Introduction to single crystal X-ray analysis X. Protein expression for X-ray structure analysis

Takashi Matsumoto*

1. Introduction

In order to elucidate various biological phenomena occurring in vivo, it is essential to determine the structure of proteins. This article will focus on expression of proteins for X-ray analysis.

For performing structure analysis, the first challenge is to establish an expression system. Structure analysis requires large amounts of proteins with high purity, so it is essential to establish a large-scale protein expression and purification systems. Even target proteins that are only present in trace amounts in vivo can be overexpressed as recombinant proteins using genetic engineering techniques and appropriate hosts. In such cases, it is crucial to select an appropriate expression system from numerous existing protein expression systems. Some expression systems cannot be used or are more difficult to use, depending on whether the protein of interest is of prokaryotic or eukaryotic origin. For multi-domain proteins or proteins expected to undergo large structure change, regions containing only certain domains can be selectively expressed to suppress structure change. For rapid purification of the expressed proteins, in most cases, purification tags are generally added to the N- and/or C-terminals of proteins. Sometimes, the expressed proteins do not fold properly and are expressed as "inclusion bodies", which cannot be directly used as starting materials for purification. As can be seen, there are many challenges to be overcome to obtain proteins that can be used as starting materials for structure analysis.

2. Characterization of individual expression systems

There are several types of expression systems using different prokaryotic or eukaryotic hosts (Table 1).

The most widely used type is the prokaryotic expression system (*Escherichia coli*). Various tools are commercially available for readily establishing the *E. coli* expression system. The system offers high expression levels and is thus suitable for obtaining large amounts of proteins at low cost. Meanwhile, it also has some disadvantages. Since the resulting recombinant proteins are not intrinsic to the host cells, depending on the expression levels, expression rates and the nature of the proteins, they may fail to fold properly and form insoluble aggregates called "inclusion bodies". Inclusion bodies do not have proper folding and thus cannot be directly used in purification; the expression levels or rates need to be controlled or the host needs to be changed. An alternative methodis to perform a "refolding" step, in which the inclusion bodies are denatured once and subsequently folded again to facilitate formation of properly structured proteins. Moreover, expression of eukaryotic proteins may be associated with problems involving rare codons or limitations on post-translational modification. When there are several codons encoding the same amino acid, those less frequently used in a particular organism are

Table 1. Characterization of individual expression systems.

Expression system	Advantages	Disadvantages
E. coli	Low cost and high expression level Abundance of tools (vectors, hosts) Easy handling	Inclusion body formation Limited post-translational modification
Yeast	High expression level Successful glycosylation in some cases	Different glycosylation
Insect cells	High expression level Successful glycosylation	High cost Protein degradation Limited glycosylation
Animal cells	High chance of forming active structures Successful post-translational modification	High cost
Cell-free	Successful expression of toxic proteins Successful expression in the presence of additives (*e.g.* ribosome)	Limited expression level High cost

* Application Laboratories, Rigaku Corporation.

called rare codons. Since rare codons differ between prokaryotes and eukaryotes, when eukaryotic proteins are expressed in prokaryotic hosts, translation may be stopped at the rare codons, resulting in a significant decrease in their expression levels. There are also some difficulties involving post-translational modification. Due to its limited post-translational modification functions, *E. coli* cannot perform glycosylation to add sugar chains to glycoproteins. For such proteins, eukaryotic expression systems must be used.

Eukaryotic systems using yeast or insect cells are capable of performing glycosylation, but they can only perform glycosylation in different or limited manners compared to animal cells. Nevertheless, due to their relatively high expression levels and low expression costs compared to those of the animal cell system described below, such systems are sometimes used, although their use is limited to cases where the difference in glycosylation does not affect protein activity. The expression system based on insect cells is associated with the protein degradation problem, so the progress of protein expression and degradation must be monitored and adequately controlled.

Meanwhile, the animal cell expression system, which directly uses animal cells, has the advantages of performing the required post-translational modification and offering proteins having active structures. However, extremely high expression costs and requirement of special facilities are posing great obstacles to its use in large-scale protein preparation.

When the proteins to be expressed are toxic, they may either form inclusion bodies, fail to express or induce lysis of the host cells, perhaps for the protection of host cells. In such cases, the cell-free expression system may be effective, which uses enzymes derived from wheat germs, *E. coli*, etc. instead of using cells such as *E. coli* cells to express recombinant proteins. Since no cells are used, there are higher chances for the toxic proteins to be expressed, but higher costs may be required for scaling-up the expression system.

3. Procedures for establishing *E. coli* expression system

This section describes the most popular expression system using *E. coli*. The *E. coli* expression system utilizes expression vectors available from various suppliers. Gene sequences (target gene sequence) encoding the target proteins are incorporated into the expression vectors and used.

3.1. Establishment of vectors for *E. coli* expression system

The first step is to select a target region to be expressed. If the target of structure analysis is a functional domain whose position can be predicted by amino acid sequence alignment of analogous proteins, this domain alone may be selected as the expression target. With regard to crystallization, the presence of highly flexible regions or regions not folding into fixed structures within the protein molecules can interfere with their crystallization. Thus, if the presence of such highly flexible regions is suggested by secondary structure prediction or analogous protein structures, successful crystallization may be facilitated by expressing recombinant proteins lacking these highly flexible regions. When large, potentially unfolding regions or loop structures are found between domains, one solution would be to remove these regions and connect the domains with appropriate linkers. Once an expression target region is selected, the target gene sequence is amplified by PCR using primers containing restriction site sequences and subsequently incorporated into an expression vector using the multicloning site (MCS) of the relevant vector. Among the multiple restriction site sequences contained in the MCS, sequences that are the same as the ones used for PCR are used to ligate the expression vector and the target gene sequence.

"Affinity tags (tags)" are often added to the N- or C-terminals of the recombinant proteins to enhance rapid purification. The addition of such tags enables the use of affinity purification, which significantly increases the degree of purity in single-step purification. Although there are many types of tags, two of them are mainly used, namely His (Histidine) and GST (Glutathione S-transferase) tags. The His tag, generally consisting of six consecutive His residues, is added to the N- or C-terminal. The GST tag consists of GST, which is about 23 kDa and is fused to the N-terminal of the recombinant protein. Since the GST tag is a protein, a protease cleavage site is usually introduced between the GST tag and the recombinant protein to cleave the GST tag after the purification is completed. Expression vectors having protease cleavage sequences incorporated between the GST tags and the cloning sites of the target protein-encoding gene sequences are commercially available and can be readily used. Most His tags are not cleaved, as they consist of only about six amino acids and seldom affect crystallization without being removed. Nevertheless, His tags are removed in some cases to avoid potential deleterious effects on crystal packing (reduction of resolution). Commercial expression vectors having protease cleavage sequences incorporated for His tags are also available and can be readily used.

3.2. Selection of *E. coli* and vector introduction

Various *E. coli* strains are available for the *E. coli* expression system. In addition to the widely used BL21 strain, commercially available strains include those resistant to toxic proteins and those supplementing rare codons. To begin with, genes encoding the target proteins should be introduced into the standard BL21 strain to check their expression. If sufficient expression levels cannot be achieved or if too many inclusion bodies are formed, attempts should be made to introduce the target genes into other *E. coli* strains. Introduction of the expression vectors into *E. coli* is performed by

transformation using competent cells. Competent cells are commercially available from many suppliers but can also be self-made without much difficulty, so the choice should be made depending on the purpose and situation. After the prepared expression vector is introduced into the competent cells, they are spread and incubated on antibiotic agar plates corresponding to the resistance genes carried by the expression vector to select *E. coli* cells carrying or not carrying the vector. *E. coli* cells carrying the expression vector can grow due to their drug resistance, while those not carrying the vector cannot.

3.3. Confirmation of expression

Single colony on the agar plate is incubated in liquid media (*e.g.* LB media) in the presence of antibiotics. First, pre-incubation is performed using about 5 ml of liquid medium in a test tube, followed by main incubation using about 50 ml of liquid medium in a conical flask. About 10 ml of *E. coli* culture is removed from the culture before inducing expression to compare the expression level of the target protein before and after the induction. An expression inducer (*e.g.* isopropyl beta-D-thiogalactopyranoside) is added to the culture in the logarithmic phase to induce protein expression. First, the culture and induction should be performed at 37°C. A few hours after starting the induction, the bacterial culture is centrifuged to harvest the *E. coli* cells, which are subsequently washed several times with buffer solutions and homogenized using an ultrasonic homogenizer. After homogenizing, the bacterial solution is centrifuged to separate insoluble and soluble fractions. SDS-PAGE is performed before and after the expression induction using the insoluble and soluble fractions to check whether the target protein has been successively expressed as a soluble protein (Fig. 1).

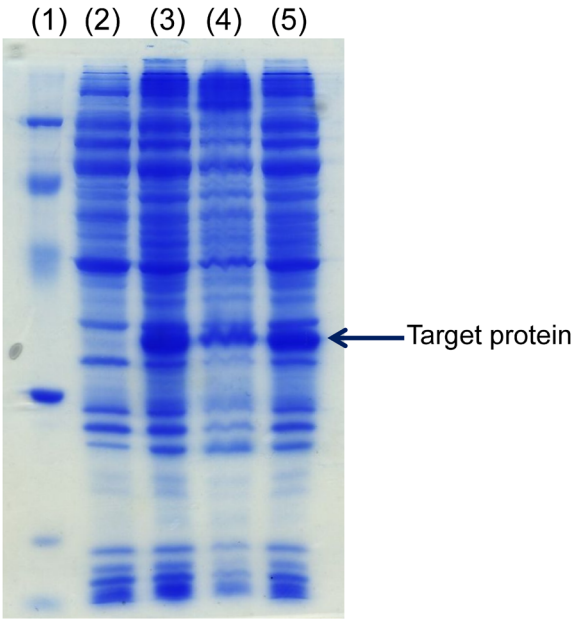

Fig. 1. SDS-PAGE for confirming protein expression.
(1) Molecular weight marker
(2) Before induction
(3) After induction
(4) Insoluble fraction
(5) Soluble fraction

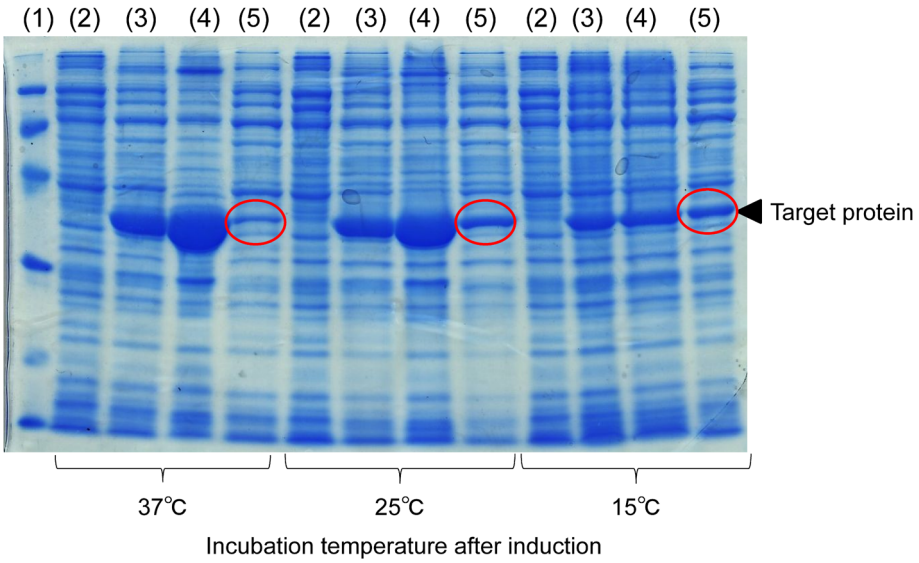

Fig. 2. SDS-PAGE of samples representing different expression induction temperatures.
Lower induction temperatures resulted in reduced expression levels but increased target protein levels in the soluble fraction.
(1) Molecular weight marker
(2) Before induction
(3) After induction
(4) Insoluble fraction
(5) Soluble fraction

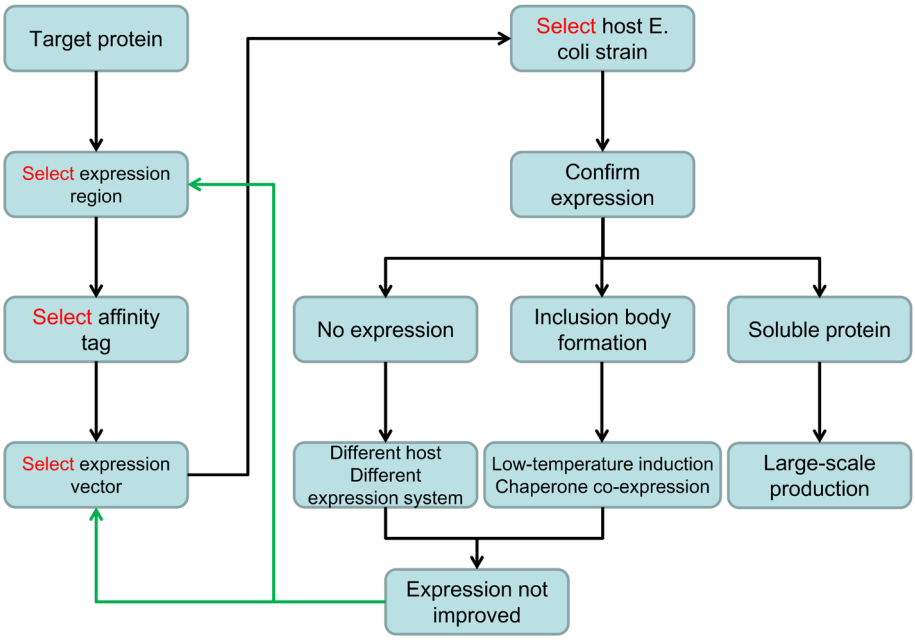

Fig. 3. Flow diagram for establishment of *E. coli* expression system and expression confirmation.

3.4. Optimization of expression conditions

Depending on the result of the expression confirmation, expression conditions should be optimized. If the target protein is expressed in the soluble fraction at a sufficient level, the process should proceed to large-scale production for crystallization. If the target protein is expressed in the soluble fraction but not at a sufficient level, the culture medium should be changed to one with a higher nutrient content to increase the cell density of *E. coli* and reconfirm protein expression. If the target protein is expressed but forms inclusion bodies, the incubation temperature during induction should be lowered. Induction is usually performed at an incubation temperature of 37℃, but when expression levels become extremely high, proteins may not fold properly and form inclusion bodies. In such cases, the incubation temperature at induction should be lowered from 37℃ to 25℃, 15℃, etc. A lower incubation temperature at induction results in a reduced expression level, but some portions of the target protein may fold properly and appear in the soluble fraction (Fig. 2). In cases where lower expression temperature does not produce any improvement, target proteins can be co-expressed with chaperones using commercial vectors encoding chaperone proteins. The abundance of chaperones may facilitate proper folding of the target proteins, leading to their appearance in the soluble fraction. When expression of the target protein cannot be detected, its high toxicity to the bacterial cells may be hampering its expression. The expression can be potentially improved by changing the host to an *E. coli* strain resistant to the toxicity of the protein. If the target protein is of a human or eukaryotic origin, rare codons may be hampering its expression. Expression of eukaryotic proteins may be improved by changing the hosts to E. coli strains supplementing rare codons.

If the expression levels of the target proteins cannot be increased or their solubility cannot be improved (i.e. inclusion bodies are not converted into soluble proteins) by the above optimization approaches, expression vectors should be changed or the choice of expression regions should be reconsidered.

4. Conclusion

For the *E. coli* expression system, various tools, vectors and host strains are available from many suppliers, and high expression levels can be readily achieved at low cost. *E. coli* strains capable of solving the problems associated with eukaryotic rare codons are commercially available, contributing to a number of successful reports on the expression of eukaryotic proteins in *E. coli*. Whether the target protein is of prokaryotic or eukaryotic origin, a recommended procedure is to first try the *E. coli* expression system (Fig. 3) and, if the expression level or formation of inclusion bodies cannot be improved, then try a eukaryotic expression system.