Chapter 4.4. High-throughput X-ray crystallography

K. H. Choi

4.4.1. Introduction

Structural-genomics projects have contributed to major developments in automation, miniaturization and process integration in X-ray crystallography. In the structural-genomics approach, multiple open reading frames from a genome are separately cloned and expressed without prior knowledge of the structure or function of the encoded proteins. Essential steps in a typical project include: (1) bioinformatics analysis of genome sequences for potential targets, (2) gene amplification by PCR (polymerase chain reaction) and subcloning into an appropriate expression vector, (3) protein expression and purification, (4) protein crystallization, (5) X-ray data collection, and (6) data processing and structure determination. Structural-genomics projects typically operate under a 'lowest hanging fruit' philosophy, pursuing structural targets that prove to be the most amenable to crystallization. However, for individual investigators working on specific problems, it is often necessary to focus on a particular protein. Many of the protocols developed in high-throughput (HT) structural genomics can also be adapted for parallel production of multiple constructs of a single protein target which is difficult to crystallize. This chapter summarizes recent developments in HT X-ray crystallography approaches and their application to parallel production of a single protein (Gräslund, Nordlund *et al.*, 2008; Manjasetty *et al.*, 2008; Sharff & Jhoti, 2003).

4.4.2. Design of multiple constructs: bioinformatics analysis of genome sequences

Often, the two primary bottlenecks in crystal structure determination are obtaining a soluble protein and the initial crystallization. If initial crystallization trials do not lead to successful crystallization, diffraction-quality crystals or structure determination, modification of the protein itself is often more successful in producing useful crystals than exhaustive screening/optimization of protein crystallization conditions. In the HT approach, generation of stable domains having variable N or C termini can be performed in parallel, vastly reducing time and resource requirements. The clones can be screened systematically for soluble, active proteins amenable to crystallization. This approach is particularly suited to problems in which one has no estimate, *a priori*, of which constructs are likely to be most successful at the expression, purification or crystallization steps.

It may be difficult to predict which modifications will influence protein solubility and the protein's ability to crystallize. A common and often useful strategy is to crystallize a similar protein from an alternative biological source. If an alternative source is either unavailable or equally problematic, gene fusion with large affinity tags can be used for increasing protein expression and solubility. Several fusion proteins, including thioredoxin, maltose binding protein, glutathione S-transferase, intein, SUMO (small ubiquitin-like modifier) protein and calmodulin-binding protein, have been used to generate soluble proteins. However, the fused tag may prevent crystallization because of conformational heterogeneity resulting from a flexible linker; hence, the tag must often be removed in an additional purification step. Protein engineering by amino-acid substitutions and chemical modifications (such as methylation of exposed lysines) have been shown to improve crystallization of some proteins (Rayment, 1997; Walter *et al.*, 2006; Wingren *et al.*, 2003). Large, flexible solvent-exposed residues (*e.g.* Lys or Glu) on the surface of proteins are substituted with smaller residues (*e.g.* Ala) to facilitate the formation of intermolecular contacts which stabilize the crystal lattice (the surface-entropy reduction approach; Cooper *et al.*, 2007; Derewenda, 2004). Random mutagenesis following selection of a desired phonotype (*e.g.* folding ability using a GFP (green fluorescent protein) reporter or solubility assays) has been used to produce soluble proteins from insoluble wild-type proteins (the directed-evolution approach; Pédelacq *et al.*, 2002). Since symmetric molecules such as homodimers may crystallize more readily than monomeric protein, new Cys residues can be introduced in a monomeric protein to induce homodimer formation *via* intermolecular disulfide bond formation (the synthetic symmetrization approach; Banatao *et al.*, 2006).

Bioinformatics tools, including multiple sequence alignment and sequence motif searches, as well as prediction of secondary structures, domain boundaries, membrane spanning and disordered regions, can be used to aid the rational design of constructs. General considerations in designing truncation mutations are to avoid truncations in the middle of predicted secondary structural elements, to avoid hydrophobic residues at the termini, and to eliminate membrane-spanning regions in the construct design. In the case of multi-domain proteins, truncation of the protein to smaller functional domains can be effective. Exact locations of the beginning and ending of a domain are still difficult to predict even with domain search programs, and thus biochemical approaches such as limited proteolysis can aid the determination of the domain boundaries. The optimal step size for truncation mutations cannot be predicted, but protein constructs varying by approximately five residues in length often show significant differences in solubility and crystallization behaviour (Choi *et al.*, 2004; Gräslund, Sagemark *et al.*, 2008).

4.4.3. Cloning

Development of robotics that utilize a 96-well format has changed traditional sequential cloning of individual proteins such that parallel HT cloning and protein preparation are now possible. The conventional steps in cloning are PCR amplification, restriction enzyme digestion, ligation, transformation, selection of transformers and protein expression tests. Many procedures, including amplification of target genes, cloning and screening for expression, can be achieved in parallel (96 samples at a time) using either a single liquid-handling robot or multichannel pipettors. In 96-well parallel cloning, reaction steps for individual wells cannot be optimized, and thus care should be taken to synchronize all reactions and to minimize the number of steps to increase efficiency. For this reason, certain cloning strategies are more popular for HT cloning. Recombinant protein

references