Published in final edited form as:
Curr Opin Struct Biol. 2013 June ; 23(3): 403–408. doi:10.1016/j.sbi.2013.03.004.

Library methods for structural biology of challenging proteins and their complexes

Darren J. Hart and
EMBL Grenoble Outstation and Unit of Virus Host-Cell Interactions, UMI3265 UFJ-EMBLCNRS, Grenoble, France

Geoffrey S. Waldo
Bioscience Division, MS-M888, Los Alamos National Laboratory, Bikini Atoll Rd, SM30, Los Alamos, NM 87545

Abstract

Genetic engineering of constructs to improve solubility or stability is a common approach, but it is often unclear how to obtain improvements. When the domain composition of a target is poorly understood, or if there are insufficient structure data to guide sited directed mutagenesis, long iterative phases of subcloning or mutation and expression often prove unsuccessful despite much effort. Random library approaches can offer a solution to this problem and involve construction of large libraries of construct variants that are analysed via screens or selections for the desired phenotype. Huge improvements in construct behaviour can be achieved rapidly with no requirement for prior knowledge of the target. Here we review the development of these experimental strategies and recent successes.

Introduction

Obtaining milligrams of well-behaving, monodisperse soluble protein is a common limiting step in structural biology; this also applies to vaccinology, many biophysical methods and high throughput screening. Recombinant proteins often express insolubly, as soluble aggregates, may be proteolysed or are undetectable in cell extracts. A common strategy for improving the expression of a problematic target is to modify the target gene sequence by i) PCR subcloning at putative domain boundaries predicted from alignments of similar sequences, or other information e.g. disorder predictions or deletion studies; ii) introduction of stabilising or solubilising mutations by site-directed mutagenesis, usually guided by some preexisting structural data. These two construct engineering approaches, combined with a quick expression and purification test in *E. coli*, are the basic molecular biology tools of structural biologists.

The limitations of rational construct engineering are obvious to most people who have attempted it. Subcloning of domains (or multi-domains) requires *a priori* the prediction of domain locations, but furthermore that constructs that are well translated and folded in the recombinant host. Domain boundaries are usually identified via alignment of similar sequences; however some proteins show little or no similarity to any other known target. Even when similar sequences are available, well designed constructs may still not yield soluble protein, perhaps due to unmet requirements for binding partners, chaperones, redox environments, or stabilising flanking sequences beyond the conserved domain cores. Introduction of solubilising or stabilising point mutations is even more challenging and requires pre-existing structural data e.g. similar structures or homology models. Most significantly, the properties that influence solubility, stability and efficient folding are not well understood, making their rational design a hit-and-miss affair.

Random library methods offer an alternative approach to identifying better behaving constructs. A genetically diverse pool of gene variants is constructed from which improved clones are identified via a screen or selection process (Fig. 1). This well established workflow resembles evolution by natural selection and has been termed ‘directed evolution’ (also *in vitro* evolution) [1]. It has underpinned numerous impressive successes including fully human antibodies by phage display used currently against a number of diseases [2], DNA binding proteins with genome specificity [3,4] and enzymes with improved activity, stability and substrate profiles used in industrial processes [5] and domestic products e.g. laundry powder proteases [6]. Structural biology and protein engineering are closely related disciplines; the former is used to design the input and analyse the output of the latter. Recently, random library methods have been applied to the structural biology process itself with advances in the experimental definition of well behaving soluble protein domains, improved protein stability and consequent crystallisation likelihood, and improvement of yield via more efficient folding pathways.

Features of random library strategies for improving protein quality

Random library approaches comprise a mutation strategy appropriate for the problem in hand, and a screen or selective process powerful enough to isolate rare improved clones in a vast background of neutral or detrimental mutants. The choice of mutation strategy is critical: varying one or both construct termini is appropriate for domain identification (analogous to PCR cloning), whilst optimisation of domain solubility or stability might be best achieved by point mutagenesis. For the screen or selection, the phenotypes of solubility and stability are generic and protein independent; in principle once established, these technologies can be applied to many different unrelated targets. This is in contrast to classical directed evolution (e.g. ligand binding or enzyme engineering) where it is normally necessary to establish specific screens/selections for each system studied. To date, random library methods have mostly been applied to protein production where screens/selections can be been developed, with only limited application to crystallisation [7] and none to improving diffraction data quality; here the low throughput nature of potential screens (crystallisation trials, or testing crystals on an X-ray source) prevents analysis of a genetically diverse mutant library at any useful throughput.

Mutation strategies for improved protein expression

Libraries of unidirectionally truncated expression constructs can be generated using the 3' to 5' activity of exonuclease III on dsDNA that has first been cut to generate 5' resistant and 3' substrate overhangs [8,9]. By careful timing of reactions, linear distributions of fragment sizes can be obtained over DNA lengths of at least 3 kb; this has been used on both single and pooled targets e.g. a pooled viral ORFeome [10]. Strikingly, one study on a panel of different human targets showed that short 5' deletions were often sufficient to improve protein solubility [9], perhaps through modification of problematic N-terminal leader sequences or 5' mRNA structure. These libraries are relatively small, comprising a few hundred to a few thousand constructs since the total diversity is defined by the gene length, with one in three constructs cloned being in the correct reading frame.

Libraries of constructs varying at both termini are required if targeting internal domains. These can be prepared by sequential exonuclease III truncations of a precloned target gene, first at one end and then the other [11]. In this way, the insert orientation can be controlled leading to one in nine constructs being in frame with flanking protein or peptide tags. Similarly, dUTP can be incorporated by PCR amplification of the target gene, with subsequent fragmentation by endonuclease V [12], or endonuclease IV coupled to uracil-DNA glycosylase [13]. Other methods for randomly fragmenting genes include physical point sink shearing [14] and random primed PCR [15]. These latter methods generate fragments that must be cloned with no control over insert orientation resulting in 1 in 18 clones being correctly oriented and in frame. These reading frame inefficiencies are usually ignored; however genetic selections to eliminate incorrect reading frames have been described including ORF-selector ESPRIT [16] based upon an earlier split intein technology [17], or through insertion into dihydrofolate reductase (DHFR) [18].

The random point mutagenesis of a gene sequence can be achieved by plasmid amplification in an *E. coli* mutator strain [19] or by error prone PCR [20], both of which introduce simple point mutations at a tuneable density. *In vitro* recombination techniques such as the staggered extension process (StEP) [21,22] or DNA shuffling [23] produce more efficient libraries since they permit exchange of mutations between lineages of amplicons, simulating the natural process of genetic recombination. Chimaeric genes can be generated by family shuffling [24] of pooled, closely related sequences. These protocols are long established for protein engineering applications; comparatively recent is the incorporation of controlled sequence variants during gene synthesis, e.g. the Slonomics technology [25], which provides the advantage of controlling the identity and frequency of amino acids at each position and avoiding unwanted stop codons.

Screening and selection strategies

Library clones that express in-frame constructs (*vide supra*) can be further winnowed to identify soluble, stable protein variants. Unstable variants can interfere with the folding (and activity) of a fused 'reporter protein' domain. GFPs [18,26,27], are used as fused 'folding reporters' to *screen* libraries of 10^5-10^6 clones on plates or by FACS (Fig. 2). Soluble clones may be *selected* from larger libraries (>10^7 clones) via fusion to antibiotic resistance

proteins such as chloramphenicol acetyl transferase [28] or murine dihydrofolate reductase [12,29] and plating on a selective growth medium. Unstable proteins or those with internal ribosome binding sites can give false positives if only one tag is used. To overcome these limitations, libraries can be inserted between two pieces of a reporter such as GFP [27], or between a leader sequence (TAT, SRP, sec) and beta-lactamase where the marker phenotype (fluorescence, ampicillin resistance) is only achieved when both ends are present. Reporter protein fusions are convenient but can affect solubility due to passenger solubilisation effects where the behaviour of the target is modified by the presence of a large soluble fusion partner. This can be eliminated by using short peptide tag fusions instead. Proteins tagged with the 15 amino acid GFP beta strand 11 can be detected by *in vivo* or *in vitro* complementation with a truncated GFP form comprising beta strands 1-10 [30]. In the ESPRIT method (Fig. 3), *in vivo* biotinylation of a 15 amino acid peptide appended to the C terminus of truncated protein inserts is used as an indicator of solubility in a printed colony array format. Simultaneous detection of an N-terminal hexahistidine tag indicates that proteins are in frame and undegraded [8,31]. Physical screens test whether proteins can be purified and bind to membranes or bead affinity media using small cultures [13,32] or colonies on plates (CoFi blot method, [33]). Phage display has been used to screen for protease resistance, solubility and capture [34] or by selective infectivity of the phage tip [35]. The use of the TAT transport pathway should extend these methods to larger proteins that fold in the cytoplasm [36].

Recent developments

Interest in protein complexes has driven the development of enabling technologies for identifying suitable well behaving candidates for structural studies. In a proof-of-principle demonstration of a variation of the combinatorial domain hunting (CDH) method of the Domainex company termed CDH^2, a bait protein was immobilised to resin using one tag and co-expressed prey proteins binding to the bait were then detected using a second tag [37]. In CoESPRIT, libraries of a truncated viral “prey” target were transformed into an *E. coli* strain that coexpresses a human interacting “bait” with expression analysis in a printed colony array format [38]. After solubility analysis of the prey, hits were purified and the presence of interacting bait proteins determined using a bait-specific tag. In a similar approach, Waldo & co-workers co-expressed hexahistidine-tagged bait and split-GFP tagged prey proteins from *E. coli* on permeable membranes. Soluble fluorescent complexes were captured on IMAC resin in an underlying agarose layer [39]. To tackle membrane proteins, Nordlund & co-workers combined the CoFi blot with detergent solubilisation [40,41], while others have used membrane protein-GFP fusions to screen expression, scale-up and production [42,43]. Plückthun & co-workers used FACS to measure labelled antagonist binding to inner-membrane expressed GPCRs in permeabilised *E. coli*. GPCR stability was dramatically improved and expression increased up to 50-fold, and variants with modified substrate specificity were generated [44,45].

Notable case studies

Hart & co-workers used ESPRIT to identify a folded domain of influenza polymerase PB2 with specific cap binding activity. The X-ray structure with bound cap analogue m^{7}GTP at

2.3 Å resolution revealed the mode of ligand binding [11]. Similarly, other domains were discovered and crystallised with roles in viral host adaptation [46] and nuclear transport [47]. All were novel folds and unpredicted from sequence. The terminase from HCMV was also discovered and crystallised via ESPRIT [48], as was the bacterial phosphatase SpoIIE [49,50]. Pedelacq et. al used a split GFP assay to identify several compact domains from each module of human p85α. X-ray quality crystals were also obtained for the acyl-transferase, dehydratase, and enoyl-reductase domains of *Mycobacterium tuberculosis* PpsC [18]. A better expressing and crystallisable form of MEK-1 kinase was obtained using CDH [51]. Structures have been solved for several domains delineated using the GFP folding assay including telomerase reverse transcriptase [52]; a new fold from the pore-targeting domain of nucleoporin Nup98 [53]; an active GTP binding domain of P element transposase precisely delineating the DNA-binding and dimerisation elements of the primary sequence [54]. Dyson & co-workers used eukaryotic DHFR as a folding reporter to discover domains clustering in the ETS module of the transcription factor Fli1, as well as domains of Pecam1 [12]. The Winter lab used phage display to screen an *E. coli* genome fragment library and identified 124 protease-resistant globular domains with unfolding energies ΔG_u ranging from 3.8-6.6 kcal/mol. Boundaries correlated with bioinformatic predictions [34]. Seitz et. al used the GFP reporter and engineered a stable version of the human glucocorticoid receptor ligand-binding domain, an important drug target for the treatment of several diseases. A four-point mutant increased thermal stability by more than 8°C and yield after expression in *E. coli* by 26-fold. In all, the structures of 3 variants were solved at resolutions as high as 1.5 Å [55]. The mammalian paraoxonases PON1 and PON3 were expressed in soluble form via family shuffling of human, mouse, rat and rabbit homologues [56] leading eventually to a structure of PON1 [57].

Future perspectives

As structural genomics matures, these versatile screening strategies are starting to bridge the gap between structural biology and cellular biology. For example, split GFP can be used for tracking pathogen effector proteins in host cells [58], mapping cell-cell contacts [59], and viral/cell membrane fusion [60]. Future applications may include tagging membrane proteins on either side of the cellular lumen. Domain screening technologies coupled with deep sequencing will likely play increasingly important roles in antigen generation for phage based antibody development and vaccinology [61]. One can expect hybrid approaches to increasing protein stability of individual proteins and protein complexes combining computational design to create ‘smart’ libraries towards stability or activity (Rosetta3, [62]), in-house microfluidic gene synthesis [63,64] of corresponding constrained diversity DNA libraries, and microfluidic screens or selections for protein stability and activity [65,66].

Acknowledgments

DJH acknowledges EU FP7 contracts P-CUBE (227764) and BioStruct-X (283570) for financial support of library methods research. GSW wishes to acknowledge NIH GM 98177

References

1. Goldsmith M, Tawfik DS. Directed enzyme evolution: beyond the low-hanging fruit. Curr. Opin. Struct. Biol. 2012; 22:406–412. [PubMed: 22579412]
2. Lonberg N. Fully human antibodies from transgenic mouse and phage display platforms. Curr. Opin. Immunol. 2008; 20:450–459. [PubMed: 18606226]
3. Pabo CO, Peisach E, Grant RA. Design and selection of novel Cys2His2 zinc finger proteins. Annu. Rev. Biochem. 2001; 70:313–340. [PubMed: 11395410]
4. Perez-Pinera P, Ousterout DG, Gersbach CA. Advances in targeted genome editing. Current Opinion in Chemical Biology. 2012; 16:268–277. [PubMed: 22819644]
5. Dalby PA. Strategy and success for the directed evolution of enzymes. Current Opinion in Structural Biology. 2011; 21:473–480. [PubMed: 21684150]
6. Ness JE, Kim S, Gottman A, Pak R, Krebber A, Borchert TV, Govindarajan S, Mundorff EC, Minshull J. Synthetic shuffling expands functional protein diversity by allowing amino acids to recombine independently. Nature Biotechnology. 2002; 20:1251–1255.
7. Keenan RJ. DNA shuffling as a tool for protein crystallization. Proceedings of the National Academy of Sciences. 2005; 102:8887–8892.
8. Yumerefendi H, Tarendeau F, Mas PJ, Hart DJ. ESPRIT: an automated, library-based method for mapping and soluble expression of protein domains from challenging targets. J Struct Biol. 2010; 172:66–74. [PubMed: 20206698]
9. Cornvik T, Dahlroth S-L, Magnusdottir A, Flodin S, Engvall B, Lieu V, Ekberg M, Nordlund P. An efficient and generic strategy for producing soluble human proteins and domains in E. coli by screening construct libraries. Proteins. 2006; 65:266–273. [PubMed: 16948159]
10. Dahlroth S-L, Lieu V, Haas J, Nordlund P. Screening colonies of pooled ORFeomes (SCOOP): a rapid and efficient strategy for expression screening ORFeomes in Escherichia coli. Protein Expr. Purif. 2009; 68:121–127. [PubMed: 19635569]
11. Guilligay D, Tarendeau F, Resa-Infante P, Coloma R, Crepin T, Sehr P, Lewis J, Ruigrok RWH, Ortin J, Hart DJ, et al. The structural basis for cap binding by influenza virus polymerase subunit PB2. Nat. Struct. Mol. Biol. 2008; 15:500–506. [PubMed: 18454157]
12. Dyson MR, Perera RL, Shadbolt SP, Biderman L, Bromek K, Murzina NV, McCafferty J. Identification of soluble protein fragments by gene fragmentation and genetic selection. Nucleic Acids Res. 2008; 36:e51. [PubMed: 18420658]
13. Reich S, Puckey LH, Cheetham CL, Harris R, Ali AAE, Bhattacharyya U, Maclagan K, Powell KA, Prodromou C, Pearl LH, et al. Combinatorial Domain Hunting: An effective approach for the identification of soluble protein domains adaptable to high-throughput applications. Protein Sci. 2006; 15:2356–2365. [PubMed: 17008718]
14. Thorstenson YR, Hunicke-Smith SP, Oefner PJ, Davis RW. An automated hydrodynamic process for controlled, unbiased DNA shearing. Genome Res. 1998; 8:848–855. [PubMed: 9724331]
15. Kawasaki M, Inagaki F. Random PCR-based screening for soluble domains using green fluorescent protein. Biochem. Biophys. Res. Commun. 2001; 280:842–844. [PubMed: 11162598]
16•. An Y, Yumerefendi H, Mas PJ, Chesneau A, Hart DJ. ORF-selector ESPRIT: A second generation library screen for soluble protein expression employing precise open reading frame selection. J. Struct. Biol. 2011; 175:189–197. [PubMed: 21515383] [Combination of domain screening with pre-selection against out of frame constructs resulting in a very high degree of sampling.]
17. Gerth ML, Patrick WM, Lutz S. A second-generation system for unbiased reading frame selection. Protein Eng. Des. Sel. 2004; 17:595–602. [PubMed: 15331775]
18. Pedelacq J-D, Nguyen HB, Cabantous S, Mark BL, Listwan P, Bell C, Friedland N, Lockard M, Faille A, Mourey L, et al. Experimental mapping of soluble protein domains using a hierarchical approach. Nucleic Acids Res. 2011; 39:e125. [PubMed: 21771856]
19. Muteeb G, Sen R. Random mutagenesis using a mutator strain. Methods Mol. Biol. 2010; 634:411–419. [PubMed: 20677000]
20. Cadwell RC, Joyce GF. Mutagenic PCR. Genome Res. 1994; 3:S136–S140.

21. Aguinaldo AM, Arnold F. Staggered extension process (StEP) in vitro recombination. Methods Mol. Biol. 2002; 192:235–239. [PubMed: 12494655]
22. Zhao H, Zha W. In vitro “sexual” evolution through the PCR-based staggered extension process (StEP). Nature Protocols. 2006; 1865–1871; 1
23. Stemmer WP. Rapid evolution of a protein in vitro by DNA shuffling. Nature. 1994; 370:389–391. [PubMed: 8047147]
24. Crameri A, Raillard SA, Bermudez E, Stemmer WP. DNA shuffling of a family of genes from diverse species accelerates directed evolution. Nature. 1998; 391:288–291. [PubMed: 9440693]
25. Van den Brulle J, Fischer M, Langmann T, Horn G, Waldmann T, Arnold S, Fuhrmann M, Schatz O, O'Connell T, O'Connell D, et al. A novel solid phase technology for high-throughput gene synthesis. BioTechniques. 2008; 45:340–343. [PubMed: 18778261]
26. Heddle C, Mazaleyrat SL. Development of a screening platform for directed evolution using the reef coral fluorescent protein ZsGreen as a solubility reporter. Protein Eng. Des. Sel. 2007; 20:327–337. [PubMed: 17584755]
27. Cabantous S, Rogers Y, Terwilliger TC, Waldo GS. New Molecular Reporters for Rapid Protein Folding Assays. PLoS ONE. 2008; 3:e2387. [PubMed: 18545698]
28. Maxwell KL, Mittermaier AK, Forman-Kay JD, Davidson AR. A simple in vivo assay for increased protein solubility. Protein Sci. 1999; 1908–1911; 8
29. Liu J-W, Boucher Y, Stokes HW, Ollis DL. Improving protein solubility: The use of the Escherichia coli dihydrofolate reductase gene as a fusion reporter. Protein Expression and Purification. 2006; 47:258–263. [PubMed: 16403649]
30. Cabantous S, Terwilliger TC, Waldo GS. Protein tagging and detection with engineered self-assembling fragments of green fluorescent protein. Nat. Biotechnol. 2005; 23:102–107. [PubMed: 15580262]
31. Yumerefendi H, Desravines DC, Hart DJ. Library-based methods for identification of soluble expression constructs. Methods. 2011; 55:38–43. [PubMed: 21723393]
32. Listwan P, Pédelacq J-D, Lockard M, Bell C, Terwilliger TC, Waldo GS. The optimization of in vitro high-throughput chemical lysis of Escherichia coli. Application to ACP domain of the polyketide synthase ppsC from Mycobacterium tuberculosis. Journal of Structural and Functional Genomics. 2010; 11:41–49. [PubMed: 20069378]
33. Cornvik T, Dahlroth S-L, Magnusdottir A, Herman MD, Knaust R, Ekberg M, Nordlund P. Colony filtration blot: a new screening method for soluble protein expression in Escherichia coli. Nat. Methods. 2005; 2:507–509. [PubMed: 15973420]
34. Christ D, Winter G. Identification of Protein Domains by Shotgun Proteolysis. Journal of Molecular Biology. 2006; 358:364–371. [PubMed: 16516923]
35. Sieber V, Plückthun A, Schmid FX. Selecting proteins with improved stability by a phage-based method. Nature Biotechnology. 1998; 16:955–960.
36. Speck J, Arndt KM, Muller KM. Efficient phage display of intracellularly folded proteins mediated by the TAT pathway. Protein Engineering Design and Selection. 2011; 24:473–484.
37. Maclagan K, Tommasi R, Laurine E, Prodromou C, Driscoll PC, Pearl LH, Reich S, Savva R. A combinatorial method to enable detailed investigation of protein-protein interactions. Future Med Chem. 2011; 3:271–282. [PubMed: 21446842]
38••. An Y, Meresse P, Mas PJ, Hart DJ. CoESPRIT: a library-based construct screening method for identification and expression of soluble protein complexes. PLoS ONE. 2011; 6:e16261. [PubMed: 21364980] [Random library construct screening in the presence of bait proteins for the formation of soluble purifiable complexes at multimilligram yields.]
39•. Lockard MA, Listwan P, Pedelacq J-D, Cabantous S, Nguyen HB, Terwilliger TC, Waldo GS. A high-throughput immobilized bead screen for stable proteins and multi-protein complexes. Protein Eng. Des. Sel. 2011; 24:565–578. [PubMed: 21642284] [Using beads immobilized in agarose under colonies on membranes, the split GFP method is extended to monitor both expression and in vitro solubility of proteins and protein complexes, and the results correlated with published crystal structures of the targets.]
40. Eshaghi S. An efficient strategy for high-throughput expression screening of recombinant integral membrane proteins. Protein Science. 2005; 14:676–683. [PubMed: 15689514]

41. Molina DM, Cornvik T, Eshaghi S, Haeggström JZ, Nordlund P, Sabet MI. Engineering membrane protein overproduction in Escherichia coli. Protein Science. 2008; 17:673–680. [PubMed: 18305199]

42. Drew D, Lerch M, Kunji E, Slotboom D-J, De Gier J-W. Optimization of membrane protein overexpression and purification using GFP fusions. Nature Methods. 2006; 3:303–313. [PubMed: 16554836]

43. Hammon J, Palanivelu DV, Chen J, Patel C, Minor DL Jr. A green fluorescent protein screen for identification of well-expressed membrane proteins from a cohort of extremophilic organisms. Protein Sci. 2009; 18:121–133. [PubMed: 19177357]

44. Sarkar CA, Dodevski I, Kenig M, Dudli S, Mohr A, Hermans E, Pluckthun A. Directed evolution of a G protein-coupled receptor for expression, stability, and binding selectivity. Proceedings of the National Academy of Sciences. 2008; 105:14808–14813.

45. Schlinkmann KM, Hillenbrand M, Rittner A, Künz M, Strohner R, Plückthun A. Maximizing Detergent Stability and Functional Expression of a GPCR by Exhaustive Recombination and Evolution. Journal of Molecular Biology. 2012; 422:414–428. [PubMed: 22683350]

46. Tarendeau F, Crepin T, Guilligay D, Ruigrok RWH, Cusack S, Hart DJ. Host determinant residue lysine 627 lies on the surface of a discrete, folded domain of influenza virus polymerase PB2 subunit. PLoS Pathog. 2008; 4:e1000136. [PubMed: 18769709]

47. Tarendeau F, Boudet J, Guilligay D, Mas PJ, Bougault CM, Boulo S, Baudin F, Ruigrok RWH, Daigle N, Ellenberg J, et al. Structure and nuclear import function of the C-terminal domain of influenza virus polymerase PB2 subunit. Nat. Struct. Mol. Biol. 2007; 14:229–233. [PubMed: 17310249]

48••. Nadal M, Mas PJ, Blanco AG, Arnan C, Solà M, Hart DJ, Coll M. Structure and inhibition of herpesvirus DNA packaging terminase nuclease domain. Proc. Natl. Acad. Sci. U.S.A. 2010; 107:16078–16083. [PubMed: 20805464] [Random library construct screening revealed protein boundaries of the terminase domain leading directly to a crystal structure of an antiviral drug target.]

49. Rawlings AE, Levdikov VM, Blagova E, Colledge VL, Mas PJ, Tunaley J, Vavrova L, Wilson KS, Barak I, Hart DJ, et al. Expression of soluble, active fragments of the morphogenetic protein SpoIIE from Bacillus subtilis using a library-based construct screen. Protein Eng. Des. Sel. 2010; 23:817–825. [PubMed: 20817757]

50. Levdikov VM, Blagova EV, Rawlings AE, Jameson K, Tunaley J, Hart DJ, Barak I, Wilkinson AJ. Structure of the Phosphatase Domain of the Cell Fate Determinant SpoIIE from Bacillus subtilis. Journal of Molecular Biology. 2012; 415:343–358. [PubMed: 22115775]

51•. Meier C, Brookings DC, Ceska TA, Doyle C, Gong H, McMillan D, Saville GP, Mushtaq A, Knight D, Reich S, et al. Engineering human MEK-1 for structural studies: A case study of combinatorial domain hunting. Journal of Structural Biology. 2012; 177:329–334. [PubMed: 22245778] [Interesting case study on a pharmaceutically important kinase.]

52. Jacobs SA, Podell ER, Wuttke DS, Cech TR. Soluble domains of telomerase reverse transciptase identified by high-throughput screening. Protein Sci. 2005; 14:2051–2058. [PubMed: 16046627]

53. Hodel AE, Hodel MR, Griffis ER, Hennig KA, Ratner GA, Xu S, Powers MA. The three-dimensional structure of the autoproteolytic, nuclear pore-targeting domain of the human nucleoporin Nup98. Mol. Cell. 2002; 10:347–358. [PubMed: 12191480]

54. Sabogal A, Rio DC. A green fluorescent protein solubility screen in E. coli reveals domain boundaries of the GTP-binding domain in the P element transposase. Protein Science. 2010; 19:2210–2218. [PubMed: 20842711]

55. Seitz T, Thoma R, Schoch GA, Stihle M, Benz J, D'Arcy B, Wiget A, Ruf A, Hennig M, Sterner R. Enhancing the Stability and Solubility of the Glucocorticoid Receptor Ligand-Binding Domain by High-Throughput Library Screening. Journal of Molecular Biology. 2010; 403:562–577. [PubMed: 20850457]

56. Aharoni A, Gaidukov L, Yagur S, Toker L, Silman I, Tawfik DS. Directed evolution of mammalian paraoxonases PON1 and PON3 for bacterial expression and catalytic specialization. Proc Natl Acad Sci U S A. 2004; 101:482–7. [PubMed: 14695884]

57. Harel M, Aharoni A, Gaidukov L, Brumshtein B, Khersonsky O, Meged R, Dvir H, Ravelli RBG, McCarthy A, Toker L, et al. Structure and evolution of the serum paraoxonase family of detoxifying and anti-atherosclerotic enzymes. Nat. Struct. Mol. Biol. 2004; 11:412–419. [PubMed: 15098021]

58. Van Engelenburg SB, Palmer AE. Imaging type-III secretion reveals dynamics and spatial segregation of Salmonella effectors. Nature Methods. 2010; 7:325–330. [PubMed: 20228815]

59. Feinberg EH, VanHoven MK, Bendesky A, Wang G, Fetter RD, Shen K, Bargmann CI. GFP Reconstitution Across Synaptic Partners (GRASP) Defines Cell Contacts and Synapses in Living Nervous Systems. Neuron. 2008; 57:353–363. [PubMed: 18255029]

60. Ishikawa H, Meng F, Kondo N, Iwamoto A, Matsuda Z. Generation of a dual-functional split-reporter protein for monitoring membrane fusion using self-associating split GFP. Protein Engineering Design and Selection. 2012; 25:813–820.

61•. D'Angelo S, Velappan N, Mignone F, Santoro C, Sblattero D, Kiss C, Bradbury AR. Filtering “genic” open reading frames from genomic DNA samples for advanced annotation. BMC Genomics. 2011; 12:S5. [PubMed: 21810207] [ORFs from a DNA fragment library from Clostridium thermocellum were selected for frame and folding using a variety of signal peptides, and the pools deep sequenced to reveal interesting patterns in specificity or promiscuity for transit, export, and scale-up for protein production.]

62•. Richter F, Leaver-Fay A, Khare SD, Bjelic S, Baker D. De Novo Enzyme Design Using Rosetta3. PLoS ONE. 2011; 6:e19230. [PubMed: 21603656] [Authors summarize the state of the art in enzyme redesign using real-world examples. A must for crystallographers interested in the used of computational methods for improving protein function.]

63. Borovkov AY, Loskutov AV, Robida MD, Day KM, Cano JA, Le Olson T, Patel H, Brown K, Hunter PD, Sykes KF. High-quality gene assembly directly from unpurified mixtures of microarray-synthesized oligonucleotides. Nucleic Acids Research. 2010; 38:e180–e180. [PubMed: 20693531]

64. Kong DS, Carr PA, Chen L, Zhang S, Jacobson JM. Parallel gene synthesis in a microfluidic device. Nucleic Acids Research. 2007; 35:e61–e61. [PubMed: 17405768]

65. Fallah-Araghi A, Baret J-C, Ryckelynck M, Griffiths AD. A completely in vitro ultrahigh-throughput droplet-based microfluidic screening system for protein engineering and directed evolution. Lab on a Chip. 2012; 12:882. [PubMed: 22277990]

66••. Agresti JJ, Antipov E, Abate AR, Ahn K, Rowat AC, Baret J-C, Marquez M, Klibanov AM, Griffiths AD, Weitz DA. Ultrahigh-throughput screening in drop-based microfluidics for directed evolution. Proc. Natl. Acad. Sci. U.S.A. 2010; 107:4004–4009. [PubMed: 20142500] [Authors demonstrate a microfluidic directed evolution platform that should have a strong impact on future efforts to improve proteins through evolution. They improve kinetics of horse radish peroxidase 10-fold using only 150 μL of sample, achieving a 1,000-fold increase in speed and a 1-million-fold reduction in cost relative to robotic methods.]

Highlights

- Random library or directed evolution strategies in structural biology
- Mutation and screening for soluble protein expression
- Review of mutation and screening strategies with case studies

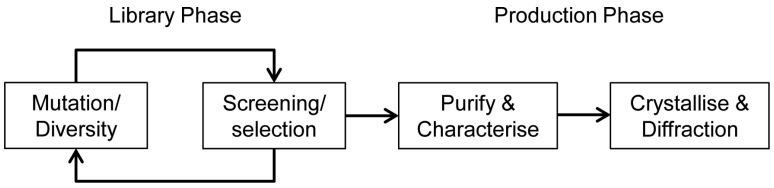

Figure 1.

The logic of random library methods for structural biology. An initial library phase comprises cycles of gene mutation by point or truncation mutagenesis protocols, and a phenotypic screening for solubility. Promising clones are isolated from the library and, after validation with a small-scale test, are purified, characterised and crystallised.

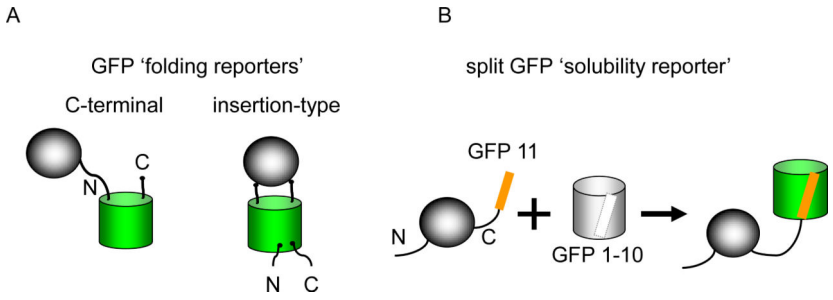

Figure 2.

Three major iterations of the GFP reporters are depicted: (A) So-called 'GFP folding reporters' comprising C-terminal and insertion type fusions where misfolding of the fusion protein (grey) results in misfolding of the fused GFP domain. The insertion type GFP folding reporter uses a circular permutant GFP starting at amino acid 172, while test proteins are inserted between the native GFP N and C termini. This topology reduces false positives from internal start sites or unstable truncated proteins that may otherwise plague C-terminal folding reporters. (B) Split GFP uses complementation of two pieces of GFP, GFP S11 tag and 'detector' GFP 1-10, to signal if fusion is soluble. The GFP fragments spontaneously assembly on if the GFP S11 tag is accessible.

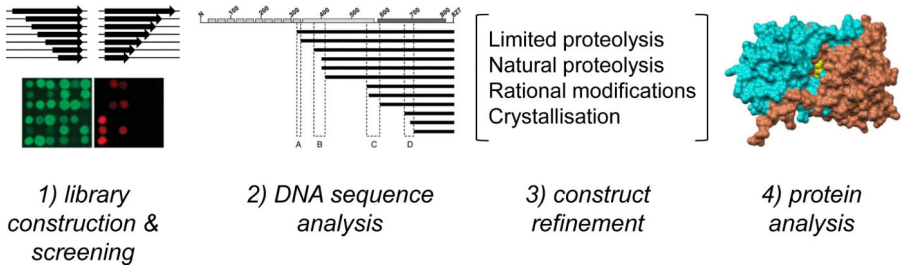

Figure 3.

In the ESPRIT method [8], construct libraries are synthesised using exonuclease III and mung bean nuclease to truncate target genes unidirectionally or bidirectionally at one or both termini respectively. A colony array screen of, typically, 28 thousand clones is performed with fluorescent streptavidin to detect levels of *in vivo* C-terminal biotinylation of the target as an indicator of solubility (green spots), and the presence of a N-terminal hexahistidine tag (red spots). Constructs with both signals are tested in small scale expression trials and sequenced to identify construct boundaries yielding soluble material. These constructs may be used in structural studies directly, or may require further optimisation e.g. using limited proteolysis to remove flexible termini. Figure adapted from a study on SpoIIE in which N-terminal deletion screening yielded soluble protein [49] that, following further optimisation, yielded a 2.6 Å structure of the C-terminal phosphatase domain [50].