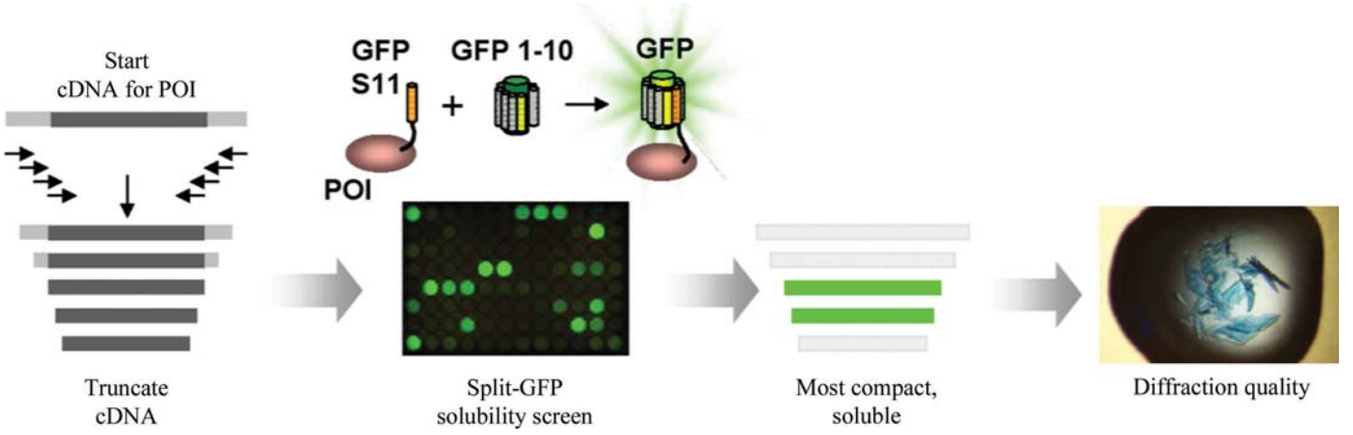


**Figure 4.3.4.1**
A domain-trapping strategy to engineer soluble variants of a protein of interest (POI) for crystallization using the split-GFP complementation methodology. (Figure courtesy of Dr Geoff Waldo, LANL.)

multiple free cysteines. In She2p, an RNA-binding protein, four cysteines (Cys14, Cys68, Cys106 and Cys180) were mutated to serines in order to overcome oxidation and aggregation (Niessing *et al.*, 2004). In an extreme case, that of human maspin, which is a serpin with antitumour activities, all unpaired cysteines were mutated (C20S, C34A, C183S, C205S, C214S, C297S, C373S) in an effort to obtain a soluble crystallizable variant (Al-Ayyoubi *et al.*, 2004).

### 4.3.4. Optimization of target constructs

The N- and C-termini of proteins are often flexible and unstructured (Thornton & Sibanda, 1983), creating a potential entropic impediment to crystallization. Initially, the preferred way to circumvent this problem was to use limited proteolysis to trim off the ends, leaving the stable core of the target protein. This strategy remains useful, particularly in its *in situ* version, in which trace amounts of proteases are added directly to crystallization screens (Dong *et al.*, 2007; Wernimont & Edwards, 2009). However, on the downside it introduces the possibility of heterogeneity in the sample owing to incomplete proteolysis. An alternative route is to first identify the smallest functional fragment of the target protein and to then design and overexpress an appropriately modified gene. A number of options are possible. The simplest is the direct prediction of intrinsically disordered regions from the amino-acid sequence alone (Obradovic *et al.*, 2003; He *et al.*, 2009). The functional core units can also be identified experimentally by mass spectrometry following limited proteolysis (Cohen *et al.*, 1995). Alternatively, deuterium–hydrogen exchange coupled to mass spectrometry (DXMS) may be used to identify fast-exchanging amides that map to unstructured fragments (Hamuro *et al.*, 2003; Pantazatos *et al.*, 2004; Sharma *et al.*, 2009).

Importantly, the choice of optimal N- and C-termini may also critically influence the solubility of the target protein. For example, in the case of MAPKAP kinase 2, 16 truncation variants were assayed, all of which contained the catalytic domain, and were shown to have dramatically differing solubilities and propensities for crystallization (Malawski *et al.*, 2006). Similarly, a series of truncations were screened in order to identify a soluble and crystallizable variant of a three-domain fragment of the Vav1 guanine nucleotide-exchange factor (Brooun *et al.*, 2007). In both these cases only a limited number of rationally designed constructs were screened. However, to increase the prospects of success it is also possible to utilize much larger libraries of variants and screen them *in vivo* using the high-throughput split-GFP complementation assay (Fig. 4.3.4.1; Cabantous & Waldo, 2006).

Another troublesome problem associated with flexible termini is their occasional propensity to form multiple intermolecular contacts, leading to crystal forms that contain multiple copies of the target protein in the asymmetric unit. This has been observed, for example, for *Plasmodium falciparum* peptide deformylase, in which removal of three residues from the N-terminus reduced the number of subunits in the asymmetric unit from ten to two (Robien *et al.*, 2004).

In addition to the disordered N- and C-termini, target proteins may contain internal unstructured regions such as subdomains or loops which can be removed or shortened to reduce conformational heterogeneity. For example, the construct used in the successful crystallization of the HIV gp120 envelope glycoprotein had two flexible loops which were replaced with Gly-Ala-Gly linkages to obtain a crystallizable variant (Kwong *et al.*, 1998, 1999). In the case of 8*R*-lipoxygenase the replacement of a flexible $Ca^{2+}$-dependent membrane-insertion loop consisting of five amino acids by a Gly-Ser dipeptide resulted in crystals that diffracted to a resolution 1 Å higher than the wild-type protein (Neau *et al.*, 2007). An interesting variation of this approach was introduced for the preparation of crystals of the $\beta$-subunit of the signal recognition particle receptor. A 26-residue flexible loop was removed, but instead of replacing it with a shorter sequence the authors connected the native N- and C-termini of the protein using a heptapeptide GGGSGGG, thus creating a circular permutation of the polypeptide chain (Schwartz *et al.*, 2004).

Given that the majority of eukaryotic proteins contain at least one stretch of 40 or more disordered residues (Vucetic *et al.*, 2003), optimization of crystallization targets by removal of these sequences is likely to become a routine strategy.

### 4.3.5. The use of fusion proteins for crystallization

Tags are routinely used in heterologous protein expression in order to enhance folding and solubility and to facilitate purification (Uhlen *et al.*, 1992; Malhotra, 2009). They are either short oligopeptides, such as a hexahistidine, with unique affinity properties or well expressed and highly soluble proteins, such as GST (glutathione *S*-transferase), MBP (maltose-binding protein) or thioredoxin. The tags are inserted into the expression vectors

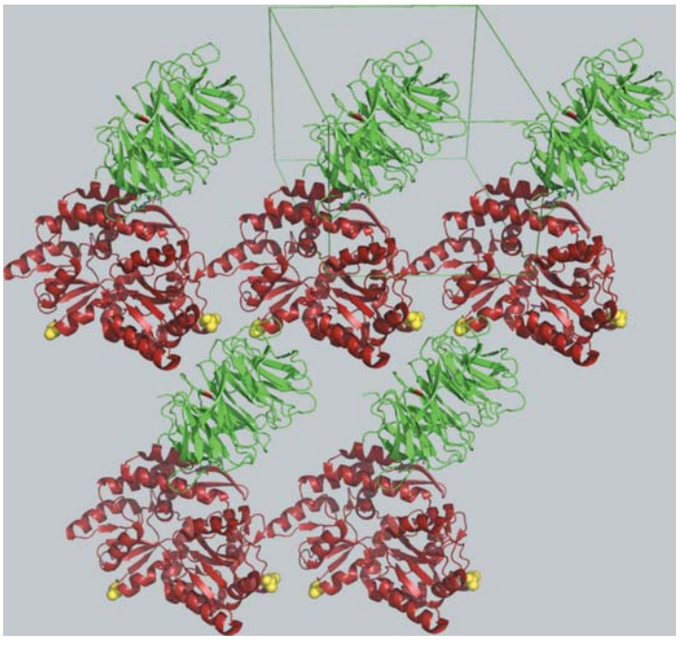

**Figure 4.3.5.1**
An example of the use of a fused carrier protein in crystallization: the crystal structure of the RACK1 protein (green) crystallized in fusion with an engineered variant of the maltose-binding protein (MBP; red); the major crystal contacts are mediated by MBP (PDB code 3dm0; Ullah *et al.*, 2008). The yellow spheres show alanines introduced by site-directed mutagenesis (see text for further details). Figs. 4.3.5.1 and 4.3.9.1 were generated using *PyMOL* (http://www.pymol.org).

downstream or upstream of the target protein and are often separated from it by a protease-sensitive linker sequence. They are cleaved proteolytically following expression and partial purification of the fusion protein and removed, leaving the isolated target ready for crystallization. However, in some cases the target protein may not be adequately soluble after cleavage or may resist crystallization. One of the possible solutions is to use the intact fusion protein in the crystallization screens in the hope that the carrier protein will both confer solubility on the construct and mediate crystal contacts. Not surprisingly, the canonical carrier proteins, all of which crystallize fairly easily on their own, constitute the obvious first choice. Using this strategy, the DNA-binding domain of DNA replication-related element-binding factor (DREF) was crystallized in fusion with *Escherichia coli* GST (Kuge *et al.*, 1997) and the U2AF homology motif (UHM) domain of splicing factor Puf60 was crystallized as a fusion with thioredoxin (Corsini *et al.*, 2008). A key problem limiting the utility of this technique is the inherent flexibility of a two-domain fusion protein, which is detrimental to its crystallizability. A possible solution to this problem is shortening the linker between the two proteins until a relatively rigid construct is identified (Smyth *et al.*, 2003). This approach was successfully pioneered for maltose-binding protein (MBP), which was used as a fusion chaperone to crystallize the human T-cell leukemia virus type 1 gp21 ectodomain fragment (Center *et al.*, 1998). The same strategy was employed in the crystallization of the ZP-N domain of ZP3 (Monne *et al.*, 2008), the islet amyloid polypeptide (IAPP; Wiltzius *et al.*, 2009) and the MAT$\alpha$1 homeodomain (Ke & Wolberger, 2003). Recently, a genetically modified version of MBP (see below) was used as an N-terminal fusion chaperone to crystallize the signal transduction regulator RACK1 from *Arabidopsis thaliana* (Ullah *et al.*, 2008; Fig. 4.3.5.1). Thus, MBP remains the most successful fusion chaperone for protein crystallization, even though the absolute number of proteins crystallized in this way is still limited.

In addition to the canonical fusion chaperones, which were originally designed as affinity tags, other carrier proteins can be used to assist crystallization. For example, a module made up of two sterile $\alpha$ motif (SAM) domains has been engineered to polymerize in response to a pH drop and was shown to drive the crystallization of 11 target proteins in a pilot study (Nauli *et al.*, 2007). In another example, barnase, a secreted ribonuclease from *Bacillus amyloliquefaciens*, was recently used as a carrier protein for crystallization of the disulfide-rich protein McoEeTI (Niemann *et al.*, 2006).

An alternative to N- or C-terminal fusions is an insertion fusion, in which a carrier protein is inserted into a loop in the sequence of a poorly soluble target. To date, this approach has exclusively been used in membrane-protein crystallization and was initially pioneered for the *E. coli* lactose permease, in which cytochrome $b_{562}$, flavodoxin and T4 lysozyme were tested as carrier proteins inserted into one of the loops (Privé *et al.*, 1994; Engel *et al.*, 2002). In this specific case none of these variants actually yielded useful crystals and the structure of lactose permease was eventually solved using crystals obtained using a variant containing the C154G mutation which stabilized a single conformation in complex with a lactose analogue (Abramson *et al.*, 2003). In contrast, a similar insertion fusion with T4 lysozyme replacing the third intracellular loop of the $\beta$2-adrenergic receptor was highly successful and yielded good-quality crystals that allowed structure determination at 2.4 Å resolution (Cherezov *et al.*, 2007; Rosenbaum *et al.*, 2007). This spectacular result attests to the potential of insertion-fusion proteins, but the method is not trivial as the constructs must be carefully evaluated for both structural and functional consequences of the insertion and a number of variants may have to be screened before a suitable one is identified.

### 4.3.6. Noncovalent crystallization chaperones

Noncovalent crystallization chaperones, *i.e.* engineered binding proteins that produce noncovalent complexes with target macromolecules, constitute an exciting alternative to fusion carrier proteins. Complexes with such chaperones often exhibit enhanced solubility and/or crystallizability in comparison to the isolated targets. The Fab and Fv fragments of antibodies are most commonly used for this purpose (Kovari *et al.*, 1995; Hunte & Michel, 2002; Prongay *et al.*, 1990; Ostermeier *et al.*, 1995; Jiang *et al.*, 2003; Dutzler *et al.*, 2003; Lee *et al.*, 2005). In its canonical version, the technique requires animal immunization with subsequent purification of hybridoma-derived antibodies and their proteolytic digestion to obtain pure homogeneous Fab fragments (Karpusas *et al.*, 2001; Kovari *et al.*, 1995). Alternatively, the Fab fragment can be directly sequenced and a synthetic gene can be used for *E. coli* expression, although this is not trivial owing to the presence of disulfides and two separate polypeptide chains in an Fab molecule. To overcome this bottleneck, a more efficient method of recombinant production of antibody fragments using mammalian HEK 293T has recently been proposed (Nettleship *et al.*, 2008). Another possibility is the use of so-called nanobodies, *i.e.* single-chain fragments derived from camelid antibodies (Koide, Tereshko *et al.*, 2007; Lam *et al.*, 2009; Korotkov *et al.*, 2009). However, this strategy requires immunization of camels or llamas, which is not technically easy.

Regardless of the specific strategy, the use of hybridoma technology and animal immunization is always time-consuming