

## Bioinformatics Analysis and Optimization of Cell-Free Protein Synthesis

Alexander A. Tokmakov, Atsushi Kurotani, Mikako Shirouzu, Yasuo Fukami, and Shigeyuki Yokoyama

### Abstract

Cell-free protein synthesis offers substantial advantages over cell-based expression, allowing direct access to the protein synthetic reaction and meticulous control over the reaction conditions. Recently, we identified a number of statistically significant correlations between calculated and predicted properties of amino acid sequences and their amenability to heterologous cell-free expression. These correlations can be of practical use for predicting expression success and optimizing cell-free protein synthesis. In this chapter, we describe our approach and demonstrate how computational and predictive bioinformatics can be used to analyze and optimize cell-free protein expression.

**Key words** Cell-free protein synthesis, Heterologous expression, Rationalization, Optimization, Physicochemical and structural protein properties, Bioinformatics analysis

---

### 1 Introduction

Eukaryotic proteins and their domains are commonly expressed in recombinant form in *Escherichia coli* bacteria [1–3] and cell-free extracts [3–6]. However, obtaining the correct folding of eukaryotic proteins expressed in the bacterial host remains a great challenge. Inability of heterologous protein synthetic machinery to support correct protein folding is considered to be a major factor behind low expression yield and poor solubility of many recombinant proteins.

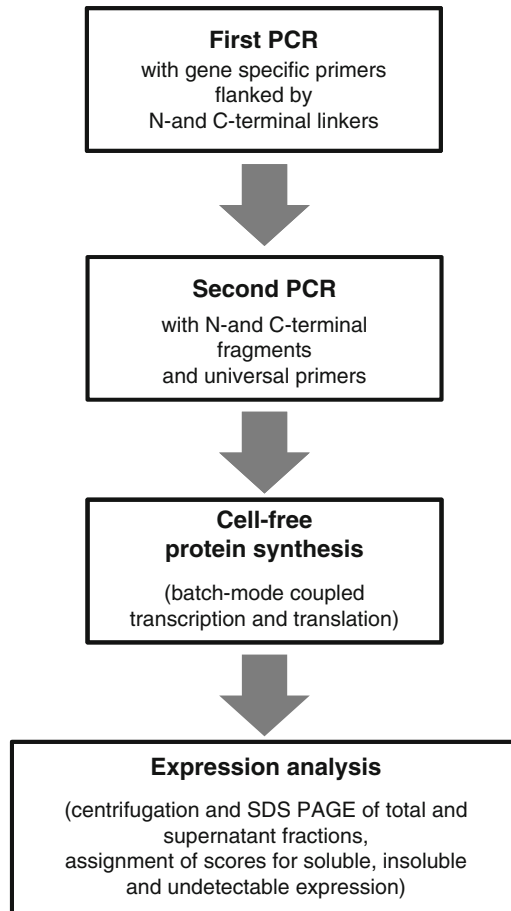
Various physicochemical properties of a polypeptide sequence have been correlated with soluble expression in bacteria [2, 7–10]. Recently we revealed a number of statistically significant correlations between the yield of heterologous cell-free protein synthesis and multiple calculated and predicted parameters of amino acid sequences [11]. They include protein length, hydrophobicity,  $pI$ , content of charged, nonpolar and aromatic residues, cysteine

content, solvent accessibility, presence of coiled coil, content of intrinsically disordered and structured ( $\alpha$ -helix and  $\beta$ -sheet) sequence, number of disulfide bonds and functional domains, presence of transmembrane regions, PEST motifs, and signaling sequences.

In addition, many eukaryotic proteins require multiple post-translational modifications (PTMs) to reach a native, biologically active conformation. However, the bacterial expression systems have only a limited capacity for PTMs. Most recently amenability of human polypeptide sequences to prokaryotic cell-free expression has been demonstrated to correlate with the presence of multiple bioinformatically predicted PTM sites [12].

Importantly, cell-free protein expression is well compatible with high-throughput protein production and optimization. It enables the use of PCR-generated linear DNA templates for programming protein synthesis without the need for their cloning into expression vectors. The protein synthesis in the cell-free system is fast and highly processive; its productivity reaches several milligrams of protein per milliliter of reaction mixture [13]; and it is amenable to efficient scaling. Coupling transcription and translation using DNA templates and bacteriophage RNA polymerases was shown to achieve the highest protein yields in a cell-free environment [14, 15]. The above features of cell-free protein synthesis allowed us to set up the protein expression pipeline described below. The developed protocol was used to screen for well-expressed and highly soluble polypeptide sequences from a large collection of candidate targets. This effort was carried out in the framework of the Japan's national structural genomic project "Protein 3000" launched in 2002 for the purpose of determining the structures of 3,000 proteins using X-ray and NMR methods [16–18].

This chapter does not cover the expression pipeline comprehensively, it is described in detail elsewhere [11, 19–21]. Instead the focus is set on the expression evaluation, data processing, bioinformatics analysis, and optimization of cell-free protein synthesis. In brief, the major steps of the screening-scale cell-free protein expression protocol are summarized in Fig. 1. They included linear template generation by the two-step PCR from the source human cDNA clones (*see* **Notes 1** and **2**), small-scale (20–50  $\mu$ l) batch-mode coupled transcription/translation protein synthesis in the cell-free extract of *E. coli* (*see* **Note 3**), separation of soluble and insoluble reaction products by centrifugation, and estimation of protein yields and solubility by SDS PAGE and protein staining (*see* **Note 4**). DNA template generation and cell-free protein production were carried out in a 96-well format to allow simultaneous processing of multiple samples. The complete dataset of human proteins and their domains expressed in our project under the same uniform set of conditions according to the developed protocol comprised 3066 non-redundant amino acid sequences.



**Fig. 1** Cell-free protein expression pipeline. Main steps of the small-scale protein production for the screening of well-expressed and highly soluble polypeptide sequences are presented

## 2 Materials

### 2.1 First-Step PCR

1. Expand High Fidelity PCR System (Roche, Basel, Switzerland).
2. dNTPs mixture (TOYOBO, Osaka, Japan).
3. Gene-specific forward and reverse primers (Invitrogen, Carlsbad, CA).
4. cDNA clones (*see Note 1*).
5. 96-Well PCR plates and strip caps.
6. A PCR thermal cycler.

### 2.2 Second-Step PCR

1. Expand High Fidelity PCR System (Roche, Basel, Switzerland).
2. dNTPs mixture (TOYOBO, Osaka, Japan).
3. T7 promoter and T7 terminator fragments (*see Note 2*).

4. Universal primer (*see* **Note 2**).
5. First-step PCR products.
6. 96-Well PCR plates and strip caps.
7. A PCR thermal cycler.

### **2.3 Cell-Free Reaction of Protein Synthesis**

1. Bacterial cell-free S30 extract prepared as described previously ([22], *see* also **Note 3**).
2. Second-step PCR product.
3. Reaction buffer: HEPES–KOH, pH 7.5, containing PEG 8000, potassium glutamate, creatine phosphate, calcium folinate, NH<sub>4</sub>OAc, cAMP, DTT, ATP, GTP, CTP, UTP.
4. Total *E. coli* tRNA (Roche, Basel, Switzerland).
5. Solution of Mg(OAc)<sub>2</sub>.
6. Mixture of 20 amino acids in 10 mM DTT.
7. Creatine kinase (Roche, Basel, Switzerland).
8. T7 RNA polymerase prepared as reported previously [23–25].
9. 96-Well PCR plates and strip caps.
10. A thermostat.

---

## **3 Methods**

The following methods are intended for analysis of output from an existing *E. coli*-based cell-free protein production pipeline. They include (1) categorical, rather than continuous quantification of protein expression; (2) identification of physicochemical and structural parameters of amino acid sequences and multiple PTM sites using Internet-based computational and predictive bioinformatics tools; (3) processing and presentation of correlation data for the continuous and discrete variable parameters; (4) statistical analysis of the observed correlations between calculated and predicted properties of proteins and their amenability to heterologous cell-free expression.

### **3.1 Estimation of Protein Yield and Solubility**

After the completion of protein synthetic reaction, soluble and insoluble protein products are separated by centrifugation at 10,000 × *g* for 10 min (*see* **Note 4**). Five-microliter aliquots of total and supernatant fractions are subjected to SDS PAGE on 12.5 % gels and protein bands are visualized with Coomassie Blue staining. The expression yield and solubility are estimated by the intensities of specific bands in the total and supernatant fractions. The bands are quantified using image analyzing software, such as Image Gauge software (Fuji Film, Tokyo, Japan). For quantity calibrations, bovine serum albumin (0.2–2.0 µg/lane) can be used. Typically,

proteins that are expressed at the levels of less than 0.1 mg/ml are difficult to reliably visualize on the Coomassie-stained gels, because the specific protein bands are masked by the endogenous *E. coli* proteins.

Based on the quantification results, the scores A, C, and N are assigned to all experimentally expressed proteins as follows: A, soluble proteins expressed at the levels of more than 0.1 mg/ml; C, expressed but insoluble proteins; and N, non-expressed proteins (expression level below 0.1 mg/ml) (*see* **Notes 5** and **6**). Analysis of protein expression in our dataset showed that the proteins of group A represented 25.7 %, the proteins of group C—46.7 %, and the proteins of group N—27.6 % of all proteins analyzed. Similar success rate of soluble expression has been reported for another subset of human proteins expressed in *E. coli* [26].

For the purpose of following bioinformatics analysis, it is important that all investigated sequences are initially expressed under the same uniform set of conditions, minimizing the influence of sequence-independent factors. The affinity purification tags, which represent the additions of a polypeptide fragment at C- or N-terminus, must be either avoided or should exert minimal effects on protein folding (*see* **Note 7**). Some tags, such as maltose-binding protein, glutathione-S-transferase, etc., are highly soluble, increasing overall solubility of the fused target proteins. This may hinder the analysis of expression correlations by diminishing the role of sequence-specific determinants of protein targets. In our dataset, all synthesized polypeptide products universally comprised the N-terminal poly-His tag to allow their purification at the step of large-scale production.

### **3.2 Calculation and Prediction of Multiple Parameters of Polypeptide Sequences**

The physicochemical parameters of amino acid sequences, such as *pI*, charge, hydrophobicity, can be calculated using the free ProtParam tool available at the Expasy server (<http://web.expasy.org/protparam/>). Solvent accessibility is calculated with the ACCpro 4.0 software downloaded from the SCRATCH Protein Predictor server ([27], <http://scratch.proteomics.ics.uci.edu/explanation.html>) and content of secondary structure is calculated with the PREDATOR 2.1.2 tool [28] provided online (<http://mobyli.pasteur.fr/cgi-bin/portal.py?#forms::predator>). Content of disordered structure is predicted with the RONN software [29] available online (<http://www.strubi.ox.ac.uk/RONN>). Coiled coil structures, signal sequences, transmembrane domains, and PEST regions are predicted with the tools provided online (<http://emboss.sourceforge.net/apps/cvs/emboss/apps/pepcoil.html>, <http://www.cbs.dtu.dk/services/SignalP/>, <http://bp.nuap.nagoya-u.ac.jp/sosui/sosuisignal/>, <http://emboss.bioinformatics.nl/cgi-bin/emboss/pestfind>, respectively).

### 3.3 Prediction of Posttranslational Modifications

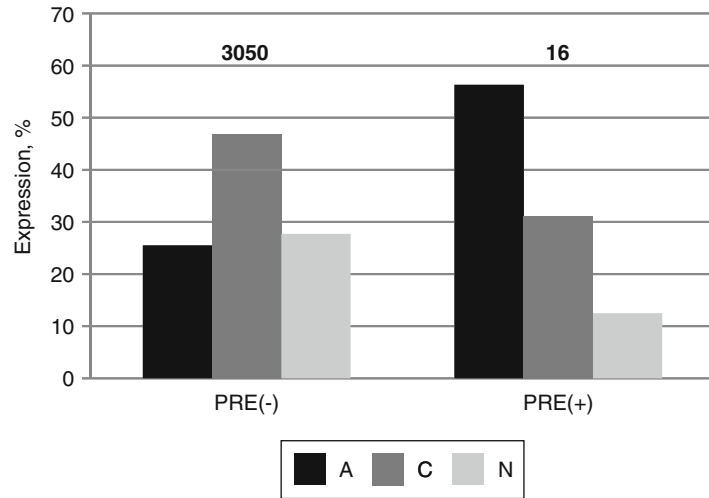
The sites of prenylation, asparagine glycosylation, phosphorylation, etc. can be predicted with the PROSITE scanning tool PS\_SCAN provided online ([http://www.hpa-bioinfotools.org.uk/cgi-bin/ps\\_scan/ps\\_scanCGI.pl](http://www.hpa-bioinfotools.org.uk/cgi-bin/ps_scan/ps_scanCGI.pl)). The sites of S-palmitoylation are predicted with the CSS-Palm tool [30] available online (<http://csspalm.biocuckoo.org>). Disulfide bonds are calculated with the Dipro tool [31] downloadable free for scientific use (<http://download.igb.uci.edu/intro.html>). Ubiquitination sites can be predicted using the predictor of protein ubiquitination UbPred [32] downloaded from <http://ubpred.org/> and SUMOylation sites are predicted with the site-specific predictor SUMOsp 2.0 [33] freely downloadable for academic research (<http://sumosp.biocuckoo.org/>).

### 3.4 Data Processing and Presentation

At the step of bioinformatics analysis of the expression data, the variable parameters or features of the two types, continuous and discrete, are used to characterize physicochemical and structural properties of polypeptide sequences. Among them, the Yes/No type of discrete variables are the features that can be either absent from or present in proteins. Localization signals, single N- and C-terminal protein modifications and other single-event protein modifications are the examples of such features. To present the expression data associated with the Yes/No type variables, the bar graphs can be used, which show the percentage of protein targets in the expression groups A, C, and N. The graphs should be built for the two datasets of proteins (i.e., excluding and including the analyzed feature).

As an example of the Yes/No data type processing, a case of protein prenylation can be considered. Prenylation was found to be a low-abundant modification—only 16 proteins in the analyzed dataset have been predicted to contain potential prenylation sites, and only single sites of prenylation could be predicted in these proteins. This is largely consistent with the previous estimates that put the number of possible prenylated proteins in the mammalian proteome to less than 2 %, corresponding to total ~100–200 proteins being potential prenylation substrates [34, 35]. Relative rates of soluble (A), insoluble (B), and non-expressed (C) proteins with (+) or without (–) the predicted sites of prenylation are shown in Fig. 2. The total number of protein targets in the (+) and (–) subsets is indicated above the bars. Using this graph, it is easy to make a side-by-side comparison of the data for the two subsets of sequences and deduce the tendencies in the protein cell-free expression amenability associated with prenylation. The evaluation of statistical significance of the observed tendencies is described further (see Subheading 3.5).

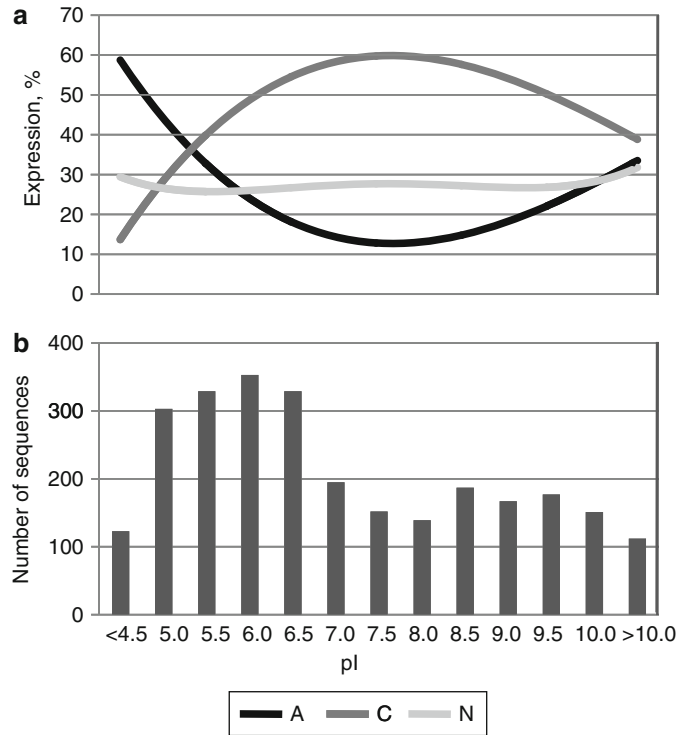
To present the expression data associated with the continuous variables, such as sequence hydrophobicity, *pI*, solvent accessibility, etc., another type of data presentation is more convenient. In this case, the percentage of protein targets in the expression groups A,



**Fig. 2** Correlation between protein amenability to cell-free expression and predicted presence of prenylation. Relative rates of soluble (A), insoluble (B), and non-expressed (C) proteins with different probability of prenylation (+ or –) are shown

C, and N is plotted at different values of the studied parameter (Fig. 3a). The plot should cover the entire range of parameter values observed in the analyzed dataset. Curve smoothing is recommended, considering the continuous nature of these protein features. In addition, the distribution graph of dataset proteins according to parameter values should also be presented (Fig. 3b). The distribution graph provides the important information about the abundance of a studied protein feature in the analyzed dataset and its relation to expression amenability. For instance, distribution of the 3,066 dataset proteins according to their *pI* was found to be bimodal with the minimal representation of proteins at *pI* 7.0–7.5 (Fig. 3b). At the same time, the lowest rate of soluble expression and the highest rate of insoluble expression were observed for the proteins with *pI* 7.0–7.5 (Fig. 3a). These data propose that the proteins with neutral *pI* values may be underrepresented in the dataset due to their low solubility.

Data development for the discrete protein features repeatedly observed in the analyzed sequences, such as multiple S–S bonds, transmembrane regions, functional domains, abundant multi-site PTMs, is similar to the processing of the expression data associated with continuous variables. The graphs of relative soluble (A), insoluble (C), and undetectable expression (N), as well as the distribution graph should be provided in the complete range of discrete feature values. As an example, the correlations of cell-free expression amenability with the predicted number of disulfide bonds are shown in Fig. 4.



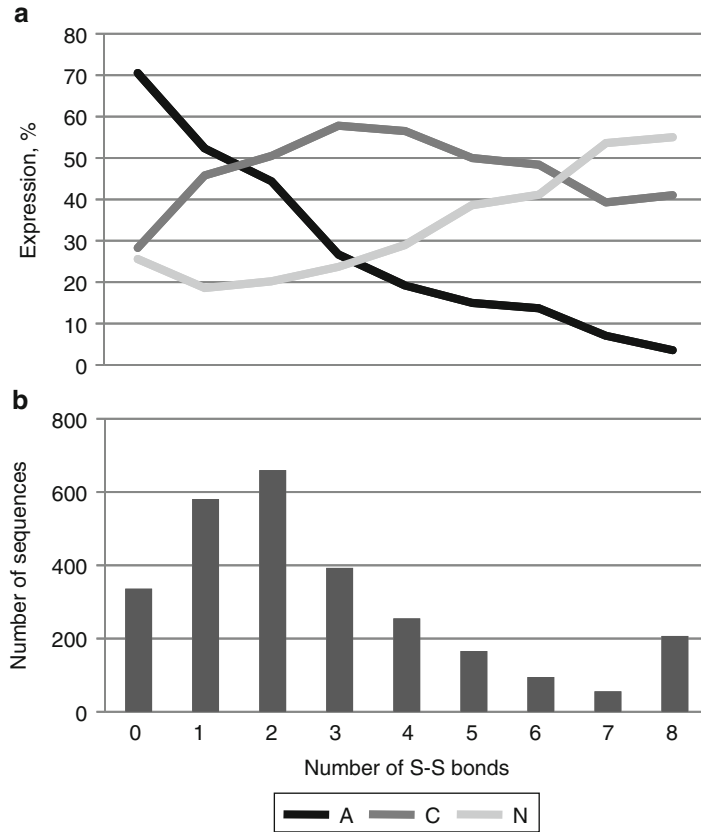
**Fig. 3** Correlation between protein amenability to cell-free expression and pI. Relative rates of soluble (curve A), insoluble (curve C), and non-expressed (curve N) proteins with different pI values are presented in (a). Distribution of the dataset proteins according to their pI is shown in (b)

Finally, it is beneficial, for comprehensiveness of presentation, to summarize the observed correlations between expression amenability and multiple protein properties in one table. The major correlations revealed by the described analysis in our studies are presented in Table 2.

### 3.5 Statistical Significance of the Observed Correlations

At the stage of expression evaluation, all investigated proteins are categorized into three mutually exclusive classes—soluble (A), insoluble (C), and non-expressed (N) proteins, with the sum of these data (i.e., the total number of studied proteins) equaling 100 % (*see Note 8*). In other words, the data can only be placed in one class and not into another. Similarly, during the following bioinformatics analysis, the expressed protein targets are categorized into these three classes at different values of calculated or predicted parameters. Thus, the expression data represent, in essence, categorical datasets. To evaluate the statistical significance of the observed correlations, the categorical data analysis should be applied. This type of analysis is used to tell whether the difference between the sets of results is significant or not when the datasets contain the entries categorized into several mutually exclusive classes [36]. The estimation of statistical significance for each





**Fig. 4** Correlation between protein amenability to cell-free expression and predicted presence of disulfide bonds. Relative rates of soluble (curve A), insoluble (curve C), and non-expressed (curve N) proteins with different number of predicted S-S bonds are presented in (a). Distribution of the dataset proteins according to the number of S-S bonds is shown in (b)

expression group (A, C, and N) should be provided. In our project, to deduce the statistical differences between the expression datasets, the two-way contingency table test has been applied. The Fisher's exact  $p$ -values were computed using the tool available online (<http://statpages.org/ctab2x2.html>). A confidence level of 95 % was set up as the null hypothesis rejection threshold.

As an example, the evaluation of statistical significance of the data presented in Figs. 2, 3, and 4 is shown in Table 1. The categorical data analysis confirms a statistically significant difference in the ratios of soluble-expressed proteins in the two subsets, Pre (+) and Pre (-). On the other hand, although the tendencies towards the decrease in ratios of insoluble and undetectable protein expression can be observed for the sequences containing predicted sites of prenylation (Fig. 2), they cannot be validated by the statistical analysis (Table 1). Notably, the low statistical significance of these tendencies is certainly related to the small number of protein

**Table 1**  
**Statistical significance of correlations between calculated and predicted protein properties and cell-free expression amenability**

Parameter	Expression		
	Soluble	Insoluble	Undetectable
Prenylation (number of sequences, +/-)	9/778	5/1427	2/845
pI (number of sequences, 5.5/7.5)	94/21	157/96	83/40
Disulfide bonds (number of sequences, +/-)	41/155	243/95	257/86
Prenylation (Fisher's exact <i>p</i> -value, +/-)	0.009	0.315	0.262
pI (Fisher's exact <i>p</i> -value, 5.5/7.5)	<0.001	0.004	0.911
Disulfide bonds (Fisher's exact <i>p</i> -value, +/-)	<0.001	<0.001	<0.001

The numbers of soluble, insoluble, and non-expressed polypeptide sequences with (+) or without (-) predicted prenylation sites and disulfide bonds, as well as the numbers of sequences with different values of the pI parameter ( $5.0 < X < 5.5$  and  $7.0 < X < 7.5$ ) are presented in the upper rows of the table. Analyzed subset of disulfide bond containing polypeptides comprised the sequences with more than five predicted S-S bonds

The Fisher's exact *p*-values calculated by the two-way contingency table analysis are presented in the lower rows

Boxes highlighted in *grey* denote the differences that are statistically significant at more than 95 % confidence level

sequences in the Pre (+) dataset because the confidence interval of categorical data is greatly affected by the sample size (*see Note 9*). The analysis of a more extended dataset of prenylated proteins is necessary to validate the observed tendencies. On the other hand, the lack of statistically significant difference demonstrated for the undetectable expression of proteins with  $5.0 \leq pI < 5.5$  and  $7.0 \leq pI < 7.5$  (Table 1) should reflect *de facto* absence of a tendency, considering the number of protein sequences in the analyzed datasets. Also, a high confidence level of the correlations between protein amenability to cell-free expression and predicted presence of disulfide bonds leans upon the large numbers of entries in the corresponding datasets (Table 1).

### 3.6 Optimizing the Conditions of Protein Synthetic Reaction

The purpose of cell-free synthesis in structural genomics and proteomics projects is to produce properly folded, functionally active proteins in the amounts sufficient for functional and structural studies. Thus, the optimization of cell-free protein synthesis

concerns, as a rule, the yield of soluble (category A) expression. After the initial evaluation of protein expression using the described cell-free platform, it is often possible to optimize the reaction mixture for enhanced production of the selected proteins based on their individual physicochemical and structural properties. These properties can be calculated and predicted using bioinformatics algorithms and tools presented in Subheadings 3.2 and 3.3. Once the feature that may hinder soluble expression of a given protein is pinpointed, it becomes possible to tailor the conditions of cell-free protein synthesis specifically for this protein.

For instance, the high overall hydrophobicity of amino acid sequences is associated with worse protein amenability to soluble expression (Table 2). The poor solubility of hydrophobic proteins can be explained by their susceptibility to aggregation due to intermolecular hydrophobic interactions. It is often possible to neutralize these interactions and to increase protein solubility by using weak nonionic detergents that bind to the hydrophobic regions of proteins. This approach also allows the synthesis of membrane proteins that can be stabilized in the presence of detergent micelles. Importantly, the presence of mild nonionic detergents at the low solubilizing concentrations does not inactivate the protein synthetic activity of the bacterial extract. It was found that only nonionic detergents with low critical micelle concentration (c.m.c.) can be used in cell-free expression systems without decreasing protein yield [37]. In our project, G-protein-coupled receptors have been successfully expressed using the weak nonionic detergents Brij35 and digitonin [38].

Another example of reaction optimization concerns the synthesis of correctly folded eukaryotic proteins containing multiple disulfide bonds in the bacterial cell-free system. The predicted presence of disulfide bonds in a polypeptide negatively correlates with soluble protein expression and positively correlates with the insoluble expression (Fig. 4, Table 2). Among all protein features analyzed, the occurrence of disulfide bonds was found to be one of the most discriminative for expression propensity. Probability of soluble expression for the proteins that are predicted to bear more than six disulfide bonds per molecule falls below 10 % [11]. The formation of S–S bonds in eukaryotic proteins is greatly compromised in the bacterial extracts, as their reducing conditions differ from those in eukaryotic cells [39]. Pretreatment with iodoacetamine, which blocks the free sulfhydryl groups of endogenous bacterial proteins, was shown to abolish the disulfide-reducing activity of the extracts [40]. The use of the iodoacetamine-treated bacterial extracts complemented with a glutathione redox buffer and the DsbC disulfide isomerase allowed efficient production of complex mammalian proteins containing multiple disulfide bonds [41, 42].

To increase the rate of soluble expression of neutral proteins (i.e., the proteins with *pI* values close to 7.0), pH of the reaction

**Table 2**  
**Correlations of soluble, insoluble, and undetectable cell-free expression with calculated and predicted properties of proteins**

Property	Expression		
	Soluble	Insoluble	Undetectable
Length	↓	—	↑
pI	↕	↕	—
Charge	↑	↕	↓
Hydrophobicity	↓	↑	—
Solvent accessibility	↑	↓	↑
Secondary structure	↑	↕	↓
Intrinsic disorder	↑	↓	↑
Disulfide bonds	↓	↑	↑
Coiled coil	↑	↓	↓
Transmembrane domains	↓	↓	↑
Localization signal sequences	↓	↑	—
PEST regions	↑	↓	↑
Prenylation	↑	—	—
Phosphorylation	↑	↓	↓
Asn glycosylation	↓	↑	—
Palmitoylation	↓	↕	↑
Ubiquitination	↑	↓	—
SUMOylation	↑	↓	↕

*Upward and downward arrows* indicate positive and negative correlations, respectively; *straight horizontal lines* denote the lack of correlation; and *bidirectional arrows* refer to the opposite tendencies of expression estimates at different values of calculated parameters

mixture can be optimized. The proteins with neutral pI values are not charged at the physiological pH of the original expression system and are prone to aggregation. Accordingly, the lowest rate

of soluble expression has been observed for the proteins with  $pI$  7.0–7.5 (Fig. 3). Shifting pH to more acidic or basic values will make these proteins charged, preventing their intermolecular interactions and resulting in better solubility. As a rule, setting the reaction mixture pH within the interval of 6.5–8.5 does not affect dramatically the total yield of protein synthetic reaction; however, often it can significantly increase the yield of soluble expression. Notably, pH also affects the oxidizing environment of the reaction mixture, because the thiol-disulfide exchange is inhibited at low pH (effectively, at pH below 8.0). It was found that the normal reaction pH was not high enough to allow the efficient formation and isomerization of disulfide bonds and both the oxidizing environment and an increased pH were required to produce the active protease domain of mammalian urokinase in a cell-free bacterial system [40].

Other conditions that can be optimized in the original reaction mixture for the enhanced production of the selected proteins include concentration of magnesium, DTT, DNA template, duration and temperature of protein synthesis, the energy substrate and amino acid composition. Additions of FAD, NAD, CoA, malic acid, 2-glutaric acid, succinic acid and introduction of chaperones have also been reported to stimulate cell-free protein synthesis. Usually, the positive effects of these compounds on the expression yields are not additive and it is quite difficult to explain their individual contributions in the particular cases.

### **3.7 Protein Engineering Aimed at Increasing Expression Success**

In addition to easily mastered optimization of the reaction conditions, the cell-free gene expression technology combined with bioinformatics analysis allows feasible protein engineering and screening with the aim of increasing expression success. Several approaches have been developed to engineer the proteins that have enhanced amenability to heterologous cell-free expression.

For instance, the solubility of signal sequence-containing proteins can often be improved by the N-terminal truncation of signal sequence regions. These sequences direct proteins to certain cellular compartments and organelles, making their existence in the cytoplasm (i.e., in a free soluble form) implausible. Similarly, truncation of N-terminal or C-terminal membrane anchoring sequences in the single-pass transmembrane proteins can improve their solubility. The presence of the signal and anchoring (transmembrane) sequences was found to worsen protein amenability to cell-free soluble expression ([11], Table 2). As an example from our project, the human pyruvate dehydrogenase kinase 4 could be successfully expressed and crystallized for X-ray studies after truncating its mitochondrial targeting sequence [43]. The signal and anchoring sequences can be reliably identified in the analyzed proteins using the existing prediction algorithms (*see* Subheading 3.2).

Other common approach in the structural/functional protein analysis is to express the truncated forms of the complex multi-domain proteins. The worsening of protein expression with the number of functional domains and protein length has been demonstrated ([11], Table 2). The removal of domains is often able to improve protein propensity for soluble expression (*see* **Note 10**). The resolved protein structures deposited in the Protein Data Bank and 3D homology modeling can be used to define the functional domains in a protein. In the case when protein structure is largely unknown, it is possible to assign functional domains using prediction algorithms based on difference in the amino acid composition between domain and linker regions [44]. The domain truncation approach is exceptionally efficient when the domains to be expressed are well defined and display high intrinsic solubility.

Alternatively, fusing target proteins with other highly soluble polypeptides often helps to increase overall solubility of the fused protein products. The addition of the solubility tags, such as GST, MBP, SUMO, ubiquitin, at the N- or C-termini of highly hydrophobic sequences can be considered to increase the yield of their soluble expression. Interestingly, it was found that the presence of predicted sites of SUMOylation and ubiquitination in a protein directly correlates with its propensity for cell-free soluble expression ([12], Table 2).

Although, in general, single amino acid replacements cannot change significantly the integral characteristics of protein molecule related to its overall solubility, the dramatic global realignments of protein structure due to a single substitution event have been documented too. Thus, in some cases, introducing even a single mutation can significantly change the outcome of protein synthetic reaction, increasing the yield of soluble protein. Experimentally confirmed or bioinformatically predicted sites of PTMs may be the primary targets for site-directed mutagenesis. For example, phosphorylatable Ser, Thr, and Tyr residues in the amino acid sequences can be mutated into Glu or Asp residues to mimic the negative charge associated with the phospho group. It is not uncommon that the phosphorylated forms of proteins gain increased solubility. The presence of predicted phosphorylation sites in the polypeptide sequences was found to be associated with the increased production of properly folded soluble protein ([12], Table 2). Admittedly, the use of site-directed mutagenesis for the enhancement of cell-free protein expression is far from being established and it requires further development. A simple and widely used technique of alanine-scanning mutagenesis can be employed to determine the contribution of specific residues to the stability and solubility of a given protein.

---

## 4 Notes

1. The template human cDNA clones can be obtained from several commercial sources, such as Invitrogen, Carlsbad, CA, USA; OriGene Technologies, Rockville, MD, USA; Kazusa DNA Research Institute, Kisarazu, Chiba, Japan; Institute of Medical Science of Tokyo University, Tokyo, Japan; GeneCopoeia, Inc., Rockville, MD, USA; and Toyobo Engineering, Osaka, Japan.
2. The forward and reverse primers for the first PCR, the T7 promoter and T7 terminator fragments, and the universal primer for the second PCR have been constructed as described previously [19, 21].
3. To circumvent the low expression yields due to rare codons, the extracts prepared from the bacterial strains containing additional copies of the genes for low-abundant tRNAs such as BL21 codon plus (Stratagene, USA), BL21-Star (DE3), or Rosetta strains (Novagen, USA) should be used.
4. At the screening stage, the term “solubility” refers to the protein solubility in the cell-free extract that may differ from solubility of purified proteins. The steps should be taken to maintain an extract-soluble protein in the soluble state during the purification procedure normally following large-scale cell-free protein synthesis.
5. The score A provides the upper estimation of soluble protein expression, because the solubility status is evaluated after separation of soluble and insoluble protein products by centrifugation at  $10,000\times g$  for 10 min. This procedure cannot discriminate between small protein aggregates and truly soluble proteins.
6. Proteins that are expressed at a lower than expected molecular size should be considered as non-expressed, because they cannot attain proper structure and function when synthesized in the employed expression system.
7. The presence of a C- or N-terminal affinity tag makes impossible the analysis of correlations between the predicted modifications of the tagged terminus and cell-free protein synthesis. For instance, the presence of an N-terminal tag hinders the effects of the N-terminal protein modifications, such as N-terminal methionine excision, myristoylation, acetylation, etc., on the protein amenability to heterologous cell-free expression. Thus, the expression of untagged or C-terminally tagged protein targets should be considered in order to deduce the correlations between N-terminal modifications and heterologous protein synthesis.

8. Often, expressed proteins can be found in both soluble and insoluble fractions of the cell-free reaction mix. However, the applied categorical data analysis allows their classification into only one expression category. Usually, lane-to-lane comparison of total and supernatant fractions of the reaction in PAGE gels is sufficient to determine the preferential pattern of protein expression.
9. The confidence level of categorical data analysis increases dramatically with the number of sequences in the datasets of expressed proteins. To deduce statistically significant tendencies, rather small datasets may be used when analyzing robust correlations. Practically, for the discussed approach, the counts in every cell of the contingency table should be at least 5 [45].
10. It is important not to cut into domains when expressing truncated forms of proteins. The truncations should be made several residues upstream and downstream of the last major structural elements ( $\alpha$ -helix or  $\beta$ -sheet) of a domain. It is recommended to produce several variants that are cut out at intervals of several residues at both N- and C-termini and examine their yield and solubility.

## References

1. Yokoyama S (2003) Protein expression systems for structural genomics and proteomics. *Curr Opin Chem Biol* 7:39–43
2. Marsden RL, Orengo CA (2008) Target selection for structural genomics: an overview. *Methods Mol Biol* 426:3–25
3. Farokki N, Hrmova M, Burton RA et al (2009) Heterologous and cell-free protein expression systems. *Methods Mol Biol* 513:175–198
4. Spirin AS (2004) High-throughput cell-free systems for synthesis of functionally active proteins. *Trends Biotechnol* 22:538–545
5. Katzen F, Chang G, Kudlicki W (2005) The past, present and future of cell-free protein synthesis. *Trends Biotechnol* 23:150–156
6. He M (2008) Cell-free protein synthesis: applications in proteomics and biotechnology. *Nat Biotechnol* 23:126–132
7. Goh CS et al (2004) Mining the structural genomics pipeline: identification of protein properties that effect high-throughput experimental analysis. *J Mol Biol* 336:115–130
8. Bertone P, Kluger Y, Lan N et al (2001) SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic Acids Res* 29:2884–2898
9. Dyson MR, Shadbolt SP, Vincent KJ et al (2004) Production of soluble mammalian proteins in *Escherichia coli*: identification of protein features that correlate with successful expression. *BMC Biotechnol* 4:32
10. Idicula-Thomas S, Balaji P (2005) Understanding the relationships between the primary structure of proteins and its propensity to be soluble on overexpression in *Escherichia coli*. *Protein Sci* 14:582–592
11. Kurotani A, Takagi T, Toyama M et al (2010) Comprehensive bioinformatics analysis of cell-free protein synthesis: identification of multiple protein properties that correlate with successful expression. *FASEB J* 24:1095–1104
12. Tokmakov AA, Kurotani A, Takagi T et al (2012) Multiple post-translational modifications affect heterologous protein synthesis. *J Biol Chem* 287:27106–27116
13. Kigawa T, Yabuki T, Yoshida Y et al (1999) Cell-free production and stable-isotope labeling of milligram quantities of proteins. *FEBS Lett* 442:15–19
14. Kigawa T, Yokoyama S (1991) A continuous cell-free protein synthesis system for coupled transcription-translation. *J Biochem* 110:166–168
15. Kudlicki W, Kramer G, Hardesty B (1992) High efficiency cell-free synthesis of proteins: refinement of the coupled transcription/translation system. *Anal Biochem* 206:389–393
16. Yokoyama S, Hirota H, Kigawa T et al (2000) Structural genomics projects in Japan. *Nat Struct Biol* 7(Suppl):943–945



17. Yokoyama S (2005) Large-scale structural proteomics project at RIKEN: present and future. *Tanpakushitsu Kakusan Koso* 50:836–845
18. Yokoyama S, Kigawa T, Shirouzu M et al (2008) RIKEN structural genomics/proteomics initiative. *Tanpakushitsu Kakusan Koso* 53:632–637
19. Yabuki T, Motoda Y, Hanada K et al (2007) A robust two-step PCR method of template DNA production for high-throughput cell-free protein synthesis. *J Struct Funct Genom* 8:173–191
20. Kigawa T, Matsuda T, Yabuki T et al (2008) Bacterial cell-free system for highly efficient protein synthesis. In: Spirin AS, Swartz JR (eds) *Cell-free protein synthesis*. Wiley, Weinheim, pp 83–97
21. Kigawa T (2010) Analysis of protein functions through a bacterial cell-free protein expression system. *Methods Mol Biol* 607:53–62
22. Kigawa T, Yabuki T, Matsuda N et al (2004) Preparation of *Escherichia coli* extract for highly productive cell-free protein expression. *J Struct Funct Genom* 5:63–68
23. Davanloo P, Rosenberg AH, Dunn JJ et al (1984) Cloning and expression of the gene for bacteriophage T7 RNA polymerase. *Proc Natl Acad Sci U S A* 81:2035–2039
24. Grodberg J, Dunn JJ (1988) ompT encodes the *Escherichia coli* outer membrane protease that cleaves T7 RNA polymerase during purification. *J Bacteriol* 170:1245–1253
25. Zawadski V, Gross HJ (1991) Rapid and simple purification of T7 RNA polymerase. *Nucleic Acids Res* 19:1948
26. Ding HT, Ren H, Cheng Q et al (2002) Parallel cloning, expression, purification, and crystallization of human proteins for structural genomics. *Acta Crystallogr D Biol Crystallogr* 58:2102–2108
27. Cheng J, Randall AZ, Sweredoski MJ et al (2005) SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res* 33(Web Server issue):72–76
28. Frishman D, Argos P (1997) Seventy-five percent accuracy in protein secondary structure prediction. *Proteins* 27:329–335
29. Yang ZR, Thomson R, McMeil P et al (2005) RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics* 21:3369–3376
30. Ren J, Wen L, Gao X et al (2008) CSS-Palm 2.0: an updated software for palmitoylation sites prediction. *Protein Eng Des Sel* 21:639–644
31. Cheng J, Saigo H, Baldi P (2006) Large-scale prediction of disulfide bridges using kernel methods, two-dimensional recursive neural networks, and weighed graph matching. *Proteins* 62:617–629
32. Radivojac P, Vacic V, Haynes C et al (2010) Identification, analysis, and prediction of protein ubiquitination sites. *Proteins* 78:365–380
33. Ren J, Gao X, Jin C et al (2009) Systematic study of protein sumoylation: development of a site-specific predictor of SUMPsp 2.0. *Proteomics* 9:3409–3412
34. Gao J, Liao J, Yang GY (2009) CAAX-box protein, prenylation process and carcinogenesis. *Am J Transl Res* 25:312–325
35. Amaya M, Baranova A, van Hoek ML (2011) Protein prenylation: a new mode of host-pathogen reaction. *Biochem Biophys Res Commun* 416:1–6
36. Xu B, Feng X, Burdine RD (2010) Categorical data analysis in experimental biology. *Dev Biol* 348:3–11
37. Betton JM, Miot M (2008) Cell-free production of membrane proteins in the presence of detergents. In: Spirin AS, Swartz JR (eds) *Cell-free protein synthesis: methods and protocols*. Wiley, Weinheim, pp 165–178
38. Ishihara G, Goto M, Saeki M et al (2005) Expression of G-protein coupled receptors in a cell-free translational system using detergents and thioredoxin-fusion vectors. *Protein Expr Purif* 41:27–37
39. Tu BP, Weissman JS (2004) Oxidative protein folding in eukaryotes. *J Cell Biol* 164:341–346
40. Kim DM, Swartz JR (2004) Efficient production of a bioactive, multiple disulfide-bonded protein using modified extracts of *Escherichia coli*. *Biotechnol Bioeng* 85:122–129
41. Yin G, Swartz JR (2004) Enhancing multiple disulfide bonded protein folding in a cell-free system. *Biotechnol Bioeng* 86:188–195
42. Yang J, Kanter G, Voloshin A et al (2004) Expression of active murine granulocyte-macrophage colony-stimulating factor in an *Escherichia coli* cell-free system. *Biotechnol Prog* 20:1689–1696
43. Kukimoto-Niino M, Tokmakov A, Terada T et al (2011) Inhibitor-bound structures of human pyruvate dehydrogenase kinase. *Acta Crystallogr D Biol Crystallogr* 67:763–773
44. Suyama M, Ohara O (2003) DomCut: prediction of inter-domain linker regions in amino acid sequences. *Bioinformatics* 19:673–674
45. Norman GR, Streiner DL (2000) *Biostatistics: the bare essentials*. B.C. Decker, Hamilton

Cell-Free Protein Synthesis

Methods and Protocols

Alexandrov, K.; Johnston, W.A. (Eds.)

2014, XI, 313 p. 66 illus., 31 illus. in color., Hardcover

ISBN: 978-1-62703-781-5

A product of Humana Press