International Tables for Crystallography (2012). Vol. F, Section 3.1.5, pp. 85–88.

mammalian cells (metallothionein, glucocorticoid responsive *etc.*) tend to be leaky in the absence of the inducer. If cell lines were chosen in which the desired protein was not synthesized in the absence of the inducer, the level of the recombinant protein that could be made in the presence of the inducer was usually, but not always, low.

There has been progress in the development of more efficient and reliable inducible promoters for cultured mammalian cells. These systems are complex and require cell lines that express regulatory proteins not normally found in cultured mammalian cells. In this sense they are the logical counterparts of the T7 RNA polymerase/*lac* expression systems for *E. coli* already discussed in this chapter. The best developed of the engineered systems designed to permit the inducible expression of genes in mammalian cells are (1) the tetracycline system, (2) the F506/rapamycin system, (3) the RU486 system and (4) the ecdysone system (Saez *et al.*, 1997; Rossi & Blau, 1998).

Although these four inducible systems differ in important ways, there are common themes. Firstly, in all cases, the small molecule used as the inducer is not normally a regulator of gene expression in mammalian cells. This means that application of the inducer to cells should not substantially perturb the normal pattern of gene expression and, by implication, the health of the cells. Secondly, the DNA target sequences used to activate the expression of the recombinant gene/protein are not sequences known to be associated with the expression of normal cellular genes. This should also help prevent the activation of normal cellular genes when these systems are used.

In all of these systems, the specific regulation of an introduced gene requires a special regulatory protein that interacts with the appropriate small-molecule inducer and recognizes the requisite DNA target sequence that is linked to the gene of interest. These regulatory proteins, which were derived, at least in part, from regulatory proteins from nonmammalian hosts, must be present in the cell line for induction/regulation to occur. This means that either the researcher must choose from a relatively limited set of cells that already express the desired regulatory factor or face the problem of introducing (and carefully monitoring the proper expression and function of) both the regulatory factor and the desired recombinant protein. Considerable effort has been put into the development of each of these systems and significant progress has been made. At the moment, the tetracycline inducible system is probably the most fully developed; however, this is a fast moving area of research, and it is not now certain which of these systems will ultimately prove to be the most useful for the high-level expression of recombinant proteins in cultured mammalian cells.

Suffice it to say, however, that despite all the efforts of a large group of talented researchers, the systems available for use in cultured mammalian cells are much less well defined and much more difficult to use than the corresponding *E. coli* and yeast expression systems, and anyone who is not well versed in the problems associated with using expression systems designed for cultured mammalian cells should be most cautious about using them for the large-scale production of recombinant protein.

Despite these problems, mammalian (and, less frequently, insect cell) expression systems have been used to prepare proteins for crystallography. For example, in the recent determination of the X-ray structure of a complex between a portion of CD4, a modified version of HIV-1 gp120 and the Fab fragment of a monoclonal antibody, each of the proteins was made in cultured cells, but three different types of cultured cells were used. The two-domain segment of CD4 was made in Chinese hamster ovary cells. The monoclonal antibody used to prepare the Fab was made in an immortalized human B cell clone, and the core of gp120 in *Drosophila* Schneider 2 cells under the control of a metallothionein promoter (Kwong *et al.*, 1998).

Tissue culture cells are much more difficult to grow than either yeast or *E. coli.* As has already been discussed in Section 3.1.4.3, there is the issue of using calf (or fetal calf) serum. A relatively small number of mammalian cell lines have been developed that will grow on defined media without serum; this is an advantage, but the media are still relatively costly. Mammalian cell lines expressing recombinant proteins must be maintained for long periods under carefully controlled conditions, both to ensure that the expression of the recombinant protein is maintained and to avoid contamination of the cultures with bacteria, yeast or moulds. Because the cells grow relatively slowly (doubling times are commonly 24–48 hours), it is usually not a simple task to produce 10–20 g (wet weight) of cells – something that can be done overnight with *E. coli.* If a useful cell line is obtained, it should be carefully stored in multiple aliquots. Cultured cells are routinely stored (in the presence of cryoprotectants) in liquid nitrogen. Short-term storage at −70 °C is an acceptable practice; however, long-term storage will be much more successful if lower temperatures are used.

3.1.5. Protein purification

3.1.5.1. Conventional protein purification

Those of us old enough to remember the task of purifying proteins from their natural sources, using conventional (as opposed to affinity) chromatography, where a 5000-fold purification was not unusual and the purifications routinely began with kilogram quantities (wet weight) of *E. coli* paste or calves' liver, are most grateful to those who developed efficient systems to express recombinant proteins. In most cases, it is possible to develop expression systems that limit the required purification to, at most, 20- to 50-fold, which vastly simplifies the purification procedure and concomitantly reduces the amount of starting material required to produce the 5–10 mg of pure protein needed to begin crystallization trials. This does not mean, however, that the process of purifying recombinant proteins is trivial. Fortunately, advances in chromatography media and instrumentation have improved both the speed and ease of protein purification. A wide variety of chromatography media (and prepacked columns) are commercially available, along with technical bulletins that provide detailed recommended protocols for their use. Purification systems (such as Pharmacia's FPLC and ÄKTA systems, PerSeptive Biosystems' BioCAD workstations and BioRad's BioLogic systems) include instruments for sample application, pumps for solvent delivery, columns, sample detection, fraction collection and information storage and output into a single integrated system, but such systems are relatively expensive. Several types of high capacity, high flow rate chromatography media and columns (for example, Pharmacia's HiTrap products and PerSeptive Biosystems' POROS Perfusion Chromatography products) have been developed and are marketed for use with these systems. However, the use of these media is not restricted to the integrated systems; they can be used effectively in conventional chromatography without the need for expensive instrumentation.

In designing a purification protocol, it is critically important that careful thought be given to the design of the protocol and to a proper ordering of the purification steps. In most cases, indi-

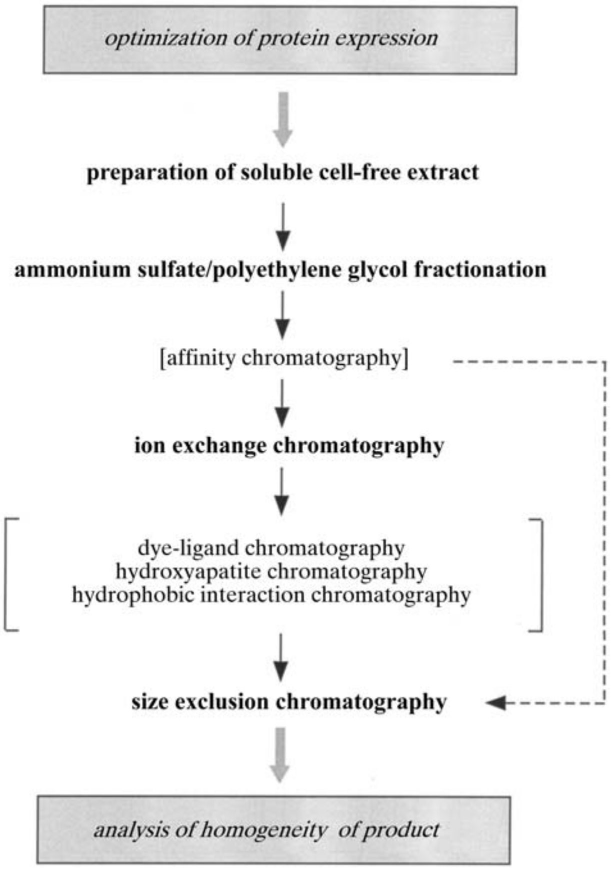
optimization of protein expression
preparation of soluble cell-free extract
ammonium sulfate/polyethylene glycol fractionation
[affinity chromatography]
ion exchange chromatography
dye-ligand chromatography
hydroxyapatite chromatography
hydrophobic interaction chromatography
size exclusion chromatography
analysis of homogeneity of product

Figure 3.1.5.1
Protein purification strategy. Purification of proteins expressed at reasonably high levels typically requires only a limited number of chromatographic steps. Additional chromatography columns (indicated in brackets) can be included as necessary. Affinity chromatography can allow efficient purification of fusion proteins or proteins with well defined ligand-binding domains.

vidual purification steps are worked out on a relatively small scale, and an overall purification scheme is developed based on an ordering of these independently developed steps. However, the experimentalist, in planning a purification scheme, should keep the amount of protein needed for the project firmly in mind. In general, crystallography takes a good deal more purified protein than conventional biochemical analyses. Scaling up a purification scheme is an art; however, it should be clear that purification steps that can be conveniently done in batch mode (precipitation steps) should be the earliest steps in a large-scale purification, chromatographic steps that involve the absorption and desorption of the protein from columns (ion-exchange, hydroxyapatite, hydrophobic interaction, dye-ligand and affinity chromatography) should be done as intermediate steps, and size exclusion, which requires the largest column volumes relative to the amount of protein to be purified, should generally be used only as the last step of purification. If reasonably good levels of expression can be achieved, most recombinant proteins can be purified using a relatively simple combination of the previously mentioned procedures (Fig. 3.1.5.1), requiring a limited number of column chromatography steps (generally two or three).

All protein purification steps are based on the fact that the biochemical properties of proteins differ: proteins are different sizes, have different surface charges and different hydrophobicity. With the exception of a small number of cases involving proteins that have unusual solubility characteristics, batch precipitation steps usually do not provide substantial increases in purity. However, precipitation is often used as the first step in a purification procedure, in part because it can be used to separate protein from nucleic acids. Nucleic acids are highly charged polyanions; the presence of nucleic acid in a protein extract can dramatically decrease the efficiency of column chromatography, for example by saturation of anion-exchange resins. If the desired protein binds to nucleic acids and the nucleic acids are not removed, ion-exchange chromatography can be compromised by the interactions of the protein and the nucleic acid and by the interactions of the nucleic acid and the column. The most commonly used precipitation reagents are ammonium sulfate and polyethylene glycols. With little effort, the defined range of these reagents needed to precipitate the protein of interest can be determined. However, if the precipitation range is broad, it may be only marginally less efficient simply to precipitate the majority of proteins by addition of ammonium sulfate to 85% saturation or 30% polyethylene glycol 6000. Precipitation can be a useful method for concentrating proteins at various steps during purification and for storing proteins that are unstable upon freezing or upon storage in solution.

Column chromatography steps in which the protein is absorbed onto the resin under one set of conditions and then eluted from the column under a different set of conditions can produce significant purification. Anion-exchange chromatography is usually a good starting point. Most proteins have acidic pIs, and conditions can often be found that allow binding of the protein to anion-exchange matrices. Elution of the protein in an optimized gradient often yields greater than tenfold purification. If conditions cannot be found under which the protein binds to an anion-exchange resin, a reverse strategy can be advantageous. Conditions can be adjusted to promote the binding of most proteins, yielding a flow-through fraction enriched for the protein of interest. Fewer proteins interact with cation-exchange resins; if the desired protein binds, this can be a powerful step. Use of an anion exchanger does not necessarily preclude use of a cation-exchange column; under appropriately chosen sets of conditions (most notably adjustment of pH), a single protein can bind to both resins. Hydroxyapatite resins provide a variation of ion-exchange chromatography that can be extremely powerful for some proteins. While hydroxyapatite columns (traditionally just a modified form of crystalline calcium phosphate) have the reputation of slow flow rates, alternative matrices exhibiting improved flow properties have made hydroxyapatite chromatography significantly less tedious. Hydrophobic interaction chromatography can also provide significant purification and has the advantage that the protein is loaded onto the resin in a high ionic strength buffer, making it a good step following ammonium sulfate precipitation. Proteins can behave very differently with different hydrophobic matrices, and an exploration of a variety of different resins is often a worthwhile exercise. Several tester kits containing an assortment of resins are commercially available. Dye-ligand chromatography can also be explored using an assortment of test columns. Several of the dyes, most notably Cibacron Blue F3GA, have structures that resemble nucleotides and have been useful in purifying kinases, polymerases and other nucleotide-binding proteins. However, many proteins have significant affinity for various dyes, independent of nucleotide-binding activity, and the usefulness of dye-ligand chromatography for any specific protein needs to be determined empirically.

Size-exclusion chromatography, which does not involve absorption of the protein onto the matrix, rarely provides as much purification as the chromatography steps described above.

However, this can be a good step to include at the end of a purification scheme. Isolation of a well defined peak in the included volume separates intact, properly folded protein from any damaged/aggregated species that may have been generated during the purification procedure. Furthermore, size-exclusion chromatography can provide a useful indication of whether the protein is a well defined, folded, compact, monodisperse population, or whether it is oligomerizing, aggregating or exists in an unfolded or extended form. Although size-exclusion chromatography does not provide a definitive analysis of such behaviour, migration of the protein consistent with its expected molecular weight is generally a good sign; elution of a relatively small protein in the void volume suggests a need for further analysis. Size-exclusion-chromatography media are available for the fractionation of proteins in many different size ranges. Substantial improvement in purification can be achieved by choosing a size range that is optimal for the protein of interest. However, the ability of size-exclusion columns to separate proteins of different molecular weights is dependent on the amount of protein loaded on to the column. Better purification is obtained when relatively small volumes of protein (generally 1–2% of the column bed volume) are loaded on size-exclusion columns. If really large amounts of protein are needed for a crystallography project, it can be difficult (and expensive) to set up size-exclusion columns large enough to fractionate the desired amount of protein.

3.1.5.2. Affinity purification

The most powerful purification steps are those that most clearly differentiate the desired protein from the other proteins present. Many proteins bind specifically to substrates, products and/or other proteins. In some cases, it is possible to use specific ligands to design columns to which the desired protein will bind selectively. For example, it may be possible to chemically link the substrate or product of a particular enzyme to an inert support. If the modification to the small molecule needed to link it to the support is chosen so that it does not interfere with the binding of the enzyme, the modified resin can be used to purify the protein by affinity chromatography. If, as expected, the desired protein binds selectively, it can usually be eluted by washing the column with the same substrate used to prepare the column. This is a powerful procedure and can produce greater than 100-fold purification in a single step. Although this is a fairly well developed field, and there is sufficient experience to show that the process is often fruitful, it must be said that the development of an efficient and effective affinity column and an attendant purification procedure can be long, difficult and, depending on the ligand and/or activated resin, sometimes expensive. In addition, the preparation of the column usually involves some moderately sophisticated chemistry; if such a step is contemplated, it is helpful to have the requisite chemical sophistication.

Immuno-affinity chromatography is a classic affinity method that uses affinity media created by coupling antibodies (either monoclonal or polyclonal) specific for the protein of interest to an activated resin. Theoretically, if good antibodies are available in sufficient quantity, this should be a powerful and widely applicable method. However, immuno-affinity chromatography has two severe limitations. In most cases, the interaction between the antibody and antigen is so tight that harsh conditions are necessary to elute the bound protein, potentially resulting in denaturation of the protein. Additionally, scaling up the procedure for isolation of 5–10 mg of protein is usually not feasible because of the large quantities of antibody required for column preparation.

Because the process of affinity chromatography is so powerful, and the development of a specific affinity column is difficult, considerable effort has been expended on the development of general procedures for affinity chromatography. As discussed previously, it is possible to modify the recombinant protein so that it contains a sequence element that can be used for affinity chromatography. Numerous systems are being marketed that pair vectors for creation of fusion proteins with appropriate resins for affinity purification. Examples of these fusion element–affinity resin pairs include His_6–Ni^{2+}-nitrilotriacetic acid, biotinylation-based epitopes–avidin, calmodulin-binding peptide–calmodulin, cellulose or chitin-binding domains–cellulose or chitin, glutathione S-transferase–glutathione, maltose-binding domain–amylose, protein A domains–IgG, ribonuclease A S-peptide–S-protein, streptavidin-binding peptides–streptavidin and thioredoxin–phenylarsine oxide.

Several considerations are important in choosing a strategy for expression and purification of a fusion protein. Some of these issues have already been discussed (see Section 3.1.3.3). The most fundamental, and unfortunately least predictable, is what construct will produce large amounts of the recombinant protein. The presence of fusion proteins and/or purification tags perturbs the recombinant protein to a greater or lesser degree. Perturbation can in some cases be beneficial, with the fusion protein aiding *in vivo* folding or *in vitro* refolding. There is also the issue of whether or not to remove the tag or fusion protein. Removal of the tag usually involves engineering a site for a specific protease, digestion with that protease and subsequent purification to isolate the final cleaved product. Additional issues should also be addressed. Most of the well developed systems allow for the elution of the fusion protein from the affinity resin under relatively mild conditions that should not harm most proteins. However, the method of elution should be considered with respect to the specific requirements of the protein of interest. Since the costs of using the different systems on a large scale varies significantly, it is wise to calculate the expense associated with scaling up, allowing for the cost and lifetime of the affinity resin, the cost of the reagent used for elution and the cost of the protease if the tag is to be removed. Finally, the nature of the fusion element–affinity resin interaction should be considered. Some of these systems, such as the His_6 tag, can be used for purification under denaturing conditions, which is a considerable advantage if the desired recombinant protein is found in inclusion bodies.

3.1.5.3. Purifying and refolding denatured proteins

As we have already discussed, expressing high levels of recombinant prokaryotic or eukaryotic proteins in *E. coli* can lead to the production of improperly folded material that aggregates to form insoluble inclusion bodies (Marston, 1986; Krueger *et al.*, 1989; Mitraki & King, 1989; Hockney, 1994). Inclusion bodies can usually be recovered relatively easily, following lysis of cells by low-speed centrifugation (5 min at 12 000 *g*); inclusion bodies are larger than most macromolecular structures found in *E. coli* and denser than *E. coli* membranes. Care should be taken to achieve complete lysis, since an intact bacterial cell that remains after lysis will co-sediment with the inclusion bodies. In most (but not all) cases, the inclusion bodies contain the desired recombinant protein in relatively pure form.

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 electrophoresis. 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

references