

3. TECHNIQUES OF MOLECULAR BIOLOGY

cular 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

3.1. PREPARING RECOMBINANT PROTEINS FOR X-RAY CRYSTALLOGRAPHY

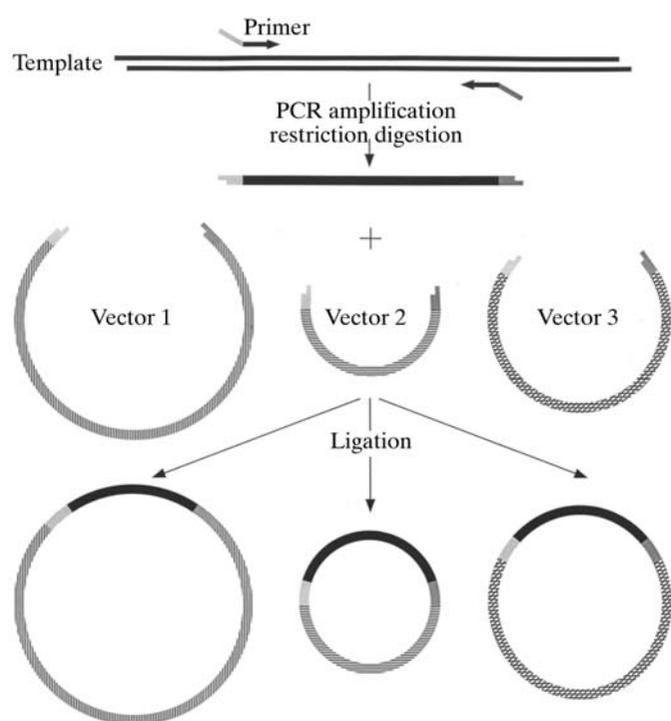


Figure 3.1.3.1

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 (His₆), 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