4. CRYSTALLIZATION

the complex and those of the protein alone. One of these mutations, together with an additional mutation resulting from a polymerase chain reaction (PCR) artefact, yielded crystals that diffracted to 3.5 Å. Additional variation of the length and composition of the RNA hairpin led to a new crystal form of this double mutant in the presence of a 21-base RNA that diffracted to 1.7 Å. A further mutation, Ser29 → Cys, was made to allow mercury binding (see Section 4.3.8), also resulting in crystals that diffracted to 1.7 Å. The authors commented that 'If any principle emerges from this study, it is that the key to success is not in concentrating on exhausting any one approach, but in the diversity of approaches used.'

The relevance of this comment is illustrated by the attempts of Scott *et al.* (1998) to obtain diffraction-quality crystals of the I-A^d class II major histocompatibility complex (MHC) protein. This complex exists *in vivo* as a heterodimer, but expression in recombinant form did not lead to satisfactory dimer formation. A leucine zipper peptide was therefore added to each chain to enhance dimerization. Attempts to crystallize this heterodimer after removal of the leucine zippers and in the presence of bound peptides led to poorly diffracting crystals. To enhance the affinity of an ovalbumin peptide for the MHC dimer, the peptide was then attached through a six-residue linker to the N-terminus of the chain, tethering it in the vicinity of the binding site. This construct, in conjunction with removal of the leucine zippers from the heterodimer, resulted in crystals that diffracted to 2.6 Å.

4.3.6. Avoiding protein heterogeneity

Protein heterogeneity can arise from many sources, including proteolysis, oxidation and post-translational modifications, and can have a severe effect on crystal quality or can prevent crystallization altogether. Limited proteolysis has frequently been used to modify proteins for crystallization, in order to avoid heterogeneity from proteolysis occurring during expression and to remove relatively unstructured regions that might hinder crystallization. Some examples are given below.

Windsor *et al.* (1996) crystallized a complex of interferon γ with the extracellular domain of the interferon γ cell surface receptor. To obtain satisfactory crystals, it was necessary to re-engineer the receptor with an eight-amino-acid residue deletion at the N-terminus to avoid the observed heterogeneity owing to proteolysis, since 2–10% of the purified protein was cleaved during expression.

Crucial to the structure determination of the complex of transducin-α bound to GTPγS (Noel *et al.*, 1993) was the systematic examination of proteolysis of the intact protein (Mazzoni *et al.*, 1991). This work revealed a cluster of protease-sensitive sites near residues Lys17–Lys25. Homogeneous material consisting of residues 26–350 of activated rod transducin, $G_{t\alpha}$, was obtained by proteolysis of the full-length protein with endo-proteinase LysC; the truncated protein was subsequently used to solve the structure.

Hickman *et al.* (1997) identified a site near the C-terminus of HIV-1 integrase that was susceptible to proteolytic cleavage during protein expression, resulting in severe protein heterogeneity in which up to 30% of the purified protein was cleaved. The proteolysis site was identified by mass spectrometry analysis, and several point mutations on either side of this site were made and evaluated for their effect on proteolysis. Substitution of either Gly or Lys for Arg284 eliminated the protease sensitivity, yielding homogeneous material.

Some proteins have surface cysteines that are susceptible to oxidation and can be adventitiously cross-linked *via* a disulfide bridge that does not exist in the native protein. If there are relatively few cysteines, this problem may be circumvented by mutating the individual cysteines to determine which ones are responsible.

Conversely, cysteines can be introduced into proteins to enhance the binding of interacting molecules (see also Section 4.3.8). An elegant example of the latter case is provided by the recent structure of HIV-1 reverse transcriptase (Huang *et al.*, 1998), which was mutated to introduce a cysteine in a position near the known binding side of the double-stranded DNA substrate. Using an oligonucleotide with a modified base that contained a free thiol group, cross-links were specifically introduced between the protein and the DNA; this covalently linked complex was used to obtain crystals that contained the incoming nucleoside triphosphate, a crystallographic problem that had defied other solutions.

Post-translationally modified proteins, such as glycoproteins, present some of the most difficult problems in X-ray crystallography, since the carbohydrate side chains are usually flexible and often heterogeneous. In some cases, enzymes can be used to trim the carbohydrate and produce a protein suitable for crystallization. Alternatively, the protein sequence can be altered so that unwanted glycosylation does not occur. A combination of approaches was used by Kwong *et al.* (1998) to determine the structure of the HIV-1 envelope glycoprotein, gp120, a protein which is extensively modified *in vivo*. The N- and C-termini were truncated, 90% of the carbohydrate was removed by deglycosylation and two large, flexible loops of the protein were replaced by tripeptides. The resulting simplified version of the glycoprotein retained its ability to bind the CD4 receptor, and crystals were ultimately obtained of a ternary complex of the envelope glycoprotein, a two-domain fragment of CD4 and an antibody Fab.

Occasionally, an mRNA sequence will fortuitously result in a false initiation of translation, resulting in a truncated form co-purifying with the intended protein. In attempting to crystallize a trimethoprim-resistant form of dihydrofolate reductase, Dale *et al.* (1994) observed that a fragment of the protein was being expressed through false initiation of translation, beginning at Ala43. They also found most of the protein in inclusion bodies and recovery was poor. They noticed that there was a putative Shine–Dalgarno sequence ten nucleotides up from the AUG codon of Met42, which could result in the expression of a smaller protein. They replaced the middle base of the Shine–Dalgarno sequence, GGGAA, with GGCAA and removed unusual codons from the first 18 amino acids. These two changes resulted in a 20-fold increase in expression level, together with removal of the contaminating fragment. Similar heterogeneity problems owing to translation initiation at an internal Shine–Dalgarno sequence upstream of Met50 were observed during expression of full-length recombinant HIV-1 integrase and were also resolved by altering the DNA to eliminate the Shine–Dalgarno sequence without changing the sequence of encoded amino acids (Hizi & Hughes, 1988).

4.3.7. Engineering crystal contacts to enhance crystallization in a particular crystal form

It is often the case that the structure of some related form of a protein is known, but the protein of interest crystallizes in a different space group. There have been attempts to use this knowledge to obtain crystals in a form that could be readily analysed. However, it may not be necessary to resort to molecular engineering approaches, since molecular replacement methods can often be successfully applied to determine the protein structure.

In one of the first applications of protein engineering to obtain crystals, Lawson *et al.* (1991) reported the crystal structure of ferritin H. Ferritin has two types of chains, H and L; the structure of rat L ferritin was known. Despite high sequence identity to L ferritin, human recombinant H ferritin did not crystallize satisfactorily. To obtain the structure of a human H ferritin homopolymer, the sequence in the subunit interface was modified to give crystals that were isomorphous with the rat L ferritin. The