centrifugation and SDS–PAGE analysis were used to determine which mutant proteins were sufficiently soluble to appear in the supernatant. The initial application of this method to 30 mutants resulted in one, Phe185 → Lys, which was soluble and which was subsequently crystallized and its structure determined (Dyda *et al.*, 1994). The protein formed a dimer, and the mutated residue was observed at the periphery of the dimer interface where the introduced lysine formed a hydrogen bond with a backbone atom of the second subunit, an interaction not possible for the unmutated protein. The position of the mutation was remote from the active site, and the physiological relevance of the observed dimer interaction was later confirmed by studies on an avian retroviral integrase (Bujacz *et al.*, 1995).

In further mutational work, it was observed that the HIV-1 integrase core-domain mutant suffered from an inability to bind to Mg^{2+} in the crystal, despite the evidence that Mg^{2+} or Mn^{2+} is needed for activity. The original crystallization took place using cacodylate as a buffer and also had dithiothreitol present in the crystallization medium. Under these conditions, cacodylate can react with –SH groups, and there were two cysteines in the structure that were clearly bonded to arsenic atoms. To avoid this problem, attempts were made to crystallize in the absence of cacodylate. These were successful only when a second mutation, designed to improve solubility, was introduced, Trp131 → Glu (Jenkins *et al.*, 1995; Goldgur *et al.*, 1998). The use of this mutant led to crystals that had the desired property of binding to Mg^{2+} and, in addition, revealed the conformation of a flexible loop that had not been previously defined.

4.3.3. Use of fusion proteins

Fusion proteins have been frequently used in a variety of applications (reviewed by Nilsson *et al.*, 1992), such as preventing proteolysis, changing solubility and increasing stability. They have also been used – although less frequently – for crystallization. The disadvantage in the context of crystallography is that the length and flexibility of the linker chain often introduce mobility of one protein domain relative to the other, which can impede, rather than enhance, crystallization.

Donahue *et al.* (1994) were able to determine the three-dimensional structure of the 14 residues representing the platelet integrin recognition segment of the fibrinogen γ chain by constructing a fusion protein with lysozyme, which was then crystallized from ammonium sulfate. Kuge *et al.* (1997) successfully obtained crystals of a fusion protein consisting of glutathione S-transferase (GST) and the DNA-binding domain (residues 16–115) of the DNA replication-related element-binding factor, DREF, under crystallization conditions similar to those used for GST alone.

In many cases, a fusion protein is made to aid in the isolation and purification of the target protein, and the intervening linker is engineered to contain a proteolytically susceptible sequence. However, subsequent cleavage to separate the two proteins can introduce the possibility of accidental proteolysis elsewhere in the protein. This was observed with a fusion protein between thioredoxin and VanH, a D-lactate dehydrogenase, where attempts to remove the carrier resulted in non-specific proteolysis and VanH inactivation (Stoll *et al.*, 1998). Fortunately, cleavage was unnecessary, and conditions were identified under which the authors were able to crystallize the intact fusion protein.

A novel approach to crystallizing membrane proteins is provided by the fusion protein in which cytochrome b_{562} was inserted into a central cytoplasmic loop of the lactose permease from *Escherichia coli* (Privé *et al.*, 1994). Although crystals have not yet been reported, the cytochrome attachment provides increased solubility together with the ability to use the red colour to assay the progress of crystallization trials.

4.3.4. Mutations to accelerate crystallization

A common problem encountered in crystallization is that certain crystals appear late and grow slowly. Sometimes, the slow appearance of crystals is the result of proteolytic processing, but often the reasons are not apparent. There are several examples where protein engineering has resulted in an increase in the rate of crystallization.

Heinz & Matthews (1994) explored the crystallization of T4 phage lysozyme using a strategy based on their understanding of the structure of the enzyme and its crystallization properties. The crystallization of the wild-type protein required the presence of β-mercaptoethanol (BME), an additive which could not be replaced with dithiothreitol. It had also been observed that the oxidized form of BME, hydroxyethyl disulfide, was trapped in the dimer interface between two lysozyme molecules (Bell *et al.*, 1991). It was hypothesized that dimer formation might be the rate-limiting step in crystallization, so dimerization was enhanced by cross-linking two monomers by disulfide-bridge formation. Applying rules developed for constructing S–S bridges, they selected Asn68 → Cys and Ala93 → Cys. In the presence of oxidized BME, the rate of crystallization of these mutant proteins was substantially increased, with crystals reaching full size in two days, in contrast to two weeks for the unmutated protein. Furthermore, they were able to crystallize a previously uncrystallizable mutant. Unexpectedly, however, the dimer formed in this way was lacking in activity, despite the selection of mutation sites on the opposite side of the molecule to the active site.

Mittl *et al.* (1994) wanted to improve the resolution of their crystals of glutathione reductase. From the 3 Å map, they could see a hole in the crystal packing where two molecules within 6 Å of each other just missed forming a crystal contact; they filled this hole by mutating Ala90 → Tyr and Ala86 → His. This designed double mutant did not improve the resolution, but did increase the rate of crystallization 40-fold, *i.e.*, initial crystals were observed within 1.5 h *versus* 60 h for the wild-type enzyme.

4.3.5. Mutations to improve diffraction quality

Another commonly encountered situation is that crystals can be obtained, but they diffract poorly. There are many examples where investigators have applied protein engineering in an effort to overcome this problem.

Proteolytic trimming is one possible approach to improving diffraction quality. For example, Zhang *et al.* (1997) attempted to crystallize a homodimer of the C2 domain of adenylyl cyclase. The initial crystals diffracted poorly (to 3.8 Å), so the effects of limited proteolysis with chymotrypsin, trypsin, GluC and LysC were investigated. A stable cleavage product was observed with GluC, approximately 4 kDa smaller than the full-length protein, but in order to avoid minor products formed during GluC proteolysis, the cleavage site was re-engineered as a thrombin site. Since there was already an atypical thrombin site seven residues from this site, proteolysis resulted in a smaller protein than expected; nevertheless, this modified protein crystallized readily and diffracted to 2.2 Å.

The importance of applying a variety of strategies to improve crystal quality is exemplified by the work of Oubridge *et al.* (1995), in which initial attempts to crystallize wild-type U1A complexed with RNA hairpins resulted in cubic crystals diffracting to 7–8 Å. By mutating surface residues, changing the N-terminal sequence to reduce heterogeneity and varying the sequence of the RNA hairpin, a new crystal form which diffracted to 1.7 Å was ultimately crystallized. However, in order to achieve this result, many variants were constructed and examined. For the protein, mutations were introduced which it was believed (incorrectly) would affect the crystal packing, and which were selected based on the observed similarity of space group and cell dimensions between crystals of

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

references