Analytical Biochemistry 465 (2014) 53–62

Contents lists available at ScienceDirect

Analytical Biochemistry

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

Identifying protein domains by global analysis of soluble fragment data

Esther M.M. Bulloch *, Richard L. Kingston

School of Biological Sciences, University of Auckland, New Zealand

ARTICLE INFO

Article history:
Received 17 February 2014
Received in revised form 17 June 2014
Accepted 25 June 2014
Available online 10 July 2014

Keywords:
Protein expression
Protein domains
Gene fragmentation
Solubility screen
Domain mapping
Cluster analysis

ABSTRACT

The production and analysis of individual structural domains is a common strategy for studying large or complex proteins, which may be experimentally intractable in their full-length form. However, identifying domain boundaries is challenging if there is little structural information concerning the protein target. One experimental procedure for mapping domains is to screen a library of random protein fragments for solubility, since truncation of a domain will typically expose hydrophobic groups, leading to poor fragment solubility. We have coupled fragment solubility screening with global data analysis to develop an effective method for identifying structural domains within a protein. A gene fragment library is generated using mechanical shearing, or by uracil doping of the gene and a uracil-specific enzymatic digest. A split green fluorescent protein (GFP) assay is used to screen the corresponding protein fragments for solubility when expressed in *Escherichia coli*. The soluble fragment data are then analyzed using two complementary approaches. Fragmentation "hotspots" indicate possible interdomain regions. Clustering algorithms are used to group related fragments, and concomitantly predict domain location. The effectiveness of this Domain Seeking procedure is demonstrated by application to the well-characterized human protein p85α.

Biochemical, biophysical, and structural analysis of proteins requires significant amounts of material. Despite the continual development of heterologous expression systems [1], obtaining sufficient quantities of a protein in a correctly folded and soluble form is often difficult, particularly for complex eukaryotic proteins. Fortunately, many large proteins are modular in nature and composed of multiple structural domains: semiautonomous regions of the polypeptide that have the capacity to fold in isolation. The individual domains may be easier to express and purify than the full-length protein, and their characterization can provide critical insights into protein function. The challenge is to identify the boundaries of these structural domains.

If trace amounts of a full-length protein can be isolated, limited enzymatic proteolysis is a useful and well-validated experimental technique for identifying domain boundaries [2]. Alternatively, domain boundaries can be inferred from the protein sequence, using varying bioinformatic approaches (see e.g., [3]). Based on such *in silico* analyses, the expression of multiple constructs with slightly differing termini is typically evaluated [4,5]. This approach is embedded in the workflows of many structural genomics consortia (see e.g., [3,6]). While these methods are unquestionably successful, there remain situations where they are difficult to apply. Some proteins cannot be expressed in full-length form, even in trace amounts, or have very limited sequence similarity with previously characterized proteins, weakening the structural inferences that can be made.

Over the past decade some alternative experimental approaches for identifying structural domain boundaries have been developed. Although the exact methodology varies greatly, the basic strategy is to express a random library of protein fragments and screen these for solubility in a high-throughput manner [7–11]. This approach is successful because fragmentation within a structural domain will generally expose hydrophobic amino acids sequestered in the domain interior, giving rise to conformationally unstable fragments with limited solubility. Studying the expression, stability, and solubility of fragments can therefore yield information about the structural domains embedded within a complex protein.

Methods used to fragment the target gene include limited exonuclease and/or endonuclease digest [12–15], mechanical shearing [14,16], PCR[1] with random primers [17], and uracil-doped PCR followed by a uracil-specific enzymatic digest [10,18,19]. Each of these

* Corresponding author. Address: School of Biological Sciences, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand.
E-mail address: e.bulloch@auckland.ac.nz (E.M.M. Bulloch).

[1] *Abbreviations used:* BCR, breakpoint cluster region-homology; bp, base pairs; diethylaminoethyl EDTA; ethylenediaminetetraacetic acid; GFP, green fluorescent protein; IPTG, isopropyl β-D-1-thiogalactopyranoside; LB, Lysogeny Broth; MBP, maltose binding protein; PCR, polymerase chain reaction; Pfu, *Pyrococcus furiosus;* SH, src-homology.

http://dx.doi.org/10.1016/j.ab.2014.06.021
0003-2697/

fragmentation techniques has some degree of sequence and/or positional bias but often the effect on the fragment library composition can be minimized by careful experimental design. Depending on the methods of fragmentation and cloning used, the probability of a gene fragment being in the correct open reading frame in the subsequent expression and solubility assay varies from 1/3 to 1/18. Potential bias in the protein fragments screened due to open reading frame selection can be reduced by using a mixture of nine different frame-shift vectors [18]. In addition, methods have been developed to select only fragments that are in the correct open reading frame prior to solubility screening [14,20,21], substantially increasing the efficiency of the screening process.

There are also several methods for medium to high-throughput solubility screening of protein fragments when expressed in *Escherichia coli*. Fluorescence screens have been developed based on fusing fragments to GFP. Of particular note is the split-GFP assay in which fragments are expressed fused to the last strand of the β-barrel of GFP (GFP11) [22,23]. If the fragment is soluble then GFP11 is available to bind to the nonfluorescent remainder of GFP (GFP1–10) when it is subsequently introduced, reconstituting GFP fluorescence. In the CoFi (colony filtration) method, fragments are expressed with a short tag, cells are lysed, and soluble proteins are transferred to a membrane that is immunochemically probed for tagged protein [13,15]. The ESPRIT (expression of soluble proteins by random incremental truncation) method is similar but fragments are expressed with tags at both ends for immunochemical probing and clones are validated with high-throughput expression and purification trials [12,21]. Life/death colony assays for solubility have also been developed, in which fragments are fused to proteins that confer antibiotic resistance [19]. However, a disadvantage of any approach where fragments are fused to a large protein is that these may exert a significant carrier effect on otherwise insoluble fragments, leading to false positive results.

Although the published methodologies have been used successfully to map protein domains [7,19], these procedures have not been widely adopted. This may be due to the labor-intensive nature of some screens or the need for robotics to carry out high-throughput screening of clones. If more widely implemented, these random fragmentation and screening techniques have the potential to allow many new proteins to be studied at a molecular level.

In this study we investigated whether simple, medium-throughput and low-cost fragmentation, screening, and analysis protocols could be combined to identify domain boundaries. We adapted the split-GFP solubility assay, developed by Waldo and co-workers [22,23], to screen gene fragment libraries created with two different methods (Fig. 1). In contrast to previous studies, we did not focus on directly optimizing fragment solubility through a multistep screening process. Instead we globally analyzed the fragment solubility data to infer the domain boundaries. Here we illustrate the effectiveness of this Domain Seeking methodology by applying it to a structurally characterized protein, human p85α [24–33].

Materials and methods

Reagents

The vector pBAD-MCS was supplied by the Protein Purification and Expression Facility at the European Molecular Biology Laboratory, Heidelberg. The pET_GFP1–10 plasmid for the split-GFP assay was a kind gift from Geoff Waldo at the Los Alamos National Laboratory (USA). DNA primers were obtained from Integrated DNA Technologies. The gene for human p85α, codon optimized for expression in *Escherichia coli*, was synthesized by GeneArt (Germany) and supplied in the vector pMK. DNA was extracted/purified using Nucleospin Gel and PCR clean-up kits (Macherey Nagel, Germany). Chemical reagents were from Sigma (USA), Life Technologies (USA), or Pure Science (NZ).

Construction of the pBAD_GFP11_T7LysH17A vector for the split-GFP assay

A 165 base pair (bp) DNA cassette (Fig. 2) for expression of protein fragments with an N-terminal His_6-tag and a C-terminal GFP11 tag, and restriction enzyme sites for sticky end (SpeI and XhoI) or blunt-end (PvuII) ligation, was created by an overlap extension PCR. An existing XhoI site was removed from the multiple cloning site of the pBAD-MCS vector using QuikChange mutagenesis (Stratagene) with the mutagenic primer 5′-GCTTGC GGCCGCACTCG*T*GAGCTTGGCTGTTTTGG-3′ and its complement. The DNA cassette was then inserted between the NcoI and the HindIII sites of the mutated pBAD-MCS vector to generate the pBAD_GFP11 vector.

The pBAD_GFP11 vector was further modified to express basal levels of the T7 polymerase inhibitor, T7 lysozyme. A 636 bp fragment of the plasmid pLysS [34], encompassing the gene for T7 lysozyme, was amplified with 5′ and 3′ SphI restriction enzymes sites using PCR with the primers 5′-CTGTGCATGCGGCCCATTGGCTGCCTC-3′ and 5′-CGGCGTAGAGCATGCGGGTCCCCTTTGATAGATTAA-3′. This fragment was then ligated into a SphI site in a nonessential region of the pBAD_GFP11 vector to create pBAD_GFP11_T7Lys (Supplementary Fig. 1).

Finally, the mutation H17A was introduced into the T7 lysozyme gene to reduce its amidase activity, using a two-step megaprimer-based site-directed mutagenesis method [35]. In brief, a 557 bp megaprimer was amplified by PCR using the H17A mutagenic primer 5′-GACGCAATCTTTGTT*GCC*TGCTCGGCTACCAGG-3′ and a primer flanking the T7 lysozyme gene, 5′-GGCCCATTGGCTGCCTC-3′. The megaprimer was then employed for standard QuikChange mutagenesis, generating the plasmid pBAD_GFP11_T7LysH17A used for screening of p85α fragment libraries. This vector is available from Addgene (plasmid 59591).

Testing functionality of the pBAD_GFP11 vector series

The pBAD_GFP11 vector and its variants were tested in the split-GFP solubility assay using maltose binding protein (MBP) and an insoluble truncated form of MBP consisting of residues 1–183 (MBP1–183). The genes for MBP and MBP1–183 were amplified by PCR and directionally inserted into pBAD_GFP11 or related variants using the available SpeI and XhoI sites. The expression vectors were then transformed into BL21(DE3) Gold cells carrying pET_GFP1–10, +/− pLysS, before conducting the split-GFP solubility assay. For all experiments employing plasmid pLysS the media were supplemented with 10 μg/ml chloramphenicol to maintain positive selection for the plasmid.

Split-GFP solubility assay

The protocol for the *in vivo* split-GFP solubility screen was similar to that originally described by Waldo and co-workers [22,23]. Libraries of pBAD_GFP11_T7LysH17A vectors carrying fragments of the target gene, created as described below, were transformed into BL21(DE3) Gold/pET_GFP1–10 cells by electroporation. Cells were plated on prewetted, supported nitrocellulose membranes (Pall Corporation, USA) placed on 12 × 12 cm square LB agar plates supplemented with 100 μg/ml ampicillin and 50 μg/ml kanamycin. Plates were incubated at 15 h for 37 °C. To achieve an appropriate colony density (~500 colonies per plate) various dilutions of the transformed cells were plated on the first day. The remainder of the transformed cells were stored at 4 °C overnight and plated out on the second day at the appropriate dilution.