3. TECHNIQUES OF MOLECULAR BIOLOGY

In such cases, the problem lies not with the purification of the protein, but in finding a proper way to refold it.

Various general procedures for refolding proteins from inclusion bodies have been described (Fischer *et al.*, 1993; Werner *et al.*, 1994; Hofmann *et al.*, 1995; Guise *et al.*, 1996; De Bernardez Clark, 1998), and the literature is filled with examples of specific protocols. The insoluble inclusion bodies are usually solubilized in a powerful chaotropic agent like guanidine hydrochloride or urea. In general, detergents are not recommended. The denaturant is sequentially removed by dilution, dialysis or filtration. Both rapid dilution and slow removal of the denaturant have been used successfully. In most refolding protocols, relatively dilute solutions of the protein are used to avoid protein–protein interactions, and, if necessary, glutathione or some other thiol reagent is included in the buffer to accelerate correct pairing of disulfides. After a refolding procedure, the properly folded soluble protein must be separated from the fraction that did not fold appropriately. Improperly refolded proteins are relatively insoluble and can usually be removed by centrifugation. It is sometimes profitable to try to refold the recovered insoluble material a second time.

Once soluble protein has been obtained, conventional purification procedures may be employed. It should be noted that recovery of soluble protein is not necessarily an indication that the protein exists in a native state. Quantitative assays of protein activity should be used to characterize the protein, if such assays exist. Alternatively, the behaviour of the refolded protein should be critically assessed during subsequent purification steps; an improperly folded protein will be prone to aggregation, will generally give broad and/or trailing peaks during column chromatography and will migrate faster than expected during size-exclusion chromatography. Some proteins are more amenable to refolding than others. As has already been pointed out, if a protein has a complex array of disulfide bonds, it is usually more difficult to refold than a protein without disulfide bonds. Greater success in refolding is generally obtained with proteins composed of single domains than with multidomain proteins.

3.1.6. Characterization of the purified product

3.1.6.1. Assessment of sample homogeneity

The ultimate test of the usefulness of a purified protein for crystallization is determined by the actual crystallization trials. However, before such trials begin, the properties and purity of the recombinant protein should be carefully checked. There is some disagreement about the degree of purity required for crystallization. In the earliest days of protein purification, crystallization was used as a technique for the purification of proteins, and it is clear that absolute purity is not a requirement for the preparation of useful protein crystals. However, most practitioners of the art of crystallization prefer to use highly purified proteins for crystallization trials. There are several reasons for this. It is easier to achieve the high concentrations of protein (greater than 10 mg ml^{-1}) usually needed for crystallization if the protein is pure, and the behaviour of highly purified proteins is more reproducible. A homogeneous preparation of protein will precipitate at a specific point rather than over a broad range of solution conditions. Furthermore, degradation during storage and/or crystallization is minimized if all of the proteases have been removed.

Although there are a number of ways to check the purity of a protein, the most convenient, and widely used, involve electro-phoresis. Most experimentalists use SDS–PAGE and/or isoelectric focusing to determine the purity and homogeneity of the protein. SDS–PAGE may be slightly more convenient for the detection of unrelated proteins; isoelectric focusing is probably more useful in detecting subspecies of the recombinant protein of interest. We will consider the nature and origins of such subspecies below. Once the protein(s) is fractionated, either on an isoelectric focusing gel or on SDS–PAGE, it is detected by staining, either with silver or with Coomassie brilliant blue. Neither reagent reacts uniformly with all proteins; depending on the proteins involved, either method can overestimate or underestimate the level of a contaminant relative to the desired recombinant protein. Silver staining is the more sensitive method. However, if there is sufficient material for a serious attempt at crystallography, the sensitivity of Coomassie staining is usually more than sufficient for analytical purposes. It is often useful to fractionate a protein preparation by both isoelectric focusing and SDS–PAGE, and stain gels with silver and Coomassie brilliant blue. This increases the chance of discovery of an important contaminant and/or heterogeneity in the protein preparation.

If the preparation is relatively free of unrelated proteins, but there is concern about the presence of multiple species of the desired recombinant protein, there are several techniques that can be applied. Mass spectrometry is capable of detecting small differences in molecular weights, and for proteins up to several hundred amino acids in length it is usually able to detect differences in mass equivalent to a single amino acid. This can be useful in detecting heterogeneity in post-translational modifications, if such are present, and in detecting heterogeneity at both the amino and carboxyl termini. Amino-terminal sequencing can also be used to detect N-terminal heterogeneity, but has some limitations that are discussed below.

In *E. coli*, the methionine used to initiate translation is modified with a formyl group. The formyl group, and sometimes the amino-terminal methionine, is removed from proteins expressed in *E. coli*. Removal of the N-terminal amino acid is dependent on the identity of the second amino acid; methionines preceding small amino acids (Ala, Ser, Gly, Pro, Thr, Val) are generally removed (Waller, 1963; Tsunasawa *et al.*, 1985). However, when large amounts of a recombinant protein are made in *E. coli*, the formylase and aminopeptidase that mediate N-terminal processing are sometimes overwhelmed, and removal of the N-terminal groups is often incomplete. It is common to observe heterogeneity at the amino termini of even the most highly purified recombinant proteins. Amino-terminal sequencing can be used to detect this type of amino-terminal heterogeneity; however, the portion of the protein that retains the formyl group will not be detected by this method, and a misleading impression of the quantity and quality of the protein preparation can be obtained.

Heterogeneity at both the amino and carboxyl termini can be introduced by proteolysis, especially when the ends of the protein are extended and unstructured. This problem is frequently encountered when domains (rather than intact proteins) are expressed and can often be avoided if the boundaries of compact structural domains are precisely defined. In addition to introducing heterogeneity due to partial proteolysis, dangling ends can contribute to aggregation.

In terms of crystallization, the ability to produce a highly concentrated monodisperse protein preparation is probably more important than absolute purity. There are a number of techniques that can be used to determine whether or not the protein is aggregating. Analytical ultracentrifugation is the classical method, and size-exclusion chromatography has been widely

used, particularly by biochemists. However, many crystallographers routinely use dynamic light scattering to check concentrated protein preparations for aggregation (Ferré-D'Amaré & Burley, 1994). The method is relatively simple, very sensitive to small amounts of aggregation and has the additional advantage that it does not consume the sample. After testing, the sample (which is often precious) can still be used for crystallization trials.

If sample heterogeneity is detected, one is faced with the issue of whether it will adversely affect crystallization, and if so, how to remove it. Unfortunately, there do not seem to be general rules. Heterogeneity at the termini of proteins is a common occurrence. In many crystal structures, the termini are disordered and heterogeneity at these unstructured ends would not be expected to be a significant problem. Indeed, in a number of instances, N-terminal sequence analysis of proteins obtained by dissolving crystals has indicated substantial heterogeneity. However, in other cases, properly defined domain boundaries are thought to have been a critical factor in obtaining useful crystals. Domain boundaries can be determined by a combination of limited proteolysis, followed by identification of the fragments using mass spectrometry (Cohen *et al.*, 1995; Hubbard, 1998). Subsequent re-engineering of expression constructs with modified termini is a relatively easy task. Similar engineering can also be used to alter internal sequences, such as removal of sites of post-translational modification or introduction of mutations that improve solubility (Chapter 4.3).

3.1.6.2. Protein storage

Even when the efforts of those engaged in crystallization and those engaged in producing the desired recombinant protein are well coordinated, it is not usually appropriate or desirable to use all the available protein for crystallization at the same time. This means that some of the material must be stored for later use. Even under the best of circumstances, protein solutions are subject to a number of unwanted events that can include, but are not limited to, oxidation, racemization, deamination, denaturation, proteolysis and aggregation. As a general rule, it is better to store proteins as highly purified concentrated solutions. This reduces problems of proteolysis (since the proteases have been removed), and, in general, proteins are better behaved if they are relatively concentrated (greater than 1 mg ml^{-1}). This is not an absolute rule, however; if there are problems with aggregation, these can sometimes be minimized by storage of proteins in dilute solutions, followed by concentration of the samples immediately prior to crystallization. If the protein contains oxidizable sulfurs (free cysteines are a particular problem), reducing agents can be added (and should be refreshed as necessary), and the solutions held in a non-reducing (N_2) atmosphere. In some cases, it is easier to mutate surface cysteines to produce a more stable protein (see Chapter 4.3).

In general, proteins behave best under conditions of pH and ionic strength similar to those they would experience in the normal host. Usually this means a pH near, or slightly above, neutral and intermediate ionic strength. These conditions are often not the ideal conditions for crystallization, and dialysis or other forms of buffer exchange may be required before beginning crystallization trials. In general, protein solutions are stored either at 4 °C in a cold room or refrigerator, or at 0 °C on ice. It is essential that the protein be stored in a manner that will not allow microbial growth, usually achieved by sterilization of the protein solution by filtration through 0.2 micron filters and/or addition of

antimicrobial agents, such as NaN_3. For long-term storage (periods longer than a few weeks), protein solutions are often precipitated in ammonium sulfate or frozen at either −20 or −70 °C. Repeated freezing and thawing is not recommended; if a protein sample is to be frozen, it should be divided into aliquots small enough so that each will be thawed only once. Whenever a protein sample is frozen and thawed, some loss of quality and/or activity can be expected. Freezing samples of intermediate concentration (1–3 mg ml^{-1}) usually works better than freezing either extremely dilute or concentrated samples. Cryoprotective agents can be added to protein samples destined to be frozen; however, it should be remembered that the same reagents that are helpful when freezing a protein sample may be distinctly unhelpful when that sample is thawed and used for crystallography. Most biochemists willingly add glycerol to their protein samples before freezing; crystallographers are not usually happy to find that their protein sample is dissolved in 50% glycerol. Both pH and ionic strength can affect a protein's tolerance to freezing and thawing. In many cases, buffer exchange and concentration procedures need to be performed to convert stored protein solutions to ones suitable for crystallization.

As is so often true in science, decisions about whether to freeze a particular protein sample and, if it is to be frozen, exactly how the freezing should be done, depend on experience. If the protein in question is an enzyme, it is often useful to set up a series of trials in which small aliquots of the protein are stored under a variety of conditions. If the aliquots are tested on a fairly regular basis, how stable the protein is in solution can usually be determined, as well as how well it will tolerate a cycle of freezing and thawing, with or without an added cryoprotectant. If enzyme assays are not available, other methods of characterization, such as gel electrophoresis, mass spectrometry and light scattering, can be used to check for degradation, oxidation of cysteines and aggregation. Armed with this information, and with a plan for how the protein will be used for crystallization, it is usually a fairly simple matter to decide whether or not to freeze a particular sample, and, if the sample is to be frozen, how best to do it. It is a good idea to make such tests early in a major crystallization effort. This will avoid the awkward dilemma that occurs when a large amount of a highly purified protein is available, and the knowledge of how best to store it is not.

3.1.7. Reprise

We have reached a point where it is possible to use recombinant DNA techniques to produce most proteins in quantities sufficient for crystallography. Both high-level expression systems and methods for making defined modifications of recombinant proteins vastly simplify the process of purification. This has played a direct and critical role in the ability of crystallographers to produce an astonishing array of new and exciting protein structures. We are beginning to come to grips with the next level of the problem: using the ability to modify the sequence of proteins to improve their crystallization properties. This is a difficult problem; however, there are already notable, if hard won, successes. It would appear that the marriage of genetic engineering and crystallography – clearly a case in which opposites attract – has been a happy union. This is entirely for the good. Collaborations between specialists in these disciplines have led to the solution of problems too difficult for any individual armed only with the skills of one or the other partner. It is important that genetic engineering be fully integrated into future crystallographic efforts, either directly within the crystallography