International Tables for Crystallography (2012). Vol. F, Section 3.1.2, pp. 75–76.

PART 3. TECHNIQUES OF MOLECULAR BIOLOGY

Chapter 3.1. Preparing recombinant proteins for X-ray crystallography

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 parti-

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