluorescent streptavidin to C-ter biotin acceptor peptide (green) and N-ter antihexahistidine tag monoclonal with fluorescent secondary antibody (red). Merged image shown on right indicating a potentially interesting construct (left white box), poorly soluble or degraded phenotype (centre white box) and potentially soluble but degraded construct with no hexahistidine tag (right white box).

PCR screens on a subset of highly biotinylated clones showed them to cluster at approximately 950 bp and 1250 bp in the 3′ deletion library, and at about 700 bp and 800–1100 bp in the 5′ truncation library. The highest signalling 48 constructs from both truncation libraries were sequenced to define exact construct boundaries. In this early version of ESPRIT (Tarendeau et al., 2007), no N-ter hexahistidine tag sequence was present in the screening plasmid, so a panel of positive clones was then subcloned into a pET-derived vector for expression and purification testing (Fig. 4a–c).

It was found that soluble constructs mapped well onto the structure of NF-κB with soluble variants found for two-domain and single-domain constructs (Fig. 4d). Sequence alignment of these constructs permitted definition of the edge of the minimal forms of both one and two domain forms of the protein.

A cluster of constructs identified through 3′ truncation was found to encompass both domains with one of the soluble constructs (40–354) very closely related to the structurally ordered residues (43–353) of PDB: 1SVC (Muller et al., 1995). The polypeptide region deleted in these constructs corresponded to a disordered C-terminal peptide. From the 3′ truncation library, the C-terminal boundary was Ile354, directly adjacent to the last ordered residue in the crystal structure. Similar data were obtained for the DNA binding domain using the 3′ truncation library approach. Whilst the shortest construct identified (40–239) as a fusion with the biotin acceptor peptide exhibited partial solubility and purifiability once subcloned to remove the C-ter tag, two and four residue longer constructs exhibited very high levels of soluble expression (Fig. 4a). This effect was attributed to perhaps minor overtruncation of this domain being compensated by the presence of the tag, leaving a small destabilising cavity after subcloning. This effect will exist in classical construct testing approaches using fused N- or C-ter tags and, if encountered, is easily addressed by adding just a few amino acids back to the truncated end upon tag deletion.

Soluble fragments were also obtained from the 5′ truncation library corresponding to the dimerisation domain and these were similar to a previously expressed and crystallised fragment 248–353 (numbering system as above) (Huang et al., 2007). Moreover, the most compact construct identified from the 5′ truncation library was found to correlate exactly with the N-ter structural edge of the domain, starting at position 251 (Fig. 4d). An optimal dimerisation domain construct was obtained by combining the sequence information from both libraries. Whereas two clones obtained directly from the unidirectional libraries (5′: 251–366; 3′: 40–354) showed a moderate level of expressed protein (Fig. 4a and b), the composite 251–354 construct was better expressed and more resistant to proteolysis (Fig. 4c).

3.2. Expression of domains from the influenza polymerase PB2 subunit

The three subunit influenza polymerase is responsible for the RNA replication and transcription of the influenza virus (Krug et al., 2003). Despite its key biological role and interest as an antiviral drug target (Ruigrok et al., 2010), no overexpression of full-length proteins or domains had been achieved, preventing structural studies. A major reason for this is that, in common with

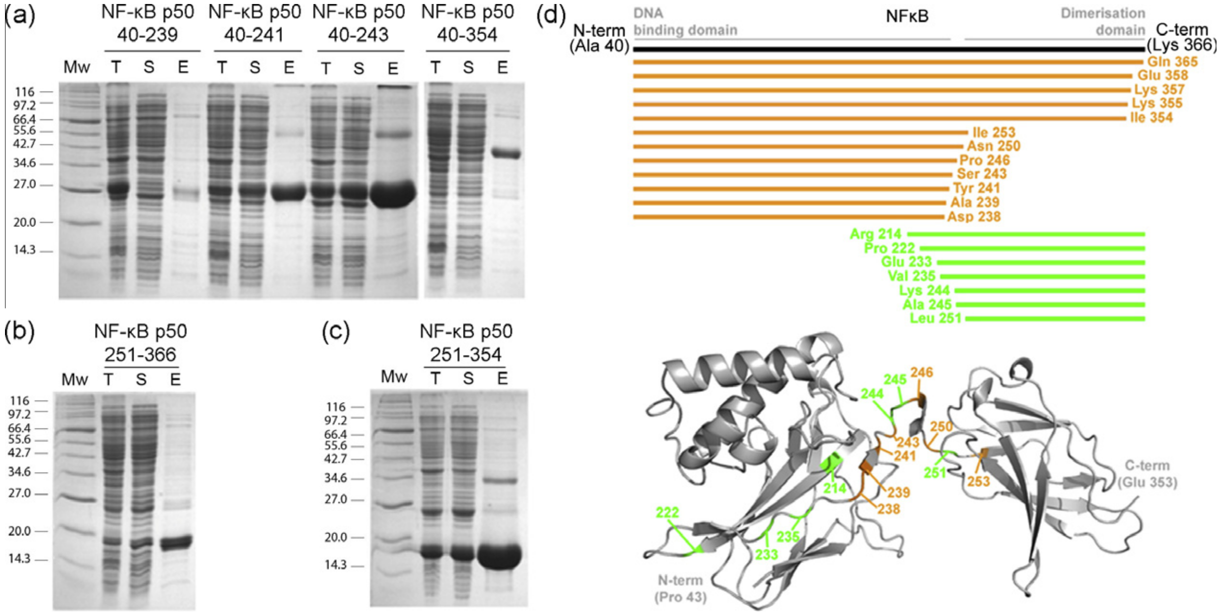

Fig. 4. Experimental mapping of structural domain boundaries of NF-κB p50. (a) SDS–PAGE of selected 3′ truncation library constructs purified by nickel affinity chromatography showing sensitivity of protein expression to small terminal variations (discussed in text). (b) SDS–PAGE of shortest construct 251–366 from the 5′ truncation library. (c) Optimal NF-κB p50 construct 251–354 generated by combining both 5′ and 3′ experimentally determined boundaries into a single clone. Mw: molecular weight marker; T: total fraction; S: soluble fraction; E: eluted protein from affinity column. (d) Construct map of purifiable domains identified from 3′ deletion library (orange) and 5′ library (green) with domain boundaries annotated on the NF-κB p50 structure (PDB: 1SVC).

many challenging targets, the protein subunits exhibit no useful sequence homology to other proteins, preventing the use of sequence alignment methods to identify domains. For the PB2 subunit, analysis of the 759 amino acid translated sequence using several order and disorder predictors suggested that the protein was folded, but no more information existed for designing constructs. The first experiment performed was an exploratory 5′ truncation library performed in parallel with NF-κB (C-ter biotin acceptor peptide, no hexahistidine tag, *tac* promoter). The colony array expressing truncated *pb2* gene fragments was probed with a horseradish peroxidase–streptavidin conjugate with detection using chemiluminescence and autoradiographic film methods. Sequencing of positive clones revealed a cluster of constructs predicted to encode a C-ter region of approximately 100 aa. Subcloning of these constructs into a pET-type vector with T7 promoter, N-ter TEV cleavable hexahistidine tag and removal of the biotin acceptor peptide identified one (678–759; Fig. 5) that expressed at over 100 mg/l of culture. Despite their size similarity, the other constructs were poorly expressed perhaps reflecting the sensitivity of protein expression to changes in the 5′ region as shown previously (Cornvik et al., 2006). Structure determination of the isolated protein by NMR, and in complex with the human nuclear import receptor importin α5 by X-ray crystallography, revealed how this so-called “NLS domain” mediates nuclear localisation of the PB2 polymerase subunit (Tarendeau et al., 2007).

To screen for internal soluble domains, the *pb2* gene was truncated at both ends simultaneously (Guilligay et al., 2008). Linearised vectors with truncated inserts were size-selected on agarose gel to generate four sublibraries. Almost 27,000 clones with an insert size range of 150–250 aa were analysed corresponding to approximately 5% of the total theoretical diversity of constructs within that sublibrary. Two major regions expressing soluble domains were identified (Fig. 5). The first cluster contained constructs between 189 and 261 aa in length falling in the central region of the protein. They expressed solubly at high levels but did not produce crystals. Further refinement by limited proteolysis was required to delineate a minimal construct (318–483; Fig. 5) that crystallised to yield an X-ray structure of the “cap-binding domain” bound to the mRNA cap analogue m7-GTP, explaining one aspect of the influenza virus cap-snatching mechanism (Guilligay et al., 2008). Why this shorter, crystallisable construct was not identified in the primary screen may be explained through the undersampling of the total diversity. Separately, a larger C-ter fragment was identified containing the NLS domain plus an additional upstream region (Fig. 5). It also expressed solubly at high levels but, like the cap-binding domain, required further refinement by limited proteolysis before successful crystallisation of both one and two domain forms (538–693 and 538–753). Structure solution of these domains revealed a compact structure with surface-borne residues known to mediate adaptation of avian influenza viruses to human hosts (Tarendeau et al., 2008).

Analysis with DALI (Holm et al., 2008) of the three new PB2 structures revealed them all to possess novel folds, consistent with there being no known sequence or structural homologues. This example of PB2 shows how experimental construct screening approaches such as ESPRIT can be highly productive for novel, poorly annotated targets.

3.3. TANK-binding kinase 1 catalytic domain screen

TANK-binding kinase 1 (TBK1) is a Ser/Thr kinase component of a tripartite complex shown to play a role in interferon production (Clark et al., 2009). A significant amount of research on this protein has been performed due to its high potential as a drug target, however, data on protein expression and structural characterisation remains unavailable. It therefore provides an example of a target class where the approximate location of the desired domain can be predicted from sequence with high confidence, distinct from the target type exemplified by PB2 where no prior knowledge exists on domain arrangement.

We selected TBK1 as part of a wider study on multi-domain kinases where the objective was to identify both catalytic and regulatory domains (unpublished data). For each target, homology sequence alignments were used to identify the core sequence of the kinase catalytic domain and the truncation library strategy was chosen to screen domain boundary variations. The catalytic

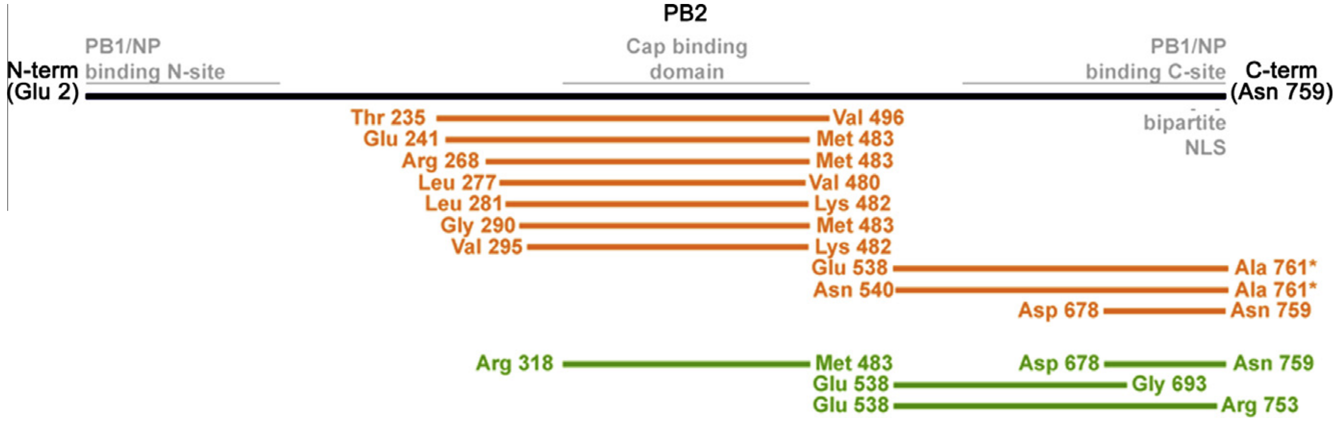

Fig. 5. Expression of previously unsuspected domains within the influenza polymerase PB2 subunit. Expressed and purified constructs from ESPRIT screening (orange) are aligned with the full-length protein sequence (black) with early functional annotations (grey). Shown in green are the domains that were eventually crystallised following limited proteolysis studies of the purified primary constructs. Asterisk indicates addition of two non-native alanine residues from cloning step.

domain of TBK1 kinase is reliably predicted at the N-ter end so a 3′ gene truncation strategy was selected permitting all C-terminal variations to be sampled within 200 amino acids of the annotated (non-expressing) domain. The resulting library was of high quality with almost all clones containing a truncated *tbk1* gene insert in the desired size range of 50–500 aa. Clones were assayed according to the colony array biotinylation assay with an approximate 3.1-fold oversample of all constructs. In contrast to NF-κB and PB2 libraries, only low intensity signals were observed on the streptavidin colony blot and no cluster of similar sized inserts was apparent in a colony PCR screen with flanking, vector-specific primers. When scaled up, the best hits exhibited only western blot detectable levels of soluble, purifiable material and, at 92–192 aa in length (Fig. 6), they were too short to form an intact kinase catalytic domain of 280–300 aa (Hanks and Hunter, 1995).

In contrast to the data presented on NF-κB and PB2, the last example demonstrates how one person in a few weeks can screen all variants of a desired domain and define a negative outcome with high confidence due the comprehensive, oversampled nature of the screening. From these data and other similar cases, we now recognize that the pattern of low expression signals in the colony screening assay and the absence of a cluster of similarly sized inserts in selected clones are early indicators of a negative outcome. Relative to the conventional approach of iterations of PCR cloning and expression testing in *E. coli*, library strategies rapidly define the outcome, thus saving considerable time and resources and hastening a change to other expression systems (e.g. baculovirus, yeast and mammalian cells).

4. Discussion

ESPRIT is a method using principles from directed evolution in which a random genetic library of truncated constructs is screened to define empirically the desired solution (soluble protein domains). Here, we have combined genetic truncation of the target using an exonuclease III/Mung bean nuclease protocol with a

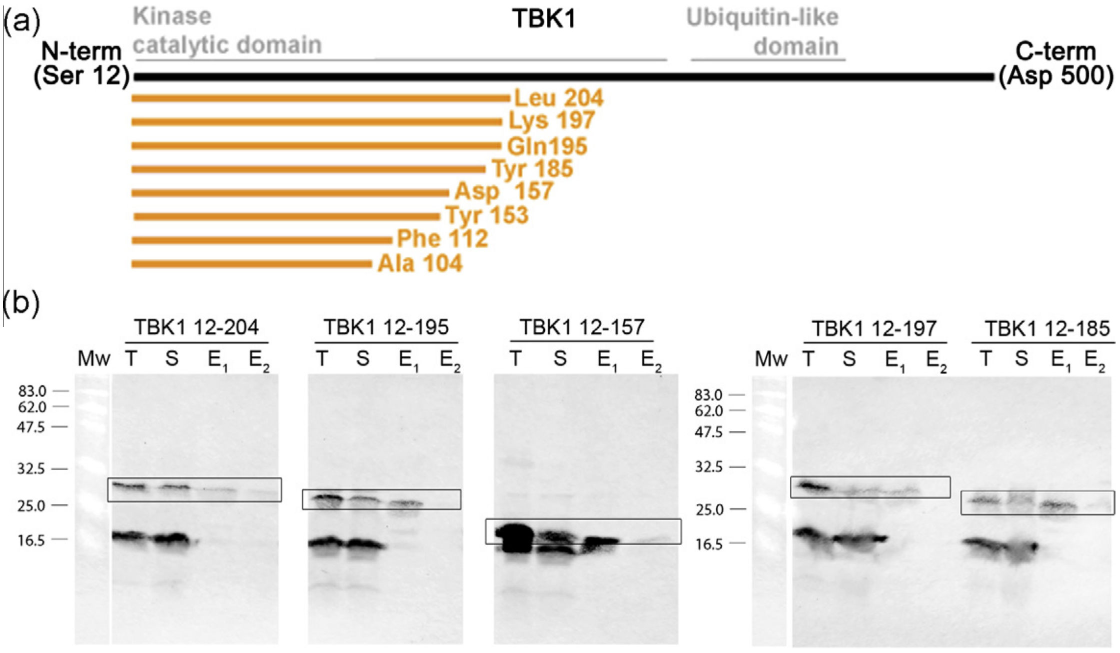

Fig. 6. Results from screening the human kinase, TBK1. (a) Construct map of best ESPRIT identified soluble constructs (orange) aligned with predictions of expected domains (grey). No clustering around a single domain size is observed (b) Streptavidin blot analysis of expression and purification samples showed that best constructs identifiable were expressed at low levels, but soluble and marginally purifiable. The lower band is the endogenous biotinylated *E. coli* BCCP protein that is removed during the purification. Lane annotations as in Fig. 4.

high-throughput automated screen capable of comparing the expression of nearly 30,000 constructs per run. Data here is presented on three types of target: a proof of principle screen on a protein of pre-existing atomic structure (NF-κB p50), a novel high value target of unknown domain architecture (PB2) and a negative experiment (TBK1) on an easily predicted domain that, nevertheless, resists expression in *E. coli*.

Screening of NF-κB p50 identified how single amino acid resolution on expressible domains can be achieved by sequence alignment of soluble constructs from the genetic screen. Conceptually this is similar to performing proteolysis experiments on a single soluble construct, but at finer resolution, and with the major advantage that the DNA is the starting material, not pre-existing soluble protein. The minimal forms of the one- and two-domain constructs identified by screening are very similar to crystallised constructs described previously (Huang et al., 1997; Muller et al., 1995). We also identified longer variants with similar expression characteristics comprising these same domains with additional disordered terminal residues that would potentially inhibit crystal packing. Therefore, more compact constructs with a better crystallisation propensity can be rapidly identified without resort to steps of purification, limited proteolysis and mass spectrometry.

The data from the screening of the influenza polymerase PB2 subunit demonstrates perhaps the greatest advantage of this approach: that a target can yield domains that are previously unpredicted due to a lack of sequence homologues for generating multiple sequence alignments. Three separate domains were identified that yielded atomic resolution structures (Guilligay et al., 2008; Tarendeau et al., 2007, 2008) which, consistent with the absence of homologues, all exhibited novel folds. One curious observation is that whilst the C-ter region comprised two independent domains that could be fused or separated, this double domain region could not be fused to the nearby, upstream cap-binding domain to form a longer, soluble construct. There is no predicted disorder in this region so it perhaps reflects the folding limitations of *E. coli* with some multi-domain constructs; it does however demonstrate clearly the power of an experimental screen over the classic construct design approach. The PB2 data also shows how an approach of isolating a soluble expressing construct via a library strategy, linked to further refinement of the construct by limited proteolysis may yield the most rapid route to a crystallisable construct, especially when a more diverse, less sampled bidirectionally truncated (or DNA fragment) library is employed. Finally, the results on TBK1 show how a firm negative answer can be obtained rapidly and with high confidence, permitting an early reallocation of time and resources towards other expression systems (or project termination).

Acknowledgments

We wish to acknowledge the contributions of all members of the Hart laboratory and of the Influenza Project to the development and validation of the ESPRIT method. HY was funded an AstraZeneca research fellowship. FT was funded by the European Commission FP5 contract 'Structural Proteomics in Europe' (SPINE, contract QLG-CT-2002-00988) and FP6 contract (FLUPOL, SP5B-CT-2007-044263). Additional funding was provided by EC FP6 Integrated Contract (3D Repertoire LSHG-CT-2005-512028) and the EMBL.

References

Aharoni, A., Gaidukov, L., Yagur, S., Toker, L., Silman, I., Tawfik, D.S., 2004. Directed evolution of mammalian paraoxonases PON1 and PON3 for bacterial expression and catalytic specialization. Proc Natl Acad Sci U.S.A. 101, 482–487.

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

Angelini, A., Tosi, T., Mas, P., Acajjaoui, S., Zanotti, G., Terradot, L., Hart, D.J., 2009. Expression of *Helicobacter pylori* CagA domains by library-based construct screening. FEBS J. 276, 816–824.

Beckett, D., Kovaleva, E., Schatz, P.J., 1999. A minimal peptide substrate in biotin holoenzyme synthetase-catalyzed biotinylation. Protein Sci. 8, 921–929.

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

Bonneau, F., Lenherr, E.D., Pena, V., Hart, D.J., Scheffzek, K., 2009. Solubility survey of fragments of the neurofibromatosis type 1 protein neurofibromin. Protein Expr. Purif. 65, 30–37.

Burley, S.K., 2000. An overview of structural genomics. Nat. Struct. Biol. Suppl. 932, 934.

Bussow, K., Cahill, D., Nietfeld, W., Bancroft, D., Scherzinger, E., Lehrach, H., Walter, G., 1998. A method for global protein expression and antibody screening on high-density filters of an arrayed cDNA library. Nucleic Acids Res. 26, 5007–5008.

Cabantous, S., Waldo, G.S., 2006. In vivo and in vitro protein solubility assays using split GFP. Nat. Methods 3, 845–854.

Cabantous, S., Pedelacq, J.D., Mark, B.L., Naranjo, C., Terwilliger, T.C., Waldo, G.S., 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.

Clark, K., Plater, L., Peggie, M., Cohen, P., 2009. Use of the pharmacological inhibitor BX795 to study the regulation and physiological roles of TBK1 and IkappaB kinase epsilon: a distinct upstream kinase mediates Ser-172 phosphorylation and activation. J. Biol. Chem. 284, 14136–14146.

Cornvik, T., Dahlroth, S.L., Magnusdottir, A., Flodin, S., Engvall, B., Lieu, V., Ekberg, M., Nordlund, P., 2006. An efficient and generic strategy for producing soluble human proteins and domains in *E. coli* by screening construct libraries. Proteins 65, 266–273.

Dahlroth, S.L., Nordlund, P., Cornvik, T., 2006. Colony filtration blotting for screening soluble expression in *Escherichia coli*. Nat. Protoc. 1, 253–258.

de Lumley, M., Hart, D.J., Cooper, M.A., Symeonides, S., Blackburn, J.M., 2004. A biophysical characterisation of factors controlling dimerisation and selectivity in the NF-kappaB and NFAT families. J. Mol. Biol. 339, 1059–1075.

Dobson, C.M., 2004. Principles of protein folding, misfolding and aggregation. Semin. Cell Dev. Biol. 15, 3–16.

Dosztanyi, Z., Csizmok, V., Tompa, P., Simon, I., 2005. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21, 3433–3434.

Dyson, M.R., Perera, R.L., Shadbolt, S.P., Biderman, L., Bromek, K., Murzina, N.V., McCafferty, J., 2008. Identification of soluble protein fragments by gene fragmentation and genetic selection. Nucleic Acids Res. 36, e51.

Ericsson, U.B., Hallberg, B.M., Detitta, G.T., Dekker, N., Nordlund, P., 2006. Thermofluor-based high-throughput stability optimization of proteins for structural studies. Anal. Biochem. 357, 289–298.

Guilligay, D., Tarendeau, F., Resa-Infante, P., Coloma, R., Crepin, T., Sehr, P., Lewis, J., Ruigrok, R.W., Ortin, J., Hart, D.J., Cusack, S., 2008. The structural basis for cap binding by influenza virus polymerase subunit PB2. Nat. Struct. Mol. Biol. 15, 500–506.

Hanks, S.K., Hunter, T., 1995. Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. FASEB J. 9, 576–596.

Hart, D.J., Tarendeau, F., 2006. Combinatorial library approaches for improving soluble protein expression in *Escherichia coli*. Acta Crystallogr. D Biol. Crystallogr. 62, 19–26.

Heddle, C., Mazaleyrat, S.L., 2007. Development of a screening platform for directed evolution using the reef coral fluorescent protein ZsGreen as a solubility reporter. Prot. Eng. Des. Sel. 20, 327–337.

Henikoff, S., 1984. Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. Gene 28, 351–359.

Hofinger, E.S., Spickenreither, M., Oschmann, J., Bernhardt, G., Rudolph, R., Buschauer, A., 2007. Recombinant human hyaluronidase Hyal-1: insect cells versus *Escherichia coli* as expression system and identification of low molecular weight inhibitors. Glycobiology 17, 444–453.

Holm, L., Kaariainen, S., Rosenstrom, P., Schenkel, A., 2008. Searching protein structure databases with DaliLite v.3. Bioinformatics 24, 2780–2781.

Huang, C.H., Mandelker, D., Schmidt-Kittler, O., Samuels, Y., Velculescu, V.E., Kinzler, K.W., Vogelstein, B., Gabelli, S.B., Amzel, L.M., 2007. The structure of a human p110alpha/p85alpha complex elucidates the effects of oncogenic PI3Kalpha mutations. Science 318, 1744–1748.

Huang, D.B., Huxford, T., Chen, Y.Q., Ghosh, G., 1997. The role of DNA in the mechanism of NFkappaB dimer formation: crystal structures of the dimerization domains of the p50 and p65 subunits. Structure 5, 1427–1436.

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

Knaust, R.K., Nordlund, P., 2001. Screening for soluble expression of recombinant proteins in a 96-well format. Anal. Biochem. 297, 79–85.

Krug, R.M., Yuan, W., Noah, D.L., Latham, A.G., 2003. Intracellular warfare between human influenza viruses and human cells: the roles of the viral NS1 protein. Virology 309, 181–189.

Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J., Russell, R.B., 2003. Protein disorder prediction: implications for structural proteomics. Structure 11, 1453–1459.

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

Muller, C.W., Rey, F.A., Sodeoka, M., Verdine, G.L., Harrison, S.C., 1995. Structure of the NF-kappa B p50 homodimer bound to DNA. Nature 373, 311–317.

O'Toole, N., Grabowski, M., Otwinowski, Z., Minor, W., Cygler, M., 2004. The structural genomics experimental pipeline: insights from global target lists. Proteins 56, 201–210.

Oefner, P.J., Hunicke-Smith, S.P., Chiang, L., Dietrich, F., Mulligan, J., Davis, R.W., 1996. Efficient random subcloning of DNA sheared in a recirculating point-sink flow system. Nucleic Acids Res. 24, 3879–3886.

Ostermeier, M., Lutz, S., 2003. The creation of ITCHY hybrid protein libraries. Methods Mol. Biol. 231, 129–141.

Pedelacq, J.D., Cabantous, S., Tran, T., Terwilliger, T.C., Waldo, G.S., 2006. Engineering and characterization of a superfolder green fluorescent protein. Nat. Biotechnol. 24, 79–88.

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. Engineering soluble proteins for structural genomics. Nat. Biotechnol. 20, 927–932.

Prodromou, C., Savva, R., Driscoll, P.C., 2007. DNA fragmentation-based combinatorial approaches to soluble protein expression. Part I. Generating DNA fragment libraries. Drug Discov. Today 12, 931–938.

Reich, S., Puckey, L.H., Cheetham, C.L., Harris, R., Ali, A.A., Bhattacharyya, U., Maclagan, K., Powell, K.A., Prodromou, C., Pearl, L.H., Driscoll, P.C., Savva, R., 2006. Combinatorial domain hunting: an effective approach for the identification of soluble protein domains adaptable to high-throughput applications. Prot. Sci. 15, 2356–2365.

Ruigrok, R.W., Crepin, T., Hart, D.J., Cusack, S., 2010. Towards an atomic resolution understanding of the influenza virus replication machinery. Curr. Opin. Struct. Biol. 20, 104–113.

Savva, R., Prodromou, C., Driscoll, P.C., 2007. DNA fragmentation based combinatorial approaches to soluble protein expression. Part II. Library expression, screening and scale-up. Drug Discov. Today 12, 939–947.

Schatz, P.J., 1993. Use of peptide libraries to map the substrate specificity of a peptide-modifying enzyme: a 13 residue consensus peptide specifies biotinylation in *Escherichia coli*. Biotechnology (NY) 11, 1138–1143.

Tarendeau, F., Crepin, T., Guilligay, D., Ruigrok, R.W., Cusack, S., Hart, D.J., 2008. Host determinant residue lysine 627 lies on the surface of a discrete, folded domain of influenza virus polymerase PB2 subunit. PLoS Pathog. 4, e1000136.

Tarendeau, F., Boudet, J., Guilligay, D., Mas, P.J., Bougault, C.M., Boulo, S., Baudin, F., Ruigrok, R.W., Daigle, N., Ellenberg, J., Cusack, S., Simorre, J.P., Hart, D.J., 2007. Structure and nuclear import function of the C-terminal domain of influenza virus polymerase PB2 subunit. Nat. Struct. Mol. Biol. 14, 229–233.

Vincentelli, R., Canaan, S., Offant, J., Cambillau, C., Bignon, C., 2005. Automated expression and solubility screening of His-tagged proteins in 96-well format. Anal. Biochem. 346, 77–84.

Waldo, G.S., 2003. Genetic screens and directed evolution for protein solubility. Curr. Opin. Chem. Biol. 7, 33–38.

Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., Jones, D.T., 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J. Mol. Biol. 337, 635–645.

Waugh, D.S., 2005. Making the most of affinity tags. Trends Biotechnol. 23, 316–320.

Yang, Z.R., Thomson, R., McNeil, P., Esnouf, R., 2005. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics 21, 3369–3376.