Latest approaches for efficient protein production in drug discovery

Toshihiko Sugiki, Toshimichi Fujiwara & Chojiro Kojima†
Institute for Protein Research, Osaka University, Osaka, Japan

Introduction: Pharmaceutical research looks to discover and develop new compounds which influence the function of disease-associated proteins or respective protein–protein interactions. Various scientific methods are available to discover those compounds, such as high-throughput screening of a library comprising chemical or natural compounds and computational rational drug design. The goal of these methods is to identify the seed compounds of future pharmaceuticals through the use of these technologies and laborious experiments. For every drug discovery effort made, the possession of accurate functional and structural information of the disease-associated proteins helps to assist drug development. Therefore, the investigation of the tertiary structure of disease-associated proteins and respective protein–protein interactions at the atomic level are of crucial importance for successful drug discovery.

Areas covered: In this review article, the authors broadly outline current techniques utilized for recombinant protein production. In particular, the authors focus on bacterial expression systems using *Escherichia coli* as the living bioreactor.

Expert opinion: The recently developed pCold-glutathione S-transferase (GST) system is one of the best systems for soluble protein expression in *E. coli.* Where the pCold-GST system does not succeed, it is preferable to change the host from *E. coli* to higher organisms such as yeast expression systems like *Pichia pastoris* and *Kluyveromyces lactis.* The selection of an appropriate expression system for each desired protein and the optimization of experimental conditions significantly contribute toward the successful outcome of any drug discovery study.

Keywords: *Escherichia coli*, Fed-batch cultivation, high-cell-density cultivation, pCold-glutathione S-transferase system, protein expression system, single protein production system

Expert Opin. Drug Discov. [Early Online]

1. Overview of recombinant protein production for drug discovery

The first step in the generation of new pharmaceuticals is to find drug molecules from a library comprising small-molecular-weight chemical compounds or a myriad of natural products which show therapeutic or disease-preventive effects by performing a huge number of *in vitro* and *in vivo* screening experiments [1]. The drugs can show their therapeutic effects by inhibiting or activating molecular signaling events such as protein–protein and/or protein–DNA interactions. Therefore, it is important to identify biomolecules such as proteins, DNA and RNA which are directly involved in disease initiation and to elucidate their interactions at the atomic level for drug discovery studies.

10.1517/17460441.2014.941801 © 2014 Informa UK, Ltd. ISSN 1746-0441, e-ISSN 1746-045X

Article highlights.

- A wide variety of recombinant protein expression methods have been developed for pharmaceutical and structural studies.
- *Escherichia coli* expression system has been most widely used to produce sufficient amount of heterologous proteins, and the development of its new technologies is still continuing.
- Protein expression systems using pCold-glutathione S-transferase and single protein production are one of the most advanced techniques using *E. coli*.
- Cultivation techniques using fed-batch, auto-induction and high cell density are useful and implicative.
- Non-*E. coli* expression systems such as yeast, insect, mammalian cells and cell-free methods have advantages and drawbacks.

This box summarizes key points contained in the article.

The most common biomolecular targets of drugs are proteins due to their functional and structural diversity or importance in many biologically significant events. In an effort to discover new small-molecule therapeutics, structure-based (or assisted) drug development techniques using computational molecular modeling/docking, structural determination by X-ray crystallography, nuclear magnetic resonances (NMRs) and electron microscopic techniques are powerful approaches that can be utilized to enhance 'hit-to-lead' or 'lead fine-tuning' processes.

Many NMR methods suitable for pharmaceutical research have been developed since NMR applications can provide insight into a number of areas related to drug discovery: i) both small drug molecules and proteins can be observed directly by NMR spectroscopy; ii) the tertiary structures of biomolecules and any conformational changes resulting from interactions with drugs can be determined at atomic resolution; and iii) the affinity and binding interface can be determined even if the molecular–molecular interaction is weak. Hence, candidate drugs that may be pharmaceutically immature but can potentially be used as building blocks for drug improvement can be identified by NMR screening [2].

In general, large amounts of protein are required for structural and protein-based pharmaceutical studies. For instance, at least 1 mg of protein is necessary for crystallographic studies [3]. However, it is almost impossible to obtain sufficient amounts of high-purity target proteins from native organic sources or by chemical synthesis. Therefore, 'recombinant' methodologies, which can produce large amounts of recombinant target protein, are indispensable for protein structural biological studies [4]. These methods rely on the use of recombinant DNA, encoding the desired protein, and utilize the molecular machinery present in host cells or organisms to invoke *de novo* protein synthesis following transcription of the desired heterologous gene. However, the production of sufficient amounts of heterologous protein with retained native structure and/or biological function/activity may be problematic, such as in the case of target proteins which are embedded within membranes, or target proteins that play a significant role in various essential biological events such as G-protein-coupled receptors (GPCRs) and kinases.

Since approaches concerning the selection and experimental purification of potential druggable protein targets in drug discovery/development have been discussed in some excellent previous reviews [5], in this article we will outline the current progress of methodologies geared toward the overproduction of recombinant proteins.

Strategies employed for the overproduction of recombinant protein can be roughly divided into two main streams: heterologous proteins are synthesized in living host cells or *in vitro*. Both approaches utilize a heterologous gene of interest and the natural molecular machinery involved in gene transcription/translation such as the ribosome in some living organisms. With the former approach, *Escherichia coli*, yeast, insect or mammalian cells are generally used as living bioreactors. Since each expression system has its own particular advantages and drawbacks (**Table 1**), it is important to select an appropriate expression system based on the physicochemical properties of the target protein of interest, and the purpose for which the target protein will be utilized. In the following sections, the characteristics of each expression system will be presented, and particular attention will be given to the *(E. coli)* expression system.

2. *Escherichia coli* expression system

In general, post-translational modifications of eukaryotic proteins such as glycosylation and phosphorylation are absent when the protein is recombinantly expressed in prokaryotic host cells (**Table 1**). In many cases where it is required that the heterologous protein form complex disulfide bonds, its overexpression in soluble form in the cytoplasm of bacterial host cells is difficult since the cytoplasm is an unsuitable environment for disulfide bond formation due to the highly reduced environment. However, the overexpression of recombinant heterologous protein is generally achieved using bacterial expression systems, and especially those employing the Gram-negative bacterial strain *E. coli*, since prokaryotic host cells offer many practical advantages: easy genetic manipulation, simple cultivation handling, no requirement of special culture rooms or specialized equipment, rapid cell growth and large amount of biomass, resilience and adaptability to a wide range of culture conditions and the availability of cost-effective sophisticated $^{2}H/^{13}C/^{15}N$ stable isotope-labeling technology. Consequently, approximately 30% of recombinant therapeutic products on the market in recent times are manufactured in *E. coli* [6]. Use of an *E. coli* expression system is the first choice when the molecular weight of the desired protein is < 100 kDa and any post-translational modifications of the target protein are not taken to be indispensable for proper structure formation and biological activity.

Table 1. Heterologous protein expression systems.

	Escherichia coli	Yeast	Insect	Mammalian	Cell-free
Gene transformation	Heat-shock Electroporation	Heat-shock Electroporation	Infection of virus	(**Table 2**)	None
Doubling time of host cells	10 – 30 min	1 – 3 h	16 – 24 h	24 h	None
Experimental period*	3 – 7 days	1 – 2 weeks	2 – 4 weeks	2 – 3 days	1 – 2 days
Expression level‡	0.1 mg – 1 g	0.1 mg – 1 g	0.1 – 100 mg	0.01 – 10 mg	1 μg – 1 mg
Total costs	Low	Low	Medium	High	High
Phosphorylation	None	Ser, Thr, Tyr	Ser, Thr, Tyr	Ser, Thr, Tyr	Strain-dependent§
Glycosylation	None	Mannan	High-mannose	complex	Strain-dependent§

*Period representing the time from plasmid construct completion to confirming recombinant protein expression levels using small-scale test cultivation.
‡Total protein expression levels per 1 L cultivation.
§If the lysates used were derived from *Escherichia coli*, insect or mammalian cells, the characteristics of each cell-free system correspond to the expression systems using the respective cells as living hosts.

2.1 Optimization of template mRNA sequence

Many amino acids are encoded by multiple codons, and codon usage frequencies differ between organisms. In several cases, it has been found that the expression level of heterologous protein was extremely low due to the particular codon usage present in the host cells. This situation may result in delayed or premature arrest of target gene transcription at a position comprising a rare codon. In such cases, codon optimization of the cDNA, where the sequence coding the target protein is modified according to the codon usage of the host organism, may be employed to improve the level of target protein expressed [7]. Many computational programs have been developed as web-based services or stand-alone software downloads to assist with codon optimization efforts [7]. An alternative approach is to employ *E. coli* cells which co-express tRNAs that provide several rare codons applicable to the gene of interest. These *E. coli* strains are commercially supplied as Rosetta2 (Novagen, Germany) and CodonPlus (Agilent Technologies, CA, USA). Further, low expression levels of target protein may result if the interaction between the target mRNA and host ribosome is impaired by steric hindrance, and which is dependent on the secondary structure of the mRNA of interest [8]. In such cases, DNA sequencing of the target gene and surrounding regions needs to be checked to determine, for example, whether sequences complementary to the ribosome binding sequence 'AGGAG' are present.

2.2 Solubility-enhancement tags

In many cases, when heterologous proteins are overexpressed in *E. coli*, a large portion of the expressed protein of interest tends to form insoluble aggregates in the cytoplasm referred to as inclusion bodies (IBs). When formed, IBs can be solubilized by treatment with chaotropic denaturants such as urea or guanidinium hydrochloride, and proteins of interest in the IBs can be purified and refolded by eliminating the denaturants. This strategy has several advantages: i) IBs can be easily isolated by centrifugation; ii) cytotoxic proteins can be overexpressed without damage to the host cell by forming IBs; and iii) protein of interest in IBs is protected from protease attack. One potential drawback with this approach is that the protein denaturation–refolding process may be unsuccessful, and subsequent exploration of optimal refolding conditions can be laborious. In an effort to obviate this drawback, polypeptides possessing predominantly high solubility and expression levels can be fused to the terminal end of the heterologous protein to improve solubility and expression levels of the protein of interest. This chimera strategy has been very successful, and chimeric partner polypeptides such as N-utilization substance A, thioredoxin (Trx), maltose-binding protein (MBP), glutathione S-transferase (GST), streptococcal protein G B1 domain (GB1), small ubiquitin modifier (SUMO) and HaloTag are well known as solubility-enhancement tags (SETs) [9-11].

Several SETs such as MBP, GST and HaloTag can also be utilized as affinity tags to facilitate protein purification or pull-down efforts since chimeric partners can bind to specific ligands or resins with high affinity. The HaloTag is particularly useful in the purification of minute amounts of target protein or immobilization of the chimeric protein onto a protein array platform since it can covalently bind to specific ligands. In general, when using the other SETs, affinity tag sequences such as hexahistidine are tandemly positioned upstream or downstream of the SETs [12].

Crystallization of the chimeric protein and X-ray crystal structure analyses are often performed without removal of the affinity tags or SETs [13]. In some cases, the homogeneity of the target protein can be improved by fusing SETs, and crystal growth is promoted by the homogeneous chemical units comprising the SETs, referred to as 'carrier-driven' effects [14]. In some NMR studies, analyses of the NMR spectra of the desired chimeric protein have been performed without elimination of the 'NMR-visible' SETs if the chemical units comprising the SETs do not interfere with analyses of the protein of interest. For example, NMR spectra of several proteins were successfully analyzed without elimination of the 'NMR-visible' GB1 tag as the NMR signals of the GB1 tag caused little signal degeneration [9,15]. Since the molecular weight of GB1 is low, the number of NMR signals

derived from GB1 is few. Further, the NMR signals are well dispersed and have been clearly assigned. In other cases, such as with GST-fused chimeric proteins, since part of the NMR signals of the GST moiety are severely broadened, NMR signals derived from the protein of interest can be exclusively detected [16]. Therefore, in the case of NMR measurements, the decision to remove the SET from the protein of interest prior to NMR analyses should be made based on the character of the target protein, the purposes for which the target protein is to be analyzed, and the general conditions of the experiments [17]. When proteins of interest possessing low thermostability or a marked propensity for aggregation are expressed as fusions with SETs in the soluble fraction of host cells, although the SETs can be correctly folded, the proteins of interest may remain partially unfolded and subsequently form insoluble aggregates when the SETs are removed by digestion [4]. This is problematic since NMR analyses ideally require that the protein of interest possess high thermodynamic stability and solubility [18]. In such cases, NMR measurements may be performed by utilizing a segmental $^{2}H/^{13}C/^{15}N$ isotope-labeling technique, whereby an isotopically unlabeled 'shadow SET' is fused to an isotope-enriched protein of interest, so that NMR signals derived only from the protein of interest are detected, in the absence of aggregation [19-21].

Since the cytoplasm of living cells is maintained in a highly reduced environment by redox regulation molecules, the overexpression of heterologous proteins that require complex disulfide bond formation and folding often leads to the formation of IBs or protein degradation. Since the reducing environment of the periplasmic space of *E. coli* is lower than that of the cytoplasm, heterologous proteins of interest have been made that are translocated from the cytoplasm to the periplasm by fusion with periplasmic signal sequences such as MBP, DsbA or DsbC [22].

As an alternative approach, the use of *E. coli* strains possessing loss-of-function mutations affecting Trx reductase B (trxB) or glutathione reductase (gor) such as the Origami series (Novagen, Germany) has been effective. Further, the SHuffle strain (commercially supplied by New England Biolabs, MA, USA) can co-express DsbC, protein disulfide isomerase (PDI) of *E. coli*, with a protein of interest in addition to possessing functional losses of trxB and gor. Use of the SHuffle strain has the potential to markedly improve the yield of heterologous proteins which require complex disulfide bonding since the reducing environment of the cytoplasm in these *E. coli* strains is markedly lowered [23]. Co-expression of other recombinant PDIs with DsbC can further promote proper folding of the heterologous protein and lead to the prevention of aggregation [24].

2.3 Strategies of peptide production

Structural biological analyses of peptides which play a role in the competitive inhibition of protein–protein interactions and the subsequent development of these small molecules as therapeutic agents are one valuable strategy [25]. This strategy has been employed as a result of insights gleaned from structural investigations. Although the therapeutic peptide itself can act as a pharmaceutical, since there are general concerns regarding peptide stability and absorption efficiency, it is important to utilize the structural insights gained from the therapeutic peptide to develop small therapeutic molecules with improved stability and absorption efficiency [26]. In several cases, overexpression of recombinant peptide can be problematic when the peptide of interest possesses many hydrophobic or aromatic amino acid residues, and it plays a biologically significant role in association with receptor protein or lipid molecules. Additionally, peptide structures are generally highly flexible, due to their short lengths, and are prone to degradation following overexpression [27]. In such cases, the ketosteroid isomerase (KSI)-peptide tandem fusion system can be employed and is commercially supplied as the pET31b(+) vector by Novagen [28]. With this strategy, a cDNA cassette encoding a tandemly repeated (usually 3 – 6 copies) target peptide is inserted downstream to the KSI cDNA in the pET31b(+) plasmid. The KSI polypeptide can be highly expressed in the cytoplasm of *E. coli* host cells and forms IBs, which protect the polypeptide of interest from protease attack. The IBs can be sufficiently purified by washing and the polypeptide of interest can then be highly purified using a simple procedure as described in the previous section. This strategy is also useful in cases where the peptide of interest can be cytotoxic to host cells, in that cytotoxicity can be avoided through the formation of IBs [29]. Further, since one methionine residue has been artificially incorporated between each concatenated peptide of interest, large amounts of the desired peptide can be obtained following one-time expression by subjecting the isolated sample to cyanogen bromide (CNBr) digestion. Clearly, this approach is unsuitable in cases where the target peptides possess methionine residues. Additionally, it should be noted that one artificial homoserine lactone moiety is generated by CNBr digestion and is located at the C terminus of the peptide of interest.

Ubiquitin and GB1 have been widely used as SETs when overexpressing peptides in soluble form in the cytoplasm of *E. coli* [12,30]. The ubiquitin tag fused on the N terminus of target polypeptides can be efficiently and cleanly cleaved by ubiquitin hydrolase, leaving the resultant polypeptide of interest intact and unmodified. This is an important point since, if cleavage of the SET generates a modified polypeptide, this could affect tertiary structure formation and/or biological function, a situation more likely to occur with shorter polypeptide sequences. Hence, employment of the ubiquitin tag and specific hydrolase system is useful in effecting the overexpression of short peptides. Recombinant ubiquitin hydrolase is also highly expressed and easily purified by the *E. coli* expression system. In a similar approach, a small ubiquitin-like modifier (SUMO)-tag and SUMO protease system have been made commercially available.

Additionally, MBP and Trx can be used as SETs for peptide expression. MBP can be successfully overexpressed in either the cytoplasm or periplasm using the pMAL vector series (commercially supplied by New England Biolabs, MA, USA), which provide the option of placing the periplasmic translocation signal sequence on the N terminus of MBP. In cases where the target polypeptide is an intrinsically disordered protein, such as kinase-inducible domain of the transcription factor c-AMP-response-element-binding protein, or where the target peptide is prone to degradation, the fusion of MBP containing a periplasmic translocation signal sequence may reduce the risk of nonspecific degradation of the target polypeptide given the relatively small number of proteases present in the periplasm [4].

2.4 Advanced techniques for production of challenging proteins by *E. coli* expression systems

The most commonly used technology for recombinant protein expression using living *E. coli* host cells is the pET vector system, which is commercially available from Novagen, Germany (**Figure 1A**). With that system, heterologous pET plasmid DNA possessing the T7 promoter and gene of interest is genetically introduced into *E. coli* host cells. Typical *E. coli* host cells for heterologous gene expression, such as the BL21(DE3) strain, possess the *lac*UV promoter-T7 RNA polymerase gene cassette within their genome, derived from λ bacteriophage DE3. During normal culture, activity of the *lac*UV promoter is essentially suppressed by the lac repressor. This suppression can be removed by addition of isopropyl β-D-1-thiogalactopyranoside (IPTG), a non-hydrolyzable lactose analog, which then leads to expression of T7 RNA polymerase (**Figure 1A**). However, since the activity of the T7 promoter is strong and suppression of the *lac*UV promoter by the lac repressor is less than complete, use of the pET system may lead to a number of undesirable outcomes during recombinant heterologous protein expression: i) large amounts of quickly overexpressed heterologous protein can easily form IBs containing incompletely folded heterologous proteins given the shortage of intrinsic molecular chaperones that are unable to cope with the high-speed heterologous protein synthesis or protein folding [31]; and ii) leaky expression of T7 RNA polymerase can occur even in the absence of IPTG, which also leads to leaky expression of heterologous protein. Consequently, especially in the case of target proteins that are cytotoxic to the host cell, transformants may be difficult to obtain since the recombinant gene becomes unstable in the host cell, and cell growth or heterologous protein expression may be significantly reduced even if transformants can be obtained.

2.4.1 *General strategies to minimize IB formation*

If the majority of the expressed heterologous protein forms insoluble aggregates, lowering the cultivation temperature to 20 – 30°C and/or replacing the culture medium with a nutritionally poor substitute can improve the solubility of target proteins, since gene transcription or translation is attenuated under these conditions, thereby permitting appropriate folding of the target protein [4]. Additionally, heterologous protein expression levels and intracellular accumulation can be controlled to some extent by altering the amount of IPTG employed when using the BL21(DE3) strain, and greater quantitative and stricter control is possible when using the BL21(DE3) Tuner strain, which is commercially available from Novagen [32]. Further, addition of a small amount of stimulant into the culture medium, such as 3% [v/v] ethanol or 1 μg/ml antibiotics which inhibit protein synthesis in *E. coli* (e.g., kanamycin, chloramphenicol, streptomycin, tetracycline, puromycin, etc.), can improve heterologous protein expression levels in the soluble fraction since the expression of intrinsic molecular chaperones can be induced by shock responses [33,34].

The pCold vector system is an advanced *E. coli* heterologous protein expression system, commercially available from Takara, Japan. Low-temperature cultivation represents one of the most effective means of improving the solubility and expression level of heterologous proteins in *E. coli* host cells [35]. When the cultivation temperature is quickly reduced to 15°C, the CspA promoter is strongly activated in concert with its cold shock response, resulting in the specific and marked expression of a series of cold shock proteins, with CspA being the predominant cold shock protein produced. In parallel, *E. coli* cell growth can be temporary arrested and *de novo* protein synthesis significantly suppressed, except for CspA, by low-temperature antibiotic truncated CspA effect [36]. With this approach, heterologous protein can be specifically overexpressed by placing the gene of interest under control of the CspA promoter in the pCold vector and cultivating the *E. coli* transformants under low-temperature conditions (**Figure 1B**) [35]. The pCold vector system can be used with *E. coli* host strains that do not possess the λ*DE3* gene within their genome, which encodes the *lac*UV promoter sequence and T7 RNA polymerase. When the expression level of the protein of interest is insufficient when using the pCold vector, a short half-life of the target mRNA and/or degradation of the expressed target protein are assumed. In such cases, the decrease in levels of the target mRNA can be prevented by using the BL21 Star strain as the host (commercially available from Life Technologies, CA, USA), since RNaseE activity in this genetically modified strain is attenuated.

2.4.2 *Techniques to suppress leaky expression*

Incomplete suppression of T7 RNA polymerase expression can result due to increased levels of intracellular cAMP, resulting from depletion of glucose from the cultivation medium and subsequent metabolism of sugars other than glucose [37]. By utilizing catabolite repression of the *lac*UV promoter, by adding 0.5 – 1.0% [w/v] of glucose into the agar medium when transforming *E. coli* host cells with recombinant plasmid, leaky expression of T7 RNA polymerase can be suppressed and the stability of the recombinant plasmid in

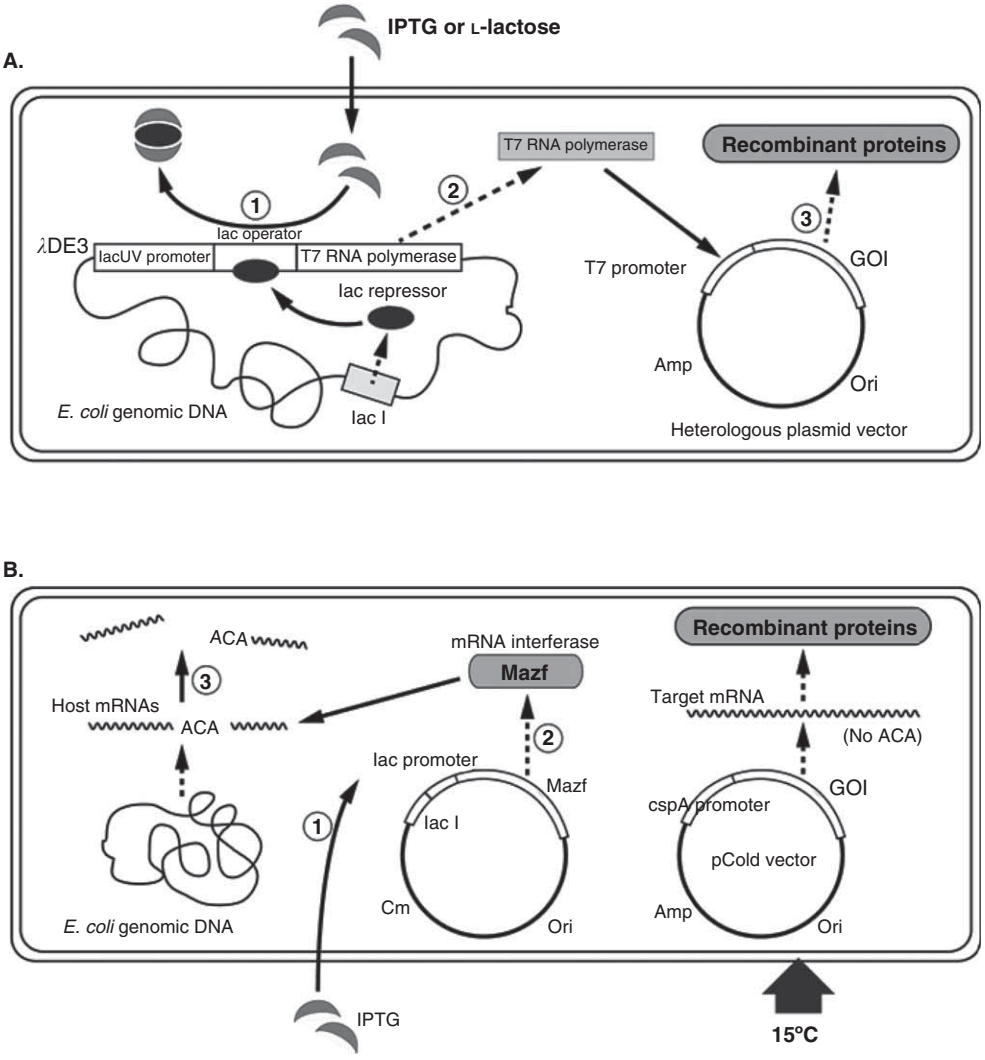

Figure 1. Schematic representations of pET and pCold vector expression systems of *E. coli*. **A.** Overview of pET vector expression system. Expression of heterologous target gene on the vector plasmid is mediated by recombinant T7 RNA polymerase. Expression of recombinant T7 RNA polymerase is suppressed by interaction with lac repressor, which is encoded in lacI and the lac operator regions. When IPTG or L-lactose are added to host *E. coli* cells, lac repressor is removed from the lac operator region following binding with IPTG, and expression of T7 RNA polymerase begins. **B.** Overview of pCold and SPP systems. Expression of heterologous target gene on the pCold vector is under control of the CspA promoter, the activity of which can be enhanced under low temperature conditions (< 15°C). Additionally, in the SPP system, heterologous protein of interest alone can be produced since heterologous mRNA of the protein of interest remains intact following co-expression with MazF, a mRNA interferase which specifically digests the 5′ side of ACA sequences present in mRNAs.

E. coli: *Escherichia coli*; IPTG: Isopropyl β-D-1-thiogalactopyranoside; SPP: Single protein production.

the host cytoplasm can also be improved [37]. Another approach is to suppress the leaky activity of T7 RNA polymerase by co-expression with T7 lysozyme, a native inhibitor of T7 RNA polymerase. BL21 pLysS and pLysE strains possess plasmids that encode T7 lysozyme. Alternatively, leaky expression of T7 RNA polymerase can be effectively suppressed by placing the T7 RNA polymerase gene under control of the araBAD promoter, the basal activity of which is strictly repressed and can be specifically activated by the addition of L-arabinose. One such strain is *E. coli* BL21-AI, commercially available from Life technologies. Hence, use of a special *E. coli* host strain together with glucose-induced catabolite repression is considered to be a very effective strategy for the expression of target proteins. Moreover, maintaining a low copy number of the recombinant plasmid is an effective means of avoiding any cytotoxic effects of the heterologous gene [38]. The pETcoco vector, commercially available from Novagen, Germany, can be maintained at 1 copy

per cell, unlike the case with general pET vectors where the copy number is usually 20 – 25 [38].

Since the natural expression level of most GPCRs is very low, the establishment of experimental protocols to express and purify large amounts of recombinant GPCRs is a significant step in the area of structural biological studies [39]. In many cases, however, it is difficult to determine optimal expression conditions due to the general low expression level of membrane proteins and insufficient biomass of *E. coli* host cells as large differences exist between bacterial and mammalian cells in terms of lipid membrane architectures and the machinery responsible for the insertion of protein into the lipid bilayer. The solubility or expression level of transmembrane, membrane-binding and highly hydrophobic proteins, or heterologous proteins prone to aggregation, can be improved by using *E. coli* host strains such as C41(DE3), C43(DE3) or Lemo21(DE3) strains [40,41]. The C41(DE3) strain was developed from BL21(DE3) by genetic modification. Since this strain has mutations in *lac*UV promoter and other component, expression levels of T7 RNA polymerase are relatively low and thereby expression rate of recombinant protein of interest can be markedly attenuated. Further, C41 (DE3) shows more strong resistance against vital stress which is caused by overexpression and accumulation of recombinant heterologous protein by some unknown effects. Due to these characters, yield of recombinant membrane protein and/or toxic protein can be improved. The C43(DE3) strain is derivative of C41(DE3), and it has more high stress tolerance than that of C41(DE3). The Lemo21(DE3) strain possesses standalone plasmid 'pLemo' which has rhaBAD promoter and follow cDNA coding T7 lysozyme. Unlike in the case of BL21(DE3)pLys strains, expression level of the T7 lysozyme can be finely controlled by adjusting the dose of L-rhamnose, and thereby researchers can optimally coordinate activity of T7 RNA polymerase. As a similar effect of the case of C41(DE3) strain, it will relieve stress led by rapid accumulation of heterologous protein, and often resulting in improvement of final yield of recombinant membrane protein. Additionally, fusing the target protein with Mistic, which is a highly hydrophilic membrane-binding protein of *Bacillus subtilis* comprising 110 amino acid residues, is a powerful strategy to augment the expression levels of target membrane proteins [42].

2.5 pCold GST vector system

When the expression level of target proteins in the soluble fraction is insufficient, even with the pCold vector system, combined use of the pCold vector and SETs has a powerful potential to increase the yield of soluble heterologous protein. By using the pCold-GST vector, which was developed by Hayashi and Kojima, many heterologous proteins that were previously difficult to express or prone to aggregation could be successfully overexpressed in *E. coli* [43]. In addition to GST, Hayashi and Kojima tested a wide variety of SETs, including MBP, GB1 and Trx in an effort to identify an optimal SET which maximizes the power of the pCold system. As a result, it was shown that fusing the GST tag led to the most successful results [44]. In addition to acting as a potent SET, GST has the advantage in that almost all of the NMR signals derived from the GST moiety of the GST-fused protein are severely broadened, so that NMR signals derived from the protein of interest can be clearly detected without the need to eliminate the GST tag, as previously described [44]. At present, in addition to the pCold-GST vector, the pCold ProS2 vector has been developed, which can utilize protein S derived from *Myxococcus xanthus* (ProS2) as a SET for heterologous protein expression by the pCold system [45]. Recently, a new segmental $^{2}H/^{13}C/^{15}N$ isotope-labeling methodology was reported where NMR-invisible ProS2 polypeptides are fused to the NMR-visible protein of interest by utilizing the pCold ProS2 system and intein–extein protein ligation technology [46].

2.6 Single protein production technology

Escherichia coli possesses an intrinsic toxin–antitoxin system that plays a role in the regulation of *E. coli* cell growth and programmed cell death in response to environmental changes. The MazF toxin protein possesses mRNA interferase activity and selectively degrades the 5' moiety of the ACA sequence in mRNA. Since most native mRNA sequences in *E. coli* possess the ACA sequence, these can be digested by MazF, thereby almost completely suppressing protein *de novo* synthesis in *E. coli*. Even under these conditions, basic vital activities of *E. coli* cells such as ATP synthesis and protein expression can continue since molecules required for these processes already exist prior to MazF-mediated mRNA clearance, although *E. coli* cell growth becomes arrested in a 'quasi-dormant state' [47]. With this approach, heterologous target protein can be almost exclusively expressed if any ACA sequences in the target gene of interest can be genetically substituted in advance with other bases so that the target protein amino acid sequence remains unaltered (**Figure 1B**). This system was developed by Inouye and his co-workers as the single protein production (SPP) method [48]. The SPP system possesses several advantages, especially in the area of $^{2}H/^{13}C/^{15}N$ isotope labeling. Isotope label can be selectively incorporated into the protein of interest, and therefore the efficiency of isotope labeling is improved compared to conventional labeling procedures. This minimizes the appearance of NMR signals derived from non-target protein chemical groups, even if the NMR sample contains, to some extent, impurities [49]. Therefore, the SPP system is especially effective in the production of isotope-enriched membrane proteins [50,51], and it is expected that NMR signals of selectively isotope-labeled membrane proteins overexpressed in the lipid membrane of *E. coli* host cells can be directly measured by solid-state NMR techniques using *E. coli* itself as the NMR sample. Recently, a new SPP system was developed that utilizes MazF-bs mRNA interferase, which specifically digests the 5' moiety of the five-base mRNA sequence

UACAU [52], and SPP technologies are on the verge of being able to successfully address a variety of issues yet to be resolved in the area of recombinant protein production.

2.7 Auto-induction

With this approach, heterologous protein expression is automatically induced in the absence of IPTG when the cell density in a culture reaches nearly saturation levels [53]. Auto-induction can be accomplished by inoculating a single colony of transformants into liquid medium containing a mixture of glycerol/D-glucose/α-lactose as a carbon source for *E. coli* and incubating the culture for 20 – 24 h. Under circumstances when the amount of D-glucose becomes limiting, D-glucose is preferentially taken up and this prevents lactose metabolism until D-glucose is depleted. Following this, the uptake of lactose is initiated, which leads to allolactose-mediated release of the lac repressor from the *E. coli* genome and subsequent induction of T7 RNA polymerase overexpression. Timing of the induction, which is critical in achieving sufficient biomass and recombinant protein expression, can be appropriately adjusted by optimizing the amount of D-glucose in the medium. This auto-induction system possesses several advantages over other approaches: cultivation procedures are simpler, reproducibility of protein induction is improved and many protein induction cultivations can be run simultaneously for the screening and selection of transformants with the highest potential to overexpress heterologous protein [53]. Further, this system can be applied to the uniform $^{2}H/^{13}C/^{15}N$ isotope labeling of heterologous protein for NMR studies [54]. One potential drawback in using this system is the increased risk of non specific degradation of the expressed heterologous protein due to the long cultivation periods employed. Additionally, cell growth rates may be attenuated by competition with induced target protein expression, an effect that may be more pronounced when using BL21(DE3) pLys strains which overexpress T7 lysozyme. Similarly, yielding sufficient amounts of biomass and heterologous protein may be more difficult compared to conventional induction systems employing IPTG when the protein of interest possesses cytotoxicity. Prior to large-scale cultivation, cell growth rates, target protein production levels and optimal culture conditions should be checked by preliminary pilot-scale cultivation.

2.8 High cell-density culture

In the case of auto-induction systems as described above, both cell growth and protein production can progress simultaneously. Therefore, yields of biomass or heterologous protein can be improved compared to manual induction methods that utilize IPTG [53]. In many cases, however, since cell growth rates are inversely proportional to target protein expression activity due to limitations in *E. coli* cell physical strength and metabolic economics, it is difficult to simultaneously achieve a large cell biomass and high target protein expression levels [4]. Consequently, in general auto-induction cultivation procedures, the amount of cells utilized is increased as much as possible prior to induction, and the medium composition is adjusted in advance to initiate auto-induction when the cell density becomes saturated [53]. In a similar manner, high cell-density cultivation and induction of protein expression is applicable to conventional manual induction systems. In such cases, a large amount of *E. coli* cells is resuspended in a small volume of fresh medium after transformants were grown by large-scale cultivation, and expression of target heterologous protein is then induced by adding inducer into the high cell-density suspension. It is important that the aforementioned *E. coli* transformants be collected at log-phase cell growth in an effort to maximize their viability and activity. When using the high cell-density induction technique for $^{2}H/^{13}C/^{15}N$ isotope labeling of heterologous protein, the amount of isotope-enriched reagents required can be reduced in accordance with the volume of the culture medium. This improves the cost-effectiveness of isotope-labeling cultivation since most isotope-enriched reagents are expensive. However, from a practical point of view, achieving high isotope incorporation efficiency is not straightforward using the cultivation method and optimal culture conditions should be investigated.

2.9 Fed-batch cultivation

Cell growth and protein expression can be concurrently boosted by the continuous feeding of fresh medium during the induction of protein expression, a technique referred to as 'fed-batch' cultivation [55]. In the case of $^{2}H/^{13}C/^{15}N$ isotope-labeling cultivation that employ poor nutritional conditions supplied by minimal medium, the fed-batch technique offers significant improvements over alternative approaches [56,57]. In general, in addition to fresh medium, acid or base solutions and air are periodically or continuously fed into the culture to maintain an optimal pH range and sufficient aeration, factors which critically affect biomass and protein expression levels. Therefore, the use of jar-fermentation equipment is preferred in an effort to ensure more effective fed-batch cultivation. If fermentation systems are unavailable, using Ultra Yield Flasks™ which have higher aeration efficiency compared to baffled Erlenmeyer flasks may be more effective [58]. Alternatively, if baffled Erlenmeyer flasks are used, good results can be obtained by constantly feeding air and/or fresh nutrients into the cultivation medium by using a peristaltic pump during shaking cultivation [57]. A modified fed-batch method was recently developed, referred to as EnBase technology. In this approach, glucose is encapsulated in sustained-release polymer gels, so that glucose can leak into the medium during cultivation in a slow and constant manner [59]. Presumably, it is possible to combine EnBase technology with the method of Miyazawa-Onami *et al*. [57]. Additionally, physiological and suitable conditions for cell activity can be maintained by continuous substrate feeding since nutrients can be supplied and it dilutes secreted waste, thus enhancing constant cell growth and protein

expression during the entire cultivation period. Recently, several advanced fed-batch techniques have been developed such as concentrated fed-batch (CFB), expanded-bed absorption (EBA) and perfusion systems [60,61]. In the CFB and EBA methods, secreted metabolic wastes were eliminated from the cultivation medium and secreted recombinant heterologous proteins is promptly collected by constantly passing the cultivation medium through an ultrafiltration membrane or chromatographic resin. These techniques allow for the efficient collection of secreted target protein and are especially advantageous in cases where the target proteins are generated in minute amounts and/or are susceptible to degradation. The perfusion system is widely used for recombinant protein production in fragile mammalian cells. In one advanced perfusion technique developed, mammalian host cells are gently immobilized on a biopolymer base with fresh liquid medium surrounding the host cells, referred to as the perfused 3D-cell culture method [62]. In addition to *E. coli*, fed-batch technology is widely utilized with yeast, insect, mammalian and cell-free (CF) expression systems [63,64].

3. Non-*E. coli* bacterial expression systems

3.1 *Lactococcus lactis*

Lactococcus lactis is a Gram-positive bacterium possessing a single-membrane envelope. This bacterium offers advantages over Gram-negative bacteria, such as *E. coli*, for the expression of large amounts of membrane protein since Gram-negative bacteria possess both inner and outer membranes and a periplasmic space with the surface of the inner membrane that is covered by a thick and rigid peptidoglycan layer [65]. Additionally, reports have indicated that the amount of membrane protein expressed in *L. lactis* was further improved by fusing Mistic to the target membrane protein [66]. Incorporation of selenomethionine into the target protein is possible with sufficient efficiency [67]. A commercial *L. lactis* expression system referred to as the NICE system is available from MoBiTec, Germany.

3.2 *Bacillus*

Bacillus also belongs to the Gram-positive group of bacteria. *Brevibacillus choshinensis* and *B. subtilis* expression systems are commercially available from Takara, Japan, and MoBiTec, Germany, respectively. *Brevibacillus* can overexpress and secrete heterologous protein into the culture medium. Therefore, it is suitable for the production of secretory proteins such as cytokines or heterologous proteins that requires complex disulfide bond formation. This system has also been used to generate $^{2}H/^{13}C/^{15}N$ isotope-labeled secreted heterologous proteins for use in solution NMR studies [68,69].

4. Yeast expression systems

Yeast is a eukaryotic microorganism. Yeast cells have the ability to express large amounts of heterologous proteins and can attain a considerably high cell-density state. Additionally, heterologous proteins expressed in yeast can be subjected to post-translational modifications such as phosphorylation and glycosylation. Since yeast host cells combine the superior characteristics of bacterial and higher-order organism host cell expression systems, it is a valuable option when heterologous proteins of interest failed to be successfully expressed in *E. coli* (**Table 1**). Although heterologous proteins can be overexpressed in the cytoplasm, in many cases, strategies are implemented whereby heterologous proteins are overexpressed and then secreted from the yeast host cell. This approach is especially useful for the isolation of proteins that require complex disulfide bonding or post-translational modifications. Further, these secretory expression methods prevent protease digestion of the heterologous protein and allow for improved isolation and purification of the secreted protein compared to approaches that require target protein to be purified from crude cell lysates. The pH of the cultivation medium is critical in the case of secretory expression systems. Although yeast cells can survive in pH environments ranging from 4 to 7, care must be taken not to induce secretion of heterologous proteins in pH environments that may lead to protein denaturation and/or aggregation. Access to a sufficient amount of carbon sources and aeration of the cultivation medium are the most critical points for successful cell growth and heterologous protein expression level [56]. The use of jar-fermentation equipment can maximize aeration efficiency with high-speed agitation and continuous air feeding. However, addition of anti-foam into the fermentation medium is essential, since bubbling generated by the strong stirring may damage the host cells.

Yeast strains generally employed for the overexpression of heterologous proteins include *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Pichia pastoris* and *Kluyveromyces lactis*. The methylotrophic yeast strain *P. pastoris* is commonly used given the strong activity of the *AOX* promoter. When a gene of interest is placed downstream of the *AOX* promoter, the expression of heterologous protein can be strongly induced by adding methanol to the cultivation medium.

In the case of *K. lactis*, heterologous protein expression is regulated by the *LAC4* promoter, the activity of which can be activated by adding galactose to the cultivation medium. Therefore, heterologous protein expression and cell growth progress simultaneously and continuously by culturing the transformants in medium containing galactose as a carbon source. This auto-induction cultivation of yeast host cells has advantages similar to that of *E. coli* expression systems described above.

Human serum albumin, MBP, GST and SUMO have been used as fusion partners of heterologous proteins in yeast expression systems [70-72]. MBP can be utilized as a SET in either cytoplasmic or secretory expression protocols since cysteine residues are absent within MBP.

Many techniques for the $^{2}H/^{13}C/^{15}N$ isotope labeling of heterologous proteins using *P. pastoris* and *K. lactis* have been developed [57,73,74].

5. Insect cells

Insect cell Sf9 or High-Five strains are generally used, and many examples have been reported detailing the successful overexpression of heterologous proteins such as membrane proteins, large molecular weight proteins and protein kinases, which are generally difficult to express using bacterial expression systems [39].

In most insect cell expression systems, the heterologous recombinant gene of interest is transfected into host insect cells using a baculovirus vector, and the cassette containing the target gene of interest is incorporated into the host genome by homologous recombination (**Table 1**). For construction of the appropriate baculovirus vector, various products are commercially available such as Bac-to-Bac (Life Technologies, MA, USA), BaculoGold (BD, NJ, USA), BacPak (Clontech, CA, USA) and BacMagic (Merck Millipore, Germany). The target gene of interest can be more stably maintained by host cells compared to methods that utilize episomal vector systems, and it is a relatively straightforward task to incorporate multiple copies of the target gene or several different genes into the host genome [75]. Therefore, baculovirus-insect cell systems can facilitate co-expression of a multiple number of target proteins. Recently, the MultiBac system was developed, which can strongly assist in the expression of multi-protein complexes in insect host cells [76].

Fundamentally, cost, time and technical skills are required for the preparation of large amounts of recombinant baculovirus particles that possess a sufficiently high titer. Recently, however, an alternative approach has been developed whereby the gene of interest can be introduced directly into insect host cells without the use of baculovirus and heterologous protein transiently overexpressed using the InsectDirect system commercially available from Merck Millipore, Germany.

SETs or soluble fusion partners widely used in insect cell expression systems include GST and SUMO [77,78]. $^{2}H/^{13}C/^{15}N$ isotope labeling of heterologous protein is also possible using insect host cells [79].

6. Mammalian cell expression systems

Mammalian cells may be employed for the expression of heterologous proteins when target protein containing complex disulfide bonding and/or post-translational modifications could not be successfully expressed using other expression systems (**Table 1**). Commonly used mammalian cell lines for heterologous protein expression include HeLa, HEK293 and CHO cells. Various heterologous gene transfection methods used for mammalian cells are summarized in **Table 2**. Selecting the most appropriate transfection method is important for successful heterologous protein production. Additionally, in cases where the protein of interest demonstrates biological activity by forming a complex with other molecules and/or post-translational modifications are required for formation of the molecular complex, the macromolecular complex can be purified intact. These complexes can also be generated when using insect host cells. The tertiary structure of the molecular complex can be determined by X-ray crystallographic or electron microscopic structure analyses. However, it is preferable in some cases to eliminate sugar chains that may be present in the molecular complex by treatment with glycosylases since heterologous sugar chains can interfere with crystallization or the subsequent structural analyses [80]. The major drawbacks in using mammalian host cells include poorer heterologous protein expression compared to other systems, and expense associated with the cultivation, especially when generating $^{2}H/^{13}C/^{15}N$ isotope-labeled heterologous proteins.

SETs that have been used with heterologous protein expression in mammalian host cells include GST, SUMOstar and HaloTag [81-83]. Selenomethionine labeling and isotope labeling of heterologous protein are possible using mammalian expression systems for crystallographic phasing and NMR measurements, respectively [84].

7. CF expression systems

In CF protein synthesis, target polypeptides are translated in a test tube by mixing cDNA or mRNA coding target protein and cell extracts which contain ribosomes and other molecular machinery required for protein biosynthesis. Sources that are used for obtaining cell extracts include *E. coli*, wheat germ, insect cells and mammalian cells (sickle-cell red cells derived from rabbit or CHO). General strategies of mainstream CF expression such as continuous exchange CF, layering and fed-batch methods provide additional fresh reaction components into the reaction mixture and eliminate by-products generated from the reaction mixture in an effort to improve yields [85-87]. In the fed-batch procedure, mRNA coding the protein of interest is continuously supplied to the reaction mixture during the course of the reaction. This is effective in improving yields, especially when the reaction scale or volume becomes large since longer reaction times are required for larger-scale reactions, and this may increase the risk of target mRNA degradation by RNase present in the lysate.

CF protein synthesis systems are useful for the production of cytotoxic proteins, membrane-integrated proteins and bioactive peptides which are generally difficult to overexpress using other expression systems that utilize living host cells. CF protein synthesis systems offer many unique advantages: i) the co-expression of a number of proteins to generate multi-protein complexes is more straightforward than with other expression systems; ii) high-throughput characteristics [88]; and iii) the possibility of full automation [89]. Further, the production of heterologous protein that requires disulfide bond formation can be more successful when using CF expression systems rather than bacterial expression systems since the redox conditions of the lysate can be arbitrarily modified using additives or changing the ratio of reduced/oxidized

Table 2. Characteristics of heterologous gene transfection methods of mammalian cells.

	Lipofection	Electroporation	Retrovirus	Lentivirus	Adenovirus
Total cost	+++	+	++	++	++
Efficiency	++	+	++	++	+++
Reproducibility	++	+	+++	+++	+++
Ease of execution	+++	+++	++	++	+
Cytotoxicity	+	++	++	++	+++
Required time*	+	+	++	++	+++
Expression levels‡	+	+	++	++	+++
Type of cells	Proliferative or non-proliferative§	Proliferative or non-proliferative§	Proliferative	Proliferative or non-proliferative¶	Proliferative or non-proliferative#
Expression types	Transient or stable**	Transient or stable**	Stable	Stable	Transient

*Period representing the time from plasmid construct completion to the completion of heterologous gene transfection into mammalian host cells.

‡Expression levels of heterologous protein.

§When the heterologous gene is transfected into host cells, the gene must translocate into the nucleus in order to work. In principle, therefore, heterologous gene transfection efficiency of the physical methods is greater when using proliferative cells rather than non-proliferative cells. Recently, however, new physical gene transfection methods showing high efficiency with non-proliferative cells are beginning to be developed.

¶Suitable for hematopoietic cells.

#Unsuitable for hematopoietic cells.

**While the heterologous gene resides in the cytoplasm within an episomal plasmid, the gene of interest is transiently expressed. However, stable heterologous gene expression transformants can be generated with small probability by incorporation of the heterologous gene into the host genome, and it can be selected by screening transformants for acquired antibiotic resistance.

reagents of the CF reaction mixture [90]. Incidentally, components of the *E. coli* lysate for the CF reaction can be readily rearranged according to the experimental purposes using PureSystem [91].

One major feature of the CF expression system is that it allows for the precise modification of heterologous proteins by incorporating unnatural or chemically modified amino acids into the desired protein position, possible due to the simple gene translation machinery and low metabolic scrambling characteristics of CF expression systems [92]. Scrambling can be further suppressed by adding an inhibitor of transaminases and amino acid synthases or $NaBH_4$ during protein expression [93]. For protein NMR studies, a wide variety of selective amino acid $^{2}H/^{13}C/^{15}N$ isotope-labeled samples can be prepared using the CF expression system, and combinatorial analyses of the target protein NMR spectra is a valuable technique in the structural biological study of large molecular weight or membrane proteins [94]. Further, a new technique of amino acid and site-specific isotope labeling has been developed by Kainosho and Güntert, referred to as the stereo-array isotope-labeling (SAIL) method, in which chemically and enzymatically synthesized amino acids possessing a unique $^{13}C/^{2}H$ labeling pattern are incorporated into target proteins using the CF expression system [95]. With SAIL technology, molecular weight limitations previously hindering many protein NMR studies have been lifted, and the SAIL approach yields simplified spectra and increased sensitivity of the NMR signals [96-98].

Additionally, CF expression systems are suitable for membrane protein production since the molecular machinery responsible for protein synthesis remains more or less intact if detergents are added to the reaction mixture [99]. Hence, many cases have been reported where solubilized membrane proteins expressed in detergent-containing CF reaction mixtures could be incorporated and reconstituted into lipid bilayers such as liposomes, bicelles and nanodiscs [100,101]. In this process, co-translational spontaneous insertion and/or post-translational reconstitution using an accompanying detergent-lipid exchange procedure are employed. Further, the easy handling and high-throughput characteristics of CF expression systems allow for screening of optimal solubilization conditions of membrane proteins of interest [102].

One drawback of the CF expression system is that it is unsuitable for the production of large amounts of protein, since protein expression levels are low compared to that found in living host cells; moreover, large-scale CF reactions are expensive.

8. Other protein expression systems

A plant cell protein expression system has been developed that utilizes tobacco BY-2 as host cells [103]. Using this system, $^{13}C/^{15}N$ isotope-labeled heterologous protein was generated and the protein tertiary structure was successfully determined by solution NMR [104]. Further, several alternative recombinant protein expression systems such as protozoa or fungi have been developed [105].

9. Expert opinion

Development of genetic-engineering technologies led to technological and conceptional revolution in production of transcriptional products of the gene of interest, that is, exactly protein. Especially in the case where the target molecule is few in a living organism, gathering sufficient amount of the target protein is one of the bottlenecks for protein structural and

biological studies. The breakthrough of recombinant technology emancipated researchers from laboriousness of gathering a large amount of protein of interest, and resulted in drastic progress of life science studies of proteins. However, empirically in many cases, it is not easy to overexpress heterologous proteins of which plays a significant role in vital activity. Therefore, development of methodology of recombinant protein overexpression system is still widely in progress in an energetic way. In this review article, we mainly outlined current protein production technologies, as many excellent reviews have already dealt with protein purification techniques. For successful drug discovery, the preparation of sufficient amounts of recombinant protein of interest possessing native structure and biological function is an important step. Needless to say, selection of an appropriate expression system can significantly improve the results of a drug discovery study. Researchers should select the best expression system based on practical realities such as cost and availability of the desired expression system. Additionally, it is important to establish assays to assess the biological functional activity of the recombinant proteins in advance. It is because the generated recombinant protein can be utilized for drug discovery studies after native structure and biological functions have been confirmed for recombinant protein using the previously established assays.

For structural biological studies, it is imperative that an expression system be chosen that can generate $^{2}H/^{13}C/^{15}N$ isotope- or selenomethionine-labeled recombinant proteins. In addition to the development of protein expression system, therefore, it is also important to promote coinstantaneous development of innovative, flexible and wide variety of labeling technologies, which can be utilized in various protein expression systems. It will not only broaden the variety and possibility of protein structural and biological studies, but it will also enable many previous hard-to-express proteins to be analyzed in atomic level.

One of the frontline research fields, which require simultaneous developments of protein expression system and protein isotope labeling, is in-cell NMR. Structural biological analyses of intermolecular interactions or protein structure determination in living cells for drug discovery and development studies are gradually being realized with improvements in in-cell NMR technology [106].

The availability of various expression systems has facilitated a plethora of molecular biological experiments using established manuals and experimental kits. However, the importance of understanding the theory and principles that underlie these experiments is critical in successfully utilizing recombinant protein expression for drug discovery studies. Since the statements made in this review are general, optimization of expression conditions may be required according to the characteristics of the target protein being examined. Recombinant protein expression technologies are rapidly advancing at present. It is important that efforts should continue to overcome present obstacles in the expression and purification of difficult but pharmaceutically significant proteins by constantly surveying the latest information concerning the experimental progress of drug discovery studies.

Acknowledgements

The authors are grateful to Drs Hideo Takahashi, Hidekazu Hiroaki, Takahisa Ikegami, Kiyoshi Ozawa, and Masayori Inouye for many useful discussions.

Declaration of interest

This work was supported by Platform for Drug Discovery, Informatics, and Structural Life Science from the Ministry of Education, Culture, Sports, Science and Technology, Japan. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

Bibliography

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Hughes JP, Rees S, Kalindjian SB, et al. Principles of early drug discovery. Br J Pharmacol 2011;162(6):1239-49
2. Hiroaki H. Recent applications of isotopic labeling for protein NMR in drug discovery. Expert Opin Drug Discov 8(5):523-36

• **A comprehensive review of modern methodology of protein isotope labeling.**

3. Canduri F, de Azevedo WF. Protein crystallography in drug discovery. Curr Drug Targets 2008;9(12):1048-53
4. Overton TW. Recombinant protein production in bacterial hosts. Drug Discov Today 2013;19(5):590-601

•• **Many interesting discussions about *Escherichia coli* expression system.**

5. Pina AS, Lowe CR, Roque AC. Challenges and opportunities in the purification of recombinant tagged proteins. Biotechnol Adv 2014;32(2):366-81
6. Ferrer-Miralles N, Domingo-Espín J, Corchero JL, et al. Microbial factories for recombinant pharmaceuticals. Microb Cell Fact 2009;8:17
7. Burgess-Brown NA, Sharma S, Sobott F, et al. Codon optimization can improve expression of human genes in Escherichia coli: a multi-gene study. Protein Expr Purif 2008;59(1):94-102
8. Tessier LH, Sondermeyer P, Faure T, et al. The influence of mRNA primary and secondary structure on human IFN-gamma gene expression in Escherichia coli. Nucleic Acids Res 1984;12(20):7663-75
9. Zhou P, Wagner G. Overcoming the solubility limit with solubility-enhancement tags: successful applications in biomolecular NMR studies. J Biomol NMR 2010;46(1):23-31

•• **A comprehensive review of solubility-enhancement tags (SETs).**

10. Young CL, Britton ZT, Robinson AS. Recombinant protein expression and purification: a comprehensive review of affinity tags and microbial applications. Biotechnol J 2012;7(5):620-34
11. Gopal GJ, Kumar A. Strategies for the production of recombinant protein in Escherichia coli. Protein J 2013;32(6):419-25
12. Kohno T, Kusunoki H, Sato K, Wakamatsu K. A new general method for the biosynthesis of stable isotope-enriched peptides using a decahistidine-tagged ubiquitin fusion system: an application to the production of mastoparan-X uniformly enriched with 15N and 15N/13C. J Biomol NMR 1998;12(1):109-21

•• **Ubiquitin-fusion system.**

13. Smyth DR, Mrozkiewicz MK, McGrath WJ, et al. Crystal structures of fusion proteins with large-affinity tags. Protein Sci 2003;12(7):1313-22
14. Corsini L, Hothorn M, Scheffzek K, et al. Thioredoxin as a fusion tag for carrier-driven crystallization. Protein Sci 2008;17(12):2070-9
15. Hiller S, Kohl A, Fiorito F, et al. NMR structure of the apoptosis- and inflammation-related NALP1 pyrin domain. Structure 2003;11(10):1199-205
16. Liew CK, Gamsjaeger R, Mansfield RE, et al. NMR spectroscopy as a tool for the rapid assessment of the conformation of GST-fusion proteins. Protein Sci 2008;17(9):1630-5
17. Chen X, Zaro JL, Shen WC. Fusion protein linkers: property, design and functionality. Adv Drug Deliv Rev 2013;65(10):1357-69
18. Fadel V, Bettendorff P, Herrmann T, et al. Automated NMR structure determination and disulfide bond identification of the myotoxin crotamine from Crotalus durissus terrificus. Toxicon 2005;46(7):759-67
19. Durst FG, Ou HD, Löhr F, et al. The better tag remains unseen. J Am Chem Soc 2008;130(45):14932-3
20. Kobashigawa Y, Kumeta H, Ogura K, et al. Attachment of an NMR-invisible solubility enhancement tag using a sortase-mediated protein ligation method. J Biomol NMR 2009;43(3):145-50
21. Xue J, Burz DS, Shekhtman A. Segmental labeling to study multidomain proteins. Adv Exp Med Biol 2012;992:17-33
22. Matos CF, Branston SD, Albiniak A, et al. High-yield export of a native heterologous protein to the periplasm by the tat translocation pathway in Escherichia coli. Biotechnol Bioeng 2012;109(10):2533-42
23. Lobstein J, Emrich CA, Jeans C, et al. SHuffle, a novel Escherichia coli protein expression strain capable of correctly folding disulfide bonded proteins in its cytoplasm. Microb Cell Fact 2012;11:56
24. Nguyen VD, Hatahet F, Salo KE, et al. Pre-expression of a sulfhydryl oxidase significantly increases the yields of eukaryotic disulfide bond containing proteins expressed in the cytoplasm of E. coli. Microb Cell Fact 2011;10:1
25. Jamieson A, Boutard N, Sabatino D, et al. Peptide scanning for studying structure-activity relationships in drug discovery. Chem Biol Drug Des 2013;81(1):148-65
26. Adessi C, Soto C. Converting a peptide into a drug: strategies to improve stability and bioavailability. Curr Med Chem 2002;9(9):963-78
27. dos Santos Cabrera MP, de Souza BM, Fontana R, et al. Conformation and lytic activity of eumenine mastoparan: a new antimicrobial peptide from wasp venom. J Pept Res 2004;64(3):95-103
28. Kuliopulos A, Walsh C. Production, purification, and cleavage of tandem repeats of recombinant peptides. J Am Chem Soc 1994;116(11):4599-607
29. Opella SJ, Ma C, Marassi FM. Nuclear magnetic resonance of membrane-associated peptides and proteins. Methods Enzymol 2001;339:285-313
30. Koenig BW, Rogowski M, Louis JM. A rapid method to attain isotope labeled small soluble peptides for NMR studies. J Biomol NMR 2003;26(3):193-202

• **GB1-fusion system.**

31. de Marco A. Recombinant polypeptide production in Escherichia coli: towards a rational approach to improve the yields of functional proteins. Microb Cell Fact 2013;12:101

• **A comprehensive review of *E. coli* expression system.**

32. Lebendiker M, Danieli T. Production of prone-to-aggregate proteins. FEBS Lett 2014;588(2):236-46
33. VanBogelen RA, Neidhardt FC. Ribosomes as sensors of heat and cold shock in Escherichia coli. Proc Natl Acad Sci USA 1990;87(15):5589-93

34. Thomas JG, Baneyx F. Divergent effects of chaperone overexpression and ethanol supplementation on inclusion body formation in recombinant Escherichia coli. Protein Expr Purif 1997;11(3):289-96

35. Qing G, Ma LC, Khorchid A, et al. Cold-shock induced high-yield protein production in Escherichia coli. Nat Biotechnol 2004;22(7):877-82

36. Jiang W, Fang L, Inouye M. Complete growth inhibition of Escherichia coli by ribosome trapping with truncated cspA mRNA at low temperature. Genes Cells 1996;1(11):965-76

37. Grossman TH, Kawasaki ES, Punreddy SR, et al. Spontaneous cAMP-dependent derepression of gene expression in stationary phase plays a role in recombinant expression instability. Gene 1998;209(1-2):95-103

38. Wild J, Szybalski W. Copy-control tightly regulated expression vectors based on pBAC/oriV. Methods Mol Biol 2004;267:155-67

39. Maeda S, Schertler GF. Production of GPCR and GPCR complexes for structure determination. Curr Opin Struct Biol 2013;23(3):381-92

•• **A front edge review of expression of recombinant G-protein-coupled receptor (GPCR).**

40. Wagner S, Klepsch MM, Schlegel S, et al. Tuning Escherichia coli for membrane protein overexpression. Proc Natl Acad Sci USA 2008;105(38):14371-6

41. Schlegel S, Löfblom J, Lee C, et al. Optimizing membrane protein overexpression in the Escherichia coli strain Lemo21(DE3). J Mol Biol 2012;423(4):648-59

42. Roosild TP, Greenwald J, Vega M, et al. NMR structure of Mistic, a membrane-integrating protein for membrane protein expression. Science 2005;307(5713):1317-21

43. Hayashi K. Kojima C. pCold-GST vector: a novel cold-shock vector containing GST tag for soluble protein production. Protein Expr Purif 2008;62(1):120-7

•• **The original article of pCold-glutathione S-transferase system.**

44. Hayashi K, Kojima C. Efficient protein production method for NMR using soluble protein tags with cold shock expression vector. J Biomol NMR 48 (3):147-55

• **This article studied the outcome of other SETs for partner of pCold system.**

45. Kobayashi H, Yoshida T, Inouye M. Significant enhanced expression and solubility of human proteins in Escherichia coli by fusion with protein S from Myxococcus xanthus. Appl Environ Microbiol 2009;75(16):5356-62

46. Kobayashi H, Swapna GV, Wu KP, et al. Segmental isotope labeling of proteins for NMR structural study using a protein S tag for higher expression and solubility. J Biomol NMR 2012;52(4):303-13

47. Suzuki M, Roy R, Zheng H, et al. Bacterial bioreactors for high yield production of recombinant protein. J Biol Chem 2006;281(49):37559-65

48. Suzuki M, Mao L, Inouye M. Single protein production (SPP) system in Escherichia coli. Nat Protoc 2007;2(7):1802-10

•• **Detailed protocol of single protein production (SPP) system.**

49. Schneider WM, Inouye M, Montelione GT, et al. Independently inducible system of gene expression for condensed single protein production (cSPP) suitable for high efficiency isotope enrichment. J Struct Funct Genomics 2009;10(3):219-25

50. Mao L, Vaiphei ST, Shimazu T, et al. The Escherichia coli single protein production system for production and structural analysis of membrane proteins. J Struct Funct Genomics 2010;11(1):81-4

• **A first paper of isotope labeling of membrane proteins using SPP system.**

51. Vaiphei ST, Tang Y, Montelione GT, et al. The use of the condensed single protein production system for isotope-labeled outer membrane proteins, OmpA and OmpX in Escherichia coli. Mol Biotechnol 2011;47(3):205-10

52. Ishida Y, Park JH, Mao L, et al. Replacement of all arginine residues with canavanine in MazF-bs mRNA interferase changes its specificity. J Biol Chem 2013;288(11):7564-71

53. Studier FW. Protein production by auto-induction in high density shaking cultures. Protein Expr Purif 2005;41(1):207-34

•• **A comprehensive article of auto-induction method using *E. coli*.**

54. Tyler RC, Sreenath HK, Singh S, et al. Auto-induction medium for the production of [U-15N]- and [U-13C, U-15N]-labeled proteins for NMR screening and structure determination. Protein Expr Purif 2005;40(2):268-78

55. Babaeipour V, Shojaosadati SA, Khalilzadeh R, et al. A proposed feeding strategy for the overproduction of recombinant proteins in Escherichia coli. Biotechnol Appl Biochem 2008;49(Pt 2):141-7

56. Sugiki T, Ichikawa O, Miyazawa-Onami M, et al. Isotopic labeling of heterologous proteins in the yeast Pichia pastoris and Kluyveromyces lactis. Methods Mol Biol 2012;831:19-36

•• **A practical article for yeast expression systems.**

57. Miyazawa-Onami M, Takeuchi K, Takano T, et al. Perdeuteration and methyl-selective (1)H, (13)C-labeling by using a Kluyveromyces lactis expression system. J Biomol NMR 2013;57(3):297-304

•• **Unique and useful fed-batch cultivation of yeast.**

58. Ukkonen K, Vasala A, Ojamo H, et al. High-yield production of biologically active recombinant protein in shake flask culture by combination of enzyme-based glucose delivery and increased oxygen transfer. Microb Cell Fact 2011;10:107

59. Panula-Perälä J, Siurkus J, Vasala A, et al. Enzyme controlled glucose auto-delivery for high cell density cultivations in microplates and shake flasks. Microb Cell Fact 2008;7:31

60. de Lamotte F. Single step purification of a series of wheat recombinant proteins with expanded bed absorption chromatography. J Chromatogr B Analyt Technol Biomed Life Sci 2005;818(1):29-33

61. Yang WC, Lu J, Kwiatkowski C, et al. Perfusion seed cultures improve biopharmaceutical fed-batch production capacity and product quality. Biotechnol Prog 2014;30(3):616-25

62. Kubo S, Nishida N, Udagawa Y, et al. A gel-encapsulated bioreactor system for NMR studies of protein-protein interactions in living mammalian cells.

Angew Chem Int Ed Engl 2013;52(4):1208-11

63. Sugiki T, Shimada I, Takahashi H. Stable isotope labeling of protein by Kluyveromyces lactis for NMR study. J Biomol NMR 2008;42(3):159-62
• **An original article regarding yeast *Kluyveromyces lactis* expression system.**

64. Sato Y, Aizawa K, Ezure T, et al. A simple fed-batch method for transcription and insect cell-free translation. J Biosci Bioeng 2012;114(6):677-9

65. Chen R. Bacterial expression systems for recombinant protein production: Escherichia coli and beyond. Biotechnol Adv 2012;30(5):1102-7

66. Xu Y, Kong J, Kong W. Improved membrane protein expression in Lactococcus lactis by fusion to Mistic. Microbiology 2013;159(Pt 6):1002-9

67. Berntsson RP, Alia Oktaviani N, Fusetti F, et al. Selenomethionine incorporation in proteins expressed in Lactococcus lactis. Protein Sci 2009;18(5):1121-7

68. Tanio M, Tanaka T, Kohno T. 15N isotope labeling of a protein secreted by Brevibacillus choshinensis for NMR study. Anal Biochem 2008;373(1):164-6
• ***Brevibacillus* expression system.**

69. Tanio M, Tanaka R, Tanaka T, et al. Amino acid-selective isotope labeling of proteins for nuclear magnetic resonance study: proteins secreted by Brevibacillus choshinensis. Anal Biochem 2009;386(2):156-60

70. Mitchell DA, Marshall TK, Deschenes RJ. Vectors for the inducible overexpression of glutathione S-transferase fusion proteins in yeast. Yeast 1993;9(7):715-22

71. Marblestone JG, Edavettal SC, Lim Y, et al. Comparison of SUMO fusion technology with traditional gene fusion systems: enhanced expression and solubility with SUMO. Protein Sci 2006;15(1):182-9

72. Wu M, Liu W, Yang G, et al. Engineering of a Pichia pastoris expression system for high-level secretion of HSA/GH fusion protein. Appl Biochem Biotechnol 2014;172:2400-11

73. Chen CY, Cheng CH, Chen YC, et al. Preparation of amino-acid-type selective isotope labeling of protein expressed in Pichia pastoris. Proteins 2006;62(1):279-87

74. Takahashi H, Shimada I. Production of isotopically labeled heterologous proteins in non-Escherichia coli prokaryotic and eukaryotic cells. J Biomol NMR 2010;46(1):3-10
•• **A comprehensive review of non-*E. coli* expression system.**

75. Cremer H, Bechtold I, Mahnke M, et al. Efficient processes for protein expression using recombinant baculovirus particles. Methods Mol Biol 2014;1104:395-417

76. Bieniossek C, Imasaki T, Takagi Y, et al. MultiBac: expanding the research toolbox for multiprotein complexes. Trends Biochem Sci 2012;37(2):49-57

77. Beekman JM, Cooney AJ, Elliston JF, et al. A rapid one-step method to purify baculovirus-expressed human estrogen receptor to be used in the analysis of the oxytocin promoter. Gene 1994;146(2):285-9

78. Liu L, Spurrier J, Butt TR, et al. Enhanced protein expression in the baculovirus/insect cell system using engineered SUMO fusions. Protein Expr Purif 2008;62(1):21-8

79. Kofuku Y, Ueda T, Okude J, et al. Efficacy of the beta2-adrenergic receptor is determined by conformational equilibrium in the transmembrane region. Nat Commun 2012;3:1045
• **Expression and methionine methyl group-selective isotope labeling of recombinant GPCR by baculovirus-insect cell expression system.**

80. Aricescu AR, Owens RJ. Expression of recombinant glycoproteins in mammalian cells: towards an integrative approach to structural biology. Curr Opin Struct Biol 2013;23(3):345-56

81. Rudert F, Visser E, Gradl G, et al. pLEF, a novel vector for expression of glutathione S-transferase fusion proteins in mammalian cells. Gene 1996;169(2):281-2

82. Peroutka RJ, Elshourbagy N, Piech T, et al. Enhanced protein expression in mammalian cells using engineered SUMO fusions: secreted phospholipase A2. Protein Sci 2008;17(9):1586-95

83. Ohana RF, Hurst R, Vidugiriene J, et al. HaloTag-based purification of functional human kinases from mammalian cells. Protein Expr Purif 2011;76(2):154-64

84. Sastry M, Bewley CA, Kwong PD. Mammalian expression of isotopically labeled proteins for NMR spectroscopy. Adv Exp Med Biol 2012;992:197-211
• **A comprehensive article of protein overexpression and isotope labeling using mammalian host cells.**

85. Kigawa T, Yabuki T, Yoshida Y, et al. Cell-free production and stable-isotope labeling of milligram quantities of proteins. FEBS Lett 1999;442(1):15-19
• **A practically instructive article of cell-free (CF) expression system.**

86. Aoki M, Matsuda T, Tomo Y, et al. Automated system for high-throughput protein production using the dialysis cell-free method. Protein Expr Purif 2009;68(2):128-36

87. Takai K, Sawasaki T, Endo Y. Practical cell-free protein synthesis system using purified wheat embryos. Nat Protoc 2010;5(2):227-38

88. Sawasaki T, Hasegawa Y, Tsuchimochi M, et al. A bilayer cell-free protein synthesis system for high-throughput screening of gene products. FEBS Lett 2002;514(1):102-5

89. Beebe ET, Makino S, Nozawa A, et al. Robotic large-scale application of wheat cell-free translation to structural studies including membrane proteins. N Biotechnol 2011;28(3):239-49

90. Michel E, Wüthrich K. Cell-free expression of disulfide-containing eukaryotic proteins for structural biology. FEBS J 2012;279(17):3176-84

91. Shimizu Y, Kuruma Y, Kanamori T, et al. The PURE system for protein production. Methods Mol Biol 2014;1118:275-84
• **PURE system.**

92. Yokoyama J, Matsuda T, Koshiba S, et al. A practical method for cell-free protein synthesis to avoid stable isotope scrambling and dilution. Anal Biochem 2011;411(2):223-9

93. Su XC, Loh CT, Qi R, et al. Suppression of isotope scrambling in cell-free protein synthesis by broadband inhibition of PLP enymes for selective 15N-labelling and production of perdeuterated proteins in H_2O. J Biomol NMR 2011;50(1):35-42

94. Hohsaka T, Muranaka N, Komiyama C, et al. Position-specific incorporation of dansylated non-natural amino acids into

streptavidin by using a four-base codon. FEBS Lett 2004;560(1-3):173-7

- **Unnatural amino acid incorporation technology using CF expression system.**

95. Kainosho M, Güntert P. SAIL–stereo-array isotope labeling. Q Rev Biophys 2009;42(4):247-300
96. Kainosho M, Torizawa T, Iwashita Y, et al. Optimal isotope labelling for NMR protein structure determinations. Nature 2006;440(7080):52-7

- **Stereo-array isotope labeling technology.**

97. Ikeya T, Takeda M, Yoshida H, et al. Automated NMR structure determination of stereo-array isotope labeled ubiquitin from minimal sets of spectra using the SAIL-FLYA system. J Biomol NMR 2009;44(4):261-72
98. Miyanoiri Y, Takeda M, Kainosho M. Stereo-array isotope labeling method for studying protein structure and dynamics. Adv Exp Med Biol 2012;992:83-93
99. Reckel S, Sobhanifar S, Durst F, et al. Strategies for the cell-free expression of membrane proteins. Methods Mol Biol 2010;607:187-212
100. Glück JM, Wittlich M, Feuerstein S, et al. Integral membrane proteins in nanodiscs can be studied by solution NMR spectroscopy. J Am Chem Soc 2009;131(34):12060-1
101. Mazhab-Jafari MT, Marshall CB, Stathopulos PB, et al. Membrane-dependent modulation of the mTOR activator Rheb: NMR observations of a GTPase tethered to a lipid-bilayer nanodisc. J Am Chem Soc 2013;135(9):3367-70
102. Isaksson L, Enberg J, Neutze R, et al. Expression screening of membrane proteins with cell-free protein synthesis. Protein Expr Purif 2012;82(1):218-25
103. Ohki S, Dohi K, Tamai A, et al. Stable-isotope labeling using an inducible viral infection system in suspension-cultured plant cells. J Biomol NMR 2008;42(4):271-7

•• **Plant BY-2 expression system.**

104. Ohki S, Takeuchi M, Mori M. The NMR structure of stomagen reveals the basis of stomatal density regulation by plant peptide hormones. Nat Commun 2011;2:512
105. Fernández FJ, Vega MC. Technologies to keep an eye on: alternative hosts for protein production in structural biology. Curr Opin Struct Biol 2013;23(3):365-73
106. Tochio H. Watching protein structure at work in living cells using NMR spectroscopy. Curr Opin Chem Biol 2012;16(5-6):609-13

Affiliation

Toshihiko Sugiki[1] PhD,
Toshimichi Fujiwara[2] PhD &
Chojiro Kojima[†3] PhD
[†]Author for correspondence
[1]Assistant Professor,
Osaka University, Institute for Protein Research,
3-2, Yamadaoka, Suita, Osaka 565-0871, Japan
[2]Professor,
Osaka University, Institute for Protein Research,
3-2, Yamadaoka, Suita, Osaka 565-0871, Japan
[3]Associate Professor,
Institute for Protein Research, Osaka University,
3-2, Yamadaoka, Suita, Osaka 565-0871, Japan
Tel: +81 6 6879 8598;
Fax: +81 6 6879 8599;
E-mail: kojima@protein.osaka-u.ac.jp