International Tables for Crystallography (2006). Vol. F, Chapter 3.1, pp. 65–80.

3. TECHNIQUES OF MOLECULAR BIOLOGY

3.1. Preparing recombinant proteins for X-ray crystallography

BY S. H. HUGHES AND A. M. STOCK

3.1.1. Introduction

Preparing protein crystals appropriate for X-ray diffraction usually requires a considerable amount of highly purified protein. When crystallographic methods were first developed, the practitioners of the art were compelled to study proteins that could be easily obtained in large quantities in relatively pure form; the first proteins whose structures were solved by crystallographic methods were myoglobin and haemoglobin. Unfortunately, some of the most interesting proteins are normally present in relatively small amounts, which, while it did not prevent crystallographers from dreaming about their structures, prevented any serious attempts at crystallization. Recombinant DNA techniques changed the rules: it is now possible to instruct a variety of cells and organisms to make large amounts of almost any protein chosen by the investigator. Not only can specific proteins be expressed in large quantities, recombinant proteins can be modified in ways that make the task of the crystallographer simpler and can, in some cases, dramatically improve the quality of the resulting crystals. It is not our intention in writing this chapter to provide either a methods manual for those interested in expressing a particular protein or a complete compendium of the available literature. The literature is vast and complex, and, as we will discuss, the problems associated with expressing a particular protein are often idiosyncratic, making it difficult to provide a simple, comprehensive, methodological guide. What we intend is to discuss issues (and problems) relevant to choosing methods appropriate for preparing recombinant proteins for X-ray crystallography. In this way, we hope to help readers understand both the extant problems and the available solutions, so that, armed with a general understanding of the issues, they can more easily confront a variety of specific projects.

Fortunately, there are a large number of additional resources available to those who are interested in expressing and purifying recombinant proteins, but lack the expertise. These include numerous methods books (*e.g.* on molecular biology: Sambrook *et al.*, 1989; Ausubel *et al.*, 1995; on protein purification: Abelson & Simon, 1990; Scopes, 1994; Bollag *et al.*, 1996), useful reviews of the literature (cited throughout), formal courses (such as those offered by Cold Spring Harbor Laboratory), meetings (*i.e.* IBC's International Conference on Expression Technologies, Washington DC, 1997) and a specialized journal (*Protein Expression and Purification*). The pace of methodological development is rapid, and company catalogues, publications and web pages can provide extensive, useful, up-to-date information. In many cases, a convenient source of information is a nearby researcher whose own research depends on expressing and purifying recombinant proteins. Those who are serious about preparing recombinant proteins for crystallography, but have little or no experience, are strongly urged to avail themselves of these resources. In many cases the help of a knowledgeable colleague is the most valuable resource. In general, the literature provides a much better guide to what will work than what will fail; quite often, in designing a good strategy to produce a recombinant protein that is suitable for crystallography, it is more important to understand the potential pitfalls. Discussion with an experienced colleague is usually the best way to avoid the most obvious errors.

Section 3.1.2 gives an overview of the problem, Section 3.1.3 discusses engineering an expression construct, Section 3.1.4 discusses expression systems, Section 3.1.5 discusses protein purification and Section 3.1.6 discusses the characterization of the purified product.

3.1.2. Overview

The idea that underlies the problem of expressing large amounts of a recombinant protein is straightforward: prepare a DNA segment that, when introduced into an appropriate host, will cause the abundant expression of the relevant protein. However, as the saying goes, 'The devil is in the details.' Not only is it necessary to design the appropriate DNA segment, but also to introduce it into an appropriate host such that the host retains and faithfully replicates the DNA. The DNA segment must contain all of the elements necessary for high-level RNA expression; moreover, the RNA, when expressed, must be recognized by the translational machinery of the host. The recombinant protein, once expressed, needs to be properly folded either by the host or, if not properly folded in the host, by the experimentalist. If the protein is subject to post-translational modifications (cleavage, glycosylation, phosphorylation *etc.*) and the experimentalist wishes to retain these modifications, the appropriate signals must be present and the chosen host must also be capable of recognizing the signals. Once the recombinant protein is expressed, assuming it is reasonably stable in the chosen host, the protein must be purified; as we will discuss, recombinant proteins can be modified to simplify purification. Once purified, the quality of the protein preparation must be evaluated to ensure it is both relatively homogeneous and monodisperse.

While this chapter will be limited to discussions of the basic strategies for creating an expression vector, expressing the protein and purifying and characterizing the product, molecular biological methods can be used in other ways that are relevant to crystallography. In some cases, a protein in its natural form is not suitable for crystallization. Crystallographers have long used proteolytic digestion and/or glycolytic digestion to produce proteins suitable for crystallization from ones that are not. Such techniques have been used to good effect on recombinant proteins; however, the ability to modify the segment encoding the protein makes it possible to alter the protein in a variety of ways beyond simple enzymatic digestions. Specific examples of such applications are described in Chapter 4.3.

Unfortunately, no single strategy for producing proteins for crystallization appears to be universally successful. Any particular protocol has the potential for displaying undesirable behaviour at any step during the process of expression, purification or crystallization. It is important to distinguish major and minor problems. If the problems are serious, it is often better to try an alternative strategy than to struggle with an inappropriate system. Because it is usually difficult to predict what will work and what will not, often the most expedient route to successful expression of a protein for crystallization is the simultaneous pursuit of several expression strategies with multiple protein expression constructs.

3.1.3. Engineering an expression construct

3.1.3.1. *Choosing an expression system*

The first step in developing an expression strategy is the choice of an appropriate expression system, and this decision is critical. As we will discuss briefly below, the rules and/or sequences necessary to express RNA and proteins in *E. coli*, yeast and insect cells (baculoviruses) differ to a greater or lesser extent from those used in higher eukaryotes, and there are considerable differences in the post-translational modifications of proteins in these different systems or organisms. Quite often the protein chosen for investigation comes from a higher eukaryote or from a virus that replicates in higher eukaryotes. The experimentalist prefers to obtain large amounts of the protein (>5–10 mg) to set up crystallization trials. In theory, one simple solution is to use a closely related host to express the protein of interest. While it is possible to produce large amounts of proteins in cultured animal cells (and in some cases in transgenic animals), the difficulties and expense of these approaches usually prevent their use for most projects that require large amounts of highly purified recombinant protein.

In general, prokaryotic (*E. coli*) expression systems are the easiest to use in terms of the preparation of the expression construct, the growth of the recombinant organism and the purification of the resulting protein. Additionally, they allow for relatively easy incorporation of selenomethionine into the recombinant protein (Hendrickson *et al.*, 1990), which is an important consideration for crystallographers intending to use multiple anomalous dispersion (MAD) phasing techniques. However, the differences between *E. coli* and higher eukaryotes means that, in some cases, the recombinant protein must be modified to permit successful expression in *E. coli*, and the available *E. coli* expression systems cannot produce many of the post-translational modifications made in higher eukaryotes. As one moves along the evolutionary path from *E. coli* to yeast, to baculovirus and finally to cultured mammalian cells, the problems associated with producing the protein in its native state are simpler, while the problems associated with expressing large amounts of material quickly, simply and cheaply in an easy-to-purify form become more difficult. In Section 3.1.4, we will consider each of these expression systems in turn; first we will briefly discuss, in a general way, how the relevant genes or cDNA strands are obtained and how an expression system is designed.

3.1.3.2. *Creating an expression construct*

The first step in preparing an expression system is obtaining the gene of interest. This is not nearly as daunting a task as it once was; an intense effort is now being directed at genome sequencing and the preparation of cDNA clones from a number of prokaryotic and eukaryotic organisms. There are also a large number of cloned viral genes and genomes. This means that, in most cases, an appropriate gene or cDNA can be obtained without the need to prepare a clone *de novo*. If the nucleic sequence is available, but the corresponding cloned DNA is not, it is usually a simple matter to prepare the desired DNA clone using the polymerase chain reaction (PCR). If the relevant genomic or cDNA clone is not available and there is no obvious way to obtain it, there are established techniques for obtaining the desired clone; however, these methods are often tedious and labour intensive. They also constitute a substantial field in their own right and, as such, lie beyond the scope of this chapter (for an overview, see Sambrook *et al.*, 1989).

In higher eukaryotes, most mRNA strands are spliced. With minor exceptions, mRNA strands are not spliced in *E. coli*. In yeast, the splicing rules do not match those used in higher eukaryotes. If one expects to express a protein from a higher eukaryote in one of these systems, a cDNA must be prepared or obtained. Because some introns are large, cDNA clones are often used as the basis of expression constructs in baculovirus systems, as well as in cultured insect and mammalian cells.

In all subsequent discussions, we will assume that the experimentalist possesses both a cDNA that encodes the protein that will be expressed and an accurate sequence. If a genomic clone is available, it can be converted to cDNA form by PCR methods or by using a retroviral vector. Retroviral vectors, by nature of their life cycle, will take a gene through an RNA intermediate, thus removing unwanted introns (Shimotohno & Temin, 1982; Sorge & Hughes, 1982). If a good sequence is not available, one should be prepared. In general, expression constructs are based, more or less exclusively, on the coding region of the cDNA. The flanking $5'$ and $3'$ untranslated regions are not usually helpful, and if these untranslated regions are included in an expression construct, they can, in some cases, interfere with transcription, translation or both. With some knowledge of the organization of the protein, it is sometimes helpful to express portions of a complex protein for crystallization. This will be discussed in more detail later in this chapter and in Chapter 4.3.

Optimizing the expression of the protein is extremely important. The amount of effort required to get an expression system to produce twice as much protein is usually less than that required to grow twice as much of the host; moreover, the effort to purify a recombinant protein is inversely related to its abundance, relative to the proteins of the host. There are specific rules for expressing a recombinant protein in the different host–vector systems; these will be discussed in the context of using various hosts (*E. coli*, yeast, baculoviruses and cultured insect and mammalian cells).

Although the precise nature of the modifications necessary to obtain efficient expression of a protein is host dependent, the tools used to produce the modified cDNA and link it to an appropriate expression plasmid or other vector are reasonably standard. In recent years, PCR has become the method of choice for manipulation of DNA; it is a relatively easy and rapid method for altering DNA segments in a variety of useful ways (Innis *et al.*, 1990; McPherson *et al.*, 1995). For most construction projects, the ends of the cDNA are modified, using PCR with appropriate oligonucleotide primers that have been designed to introduce useful restriction sites and/or elements essential for efficient transcription and/or translation. Since it can often be advantageous to try the expression of a given protein construct in a number of different vectors, it is useful to incorporate carefully chosen restriction sites that will enable the fragment to be inserted simultaneously, or transferred seamlessly, into different plasmids or other vectors (Fig. 3.1.3.1). PCR can also be used to create mutations in the interior of the cDNA. For some projects where large-scale mutagenesis is planned, other mutagenic techniques are particularly helpful (for example, site-directed cassette mutagenesis using *Bsp* MI or a related enzyme; Boyer & Hughes, 1996). Ordinarily, however, these alternative strategies are only useful if a relatively large number of mutants are needed for the project.

If PCR is used either to modify the ends of a DNA segment or to introduce specific mutations within a segment, it should be remembered that the PCR can introduce unwanted mutations. PCR conditions should be chosen to minimize the risk of introducing unwanted mutations (start with a relatively large amount of template DNA, limit the number of amplification cycles, use relatively stringent conditions for hybridization of the primers, choose solution conditions that reduce the number of errors made in copying the DNA and use enzymes with good fidelity, such as *Pfu* or others that have proofreading capabilities). It is also important to sequence all of the DNA pieces generated by PCR after they have been cloned.

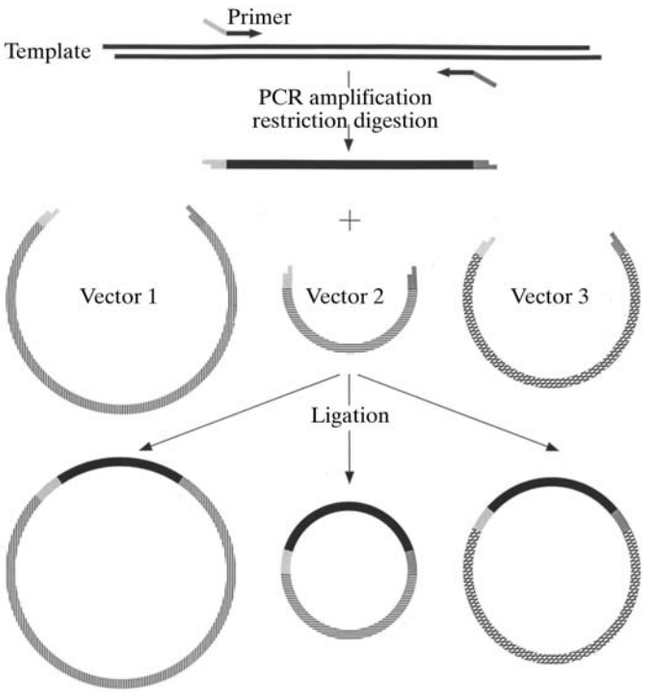

Fig. 3.1.3.1. Creating an expression construct. PCR can be used to amplify the coding region of interest, providing that a suitable template is available. PCR primers should be designed to contain one or more restriction sites that can be conveniently used to subclone the fragment into the desired expression vector. It is often possible to choose vectors and primers such that a single PCR product can be ligated to multiple vectors. The ability to test several expression systems simultaneously is advantageous, since it is impossible to predict which vector/host system will give the most successful expression of a specific protein.

3.1.3.3. *Addition of tags or domains*

In some cases it is useful to add a small peptide tag or a larger protein to either the amino or carboxyl terminus of the protein of interest (Nilsson *et al.*, 1992; LaVallie & McCoy, 1995). As will be discussed in more detail below, such fused elements can be used for affinity chromatography and can greatly simplify the purification of the recombinant protein. In addition to aiding purification, some protein domains used as tags, such as the maltose-binding protein, thioridazine, and protein A, can also act as molecular chaperones to aid in the proper folding of the recombinant protein (LaVallie *et al.*, 1993; Samuelsson *et al.*, 1994; Wilkinson *et al.*, 1995; Richarme & Caldas, 1997; Sachdev & Chirgwin, 1998). Tags range in size from several amino acids to tens of kilodaltons. Numerous tags [including hexahistidine (His_6), biotinylation peptides and streptavidin-binding peptides (Strep-tag), calmodulin-binding peptide (CBP), cellulose-binding domain (CBD), chitin-binding domain (CBD), glutathione S-transferase (GST), maltose-binding protein (MBP), protein A domains, ribonuclease A S-peptide (S-tag) and thioridazine (Trx)] have already been engineered into expression vectors that are commercially available. Additional systems are constantly being introduced. While these systems provide some advantages, there are also drawbacks, including expense, which can be considerable when both affinity purification and specific proteolytic removal of the tag are performed on a large scale.

If a sequence tag or a fusion protein is added to the protein of interest, one problem is solved but another is created, *i.e.* whether or not to try to remove the fused element. During the past year, there have been numerous reports of crystallization of proteins containing His-tags, but there are also unpublished anecdotes about cases where removal of the tag was necessary to obtain crystals. In a small number of cases, additional protein domains present in fusion proteins appear to have aided crystallization (see Chapter 4.3). Experiences with tags appear to be protein specific. There are a number of relevant issues, including the protein, the tag and the length and composition of the linker that joins the two. If the tag is to be removed, it is usually necessary to use a protease. To avoid unwanted cleavage of the desired protein, 'specific' proteases are usually used. When the expression system is designed, the tag or fused protein is separated from the desired protein by the recognition site for the protease. While this procedure sounds simple and straightforward, and has, in some cases, worked exactly as outlined here, there are a number of potential pitfalls. Proteases do not always behave exactly as advertised, and there can be unwanted cleavages in the desired product. Since protease cleavage efficiency can be quite sensitive to structure, it may be more difficult to cleave the fusion joint than might be expected. Unless cleavage is performed with an immobilized protease, additional purification is necessary to separate the protease from the desired protein product. A variation of the classic tag-removal procedure is provided by a system in which a fusion domain is linked to the protein of interest by a protein self-cleaving element called an intein (Chong *et al.*, 1996, 1997).

3.1.4. Expression systems

3.1.4.1. *E. coli*

If the desired protein does not have extensive post-translational modifications, it is usually appropriate to begin with an *E. coli* host–vector system (for an extensive review of expression in *E. coli*, see Makrides, 1996). Both plasmid-based and viral-based (M13, λ *etc.*) expression systems are available for *E. coli*. Although viral-based vector systems are quite useful for some purposes (expression cloning of cDNA strands, for example), in general, for expression of relatively large amounts of recombinant protein, they are not as convenient as plasmid-based expression systems. Although there are minor differences in the use of viral expression systems and plasmid-based systems, the rules that govern the design of the modified segment are the same and we will discuss only plasmid-based systems. We will first consider general issues related to design of the plasmid, then continue with a discussion of fermentation conditions, and finally address some of the problems commonly encountered and potential solutions.

Basically, a plasmid is a small circular piece of DNA. To be retained by *E. coli*, it must contain signals that allow it to be successfully replicated by the host. Most of the commonly used *E. coli* expression plasmids are present in the cell in multiple copies. Simply stated, in the selection of *E. coli* containing the plasmid, the plasmids carry selectable markers, which usually confer resistance to an antibiotic, typically ampicillin and/or kanamycin. Ampicillin resistance is conferred by the expression of a β-lactamase that is secreted from cells and breaks down the antibiotic. It has been found that, in typical liquid cultures, most of the ampicillin is degraded by the time cells reach turbidity (approximately 10^7 cells ml^{-1}), and cells not harbouring plasmids can overgrow the culture (Studier & Moffatt, 1986). For this reason, kanamycin resistance is being used as the selectable marker in many recently constructed expression plasmids.

There are literally dozens, if not hundreds, of expression plasmids available for *E. coli*, so a comprehensive discussion of the available plasmids is neither practical nor useful. Fortunately, this broad array of choices means that considerable effort has been expended in developing *E. coli* expression systems that are efficient and easy to use (for a concise review, see Unger, 1997). In most cases, it is possible to find expression and/or fermentation conditions that result in the production of a recombinant protein

that is at least several per cent of the total *E. coli* protein. This should result in the expression of greater than 5 mg of recombinant protein per litre of culture, making the scale of fermentation reasonable and the job of purification relatively simple.

Broadly speaking, *E. coli* expression systems are either constitutive (that is, they always express the encoded protein) or inducible, in which case a specific change in the culture conditions is necessary to induce the expression of the recombinant protein. As is often the case, both systems have advantages and disadvantages, and both systems have been successfully used to generate recombinant protein for X-ray crystallographic experiments. There is no question that constitutive systems are simple and convenient. However, the high-level expression of even a relatively benign recombinant protein usually puts the *E. coli* host at a selective disadvantage. Unless precautions are taken, the growth and repeated passage of *E. coli* carrying a constitutive expression plasmid tend to select for variants that express lower (and sometimes much lower) levels of the desired recombinant protein than were seen when the clone was first prepared. This can be avoided by storing the stock as plasmid DNA and regularly preparing fresh transformants.

If the desired protein is toxic to *E. coli* (as are a substantial number of recombinant proteins), then an inducible system is required. There are several considerations when choosing an inducible system. The method used to induce the expression of the protein should be compatible with the scale required to produce the recombinant protein. For example, inducible systems which use the bacteriophage $\lambda\, p_L$ promoter and the temperature-sensitive repressor CI857ts require a temperature shift from approximately 30 to 42 °C. This can be done quite conveniently in small cultures, but it is much more difficult to achieve a rapid shift of temperature if *E. coli* are grown in batches larger than 10 l. Inducible expression systems based on the *lac* repressor are usually induced with isopropyl-β-D-thiogalactopyranoside (IPTG). The cost of this gratuitous inducer is not an issue when *E. coli* are grown in small cultures; however, in large-scale fermentations, the costs of the inducer are nontrivial. Despite this caveat, expression systems controlled by the *lac* repressor are commonly used. In the original *lac*-based inducible expression systems, the *lac* operator/promoter was located on the plasmid, proximal to the 5′ end of the insert. Because expression plasmids are present in multiple copies in *E. coli*, the *lac* repressor must be overexpressed to a substantial degree for it to be present in sufficient quantity to control a plasmid-borne operon. Even if a highly expressed *lac* repressor gene (*lacI*^{q}, which produces approximately ten times as much repressor than does the wild type) is expressed from a single chromosomal copy (*i.e.*, provided by the host strain rather than by the vector), repression is rarely complete, and some constitutive expression is generally observed, with only moderately increased levels of expression achieved upon induction. Note that the same plasmid constructs will often give different levels of expression of the plasmid-borne gene in different host strains because of the nature of the *lac* repressor gene (wild-type or *lacI*^{q}).

Better control of induction can usually be obtained using a T7 polymerase expression system in a specifically designed vector–host strain pair (Tabor & Richardson, 1985; Studier & Moffatt, 1986; Studier *et al.*, 1990). In such systems, a *lac*-controlled operon that encodes the bacteriophage T7 RNA polymerase is embedded in the genome of the *E. coli* host and is, as a consequence, present in the cell in only one copy. Induction with IPTG leads to the synthesis of the T7 RNA polymerase, which recognizes a promoter sequence that is different from the sequence recognized by *E. coli* RNA polymerase. If the *E. coli* host that carries the T7 RNA polymerase under the control of *lac* also carries a multicopy plasmid, in which the gene of interest is linked to a T7 promoter, the T7 RNA polymerase efficiently produces mRNA from the plasmid; this usually leads to the production of a large amount of the desired recombinant protein. *E. coli* strains that carry a *lac*-inducible T7 RNA polymerase are readily available, as are the corresponding expression plasmids that carry T7 promoters. Some such *E. coli* strains have been specifically engineered so that the expression of the T7 RNA polymerase (and, by extension, the expression of the gene of interest on the plasmid) is tightly regulated (Studier *et al.*, 1990); these strains are particularly useful for expressing recombinant proteins that are toxic to the *E. coli* host. A recent variation on this system uses an *E. coli* strain in which the T7 RNA polymerase gene is under control of the NaCl-induced *proU* promoter (Bhandari & Gowrishankar, 1997). The same plasmids used for other T7 systems can be used with this *E. coli* strain. The osmo-regulated system has the advantages of requiring a much less expensive inducer and, in at least some cases where inclusion-body formation is a problem, of producing higher levels of soluble protein.

Table 3.1.4.1. *Strategies for improving expression in E. coli*

See text for details.

Factor limiting expression	Possible solution
Transcription and/or translation initiation sites	Use vectors with optimized promoter regions
Toxicity	Use inducible expression systems
	Use mutagenesis to eliminate enzymatic activity
	Express a domain of the protein
Rare codons	Use plasmids that co-express corresponding tRNA strands
	Use mutagenesis to optimize codons
Proteolysis	Use protease-deficient host strains
	Use N-end rules to avoid degradation
Inclusion-body formation	Express the protein as a fusion
	Co-express chaperone proteins
	Grow cells at lower temperatures

Adaptations of a cDNA may be necessary for high-level expression in *E. coli* (Table 3.1.4.1). Although the genetic code is universal, the signals necessary for transcription of RNA and translation of proteins are not. Most *E. coli* expression plasmids contain the recognition/regulation sites necessary for controlling RNA transcription; the signals necessary to initiate translation are not always included in expression plasmids. In *E. coli*, the initiation of translation requires not only an appropriate initiation codon (usually AUG, occasionally GUG), but also a special element, the Shine–Dalgarno sequence, just 5′ of the initiator AUG (Gold *et al.*, 1981; Ringquist *et al.*, 1992). In *E. coli*, the first step in translation involves the binding of the 30S ribosomal subunit and the initiator fMet-tRNA to the mRNA. The Shine–Dalgarno sequence is complementary to the 3′ end of the 16S RNA found in the 30S subunit. Eukaryotic mRNAs do not contain Shine–Dalgarno sequences. Some *E. coli* expression plasmids carry a Shine–Dalgarno sequence, others do not. If one is not present in the plasmid, it must be introduced when the cDNA sequence is modified for introduction into the expression plasmid.

The Shine–Dalgarno sequence needs to be positioned in close proximity to the ATG. Ideally, the nucleotide that pairs with C1535 of the 16S RNA should be positioned eight nucleotides upstream of the A of the initiation codon, although a range of 4–14 nucleotides is tolerated (Gold *et al.*, 1981). If the Shine–Dalgarno sequence is supplied by the plasmid, the restriction enzyme recognition site used to join the cDNA to the plasmid must be quite close to the

ATG. Expression systems have been developed in which the restriction site used for creating the expression system includes the initiator ATG. Many expression plasmids are available that have *Nde*I (CATATG) or *Nco*I (CCATGG) recognition sites at the initiator ATG, which makes it possible to move only the coding region of the cDNA into the expression plasmid. Note, however, that if the *Nco*I site is used, retaining the *Nco*I site in the final construction specifies the first base of the second codon. This limits the choices for the second amino acid in the recombinant protein. *Nde*I does not have this limitation.

The termini of proteins influence their susceptibility to degradation by cellular proteases, most notably ClpA. The N-end rule for bacteria is that proteins in which the N-terminal amino acid is Phe, Leu, Trp, Tyr, Arg or Lys are unusually susceptible to proteolysis (Tobias *et al.*, 1991). (The stability of proteins beginning with Pro has not been determined.) These amino acids (as well as others) seem to impart instability in eukaryotic cells as well. Thus in most cases, if one is expressing intact proteins, the N-terminal amino acid of the native sequence will not generally present a problem. Furthermore, under most circumstances, generation of proteins with N-end-rule amino acids at their N-termini are unlikely, since all proteins are initiated with methionine, and while N-terminal methionines are sometimes removed, the specificity of *E. coli* aminopeptidase is such that methionines adjacent to N-end-rule amino acids are not removed with high efficiency (Hirel *et al.*, 1989). Note that methionine removal sometimes occurs, though to a lesser degree, if the penultimate residue is Asn, Asp, Leu or Ile. Thus, Leu residues should probably be avoided at the second position. It is also possible to generate termini containing N-end-rule amino acids by endoproteolytic cleavage; thus it might be advantageous to avoid these amino acids at the beginnings of proteins where unstructured ends are suspected.

Codon usage can influence expression levels. Although it is not something that should routinely be considered in the initial stages of a project, it is a factor that should be kept in mind if no or low levels of expression are observed. Although the genetic code is universal, it is also degenerate: twenty amino acids are specified by 61 codons. Most amino acids are specified by more than one codon; in many cases, some of the codons are used more often (and translated more efficiently) than others. Unfortunately, there are substantial differences in codon preference/usage in prokaryotes and eukaryotes (Zhang *et al.*, 1991). In *E. coli*, codon usage reflects the abundance of the cognate tRNA strands, and poorly expressed genes tend to contain a higher frequency of rare codons (De Boer & Kastelein, 1986). Although a number of theories have been proposed, prediction of the adverse effects of rare codons on the expression of any given sequence is not currently feasible. Factors such as the position of the codons, their clustering or dispersity and the RNA secondary structure may all contribute to levels of expression (Goldman *et al.*, 1995; Kane, 1995). In many instances, *E. coli* do make relatively large amounts of recombinant protein from mRNA strands that contain a number of rare codons (Ernst & Kawashima, 1988; Lee *et al.*, 1992). But in other cases, optimizing codon usage (Hernan *et al.*, 1992; Mohsen & Vockley, 1995) or co-expressing low abundance tRNAs (Brinkmann *et al.*, 1989; Del Tito *et al.*, 1995) has improved the level of expression of recombinant proteins. Since oligonucleotides 50–75 bases long can be synthesized relatively easily, it is possible to create relatively large synthetic cDNA strands or genes that have optimal codon usage. An alternative strategy is to take advantage of plasmids that have been constructed for co-expression of low abundance tRNA strands [$tRNA^{Arg(AGA/AGG)}$ and $tRNA^{Ile(AUA)}$] (Schenk *et al.*, 1995; Kim *et al.*, 1998). Fortunately, these strategies are not usually necessary; before attempting to optimize codon usage, one should first ask whether the natural sequence can be expressed efficiently.

Once plasmid constructs have been created and strains have been assembled, it is important that they be properly stored. Although it is possible to persuade *E. coli* to make large amounts of recombinant protein, it should be remembered that this is an artificial situation chosen by the investigator, not the *E. coli* host. As such, it behoves the experimentalist to pay careful attention to the host; *E. coli* have no *a priori* interest in what the experimentalist wants. All strains and plasmids should be carefully maintained using sterile techniques. Passage of bacterial stocks should be minimized, and master stocks should always be prepared when an expression clone is first isolated or received. The expression system can, in many cases, be successfully stored as a plasmid-containing strain, frozen as a glycerol stock (containing 15% glycerol) at −70 °C. However, it is best to also store the components separately – the expression plasmid as a DNA preparation, ideally as an ethanol precipitate at −20 °C, and the *E. coli* host strain as a frozen glycerol stock at −70 °C. As has already been discussed, changes in the host, as well as in the plasmid, can lead to a decrease in the amount of recombinant protein produced. This problem can be reduced by producing a freshly transformed bacteria stock to start a large-scale fermentation, and this is the reason some people prefer to store plasmid DNA rather than *E. coli* expression strains. It is important to remember that freshly transformed colonies should be restreaked onto selective plates before growth in liquid culture; this avoids the small background of cells not carrying plasmids that are present on the original transformation plates and that can cause problems in liquid cultures. Cells lacking plasmids generally have a faster growth rate and can survive in liquid cultures containing plasmid-carrying cells that express enzymes that degrade the antibiotics. Contamination by cells lacking plasmids can significantly reduce the yield of recombinant proteins.

Even with these precautions, it is important to remember that the *E. coli* host can modify the plasmid. Wild-type *E. coli* contain a number of recombination systems that can act on plasmid DNA. This is a particular problem if a plasmid contains repeated sequences. Recombination between direct repeats is quite efficient in wild-type *E. coli*, but is greatly reduced in *recA* strains. Most of the *E. coli* hosts commonly used for producing recombinant proteins are *recA* deficient, and the use of such strains is strongly recommended.

Fermentation is an especially important part of protein expression. Using an identical strain and plasmid, slight alterations in growth conditions can make a substantial difference in the yield of the desired protein. Ideally, it is preferable to grow large amounts of *E. coli* that contain (relative to the host proteins) large amounts of the desired recombinant protein. In fermentation, the experimentalist controls the media, the temperature of fermentation and, in a large fermenter, the aeration and stirring. In rich media, if the culture is taken to saturation in shake flasks, it is usually possible to produce 4–8 g of *E. coli* (wet weight) per litre; substantially higher cell densities can be obtained in fermenters. The amount of *E. coli* that can be produced in actual practice and, more importantly, the amount of the recombinant protein relative to the *E. coli* host proteins, are sensitive to all of the variables. Unfortunately, there are relatively few hard and fast rules. To make matters worse, when the scale of the fermentation is changed, it is often necessary to develop new fermentation conditions; this is a particular problem when the scale is changed from shake flasks to a fermenter. Developing optimum conditions for the production of a recombinant protein in a fermenter usually requires repeated trials with the fermenter; this is both time consuming and expensive. Fortunately, with many expression systems, sufficient yields can be obtained using shake flasks, and, in cases where a fermenter is required, it is usually possible to get satisfactory (if suboptimal) results without extensive experimentation.

As a general rule, more total *E. coli* and more recombinant protein can be obtained by growing cells in rich media than in minimal media. The cells grow faster in such media, and inductions, in general, are fast and efficient. In some cases, it is necessary to choose between conditions that produce more total *E. coli* and conditions that produce a higher relative yield of the desired recombinant protein. Of these two, the relative yield of the desired protein is the more important. In designing fermentation protocols, it helps to understand how the host organism works. For example, *E. coli* are subject to catabolite repression. Given a choice of two carbon sources, *E. coli* will concentrate on the preferred carbon source to the exclusion of the second. *E. coli* prefer glucose to lactose; if a *lac*-based expression system is used, it is a good idea to avoid using growth media that contain glucose. Good results can often be obtained with media rich in amino acids ($2 \times$ YT or superbroth without glucose). In general, vigorous aeration is helpful. Begin fermentation trials by putting a relatively small amount of broth in a shake flask. For volumes of 1–1.5 l, aeration is much more efficient in wide-bottomed Fernbach flasks, and use of Fernbach flasks improves the yield of cells. In most fermenters, oxygen levels can be monitored, and air and O_2 delivery can be regulated to provide optimal levels of oxygen.

Finding an optimal temperature for maximum production of soluble recombinant protein usually requires experimentation. *E. coli* grow faster at 37 °C than at lower temperatures, and if a high-level expression of soluble protein is obtained under such conditions, there is rarely any advantage in looking further. However, in some cases, the relative yield of a recombinant protein can be substantially increased by growing *E. coli* expression strains at temperatures below 37 °C (discussed further below).

When screening expression constructs for production of recombinant protein, four scenarios are most commonly encountered:

(1) high-level expression of soluble recombinant protein;

(2) high-level expression of the recombinant protein, with a greater or lesser proportion of the protein in inclusion bodies;

(3) no expression or very low levels of expression; and

(4) lysis of cells.

The first result is usually the most welcome. Occasionally however, the expressed protein is smaller than predicted, presumably due to proteolysis. In such cases, production of a stable fragment suggests the presence of a compact, folded domain which might be worth pursuing for crystallography. However, it should be noted that not all soluble proteins are properly folded. Occasionally, misfolded proteins are expressed at high levels in soluble form. Such proteins usually exhibit aberrant behaviour during purification, such as aggregation or precipitation, migration as broad peaks during column chromatography and elution in the void volume during size-exclusion chromatography. In such cases, additional experimentation is required. Inclusion bodies are usually the result of improper protein folding, and cell lysis generally indicates severe toxicity. There are two obvious reasons for a failure to produce measurable amounts of a recombinant protein: either there is a problem at the level of transcription and/or translation, or there is proteolytic degradation of the protein. Some potential solutions to these problems are discussed below.

In some cases, the stability of the recombinant protein is related to its solubility. In general, only well folded proteins are soluble at high concentrations. In all living cells, protein concentrations are high; if a recombinant protein is expressed at a high level, it will be present inside the host cell at a high concentration. Protein folding is an active process in living cells. Molecular chaperones are used both to prevent unwanted interactions with other partially folded proteins and to promote the folding process. In some cases, when a recombinant protein is expressed at high levels, it will not fold properly in *E. coli*, either because it fails to interact properly with *E. coli* chaperones, or because it is made at such high levels that it overwhelms the available chaperones. In such cases the unfolded and/or partially folded protein may aggregate in inclusion bodies (Mitraki & King, 1989), which is both a blessing and a curse. Proteins in inclusion bodies are essentially immune to proteolytic degradation. Additionally, it is usually relatively easy to obtain the inclusion bodies in relatively pure form, making it simple to purify the recombinant protein. Unfortunately, the recombinant protein obtained from the inclusion bodies must be refolded. There are a variety of protocols for refolding proteins (discussed in Section 3.1.5.3), but few simple, universal prescriptions. Even under the most favourable conditions, with proteins that refold easily and (relatively) efficiently, the yield of properly folded material is often low. For some recombinant proteins obtained from inclusion bodies, it is the efficiency of the refolding step that limits the amount of material that can be obtained for crystallization. We will discuss this issue in more detail in Section 3.1.5.3.

The formation of inclusion bodies is the result of aggregation of non-native proteins. Factors that alter the folding pathway and/or affect the concentrations of unfolded or misfolded proteins can have a dramatic influence on the yield of soluble protein. It is not uncommon for recombinant proteins that form inclusion bodies when expressed at high levels (such as in a T7 expression system) to be present at undetectable levels when expressed at slightly lower levels (such as from a constitutive *lac* promoter). Presumably, in both cases the protein is failing to fold rapidly and efficiently. In the former case, the high levels of unfolded intermediates lead to the formation of inclusion bodies; in the latter case, the concentration of the unfolded protein is not sufficient to form inclusion bodies, and the unfolded protein is degraded. In some cases, it is relatively easy to express the protein, but variations in expression systems and/or culture conditions result in quite different yields of soluble and insoluble protein. In all of these situations, it is appropriate to try a number of different expression systems, with the hope that different kinetics of transcription and/or translation may result in concentrations of intermediates in which protein folding is favoured relative to aggregation and/or degradation. In some cases, reducing the temperature of fermentation is helpful (Schein & Noteborn, 1988). In addition to affecting rates of transcription and/or translation, temperature also affects folding. There are numerous examples in the literature where low temperature was essential to the recovery of soluble recombinant protein; however, there does not seem to be a general solution: optimal conditions seem to vary with each protein. In many cases, reducing the temperature of growth from 37 to 30 °C has improved the yield of soluble protein. However, temperatures as low as 17 °C have been reported as optimal for expression of some recombinant proteins (Biswas *et al.*, 1997), and there are anecdotal reports which indicate that protein can be successfully expressed as low as 14 °C. At low temperatures, growth of *E. coli* is quite slow. In most cases, inducible expression systems are used. Cells are grown to mid-logarithmic phase at 37 °C, and are cooled to the desired temperature just prior to induction (Yonemoto *et al.*, 1998). Significantly longer post-induction times are required for high protein yields, and soluble protein expression should be assessed over a 24-hour period to determine optimal times for maximum yields.

The rapid and proper folding of the overexpressed protein appears to be one of the most important factors in achieving high yields of recombinant proteins in *E. coli*. Attempts have been made to improve *in vivo* folding by co-expression of chaperones and other proteins that might aid the folding process (Wall & Plückthun, 1995; Cole, 1996; Georgiou & Valax, 1996). Once again, the usefulness of these strategies appears to be specific for individual recombinant proteins, although some folding components appear to be more broadly useful than others (Yasukawa *et al.*, 1995). A variety of proteins, including GroEL and GroES, DnaK and DnaJ,

chaperones cloned from the host organism for the recombinant protein, thioridazine, protein disulfide isomerases (PDIs) and disulfide-forming protein DsbA, have all been used with varying degrees of success in different systems. It is unlikely that co-expression of chaperones or other proteins will be useful in overcoming folding or stability problems in proteins that are inherently unstable (those made unstable by removal of other domains, those lacking essential post-translational modifications or those failing to form essential disulfide bonds in the reducing environment inside *E. coli*).

Proteolytic degradation is an active process in *E. coli*, and several strategies for minimizing proteolysis of recombinant proteins have been developed (Enfors, 1992; Murby *et al.*, 1996). These strategies include secretion of proteins to the periplasm or external media, engineering of proteins to remove proteolytic cleavage sites, growth at low temperature and other strategies to promote folding, such as use of fusion proteins and co-expression with chaperones. One popular strategy, which unfortunately appears to be more protein-specific than might be expected, involves the use of *E. coli* strains that have genetic defects in the known proteolytic degradation pathways (Gottesman, 1990). If the desired protein is rapidly degraded in *E. coli*, and fermentation at lower temperatures does not solve the problem, *E. coli* deg$^-$ (degradation) mutants can be tried. However, the proteolytic machinery in *E. coli* is quite complicated, and a number of deg$^-$ mutants are available. All of the deg$^-$ mutants are more difficult to work with than wild-type strains, and there is no guarantee that expressing a particular recombinant protein in any of the available deg$^-$ mutants will cause a substantial increase in the yield of the recombinant protein. For these reasons, deg$^-$ mutants are usually tried only as a last resort. In most cases, proteolysis indicates a problem with protein folding, and efforts to improve protein folding are generally more fruitful than efforts to minimize proteolysis.

We briefly touched on the issue of the potential toxicity (to the *E. coli* host) of recombinant proteins when discussing constitutive and inducible vector systems. In general, the greatest difficulties are encountered with membrane proteins and enzymes. For the most part, enzymes are a problem because their enzymatic activities derange the host cell. For example, proteases are notoriously difficult to produce in large amounts. There are several ways to address this problem. First, as has already been discussed, it is important to use a tightly controlled inducible system if the recombinant protein is likely to disturb the metabolism of the *E. coli* host profoundly. If the recombinant protein is not properly folded, and is present primarily in inclusion bodies, the degree of toxicity is less, and often much less, than if the recombinant protein is present primarily in an active, soluble form. Although it is not the preferred procedure, and is not usually necessary, it is also possible to mutate the recombinant protein to reduce (or eliminate) its toxicity. In cases where the desired product is an enzyme, the enzyme can be inactivated by altering the amino acids at the active site.

Additional problems are encountered when trying to produce recombinant proteins that would, in higher eukaryotes, either be bound to or pass through membranes. There are several problems: *E. coli* do not usually grow well if they have large amounts of foreign protein in their membrane; this problem is compounded by the fact that the rules for membrane signals and signal processing are different in *E. coli* and higher eukaryotes. In general, the solution to this issue has been to express, in *E. coli*, only the internal or the external domain of membrane proteins from higher eukaryotes. Not only does this usually solve the problem of the toxicity of the protein in *E. coli*, but domains that are not directly associated with the membrane are usually much more soluble, easier to purify and much better candidates for crystallization. There is an additional issue. In contrast to the cell interior, which is, in general, a reducing environment, the milieu outside the cell is usually an oxidizing environment. Many of the proteins found on the outside of higher eukaryotic cells, or proteins that are exported from higher eukaryotic cells, have disulfide bridges that help stabilize their secondary and/or tertiary structure. Such disulfide bonds do not ordinarily form properly inside *E. coli*, and it can be much more difficult to obtain recombinant proteins that have extensive and complex disulfide bridges in a properly folded form from *E. coli*.

3.1.4.2. *Yeast*

Yeasts are simple eukaryotic cells. Considerable effort has been expended in studying brewers' yeast, *Saccharomyces cerevisiae*, and in developing plasmid systems and expression vectors that can be used in this organism. Recently, methylotrophic yeasts, most notably *Pichia pastoris*, have been developed as alternative systems that offer several advantages over *S. cerevisiae*. Although yeast expression systems are reasonably robust, the expertise required to use these systems effectively is not as widespread as the corresponding expertise for the manipulation of *E. coli* strains. Nor are the tools, media and reagents necessary to grow yeast and select for the presence of expression plasmids as broadly available as those used for *E. coli* systems. However, the increasing commercial availability of complete kits (such as *Pichia* expression systems from Invitrogen) is making yeast systems more accessible.

While yeast systems do offer some advantages relative to *E. coli*, these advantages are, in general, modest. One primary advantage, the ability to produce large amounts of biomass using simple, inexpensive culture media, is probably more important for industrial-scale protein expression than for most laboratory applications, even those involving crystallography, which requires more protein than most simple biochemical experiments. Yeast systems do not, in general, offer solutions to some of the most difficult problems encountered when trying to express recombinant proteins in *E. coli.* Specifically, the problem of mimicking the post-translational modifications found in higher eukaryotes (particularly glycosylation), which has not been solved for *E. coli*, has not been solved in yeast either. None of the available systems recapitulates the post-translational modifications found in higher eukaryotes. Additionally, yeast systems introduce some new problems not seen with *E. coli* expression systems, specifically genetic instability and hyperglycosylation, both of which are more problematic in *S. cerevisiae* than in *Pichia*.

Yeast systems are perhaps most valued for high-level production of secreted proteins. For some naturally secreted proteins, passage through the secretory pathway is necessary for proteolytic maturation, glycosylation and/or disulfide bond formation and is essential for proper folding or function. But secretion is complex, and numerous factors, such as the signal sequence, gene copy number and host strain, can be critical for high-level expression. Secretion can significantly simplify purification, since secreted recombinant proteins can constitute as much as 80% of the protein in the culture medium. However, degradation of secreted proteins can be a major problem. In some instances, proteolysis has been minimized by alteration of the pH of the culture medium, by addition of amino acids and peptides, and by use of protease-deficient strains (Cregg *et al.*, 1993).

The rules for expression of proteins in yeast are not the same as those used either in *E. coli* or in higher eukaryotes. In yeast, as in *E. coli*, cDNA sequences from a higher eukaryote must be tailored for high-level expression, following rules that are fairly well understood. Yeast grows at 25–30 °C and has a slower growth rate than *E. coli* (under typical growth conditions, yeast has a doubling time of approximately 90 min, compared to 30 min for *E. coli*). Transfor-

mation of yeast can be achieved using competent cells, sphaeroplasts or electroporation, but by any technique it is less efficient than the transformation of *E. coli*. For these reasons, most yeast plasmids are designed to replicate both in *E. coli* and yeast; the DNA manipulations are done using an *E. coli* host, and the completed expression plasmid is introduced into yeast as the final step in the process.

Most expression vectors in *S. cerevisiae* are based on the yeast 2μ plasmid (Beggs, 1978; Broach, 1983) that is maintained as an episome, present at approximately 100 copies per cell. Plasmid instability can result in loss of expression during production, and integrating vectors have been developed that provide greater stability, albeit with levels of expression that are, in general, lower than the plasmid systems. Both constitutive and tightly regulated inducible expression systems have been developed using a variety of promoters. The most widely used systems involve galactose-regulated promoters, such as *GAL1*, which are capable of rapid and high-level induction. An extensive review of recombinant gene expression in yeast (Romanos *et al.*, 1992) is highly recommended as a resource for anyone seriously contemplating the expression of recombinant proteins in *S. cerevisiae.*

In terms of high-level expression, the *Pichia* system may ultimately prove to be more useful than *S. cerevisiae* (for reviews see Cregg *et al.*, 1993; Romanos, 1995; Hollenberg & Gellissen, 1997). There is considerable interest in developing the *Pichia* system for the expression of recombinant proteins, especially for industrial applications, and there has been sufficient progress made to support the publication of a useful monograph for specific techniques (Higgins & Cregg, 1998). *Pichia* offer several advantages over *S. cerevisiae*. Intracellular protein expression can be extremely high in *Pichia*, reaching grams per litre of cell culture. Large amounts of secreted proteins can be produced using media that are almost protein-free, although the expression levels are not quite as high as for intracellular proteins. *Pichia* can be cultured to very high cell density with good genetic stability. Additionally, hyperglycosylation is less of a problem in *Pichia*, which typically have shorter outer-chain mannose units (less than 30 outer-chain residues) than *S. cerevisiae* (greater than 50 residues) (Grinna & Tschopp, 1989).

Methylotrophic yeasts, which are able to use methanol as their sole carbon source, contain regulated methanol enzymes that can be induced to give extremely high levels of expression. In *Pichia* expression systems, the gene that encodes alcohol oxidase (*AOX1*) is most commonly used for the expression of foreign genes, but constitutive promoters are also available. Heterologous genes are inserted into vectors and then integrated into the *Pichia* genome, either duplicating or replacing (transplacement) the target gene, depending on how the linearized vector is constructed. High-level expression relies on integration of multiple copies of the foreign gene and, since this varies significantly, screening colonies to obtain clones with the highest levels of expression is required. Culture conditions and induction protocols are critical for optimal expression. Since *Pichia* are readily oxygen-limited in shake flasks, growth in fermenters is required for high-level expression (approximately five- to tenfold greater than in shake flasks).

Numerous factors make yeast expression systems significantly less straightforward than those of *E. coli*. In addition to the considerations mentioned above, it should be noted that yeast cells are surrounded by a tough cell wall and are therefore notoriously difficult to break. This makes the problem of purification of intracellular protein from yeast that much more difficult. Given the many complexities of expression in yeast, it is usually better to begin with an *E. coli* expression system and move to yeast only if the results obtained with *E. coli* systems are unacceptable. If yeast is used as an expression system, careful attention should be paid to maintaining defined stocks of the expression strain and the corresponding expression plasmids. Despite the availability of comprehensive kits, if the researcher does not have considerable experience with yeast, the enlistment of an experienced colleague is recommended.

3.1.4.3. *Baculoviruses and insect cells*

Baculovirus expression systems are becoming increasingly important tools for the production of recombinant proteins for X-ray crystallography. The insect cell–virus expression systems are more experimentally demanding than bacteria or yeast, but they offer several advantages. Expression of some mammalian proteins has been achieved in baculovirus when simpler expression systems have failed. Because insects are higher eukaryotes, many of the difficulties associated with expression of proteins from higher eukaryotes in *E. coli* do not apply: there is no need for a Shine–Dalgarno sequence, no major problems with codon usage and fewer problems with a lack of appropriate chaperones. Although glycosylation is not the same in insect and mammalian cells, in some cases it is close enough to be acceptable. In addition, for many crystallography projects, minimizing glycosylation is helpful, so that it may be more appropriate to modify the gene or protein to avoid glycosylation (or minimize it) than to try to find ways to recapitulate the glycosylation pattern found in mammalian cells. As is the general case in biotechnology, the development of baculovirus expression systems is work in progress. Progress has been made towards making recombinant proteins in insect cells with glycosylation patterns that match those in mammalian cells (reviewed by Jarvis *et al.*, 1998). Baculovirus systems allow expression of recombinant proteins at reasonable levels, typically ranging from 1–500 $\mathrm{mg\,l^{-1}}$ of cell culture. Considerable work has gone into the development of convenient transfer vectors, and baculovirus expression kits are available from more than ten different commercial sources.

Baculoviruses usually infect insects; in terms of the expression of foreign proteins, the important baculoviruses are the *Autographa californica* nuclear polyhedrosis virus (AcNPV) and the *Bombyx mori* nuclear polyhedrosis virus (BmNPV). AcNPV has been used more widely than BmNPV in cell-culture systems; the BmNPV virus is used primarily to express recombinant proteins in insect larvae. The advantage of BmNPV is that it grows well in larger insect larvae, making the task of harvesting the haemolymph easier. Proteins expressed for crystallography have all been, to the best of our knowledge, expressed using the AcNPV virus system; we will not discuss the BmNPV virus expression system here. Anyone wishing to learn more about either AcNPV or BmNPV is urged to consult two useful monographs: *Baculovirus Expression Protocols* (Richardson, 1995) and *Baculovirus Expression Vectors: A Laboratory Manual* (O'Reilly *et al.*, 1992). There are also shorter reviews that are quite helpful (Jones & Morikawa, 1996; Merrington *et al.*, 1997; Possee, 1997).

In nature, in the late stage of replication in insect larvae, nuclear polyhedrosis viruses produce an occluded form, in which the virions are encased in a crystalline protein matrix, polyhedrin. After the virus is released from the insect larvae, this proteinaceous coat protects the virus from the environment and is necessary for the propagation of the virus in its natural state. However, replication of the virus in cell culture does not require the formation of occlusion bodies. In tissue culture, the production of occlusion bodies is dispensable, and the primary protein, polyhedrin, is not required for replication. Cultured cells infected with wild-type AcNPV produce large amounts of polyhedrin; cells infected with modified AcNPV vectors, with other genes inserted in place of the polyhedrin gene (or in place of another highly expressed gene, *p10*, that is

dispensable in cultured cells), can express impressive amounts of the recombinant protein.

The AcNVP genome is 128 kb, which is too large for convenient direct manipulations. In most cases, novel genes are put into the AcNPV genome by homologous recombination using transfer vectors. Transfer vectors are small bacterial plasmids that contain AcNPV sequences that allow homologous recombination to direct the insertion of the transfer vector into the desired place in the AcNPV genome (often, but not always, the polyhedrin gene). Originally, the purified circular DNA from AcNPV and the appropriate transfer plasmids were simply cotransfected onto monolayers of insect cells. Plaques develop, and if the insertion is targeted to the polyhedrin gene, plaques that contain viruses that retain the ability to make polyhedrin (those that contain the wild-type virus) can be distinguished in the microscope from plaques that do not. This technique works, but has been largely replaced by systems that make it easier to obtain and/or find the recombinant plaques. The AcNPV genome is circular; if the DNA is linearized, it will not produce a replicating virus unless the break is repaired. The repair process is facilitated by the presence of homologous DNA flanking the break. Systems have been set up to exploit this property to increase the efficiency of the generation of vectors that carry the desired insert. Basically, the genome of the AcNPV vector is modified so that there is a unique restriction site at the site where the transfer vector would insert. Linear AcNPV DNA is cotransfected with a transfer vector. This can produce stocks in which greater than 90% of the virus is recombinant. Systems have also been developed in which a DNA insert can be ligated directly into a linearized AcNPV genome. This protocol also produces a high yield of recombinant virus (Lu & Miller, 1996).

There are also a number of systems that allow either the selection or, more often, the ready identification of recombinant virus. The marker most commonly used for this purpose is β-galactosidase; a number of AcNPV vectors or transfer systems that make use of β-galactosidase are commercially available. Once a recombinant plaque is identified, it should be purified through multiple rounds of plaque purification to ensure that a homogeneous stock has been prepared. Several independent isolates should be prepared and each checked for expression of the desired protein.

There are several important things to consider when setting up the cell-culture system. Although most baculoviruses have a relatively restricted host range, and AcNPV was first isolated from alfalfa looper (*Autographa californica*), for the purpose of expressing foreign proteins, it is usually grown in cells of the fall armyworm (*Spodoptera frugiperda*) or the cabbage looper (*Trichoplusia ni*). The isolation and purification of the appropriate AcNPV vectors are usually done in monolayer cultures. In contrast, the production of large amounts of recombinant protein is usually done in suspension cultures. There is also the issue of whether or not to include fetal calf serum in the culture media. In theory, since the cells can be grown in serum-free media, which saves money and makes the subsequent purification of the recombinant protein simpler, serum-free culture is the appropriate choice. However, growing cells in serum-free media is a trickier proposition, and the cells are more sensitive to minor contaminants. As a general rule, high-level production of recombinant proteins using a baculovirus vector requires host cells that are growing rapidly; this is sometimes easier to achieve with serum-containing media. It is not always a simple matter to switch cells adapted to growth on plates to suspension culture, nor is it always easy to switch cells grown in the presence of serum to serum-free culture. Since the vector is a virus, it is usually more convenient to use cells adapted to different conditions than to try to adapt the cells. However, the relative yield of the recombinant protein will not necessarily be the same in different cells grown under different culture conditions.

Although baculoviruses, particularly AcNPV, are convenient vectors, the expression of the recombinant protein is carried out by the insect cell host. Baculovirus infection kills the host cell, so it is not possible to use baculoviruses to derive insect cell cultures that continuously express a recombinant protein. It is possible, however, to introduce DNA segments directly into insect cells and derive cell lines that stably express a recombinant protein; there are constitutive and inducible promoters that can be used in insect cell systems (McCarroll & King, 1997; Pfeifer, 1998). Basically, the protocols used to introduce DNA expression constructs into cultured insect cells are similar to those used in cultured mammalian cells ($CaPO_4$, electroporation, liposomes *etc.*), and similar selective protocols are used (G418, hygromycin, puromycin *etc.*).

Expression systems have been prepared based on baculovirus immediate early promoters and on cellular promoters, including the hsp70 promoter and metallothionein (McCarroll & King, 1997; Kwong *et al.*, 1998; Pfeifer, 1998). Insect cells are, in general, easier (and cheaper) to grow in culture than mammalian cells, although many of the problems that exist in mammalian cell culture also exist in insect cell culture. Relative to the baculovirus system, the use of stable insect cell lines not only allows the continuous culture of cells that contain the desired expression system (provided the expressed protein is not too toxic), it also permits the use of *Drosophila* cell lines, which appear to have some advantages for the high-level production of recombinant proteins.

Compared to bacteria or yeast cells, cells from higher eukaryotes are quite delicate, and considerable care must be taken in cell culture. The cells are subject to shear stress, which can be a problem in stirred and/or shaken cultures; some researchers use airlift fermenters to help alleviate the problem. Compared to yeast and bacterial cells, cultured cells grow relatively slowly and require rich media that will support the rapid growth of a wide variety of unwanted organisms, so special care must be taken to avoid contaminating the cultures. Antibiotics are commonly used; however, antibiotics will not, in general, prevent contamination with yeasts or moulds, which often cause the greatest problems. If the baculovirus system is used, then the cells and viruses are kept separate, and the cells are relatively standard reagents. If there is contamination, the contaminated cultures can be discarded and replaced with fresh cells (and viruses). Stable transformed insect cells that express a recombinant protein must be kept free of all contaminants. As is always the case, both cells and viruses should be carefully stored. Any useful recombinant baculovirus can be easily stored as DNA.

3.1.4.4. *Mammalian cells*

In some cases, however, even the baculovirus and/or insect cell expression systems are not able to make the desired recombinant protein product. If the recombinant protein is sufficiently important, it can be produced in cultured mammalian cells. Although biotechnology companies have demonstrated that it is feasible to produce kilograms of pure recombinant proteins using cultured mammalian cells, the effort required to produce tissue culture cells that express high levels of recombinant protein is substantial, and the costs of growing large amounts of tissue culture cells are beyond the means of all but the best-funded laboratories. To make matters worse, there are no well defined plasmids that reliably and stably replicate in mammalian cells. It may be possible to develop reliable episomal replication systems based on viral replicons; however, even the best developed viral episomes are still not entirely satisfactory (see, for example, Sclimenti & Calos, 1998). Cell lines are usually prepared by transfection; following transfection, some of the cells (usually a small percentage) will incorporate transfected DNA into their genomes. A number of agents can be used to transfect DNA; these include, but are not limited to, $CaPO_4$,

DEAE Dextran, cationic lipids *etc.* This is a complex and poorly defined process; the transfected DNA is often incorporated into complex tandem arrays. Neither the amount of transfected DNA nor its location in the host genome is controlled in a standard transfection; as a consequence, the expression level varies substantially from one transfected cell to another. This makes the process of creating mammalian cell lines that efficiently and stably express a recombinant protein a labour-intensive process. Ordinarily, the DNA segment carrying the gene for the desired recombinant protein is linked to a selectable marker; selection for the marker is usually sufficient to cause the retention of the gene for the desired recombinant protein, provided that it is not toxic to the host cell. The tandem arrays produced by transfecting DNA into mammalian cells are often unstable. Recombination within the tandem array can decrease (or less commonly increase) the number of copies of the transfected gene. It is possible to take advantage of this instability. Selection protocols, which usually involve the DHFR gene and methotrexate, have been developed that can select for cells that have the DNA segments containing both the selectable marker and the gene for the desired recombinant protein in higher copy number (Kaufman, 1990); these can be used to develop cell lines that express high levels of the recombinant protein.

There are alternative methods that can be used to deliver an expression construct to a cultured mammalian cell. For example, the DNA can be introduced by electroporation, and homologous recombination can be used to embed an expression construct at a specific place in the host genome. However, such strategies, while in some ways more elegant than simple DNA transfection, do not appear to simplify the problem of creating a cell line that produces large amounts of a specific gene product. There are also a variety of viral vectors that can be used to introduce genes into cells either transiently or stably. At the time of writing, viral vector systems, which are extremely useful for studying the effects of expressing foreign genes in cultured mammalian cells, do not appear to offer any obvious advantages for the preparation of cultured cells that can express the relatively large amounts of recombinant protein needed for crystallography. However, this is an area where the research effort is particularly intense, so it is entirely possible that in the near future there will be a viral vector (or vectors) which will offer significant advantages for inducing high-level expression of recombinant proteins in cultured mammalian cells.

Until relatively recently, one of the primary problems in working with expression systems in cultured mammalian cells has been the lack of a tightly regulated inducible system. This has made the high-level expression of proteins that are deleterious to the growth of the cell an exceptionally difficult problem. The promoters originally used for inducible expression in cultured 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, individual 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

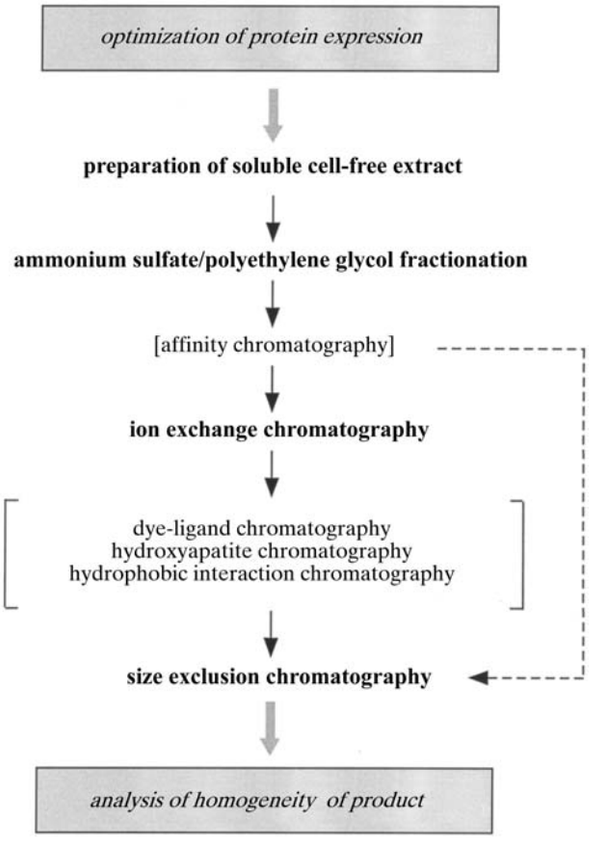

Fig. 3.1.5.1. Protein purification strategy. Purification of proteins expressed at reasonably high levels typically requires only a limited number of chromatographic steps. Additional chromatography columns (indicated in brackets) can be included as necessary. Affinity chromatography can allow efficient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

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 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 laboratory or through close collaborations. There yet remain formidable problems in protein structure and function that will require all the combined talents of the most skilled practitioners of these arcane arts.

References

Abelson, J. N. & Simon, M. I. (1990). *Guide to protein purification. Methods Enzymol.* **182**, 1–894.

Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A. & Struhl, K. (1995). *Short protocols in molecular biology: a compendium of methods from current protocols in molecular biology*, 3rd ed. New York: Greene Publishing Associates and Wiley.

Beggs, J. D. (1978). *Transformation of yeast by a replicating hybrid plasmid. Nature (London)*, **275**, 104–109.

Bhandari, P. & Gowrishankar, J. (1997). *An Escherichia coli host strain useful for efficient overproduction of cloned gene products with NaCl as the inducer. J. Bacteriol.* **179**, 4403–4406.

Biswas, E. E., Fricke, W. M., Chen, P. H. & Biswas, S. B. (1997). *Yeast DNA helicase A: cloning, expression, purification, and enzymatic characterization. Biochemistry*, **36**, 13277–13284.

Bollag, D. M., Rozycki, M. D. & Edelstein, S. J. (1996). *Protein methods.* New York: Wiley-Liss.

Boyer, P. L. & Hughes, S. H. (1996). *Site-directed mutagenic analysis of viral polymerases and related proteins. Methods Enzymol.* **275**, 538–555.

Brinkmann, U., Mattes, R. E. & Buckel, P. (1989). *High-level expression of recombinant genes in Escherichia coli is dependent on the availability of the dnaY gene product. Gene*, **85**, 109–114.

Broach, J. R. (1983). *Construction of high copy number yeast vectors using 2 μm circle sequences. Methods Enzymol.* **101**, 307–325.

Chong, S., Mersha, F. B., Comb, D. G., Scott, M. E., Landry, D., Vence, L. M., Perler, F. B., Benner, J., Kucera, R. B., Hirvonen, C. A., Pelletier, J. J., Paulus, H. & Xu, M. Q. (1997). *Single-column purification of free recombinant proteins using a self-cleavable affinity tag derived from a protein splicing element. Gene*, **192**, 271–281.

Chong, S., Shao, Y., Paulus, H., Benner, J., Perler, F. B. & Xu, M. Q. (1996). *Protein splicing involving the Saccharomyces cerevisiae VMA intein. The steps in the splicing pathway, side reactions leading to protein cleavage, and establishment of an in vitro splicing system. J. Biol. Chem.* **271**, 22159–22168.

Cohen, S. L., Ferre-D'Amare, A. R., Burley, S. K. & Chait, B. T. (1995). *Probing the solution structure of the DNA-binding protein Max by a combination of proteolysis and mass spectrometry. Protein Sci.* **46**, 1088–1099.

Cole, P. A. (1996). *Chaperone-assisted protein expression. Structure*, **4**, 239–242.

Cregg, J. M., Vedick, T. S. & Raschke, W. C. (1993). *Recent advances in the expression of foreign genes in Pichia pastoris. Biotechnology*, **11**, 905–910.

De Bernardez Clark, E. (1998). *Refolding of recombinant proteins. Curr. Opin. Biotechnol.* **9**, 157–163.

De Boer, H. A. & Kastelein, R. A. (1986). *Biased codon usage: an exploration of its role in optimization of translation.* In *Maximizing gene expression*, edited by W. S. Reznikioff & L. Gold, pp. 225–285. Boston: Butterworths.

Del Tito, B. J. Jr, Ward, J. M., Hodgson, J., Gershater, C. J. L., Edwards, H., Wysocki, L. A., Watson, F. A., Sathe, G. & Kane, J. F. (1995). *Effects of a minor isoleucyl tRNA on heterologous protein translation in Escherichia coli. J. Bacteriol.* **177**, 7086–7091.

Enfors, S.-O. (1992). *Control of in vivo proteolysis in the production of recombinant proteins. Trends Biotechnol.* **10**, 310–315.

Ernst, J. F. & Kawashima, E. (1988). *Variations in codon usage are not correlated with heterologous gene expression in Saccharomyces cerevisiae and Escherichia coli. J. Biotechnol.* **7**, 1–9.

Ferré-D'Amaré, A. R. & Burley, S. K. (1994). *Use of dynamic light scattering to assess crystallizability of macromolecules and macromolecular assemblies. Structure*, **2**, 357–359.

Fischer, B., Sumner, I. & Goodenough, P. (1993). *Isolation, renaturation, and formation of disulfide bonds of eukaryotic proteins expressed in Escherichia coli as inclusion bodies. Biotechnol. Bioeng.* **41**, 3–13.

Georgiou, G. & Valax, P. (1996). *Expression of correctly folded proteins in Escherichia coli. Curr. Opin. Biotechnol.* **7**, 190–197.

Gold, L., Pribnow, D., Schneider, T., Shinedling, S., Singer, B. S. & Stormo, G. (1981). *Translational initiation in prokaryotes. Annu. Rev. Microbiol.* **35**, 365–403.

Goldman, E., Rosenberg, A. H., Zubay, G. & Studier, F. W. (1995). *Consecutive low-usage leucine codons block translation only when near the 5′ end of a message in Escherichia coli. J. Mol. Biol.* **245**, 467–473.

Gottesman, S. (1990). *Minimizing proteolysis in Escherichia coli: genetic solutions. Methods Enzymol.* **185**, 119–129.

Grinna, L. S. & Tschopp, J. F. (1989). *Size distribution and general structural features of N-linked oligosaccharides from the methylotrophic yeast, Pichia pastoris. Yeast*, **5**, 107–115.

Guise, A. D., West, S. M. & Chaudhuri, J. B. (1996). *Protein folding in vivo and renaturation of recombinant proteins from inclusion bodies. Mol. Biotechnol.* **6**, 53–64.

Hendrickson, W. A., Horton, J. R. & LeMaster, D. M. (1990). *Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. EMBO J.* **9**, 1665–1672.

Hernan, R. A., Hui, H. L., Andracki, M. E., Noble, R. W., Sligar, S. G., Walder, J. A. & Walder, R. Y. (1992). *Human hemoglobin expression in Escherichia coli: importance of optimal codon usage. Biochemistry*, **31**, 8619–8628.

Higgins, D. R. & Cregg, J. (1998). *Methods in molecular biology*, Vol. 103. *Pichia protocols.* Totowa: Humana Press.

Hirel, P. H., Schmitter, M. J., Dessen, P., Fayat, G. & Blanquet, S. (1989). *Extent of N-terminal methionine excision from Escherichia coli proteins is governed by the side-chain length of the penultimate amino acid. Proc. Natl Acad. Sci. USA*, **86**, 8247–8251.

Hockney, R. C. (1994). *Recent developments in heterologous protein production in Escherichia coli. Trends Biotechnol.* **12**, 456–463.

Hofmann, A., Tai, M., Wong, W. & Glabe, C. G. (1995). *A sparse matrix screen to establish initial conditions for protein renaturation. Anal. Biochem.* **230**, 8–15.

Hollenberg, C. P. & Gellissen, G. (1997). *Production of recombinant proteins by methylotrophic yeasts. Curr. Opin. Biotechnol.* **8**, 554–560.

Hubbard, S. J. (1998). *The structural aspects of limited proteolysis of native proteins. Biochim. Biophys. Acta*, **1382**, 191–206.

Innis, M. A., Gelfand, D. H., Sninsky, J. J. & White, T. J. (1990). *PCR protocols: a guide to methods and applications.* San Diego: Academic Press.

Jarvis, D. L., Kawar, Z. S. & Hollister, J. R. (1998). *Engineering N-glycosylation pathways in the baculovirus-insect cell system. Curr. Opin. Biotechnol.* **9**, 528–533.

Jones, I. & Morikawa, Y. (1996). *Baculovirus vectors for expression in insect cells. Curr. Opin. Biotechnol.* **7**, 512–516.

Kane, J. F. (1995). *Effects of rare codon clusters on high-level expression of heterologous proteins in Escherichia coli. Curr. Opin. Biotechnol.* **6**, 494–500.

Kaufman, R. J. (1990). *Selection and coamplification of heterologous genes in mammalian cells. Methods Enzymol.* **185**, 537–566.

Kim, R., Sandler, S. J., Goldman, S., Yokota, H., Clark, A. J. & Kim, S.-H. (1998). *Overexpression of archaeal proteins in Escherichia coli. Biotechnol. Lett.* **20**, 207–210.

Krueger, J. K., Kulke, M. H., Schutt, C. & Stock, J. (1989). *Protein inclusion body formation and purification. BioPharm*, March issue, 40–45.

Kwong, P. D., Wyatt, R., Robinson, J., Sweet, R. W., Sodroski, J. & Hendrickson, W. A. (1998). *Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. Nature (London)*, **393**, 648–659.

LaVallie, E. R., DiBlasio, E. A., Kovacic, S., Grant, K. L., Schendel, P. F. & McCoy, J. M. (1993). *A thioredoxin gene fusion expression system that circumvents inclusion body formation in the E. coli cytoplasm. Biotechnology*, **11**, 187–193.

LaVallie, E. R. & McCoy, J. M. (1995). *Gene fusion expression systems in Esherichia coli. Curr. Opin. Biotechnol.* **6**, 501–506.

Lee, H. W., Joo, J.-H., Kang, S., Song, L.-S., Kwon, J.-B., Han, M. H. & Na, D. S. (1992). *Expression of human interleukin-2 from native and synthetic genes in E. coli: no correlation between major codon bias and high level expression. Biotechnol. Lett.* **14**, 653–658.

Lu, A. & Miller, L. K. (1996). *Generation of recombinant baculoviruses by direct cloning. Biotechniques*, **21**, 63–68.

McCarroll, L. & King, L. A. (1997). *Stable insect cell cultures for recombinant protein production. Curr. Opin. Biotechnol.* **8**, 590–594.

McPherson, M. J., Hames, B. D. & Taylor, G. R. (1995). *PCR 2: a practical approach*. Oxford, New York: IRL Press at Oxford University Press.

Makrides, S. C. (1996). *Strategies for achieving high-level expression of genes in Escherichia coli. Microbiol. Rev.* **60**, 512–538.

Marston, F. A. (1986). *The purification of eukaryotic polypeptides synthesized in Escherichia coli. Biochem. J.* **240**, 1–12.

Merrington, C. L., Bailey, M. J. & Possee, R. D. (1997). *Manipulation of baculovirus vectors. Mol. Biotechnol.* **8**, 283–297.

Mitraki, A. & King, J. (1989). *Protein folding intermediates and inclusion body formation. Biotechnology*, **7**, 690–697.

Mohsen, A.-W. A. & Vockley, J. (1995). *High-level expression of an altered cDNA encoding human isovaleryl-CoA dehydrogenase in Escherichia coli. Gene*, **160**, 263–267.

Murby, M., Uhlén, M. & Ståhl, S. (1996). *Upstream strategies to minimize proteolytic degradation upon recombinant production in Escherichia coli. Protein Expr. Purif.* **7**, 129–136.

Nilsson, B., Forsberg, G., Moks, T., Hartmanis, M. & Uhlén, M. (1992). *Fusion proteins in biotechnology and structural biology. Curr. Opin. Struct. Biol.* **2**, 569–575.

O'Reilly, D. R., Miller, L. K. & Luckow, V. A. (1992). *Baculovirus expression vectors: A laboratory manual.* New York: W. H. Freeman & Co.

Pfeifer, T. A. (1998). *Expression of heterologous proteins in stable insect culture. Curr. Opin. Biotechnol.* **9**, 518–521.

Possee, R. D. (1997). *Baculoviruses as expression vectors. Curr. Opin. Biotechnol.* **7**, 569–572.

Richardson, C. D. (1995). *Methods in molecular biology*, Vol. 39. *Baculovirus expression protocols*. Totowa: Humana Press.

Richarme, G. & Caldas, T. D. (1997). *Chaperone properties of the bacterial periplasmic substrate-binding proteins. J. Biol. Chem.* **272**, 15607–15612.

Ringquist, S., Shinedling, S., Barrick, D., Green, L., Binkley, J., Stormo, G. D. & Gold, L. (1992). *Translational initiation in Escherichia coli: sequences within the ribosome-binding site. Mol. Microbiol.* **6**, 1219–1229.

Romanos, M. (1995). *Advances in the use of Pichia pastoris for high-level gene expression. Curr. Opin. Biotechnol.* **6**, 527–533.

Romanos, M. A., Scorer, C. A. & Clare, J. J. (1992). *Foreign gene expression in yeast: a review. Yeast*, **8**, 423–488.

Rossi, F. M. & Blau, H. M. (1998). *Recent advances in inducible expression systems. Curr. Opin. Biotechnol.* **9**, 451–456.

Sachdev, D. & Chirgwin, J. M. (1998). *Solubility of proteins isolated from inclusion bodies is enhanced by fusion to maltose-binding protein or thioredoxin. Protein Expr. Purif.* **12**, 122–132.

Saez, E., No, D., West, A. & Evans, R. M. (1997). *Inducible expression in mammalian cells and transgenic mice. Curr. Opin. Biotechnol.* **8**, 608–616.

Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular cloning: A laboratory manual*, 2nd ed. New York: Cold Spring Harbor Laboratory Press.

Samuelsson, E., Moks, T., Nilsson, B. & Uhlén, M. (1994). *Enhanced in vitro refolding of insulin-like growth factor I using a solubilizing fusion partner. Biochemistry*, **33**, 4207–4211.

Schein, C. H. & Noteborn, M. H. M. (1988). *Formation of soluble recombinant proteins in Escherichia coli is favored by lower growth temperature. Biotechnology*, **6**, 291–294.

Schenk, P. M., Baumann, S., Mattes, R. & Steinbiss, H.-H. (1995). *Improved high-level expression system for eukaryotic genes in Escherichia coli using T7 RNA polymerase and rare ArgtRNAs. Biotechniques*, **19**, 196–198.

Sclimenti, C. R. & Calos, M. P. (1998). *Epstein–Barr virus vectors for gene expression and transfer. Curr. Opin. Biotechnol.* **9**, 476–479.

Scopes, R. K. (1994). *Protein purification: principles and practice*. New York: Springer-Verlag.

Shimotohno, K. & Temin, H. M. (1982). *Loss of intervening sequences in mouse α-globin DNA inserted in an infectious retrovirus vector. Nature (London)*, **299**, 265–268.

Sorge, J. & Hughes, S. H. (1982). *Splicing of intervening sequences introduced into an infectious retroviral vector. J. Mol. Appl. Genet.* **1**, 547–559.

Studier, F. W. & Moffatt, B. A. (1986). *Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. J. Mol. Biol.* **189**, 113–130.

Studier, F. W., Rosenberg, A. H., Dunn, J. J. & Dubendorff, J. W. (1990). *Use of T7 RNA polymerase to direct expression of cloned genes. Methods Enzymol.* **185**, 60–89.

Tabor, S. & Richardson, C. C. (1985). *A bacteriophage T7 RNA polymerase/promoter system for controlled exclusive expression of specific genes. Proc. Natl Acad. Sci. USA*, **82**, 1074–1078.

Tobias, J. W., Shrader, T. E., Rocap, G. & Varshavsky, A. (1991). *The N-end rule in bacteria. Science*, **254**, 1374–1377.

Tsunasawa, S., Stewart, J. W. & Sherman, F. S. (1985). *Amino-terminal processing of mutant forms of yeast iso-1-cytochrome c. The specificities of methionine aminopeptidase and acetyltransferase. J. Biol. Chem.* **260**, 5382–5391.

Unger, T. F. (1997). *Show me the money: prokaryotic expression vectors and purification systems. The Scientist*, **11**, 20–23.

Wall, J. G. & Plückthun, A. (1995). *Effects of overexpressing folding modulators on the in vivo folding of heterologous proteins in Escherichia coli. Curr. Opin. Biotechnol.* **6**, 507–516.

Waller, J.-P. (1963). *The NH_2-terminal residues of the proteins from cell-free extracts of E. coli. J. Mol. Biol.* **7**, 483–496.

Werner, M. H., Clore, G. M., Gronenborn, A. M., Kondoh, A. & Fisher, R. J. (1994). *Refolding proteins by gel filtration chromatography. FEBS Lett.* **345**, 125–130.

Wilkinson, D. L., Ma, N. T., Haught, C. & Harrison, R. G. A. (1995). *Purification by immobilized metal affinity chromatography of human atrial natriuretic peptide expressed in a novel thioredoxin fusion protein. Biotechnol. Prog.* **11**, 265–269.

Yasukawa, T., Kanei-Ishii, C., Maekawa, T., Fujimoto, J., Yamamoto, T. & Ishii, S. (1995). *Increase of solubility of foreign proteins in Escherichia coli by coproduction of the bacterial thioredoxin. J. Biol. Chem.* **270**, 25328–25331.

Yonemoto, W. M., McGlone, M. L., Slice, L. W. & Taylor, S. S. (1998). *Prokaryotic expression of catalytic subunit of adenosine cyclic monophosphate-dependent protein kinase.* In *Protein phosphorylation*, edited by B. M. Sefton & T. Hunter, pp. 419–434. San Diego: Academic Press.

Zhang, S. P., Zubay, G. & Goldman, E. (1991). *Low usage codons in Escherichia coli, yeast, fruit fly and primates. Gene*, **105**, 61–72.