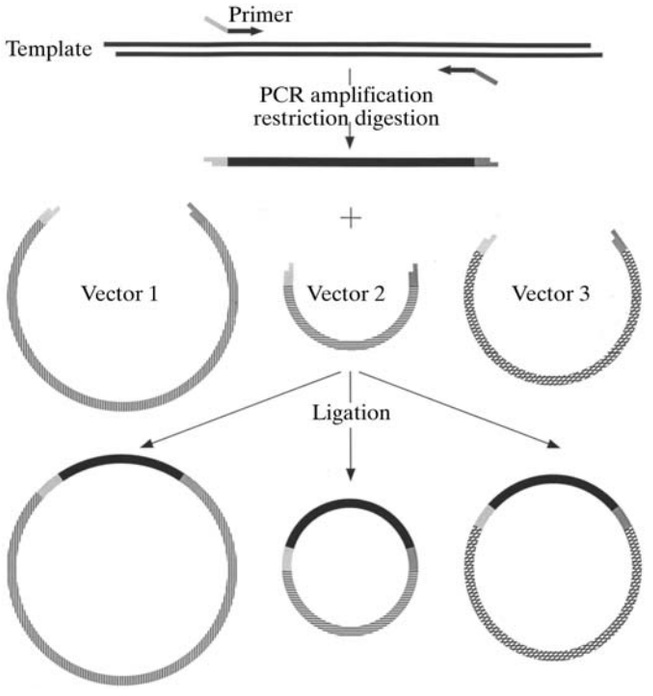

Figure 3.1.3.1
Creating an expression construct. PCR can be used to amplify the coding region of interest, providing that a suitable template is available. PCR primers should be designed to contain one or more restriction sites that can be conveniently used to subclone the fragment into the desired expression 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.

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.

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 λ 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.

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

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

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 × 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*). Transformation 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 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, indi-