stochastic process generated by mostly random contacts (Janin & Rodier, 1995; Janin, 1997; Carugo & Argos, 1997). However, more recent stringent statistical analyses using a larger database strongly suggested that crystal contacts are generated by anisotropic interactions that favour small hydrophobic residues and disfavour large polar side chains with high conformational entropy (Cieślik & Derewenda, 2009). This view is also supported by a large-scale comparison of the amino-acid sequences of crystallizable and noncrystallizable proteins, which established that crystallization propensity is negatively correlated with the prevalence of residues with high side-chain entropy (Price *et al.*, 2009). Finally, molecular-dynamics simulations of the intermolecular interactions of lysozyme in solution show that they are anisotropic and that their magnitude and nature depend on the physical chemistry of the participating interfaces, suggesting that the nucleation phenomenon is initiated in a nonstochastic fashion (Pellicane *et al.*, 2008).

Understanding the physical principles that govern crystallization at the microscopic level provides the singular underpinning to rationally engineer target proteins to enhance their crystallizability either by improving their solution properties or by increasing their propensity to engage in weak but specific interactions that organize the transformation of nascent clusters into nuclei and drive subsequent crystal growth.

4.3.3. Engineering proteins with enhanced solubility

The solubility of a protein is the primary essential prerequisite for its crystallization. It should be noted that the expression 'low solubility' is often used indiscriminately to describe quite different phenomena, including a propensity to aggregate and precipitate upon overexpression owing to misfolding, amyloid formation and finally genuine low *in vitro* solubility, *i.e.* low protein concentration in equilibrium with the solid phase, of otherwise fully folded and stable proteins (Trevino *et al.*, 2008). Here, the strategies and methods that specifically address the latter case are discussed, *i.e.* precipitation at low concentrations of properly folded proteins.

It has been well established that even single-site mutations of surface residues can dramatically affect the solubility of a protein and its crystallizability (McElroy *et al.*, 1992). Consequently, the intuitively obvious approach is to mutate solvent-exposed hydrophobic amino acids to hydrophilic residues. In this way, the low solubility of the catalytic domain of HIV-1 integrase was addressed by screening 29 mutants in which hydrophobic residues were systematically mutated to hydrophilic amino acids; of the variants tested, the single-site mutant F185K showed a dramatically improved solubility and ultimately yielded X-ray-quality crystals (Dyda *et al.*, 1994; Jenkins *et al.*, 1995). In the case of leptin, the product of the *obese* gene, the solubility-enhancing W100E mutation proved to be critical for crystallization of the protein (Zhang *et al.*, 1997). Recently, a screen of several variants of human apolipoprotein D identified a triple mutant (W99H, I118S, L120S) which was much more soluble than the wild-type protein and which was ultimately used to obtain well diffracting crystals (Nasreen *et al.*, 2006; Eichinger *et al.*, 2007).

While engineering enhanced solubility using site-directed mutagenesis is potentially a powerful approach, in the absence of structural information it is a challenge to predict which hydrophobic residues are solvent-exposed and might therefore constitute useful targets for mutagenesis. Moreover, even if structural information is available for a homologue or the target itself, it may not be clear what type of mutation actually works best, forcing the investigator to rely on extensive screening. This uncertainty arises from the fact that hydrophobicity scales for individual amino acids cannot be used directly to evaluate the increase or decrease of protein solubility as a consequence of a specific mutation. Furthermore, there have been few rigorous studies of the effects of specific mutations on protein solubility. A notable example is a study on ribonuclease SA in which the solvent-exposed Thr76 was replaced by 19 other amino acids and the solubility of all of the variants was carefully evaluated (Trevino *et al.*, 2007). Those variants that contained Asp, Arg, Glu and Ser were the most soluble. Unexpectedly, even though a lysine might be expected to confer higher solubility than a serine or alanine, the T76S mutation actually led to a significantly higher solubility than T76K, while the T76A variant was only marginally less soluble than T76K (Trevino *et al.*, 2007). The authors of the study concluded that mutating Asn and Gln to their respective acids may constitute the most robust strategy of enhancing solubility. Interestingly, one of the first examples of rational enhancement of solubility, *i.e.* the study of trimethoprim-resistant type S1 hydrofolate reductase (Dale *et al.*, 1994), used this very strategy: the amide-containing side chains were systematically substituted with carboxylic amino acids and one specific variant, a double mutant N48E, N130D, was found to exhibit markedly increased solubility and ultimately yielded crystals that were suitable for crystallographic analysis.

Somewhat ironically, large charged residues such as glutamate that confer higher solubility on the target protein may at the same time impede crystallization because they increase the total surface side-chain entropy, making the surface recalcitrant to engaging in crystal contact-mediating interactions. Thus, variants engineered for increased solubility may simultaneously show a decreased propensity to crystallize.

Some of the above uncertainties can be overcome with an alternative approach of directed evolution and phenotypic selection methods, in which soluble mutants are directly selected from vast protein libraries (Farinas *et al.*, 2001; Farinas, 2006; Pédelacq *et al.*, 2002; Waldo, 2003; Cabantous *et al.*, 2005). Several different variations of this method have been reported (Waldo, 2003). For example, the target protein may be fused to the N-terminus of a reporter protein such as the green fluorescent protein (GFP; Waldo *et al.*, 1999) or direct detection methods can be employed to identify soluble variants (Peabody & Al-Bitar, 2001). While elegant and potentially very effective, directed evolution has not yet been widely adopted for the generation of crystallizable proteins.

Solubility problems are not always caused by excessively exposed hydrophobic surfaces. In some cases, the root of the problem is aggregation caused by exposed free cysteines. Reduced cysteines can be identified by alkylation with *N*-ethylmaleimide or iodoacetamide under anaerobic conditions and subsequent electrospray mass spectrometry (Niessing *et al.*, 2004). Several examples illustrate how this approach is helpful in generating samples that are suitable for crystallization. In mitogen-activated protein (MAP) kinase p38α, a single-site mutation (C162S) prevented aggregation and yielded a crystallizable variant (Patel *et al.*, 2004). Similarly, a double mutant (C95K, C142S) of foot-and-mouth disease virus 3C protease showed none of the aggregation problems that plagued the wild-type protein and was subsequently crystallized (Birtley & Curry, 2005). It is noteworthy that in this case an alternative strategy involving mutations of the exposed hydrophobic residues Met81, Leu82 and Val140 did not eliminate aggregation (Birtley & Curry, 2005). In a number of cases aggregation problems were traced to

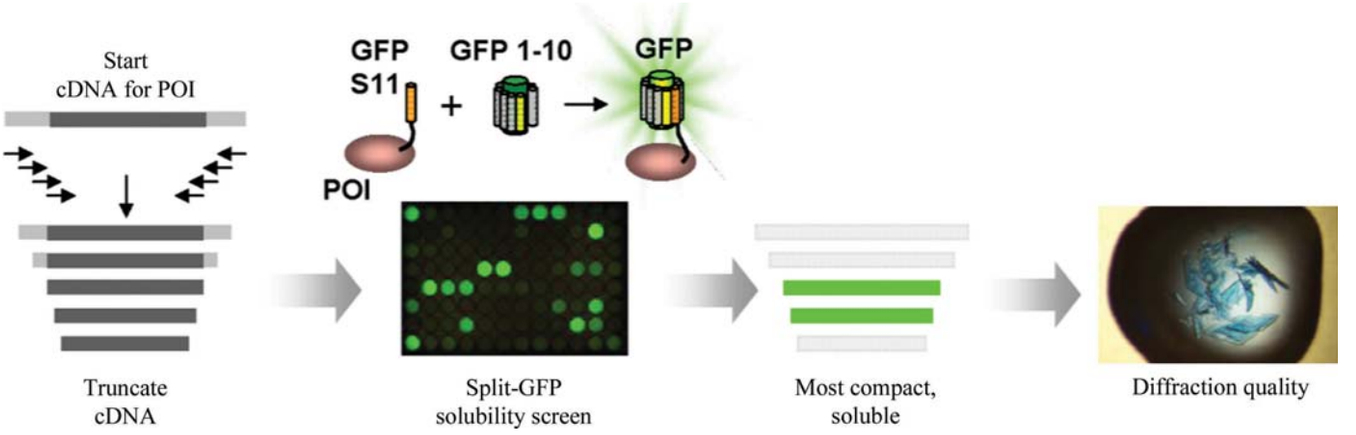

Figure 4.3.4.1
A domain-trapping strategy to engineer soluble variants of a protein of interest (POI) for crystallization using the split-GFP complementation methodology. (Figure courtesy of Dr Geoff Waldo, LANL.)

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 β-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