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