International Tables for Crystallography (2012). Vol. F, Section 4.3.2, pp. 129–130.

Chapter 4.3. Application of protein engineering to enhance crystallizability and improve crystal properties

Z. S. Derewenda

4.3.1. Introduction

Following the dawn of recombinant technology brought about by the groundbreaking overexpression of synthetic genes coding for insulin and somatostatin in *Escherichia coli* (Goeddel *et al.*, 1979; Itakura *et al.*, 1977) and the subsequent discovery of the polymerase chain reaction (PCR; Mullis *et al.*, 1986; Saiki *et al.*, 1985, 1988), macromolecular crystallography was freed of its long-standing dependence on purified native protein samples for crystallization. Heterologous expression made it possible to generate samples of proteins and complexes that are found in only small or trace amounts in living cells and to engineer large and unstable proteins so that isolated domains or modified forms can be made available for crystallization. At the same time, the effort required for protein purification was dramatically reduced by the use of fusion proteins and affinity tags (Brewer *et al.*, 1991; Sassenfeld, 1990; Malhotra, 2009). As a consequence, the overwhelming majority of samples used today for crystallization are recombinantly derived proteins. However, even though material for crystallization is more easily available, the preparation of single, well diffracting crystals of the target macromolecule is still a time-consuming challenge.

Historically, two complementary approaches to protein crystallization were developed in parallel. Firstly, natural variations in the amino-acid sequences of homologues from different species were exploited to identify a target with suitable crystallization properties during the purification procedure (Kendrew *et al.*, 1954; Campbell *et al.*, 1972). The second approach was to extensively screen a specific target protein against a range of diverse precipitating agents, buffers and additives until the right conditions for crystallization were identified (Carter & Carter, 1979; Jancarik & Kim, 1991). These strategies remain the pillars of contemporary macromolecular crystallization. However, as the palette of molecular biology techniques expanded to include site-directed mutagenesis, ligation-independent cloning and other tools, it became possible to modify proteins with relative ease with the specific purpose of enhancing their propensity to crystallize or improving the diffraction quality of the resulting crystals. The early proof-of-principle of these capabilities was the crystallization of an engineered variant of human H-ferritin in which a single-site mutation, K86Q, was introduced to duplicate a crystal contact mediated by Cd^{2+} ions in the crystal structure of the homologous rat L-ferritin (Lawson *et al.*, 1991).

In this review, current progress in the methodologies of protein engineering used to enhance the crystallizability of targets that are recalcitrant to crystallization in their wild-type form is discussed. This burgeoning field is very broad and includes both general strategies that apply to a range of targets and many diverse approaches that only apply to specific proteins or protein families. Thus, owing to space limitations, the focus is on those techniques that have either been demonstrated to be of general utility or are at a point in their development to clearly have the potential to become widely used in the future. Understandably, only representative examples are provided.

4.3.2. Microscopic aspects of protein crystallization

Proteins are inherently dynamic entities, a property that greatly hinders their crystallizability. Not surprisingly, it has been estimated that even for the stable and relatively small single-domain prokaryotic proteins fewer than one in four will yield X-ray-quality crystals when using a routine screening process (Canaves *et al.*, 2004; Price *et al.*, 2009). In order to rationally modify proteins to enhance their crystallizability, it is first necessary to understand the physical properties that make most proteins resistant to crystallization.

Protein crystals are nucleated *ab initio* at supersaturation levels in the 200–1000% range (McPherson, 1982). Nucleation is believed to proceed *via* a two-step process: clusters of solute molecules form first and upon reaching critical size reorganize into three-dimensionally ordered nuclei (Georgalis *et al.*, 1997; Vekilov, 2004; Erdemir *et al.*, 2009). Subsequent transfer of protein molecules from solution onto the growing crystal surface is driven by relatively small negative changes in Gibbs free energy (ΔG°), from approximately -10 to -100 kJ mol^{-1}, at ambient temperature (Vekilov, 2003). Interestingly, enthalpy changes are generally negligible during crystallization (Yau *et al.*, 2000; Petsev *et al.*, 2001; Gliko *et al.*, 2005), so that entropic phenomena dominate (Vekilov *et al.*, 2002; Vekilov, 2003; Derewenda & Vekilov, 2006). The microscopic effects underlying the entropy changes, both favourable and unfavourable, involve the protein itself as well as the solvent. Protein packing, which results in an ordered three-dimensional lattice and loss of translational and rotational degrees of freedom, is unfavourable and produces an energy barrier in the 30–100 kJ mol^{-1} range at room temperature (Finkelstein & Janin, 1989; Tidor & Karplus, 1994). Similarly, incorporation into the growing crystal and ordering of any intrinsically unstructured elements, such as flexible termini or loops and side chains, at the point of crystal contacts further increases the entropic cost. However, the release of ordered solvent molecules from the surfaces involved in crystal contacts, which is estimated to be in the 25–150 kJ mol^{-1} range, may sufficiently compensate for these entropy losses and ultimately provide the driving force for crystal growth (Vekilov *et al.*, 2002; Vekilov, 2003).

Based on these considerations, it is evident that a protein must satisfy certain criteria in order to crystallize. Firstly, it must have a molecular surface that confers adequate solubility under initial conditions to reach the necessary supersaturation level for nucleation. Furthermore, it should have few, if any, intrinsically unstructured fragments such as extended N- or C-termini or long and solvent-exposed loops which may impede crystallization. Finally, the protein should have distinct 'sticky' patches on the surface with a structured layer of solvent molecules, allowing the ordering of nascent nuclei by mediating thermodynamically viable specific crystal contacts.

The notion that protein crystallization involves specific and anisotropic intermolecular interactions, as opposed to random contacts, is relatively new. Early analyses of intermolecular contacts in protein crystals concluded that crystallization is a

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