J. Biochem. 2011;150(1):73–81 doi:10.1093/jb/mvr042

JB THE JOURNAL OF BIOCHEMISTRY

Statistical analysis of features associated with protein expression/solubility in an *in vivo Escherichia coli* expression system and a wheat germ cell-free expression system

Received October 31, 2010; accepted March 3, 2011; published online April 9, 2011

Shuichi Hirose[1,*], Yoshifumi Kawamura[2], Kiyonobu Yokota[1], Toshihiro Kuroita[3], Tohru Natsume[4], Kazuo Komiya[5], Takeshi Tsutsumi[5], Yorimasa Suwa[5], Takao Isogai[5], Naoki Goshima[4] and Tamotsu Noguchi[1]

[1]Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 135-0064; [2]Japan Biological Informatics Consortium (JBiC), Tokyo 135-8073; [3]Toyobo Co., Ltd., Tsuruga Institute of Biotechnology, Fukui 914-0074; [4]Biomedicinal Information Research Center (BIRC), National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 135-0064; and [5]Reverse Proteomics Research Institute, Co., Ltd., Tokyo 110-0044, Japan

*Shuichi Hirose, AIST Tokyo Waterfront Bio-IT Research Building, 2-4-7, Aomi, Koto-ku, Tokyo 135-0064, Japan. Tel: +81-3-3599-8730, Fax: +81-3-599-8081, email: hirose-shuichi@aist.go.jp

Recombinant protein technology is an important tool in many industrial and pharmacological applications. Although the success rate of obtaining soluble proteins is relatively low, knowledge of protein expression/solubility under 'standard' conditions may increase the efficiency and reduce the cost of proteomics studies. In this study, we conducted a genome-scale experiment to assess the overexpression and the solubility of human full-length cDNA in an *in vivo Escherichia coli* expression system and a wheat germ cell-free expression system. We evaluated the influences of sequence and structural features on protein expression/solubility in each system and estimated a minimal set of features associated with them. A comparison of the feature sets related to protein expression/solubility in the *in vivo Escherichia coli* expression system revealed that the structural information was strongly associated with protein expression, rather than protein solubility. Moreover, a significant difference was found in the number of features associated with protein solubility in the two expression systems.

Keywords: *Escherichia coli*/protein expression/protein solubility/statistical analysis/wheat germ cell-free.

Abbreviations: cDNA, complementary DNA; ORF, open reading frame; RF, random forest; SDS–PAGE, sodium dodecyl sulfate–poly-acrylamide gel electrophoresis; SVM, support vector machine.

Obtaining highly concentrated, soluble proteins' preparations is necessary for conducting various structural and biophysical studies or using proteins as materials for pharmaceutical or industrial products. *Escherichia coli*, which is easy to handle and manipulate genetically, is the preferred host for overexpressing recombinant proteins, since it can be cultivated rapidly and inexpensively. Moreover, it generally yields high levels of recombinant proteins (*1*). Since the proteins are expressed by the host, one reason for non-expression is a deleterious interaction with the host's metabolism. In addition, a common reason for insolubility is the formation of inclusion bodies. Therefore, the success rate for obtaining soluble proteins is relatively low. For that reason, the construction of protein overexpression systems is an important experimental challenge.

To overcome these unfavourable circumstances, several solutions have been proposed, based on the results of experimental studies: using a different strain of *E. coli*; modifying the N-terminal (*2*) and C-terminal sequences (*3*); fusion with solubility enhancing tags (*4*) and coexpression with molecular chaperones (*5*). Similarly, various alternative cell-based expression systems have been developed. Such systems utilize yeast, insect cells or murine myeloma cells as hosts (*6*). In recent years, cell-free methods for protein synthesis with extracts from prokaryotic (*7*) or eukaryotic (*8*) cells have become an alternative to cell-based methods. The distinctive feature is an *in vitro* translation system. Cell-free expression systems are popular in proteomics and biotechnology, because of their high levels of protein expression and ease of handling (*9*, *10*).

In theoretical computational science, clear sequence differences between proteins that remain soluble and those that form inclusion bodies have been reported, thereby yielding some successes in predicting protein solubility solely from amino acid sequences (*11*–*17*). The first attempt to determine the interconnection between amino acid sequences and protein solubilities was performed by Wilkinson and Harrison (*11*). They observed that protein solubility is strongly associated with the charge average and the turn-forming residue fraction. Subsequent studies revealed several factors associated with protein expression and solubility. Such knowledge under 'standard' conditions may provide a clue for determining priority targets in a large-scale proteomics analysis. However, the difference between the factors related to protein expression

and solubility has remained unclear for genome-scale analyses. In addition, no study has revealed the differences in these factors among protein expression systems.

In this study, we estimated and compared the minimal sets of sequence and structural features associated with protein expression/solubility in an *in vivo E. coli* expression system and a wheat germ cell-free expression system, from an analysis of genome-wide experimental data. The results provide useful information for proteomics analyses.

Materials and Methods

Protein expression experiments

We conducted genome-scale experiments that assessed the overexpression and the solubility of human full-length complementary DNA (cDNA) in an *in vivo E. coli* expression system and a wheat germ cell-free expression system.

In the *E. coli* expression system, the 17,739 human open reading frames [ORFs as Gateway entry clones (*9*)] were first subcloned into the pDEST17 vector (T7 promoter, amino-terminal His-tag fusion), using LR Clonase (Invitrogen). *E. coli* BL21 star (DE3) pLysS was transformed with the reaction products. The SOC expression mixture was then plated on LB agar plates containing ampicillin. For each ORF, LB medium was inoculated with a single colony and grown overnight at 37°C. The overnight culture was diluted 1:100 into SB medium, grown at 37°C for 3 h and cooled to 20°C. Protein expression was then induced by adding isopropyl 1-thio-β-D-galactopyranoside to a final concentration of 0.1 mM. After 16 h at 20°C, the cells were harvested and suspended in BugBuster (Novagen Inc.). Some of the lysate was reserved as the whole cell sample. The lysate was centrifuged at 15,000*g* for 5 min. The supernatants were prepared as the soluble fraction samples. These protein samples were denatured with SDS-sample buffer and fractionated by 12.5% sodium dodecyl sulfate–poly-acrylamide gel electrophoresis (SDS–PAGE). Proteins were visualized with CBB R-250. Protein expression and solubility were judged by visual inspection, according to whether a clear band was detectable at the expected position calculated from the molecular weight in the whole cell and soluble fraction samples. When a clear band in the whole cell sample lane was obviously observed in the presence of bands derived from *E. coli*, it was regarded as overexpression. Conversely, when a specific band was not detected at the expected position, it was regarded as no expression. In the case of a smeared band, it was regarded as low expression. On the same basis, the protein solubility was evaluated using the band in the soluble fraction sample lane.

In the wheat germ cell-free expression system, we expressed 8850 human proteins, for which the entry clones were chosen randomly from our human Gateway entry clone resources (*9*, *18*). The expressed proteins were fused with a carboxy-terminal His-tag (destination vectors: pEW-3H) and were labelled with radioactive amino acids (^{14}C-Leu or ^{35}S-Met). The wheat germ extract was purchased from Toyobo and Cell Free Sciences Co., Ltd. The expressed proteins were separated into soluble and insoluble fractions by centrifugation at 19,000*g* for 20 min. The samples were separated by SDS–PAGE, and the protein expression and solubility were measured by detecting the specific activities of the ^{14}C-Leu and ^{35}S-Met radioisotopes (RI), using a BAS 2000 scanner (Fuji). The rate of dissolution was estimated as described below

$$\text{Rate of dissolution (\%)} = (\text{signal intensity of soluble fraction}/\text{signal intensity of whole sample}) \times 100 \quad (1)$$

When the rate of dissolution was >60%, the protein was regarded as being highly soluble. In contrast, a protein with the rate of dissolution of <40% was regarded as an insoluble protein.

Data sets

We prepared two data sets, based on the number of experiments per sequence, for the statistical analysis. One group, designated as 'data set_S' ('S' means single), comprised sequences for which the protein expression and solubility were experimentally assessed one time. The other group, designated as 'data set_M' ('M' means multiple), comprised sequences for which the experiments were conducted two or more times.

The two data sets were constructed as follows (Fig. 1). First, the clones bearing the same sequence among the experimental expression data were selected (Fig. 1, Step1). The redundancy checks were executed on the nucleotide level and the amino acid level for protein expression and solubility, respectively. If the same sequences existed, they were regarded as one sequence. In the case of protein expression in the *in vivo E. coli*, 474 out of 17,739 clones were redundant, and 227 sequences remained after checking the sequence redundancy. Next, the sequences for which the experimental results were not the same were removed, to use only the reproducible data (Fig. 1, Step2). In the case of protein expression in the *in vivo E. coli*, 44 sequences were removed. The representative sequences were

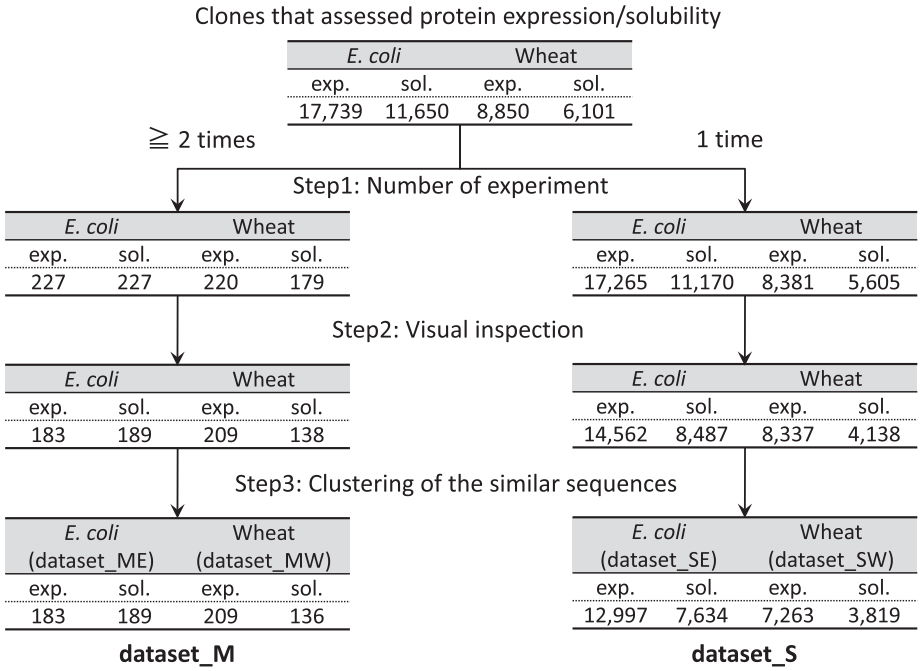

Fig. 1 Construction process of data set. Each table represents the number of data set size in each step.

Table I. Data set sizes for statistical analyses.

		Expression		Solubility	
Expression system	**Data set**	**Positive (%)**	**Negative (%)**	**Positive (%)**	**Negative (%)**
E. coli	Data set_ME	113 (61.7)	70 (38.3)	71 (37.6)	118 (62.4)
	Data set_SE	7631 (58.7)	5366 (41.3)	2725 (35.7)	4909 (64.3)
Wheat Germ	Data set_MW	208 (99.5)	1 (0.5)	86 (63.2)	50 (36.8)
	Data set_SW	7062 (97.2)	201 (2.8)	2653 (69.5)	1166 (30.5)

Numbers in parentheses signify ratios of positive data and negative data for respective data sets.

then selected in each data set, to avoid the bias of similar sequences (Fig. 1, Step3). The sequences with pair-wise sequence identity of >80%, using CD-hit (*19*), having similar length >80% were assumed to be in a cluster. The longest sequence in each cluster was selected as the representative sequence of each cluster. This collection of sequences was defined as data set_M. On the other hand, data set_S was constructed from the data from which the redundant clones from the expression data had been removed. In the case of protein expression in the *in vivo E. coli*, 17,265 (=17,739–474) sequences were used. The data that showed a smeared band were removed by visual inspection (Fig. 1, Step2) (see 'Results' section). In the case of protein expression in the *in vivo E. coli*, 2703 sequences were removed. Next, in the same manner as for data set_M, the representative sequences were selected from each cluster consisting of similar sequences (Fig. 1, Step3). This collection of sequences was defined as data set_S. The data set size is shown in Table I.

In this study, data set_M was used for estimating the features associated with the protein expression and solubility; data set_S was used for assessing whether a set of selected features corresponds to the general characteristics on a genomic scale. The initial letter of the expression system was added to the end of the data set name, to distinguish them. For example, 'data set_SE' consists of the sequences for which experimental evaluations were performed one time in the *in vivo E. coli* expression system.

Estimation of the features associated with protein expression/solubility

We defined 437 features to investigate the factors associated with protein expression/solubility in the two kinds of expression systems. The features were divided into two groups, based on the information used for producing them, except for the size of the polypeptide.

The first group was derived from sequence information, from both the nucleotides and amino acids. The nucleotide information included the occurrence frequencies of four single nucleotides, 64 codons and the GC contents. Similarly, the amino acid information contained the occurrence frequencies of 20 single amino acids and the property groups, defined by their chemical properties (eight groups: [GALVI][FYW][ST][DE][NQ][RHK][CM][P]) and physical properties (five groups: [GAVLIP][FWY][STCMNQ][DE][RKH]) (Supplementary Table SI). Additionally, the repeat was defined as the maximum number of consecutive same amino acids or property groups. The values of these features were computed for the entire chains and both terminal regions, defined as 60 bases (meaning 20 amino acid residues), because modification of the terminal regions influences protein expression and solubility (*2–4*). The use of a His-tag fusion raises the possibility that the features in the N-terminal region of the *in vivo E. coli* expression system and the C-terminal region in the wheat germ cell-free expression system may not be evaluated properly. We considered the His-tag to have the same influence on any sequences, since we conducted the protein expression experiments under the same conditions. Therefore, we evaluated them under this hypothesis. In total, the first group was composed of 396 features.

The second group was derived from structural information, obtained with several prediction using amino acid information. The structural information included the secondary structures—α-helix, β-sheet and others predicted by PHD (*20*)—along with the transmembrane regions [predicted using TMHMM (*21*)] and the disordered regions [predicted using POODLE-L (*22*)]. For the secondary structures, the ratio of each element to the entire chain was computed. For the disordered regions, their number of occurrences, lengths and proportions in relation to the entire chain were computed. For the transmembrane regions, the number of occurrences in the entire chain was computed. The structure information also included the occurrence frequencies of single amino acids and the same property groups on the protein surface. The accessible surface area was predicted using RVPnet (*23*). In total, the second group included 40 features.

We estimated which features are associated with protein expression/solubility by analyzing data set_ME and data set_MW. For all features, the statistical difference between positive and negative data was determined using the Student's *t*-test. The positive data of protein expression and solubility mean that a clear band was found in the whole cell sample and the soluble fraction sample. The negative data signify the opposite. A difference of $P<0.05$ was considered significant.

Assessing the generality of the features

To evaluate whether the set of features selected in the previous section corresponds to the general characteristics of protein expression/solubility on a genome-scale in the two expression systems, we built a statistical model that distinguishes between overexpression and low expression, using sequence information only. Similarly, a statistical model to discriminate between soluble and insoluble proteins was built as well. In this study, we applied the random forest (RF) algorithm (*24*) to produce the statistical models.

First, the sequence in the training and evaluation data set was expressed as a multi-dimensional vector that defined the selected features in the previous section as descriptors. The numbers of elements in a vector were 64 and 45 for protein expression and solubility in the *in vivo E. coli* expression system, respectively (see 'Results' section). In contrast, the sequence was expressed by 32 elements in the wheat germ cell-free expression system (see 'Results' section). The statistical models were then built by training data sets. The default values were used as the RF parameters.

The classification abilities of the statistical models for both expression systems were estimated, using two kinds of evaluation methods. One method was a 5-fold cross validation test using data set_M only. The other method was an expanded test. The statistical models were constructed using data set_M. The classification abilities of these models were then estimated, using data set_S. Finally, the classification abilities obtained from the two evaluation methods were compared. Moreover, in order to validate the features, the classification abilities of these models were compared with that of the Wilkinson and Harrison model (*11*). The model was used to predict the *in vivo* solubility of recombinant proteins in *E. coli*:

$$\mathrm{CV} = 15.43\frac{\mathrm{N}+\mathrm{G}+\mathrm{P}+\mathrm{S}}{n} - 29.56\left|\frac{(\mathrm{R}+\mathrm{K})-(\mathrm{D}+\mathrm{E})}{n} - 0.03\right| + 1.71,$$

where N, G, P, S, R, K, D and E are the numbers of asparagines, glycines, prolines, serines, arginines, lysines, aspartic acids and glutamic acids, respectively and n is the total number of residues in the sequence. If $\mathrm{CV}<0$, then the protein is predicted to be soluble. If $\mathrm{CV}>0$, then the protein is predicted to be insoluble.

Results

Comprehensive assessment of protein expression/solubility of human full-length cDNA in two expression systems

The human full-length cDNA was expressed in the two expression systems. The results were analysed

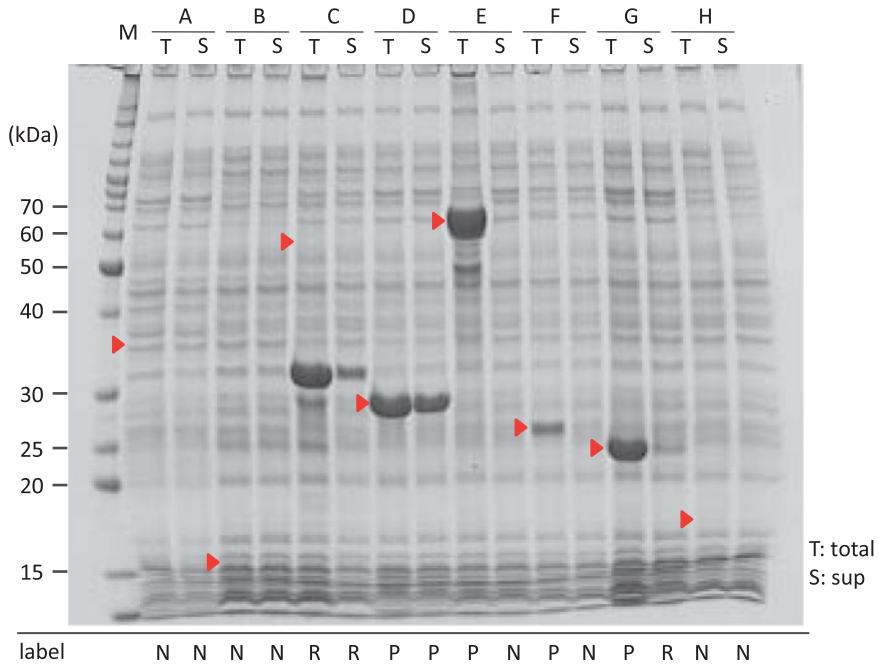

Fig. 2 Example of an SDS–PAGE analysis for eight proteins expressed in the *in vivo E. coli* expression system. M and A–H, respectively show molecular weight markers and samples. The T and S lanes, respectively show samples obtained from whole cell samples and soluble fraction samples. The red triangles represent the expected positions calculated from the molecular weights. P and N in the label signify positive and negative data, respectively. R in the label shows data removed from statistical analyses.

using SDS–PAGE (Fig. 2). The gels containing the fractionated proteins expressed in the wheat germ cell-free expression system can be seen at the site HGPD (http://riodb.ibase.aist.go.jp/hgpd/cgi-bin/index.cgi) (*18*). When a clear band is present at the expected position calculated from the molecular weight, such as in lane T of sample D in Fig. 2, the data are considered to be positive. However, when an expected band in SDS–PAGE cannot be detected, such as that in lane T of sample A in Fig. 2, the data are considered to be negative. Data were removed from the following analysis if a smeared band (lane S of sample G in Fig. 2) was observed or a clear band existed at an unexpected position (lane T of sample C in Fig. 2), in order to avoid ambiguity in the experimental data. When the SDS–PAGE results were visually inspected to check the protein expression in the *in vivo E. coli* expression system, 44 of 227 raw data sets were removed between the multiple measurements. Similarly, 16.7 and 23.0% of the raw data were excluded, respectively, from the protein solubility in the *in vivo E. coli* and the wheat germ cell-free expression systems.

The sizes of data set_ME and data set_MW are smaller than those of data set_SE and data set_SW (Table I), but data set_ME and data set_MW are more reliable experimental data. In the *in vivo E. coli* expression system, ~60 and 35% of the proteins, respectively, were expressed and soluble. In contrast, almost all of the proteins were expressed in the wheat germ cell-free expression system: ~65% of the proteins were soluble (Table I). The wheat germ cell-free expression system exhibited higher performance in obtaining soluble proteins. For the wheat germ cell-free expression system, only the protein solubility data were used in the following statistical analyses.

Estimation of the features associated with protein expression/solubility in the two expression systems

The sizes of the polypeptides used to assess the protein expression/solubility experimentally in the *in vivo E. coli* expression system were investigated (Fig. 3A). The average size of the overexpressed polypeptides was significantly longer ($P < 0.05$) than that of the polypeptides with low expression, but no statistically significant difference was found between the sizes of the soluble and insoluble polypeptides. Conversely, in the wheat germ cell-free expression system, the average size of the insoluble polypeptides was significantly longer than that of the soluble polypeptides (data not shown).

Similarly, some sequence and structural features associated with protein expression/solubility were identified from statistical analyses of data set_ME and data set_MW (Fig. 4). In this study, data set_M is not suitable for analyzing the nucleotide information associated with the protein solubility, because data set_M of solubility included sequences that are not identical on the nucleotide level. Consequently, the analysis of the nucleotide information was performed only for the protein expression.

From the perspective of nucleotide information, no GC content or single nucleotide was selected in the *in vivo E. coli* expression system, but 18 out of 61 codons were chosen to have significant contribution to protein expression. Only three rare-frequency codons in the *E. coli* genetic code, among eight tested, passed the Student's *t*-test having significant

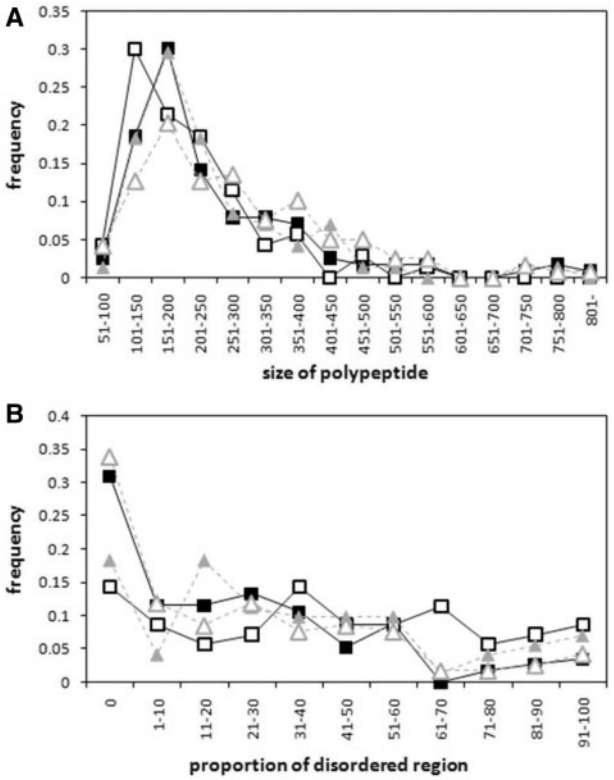

Fig. 3 Distribution of (A) polypeptide sizes and (B) disordered regions in data set_ME. The black squares and grey triangles signify protein expression and solubility, respectively. The filled symbols represent positive data, whereas the open symbols show negative data. The horizontal and vertical axis, respectively show the size and the frequency of the polypeptides.

effect on protein expression. Although it has been suggested that the codon usage influences protein expression (*25–27*), little correlation between rare codons and protein expression was detected in this study. The discrepancy might be explained by the fact that the data set does not include point mutation experiments that change low-usage codons into high-usage ones. Therefore, the estimation suggests the possibility that the influence of rare codons cannot be evaluated. In addition, many selected features are corresponding to amino acids that are encoded by several codons. These observations are the same as those reported by Welch *et al.* (*28*).

Regarding the amino acid information, in the *in vivo E. coli* expression system, the number of features that passed the Student's *t*-test is larger for protein solubility than for protein expression. Particularly, there were many features related to protein solubility in the C-terminal region. Charged residues have a positive effect on both protein expression and solubility, but aromatic residues have a negative effect. In addition, a sulfur-containing residue influences only the protein solubility. In the wheat germ cell-free expression system, the number of selected features is smaller than that in the *in vivo E. coli* expression system. Specifically, the presence of the charged and sulfur-containing residues has little effect on protein solubility. Non-polar residues show the opposite effect in the *in vivo E. coli* expression system.

Regarding the structural information, in the *in vivo E. coli* expression system, the number of features that passed the Student's *t*-test is larger for protein expression than for protein solubility. Statistical analyses revealed that the difficulty of expressing a protein tends to increase in the presence of more disordered regions (Fig. 3B). In contrast, the secondary structure has no effect on protein expression/solubility.

In the wheat germ cell-free expression system, the number of structural features that passed the Student's *t*-test is smaller than that in the *in vivo E. coli* expression system, along with the number of sequence features. In this study, we also examined the correlation between the protein expression/solubility and the number of folded domains predicted by DOMpro (*29*). No significant relation was found (data not shown). This is because more than half of the proteins in our data set have multiple domains, and it is difficult to estimate the number of domains from amino acid information. For that reason, our data set might be unsuitable for analyzing the relation between the number of domains and the protein expression/solubility.

In this study, the definition of the terminal region was 60 nt. To lend credence to the analysis, we estimated the important parameters using new definitions of the terminal region, 30 and 90 nt, and compared them. A strong relationship between protein expression and the presence of rare-frequency codons was not detected in the *in vivo E. coli* expression system, although the some of codons having statistically significant difference changed depending on the length of the terminal region. For the amino acid information, similar features passed the Student's *t*-test. Overall, the results indicated that the tendencies of the important features were the same under any conditions (Supplementary Fig. S1).

Generality of the features

To assess the generality of the features selected in the previous section, we built statistical models that classified the overexpressed proteins and the soluble proteins, based on the sequence information. Then, we compared the classification abilities of the two models produced from the different data sets.

In the *in vivo E. coli* expression system, using data set_ME, the statistical models' abilities were estimated using a 5-fold cross validation test (Table II). The accuracies, which signified the proportions of correct prediction, were 77.6 and 71.4%, respectively, for protein expression and solubility in the *in vivo E. coli* expression system. These values were almost identical to those of the models using all features, presented in the 'Materials and Methods' section. Next, we built a statistical model trained using data set_ME, and evaluated its classification ability using data set_SE (Table II). Based on the accuracy (*Acc.*), the classification ability for data set_SE was slightly lower than that for data set_ME. This difference in the ability between the two models is considered to reflect the experimental error that data set_SE includes, because it was much smaller than the experimental error rate inferred from the analysis of data set_ME. Therefore, the two kinds

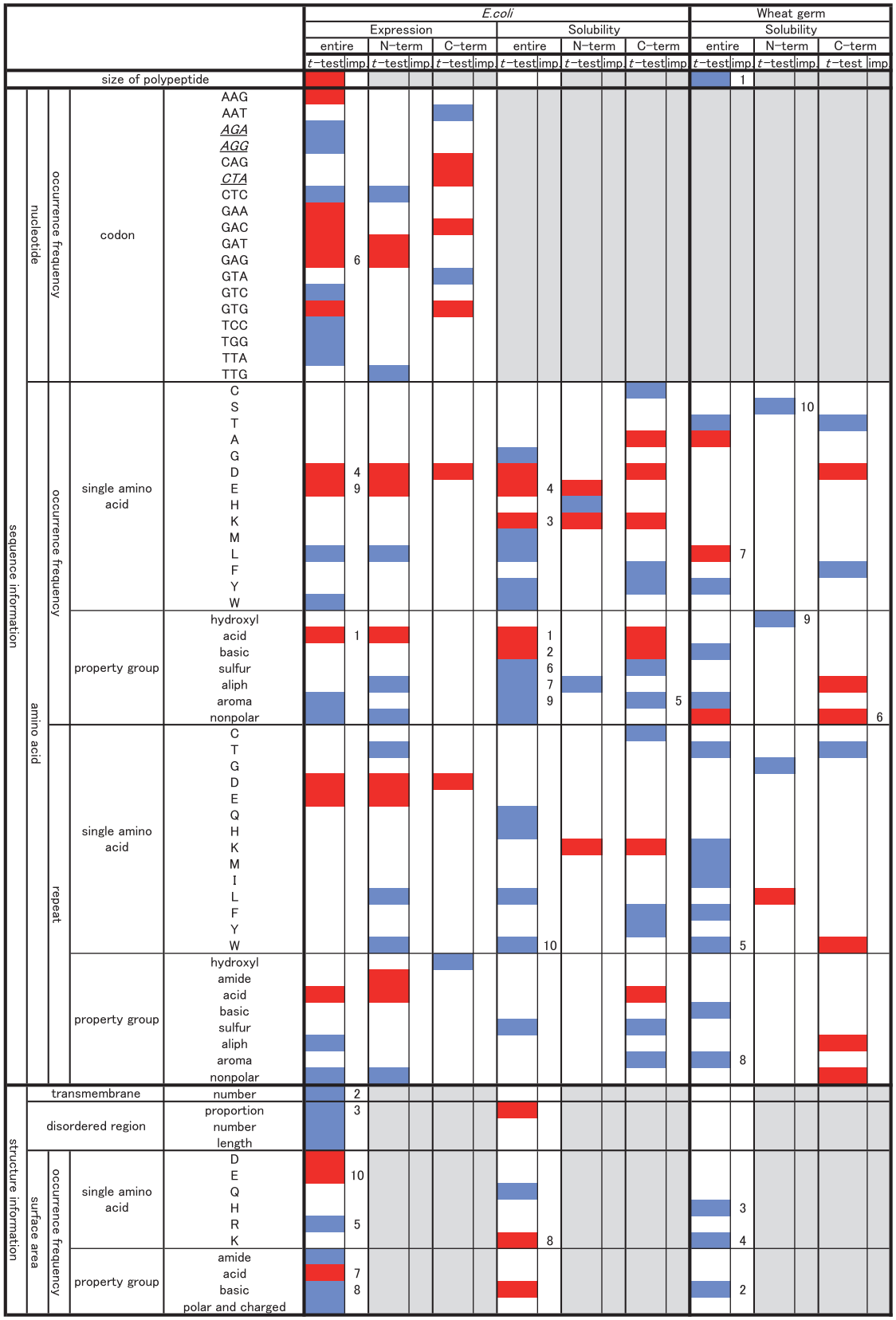

Fig. 4 Comparison of features associated with protein expression/solubility in the two expression systems. Only the features with statistically significant differences detected by the Student's *t*-test are listed in this figure. *t*-test shows the results of the Student's *t*-test. Red signifies the features that have a positive effect on protein expression or solubility, and blue shows the features that have a negative influence. White denotes features not found to have a statistically significant difference; grey shows that no statistical test was done. Entire, N-term, and C-term signify features computed using the entire chain, the N-terminal region, and the C-terminal region, respectively. 'imp.' shows the ranking of features that contribute to protein expression/solubility. The ranking was determined for three categories: protein expression in the *in vivo E. coli*, protein solubility in the *in vivo E. coli*, and protein solubility in wheat germ. The rare-frequency codons in the genetic code of *E. coli* are italicized and underlined.

Table II. Classification abilities of protein expression/solubility in the two expression systems.

		Expression			Solubility		
Expression system	**Data set**	**Recall**	**Precision**	***Acc.***	**Recall**	**Precision**	***Acc.***
Escherichia coli	Data set_ME	0.807	0.838	0.776	0.673 (0.296)	0.468 (0.429)	0.714 (0.587)
	Data set_SE	0.876	0.702	0.694	0.424 (0.295)	0.551 (0.432)	0.671 (0.610)
Wheat Germ	Data set_MW	–	–	–	0.736 (0.302)	0.853 (0.897)	0.714 (0.537)
	Data set_SW	–	–	–	0.892 (0.294)	0.718 (0.846)	0.682 (0.469)

The prediction results were classified into four categories: TP is the number of true positives, which is defined as the number of correctly predicted positives. Similarly, FP, TN and FT denote the numbers of false positives, which are defined, respectively, as: negatives that were incorrectly predicted as positives, the number of true negatives, which are defined as correctly predicted negatives, and the number of false negatives, which are defined as positives incorrectly predicted as negatives. Recall and Precision were defined as [=TP/(TP + FP)] and [=TP/(TP + FN)], respectively. *Acc.* [=(TP + TN)/(TP + TN + FP + FN)] represents the proportion of correctly identified positives plus negatives. A hyphen shows that no statistical test was done. The figures in parentheses signify the results of Wilkinson and Harrison model.

of statistical models are considered to have comparable classification abilities. A similar tendency was observed for the wheat germ cell-free expression system (Table II).

These results indicate that the characteristics of the two pairs of data sets—data set_ME and data set_SE, and data set_MW and data set_SW—are similar. Consequently, the features selected in the previous section represent the general characteristics of the protein expression/solubility in each expression system. Therefore, these features in each expression system are considered to be the minimal sets of features associated with protein expression/solubility.

The RF model can estimate the importance of features more simply than commonly-used machine learning methods, such as support vector machine (SVM) (*30*). We estimated the 10 important features based on the mean degrees of *Acc.* (Fig. 4). A comparison of the two expression systems revealed that the key features associated with protein solubility are different. The features related to charge occupied the top rank in the *in vivo E. coli* expression system, while they are hardly found in the wheat germ cell-free expression system.

Discussion

We identified a minimal set of features associated with protein expression/solubility in two expression systems, by the application of two statistical analyses. A comparison of the features associated with protein expression/solubility in the *in vivo E. coli* expression system revealed their different influences. In short, the 'structural information' has a strong influence at the protein expression stage, whereas the amino acid 'sequence information' exerts effects at the protein solubility stage (Fig. 4). These observations suggest a mechanism for yielding a soluble protein in the *in vivo E. coli* expression system. Regarding the protein expression stage, increased numbers of disordered regions and transmembrane regions act to prevent protein expression. Experiments with individual proteins have also shown that disordered regions affect protein expression (*31*). In addition, the presence of charged residues on the protein surface has a positive effect on protein expression. These are common characteristics of globular proteins. For this reason, it may be important for a protein to fold into the proper structure at the protein expression stage. In contrast, the amino acid sequence information is important for the solubility stage. The statistical analysis indicated that an abundance of charged residues in the C-terminal region leads increase of protein solubility. In a study of an individual protein, Kato *et al.* (*32*) reported that adding several arginine residues to the C-terminus of BPTI increased its solubility by preventing aggregation. Therefore, it may be important for a protein not to aggregate at the protein solubility stage.

A comparison of the two expression systems revealed two important points. One is that the number of features associated with protein solubility in the wheat germ cell-free expression system is smaller than that in the *in vivo E. coli* expression system (Fig. 4). This observation implies that the wheat germ cell-free expression system is less sensitive to the various sequence and structural features of a protein, corresponding to the fact that the wheat germ cell-free expression system has a higher success rate than the *in vivo E. coli* expression system in generating soluble proteins (Table I). The other is that the key features in the two expression systems are different. In the *in vivo E. coli* expression system, the charge is important, but it has little influence on the solubility in the wheat germ cell-free expression system. The differences between the features in the two expression systems might be related to the translation speed (*33*). In general, the speed is faster in bacteria than in eukaryotes. The charged residues are considered to be important for partial fording in the *in vivo E. coli* expression system.

The minimal sets of features associated with protein expression/solubility in the two expression systems are useful to screen targets in protein expression experiments. When the statistical model that used the minimal set of features identified in this study was compared with Wilkinson's statistical model (*11*) to predict the *in vitro* solubility of a recombinant protein in an *E. coli* expression system, the *Acc.* of our model for data set_SE was 6.1% higher than that of Wilkinson's model.

Supplementary Data

Supplementary Data are available at *JB* Online.

Acknowledgements

We thank the members of the molecular function team at the Computational Biology Research Center (CBRC) and Dr Kuroda at Tokyo University of Agriculture and Technology (TUAT) for helpful discussions and advice.

Funding

The Okawa Foundation for Information and Telecommunications.

Conflict of interest

None declared.

References

1. Clark, E.D.B. (1998) Refolding of recombinant proteins. *Curr. Opin. Biotechnol.* **9**, 157–163
2. Doray, B., Chen, C.D., and Kemper, B. (2001) N-terminal deletions and His-tag fusions dramatically affect expression of cytochrome p450 2C2 in bacteria. *Arch. Biochem. Biophys.* **393**, 143–153
3. Sati, S.P., Singh, S.K., Kumar, N., and Sharma, A. (2002) Extra terminal residues have a profound effect on the folding and solubility of a Plasmodium falciparum sexual stage-specific protein over-expressed in *Escherichia coli*. *Eur. J. Biochem.* **269**, 5259–5263
4. Kapust, R.B. and Waugh, D.S. (1999) *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Sci.* **8**, 1668–1674
5. Tresaugues, L., Collinet, B., Minard, P., Henckes, G., Aufrere, R., Blondeau, K., Liger, D., Zhou, C.Z., Janin, J., Van Tilbeurgh, H., and Quevillon-Cheruel, S. (2004) Refolding strategies from inclusion bodies in a structural genomics project. *J. Struct. Funct. Genomics* **5**, 195–204
6. Andersen, D.C. and Krummen, L. (2002) Recombinant protein expression for therapeutic applications. *Curr. Opin. Biotechnol.* **13**, 117–123
7. Kramer, G., Kudlicki, W., Hardesty, B., Higgens, S.J., and Hames, B.D. (1999) Cell-free coupled transcription-translation systems from *Escherichia coli*. In *Protein Expression: A Practical Approach* (Higgens, S.J. and Hames, B.D., eds.), pp. 201–223, Oxford University Press, Oxford
8. Clemens, M.M., Prujin, G.J., Higgens, S.J., and Hames, B.D. (1999) Protein synthesis in eukaryotic cell-free systems. In *Protein Expression. A Practical Approach* (Higgens, S.J. and Hames, B.D., eds.), pp. 129–165, Oxford University Press, Oxford
9. Goshima, N., Kawamura, Y., Fukumoto, A., Miura, A., Honma, R., Satoh, R., Wakamatsu, A., Yamamoto, J., Kimura, K., Nishikawa, T., Andoh, T., Iida, Y., Ishikawa, K., Ito, E., Kagawa, N., Kaminaga, C., Kanehori, K., Kawakami, B., Kenmochi, K., Kimura, R., Kobayashi, M., Kuroita, T., Kuwayama, H., Maruyama, Y., Matsuo, K., Minami, K., Mitsubori, M., Mori, M., Morishita, R., Murase, A., Nishikawa, A., Nishikawa, S., Okamoto, T., Sakagami, N., Sakamoto, Y., Sasaki, Y., Seki, T., Sono, S., Sugiyama, A., Sumiya, T., Takayama, T., Takayama, Y., Takeda, H., Togashi, T., Yahata, K., Yamada, H., Yanagisawa, Y., Endo, Y., Imamoto, F., Kisu, Y., Tanaka, S., Isogai, T., Imai, J., Watanabe, S., and Nomura, N. (2008) Human protein factory for converting the transcriptome into an in vitro-expressed proteome. *Nat. Methods* **5**, 1011–1017
10. He, M. (2008) Cell-free protein synthesis: applications in proteomics and biotechnology. *N. Biotechnol.* **25**, 126–132
11. Wilkinson, D.L. and Harrison, R.G. (1991) Predicting the solubility of recombinant proteins in *Escherichia coli*. *Biotechnology* **9**, 443–448
12. Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Savchenko, A., Cort, J.R., Booth, V., Mackereth, C.D., Saridakis, V., Ekiel, I., Kozlov, G., Maxwell, K.L., Wu, N., McIntosh, L.P., Gehring, K., Kennedy, M.A., Davidson, A.R., Pai, E. F., Gerstein, M., Edwards, A.M., and Arrowsmith, C.H. (2000) Structural proteomics of an archaeon. *Nat. Struct. Biol.* **7**, 903–909
13. Bertone, P., Kluger, Y., Lan, N., Zheng, D., Christendat, D., Yee, A., Edwards, A.M., Arrowsmith, C.H., Montelione, G.T., and Gerstein, M. (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
14. Goh, C.S., Lan, N., Douglas, S.M., Wu, B., Echols, N., Smith, A., Milburn, D., Montelione, G.T., Zhao, H., and Gerstein, M. (2004) Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis. *J. Mol. Biol.* **336**, 115–130
15. Luan, C.H., Qiu, S., Finley, J.B., Carson, M., Gray, R.J., Huang, W., Johnson, D., Tsao, J., Reboul, J., Vaglio, P., Hill, D.E., Vidal, M., Delucas, L.J., and Luo, M. (2004) High-throughput expression of *C. elegans* proteins. *Genome Res.* **14**, 2102–2010
16. Idicula-Thomas, S. and Balaji, P.V. (2005) Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in *Escherichia coli*. *Protein Sci.* **14**, 582–592
17. Niwa, T., Ying, B.W., Saito, K., Jin, W., Takada, S., Ueda, T., and Taguchi, H. (2009) Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of *Escherichia coli* proteins. *Proc. Natl Acad. Sci. USA* **106**, 4201–4206
18. Maruyama, Y., Wakamatsu, A., Kawamura, Y., Kimura, K., Yamamoto, J., Nishikawa, T., Kisu, Y., Sugano, S., Goshima, N., Isogai, T., and Nomura, N. (2009) Human Gene and Protein Database (HGPD): a novel database presenting a large quantity of experiment-based results in human proteomics. *Nucleic Acids Res.* **37**, 762–766
19. Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659
20. Rost, B. (1996) PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol.* **266**, 525–539
21. Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580
22. Hirose, S., Shimizu, K., Kanai, S., Kuroda, Y., and Noguchi, T. (2007) POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions. *Bioinformatics* **23**, 2046–2053
23. Ahmad, S., Gromiha, M.M., and Sarai, A. (2003) RVP-net: online prediction of real valued accessible surface area of proteins from single sequences. *Bioinformatics* **19**, 1849–1851

24. Liaw, A. and Wiener, M. (2002) Classification and Regression by randomForest. *R News* **2**, 18–22
25. Makrides, S.C. (1996) Strategies for achieving high-level expression of genes in *Escherichia coli*. *Microbiol. Rev.* **60**, 512–538
26. Drummond, D.A. and Wilke, C.O. (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**, 341–352
27. Lorimer, D., Raymond, A., Walchli, J., Mixon, M., Barrow, A., Wallace, E., Grice, R., Burgin, A., and Stewart, L. (2009) Gene composer: database software for protein construct design, codon engineering, and gene synthesis. *BMC Biotechnol.* **9**, 36
28. Welch, M., Govindarajan, S., Ness, J.E., Villalobos, A., Gurney, A., Minshull, J., and Gustafsson, C. (2009) Design parameters to control synthetic gene expression in *Escherichia coli*. *PLoS One* **4**, e7002
29. Cheng, J., Sweredoski, M., and Baldi, P. (2006) DOMpro: protein domain prediction using profiles, secondary structure relative solvent accessibility, and recursive neural network. *Data Min. Knowl. Disc.* **13**, 1–10
30. Breinman, L. (2001) Random forests. *Mach. Learn.* **45**, 5–32
31. Quevillon-Cheruel, S., Leulliot, N., Gentils, L., van Tilbeurgh, H., and Poupon, A. (2007) Production and crystallization of protein domains: how useful are disorder predictions? *Curr. Protein Pept. Sci.* **8**, 151–160
32. Kato, A., Maki, K., Ebina, T., Kuwajima, K., Soda, K., and Kuroda, Y. (2007) Mutational analysis of protein solubility enhancement using short peptide tags. *Biopolymers* **85**, 12–18
33. Siller, E., DeZwaan, D.C., Anderson, J.F., Freeman, B.C., and Barral, J.M. (2010) Slowing bacterial translation speed enhances eukaryotic protein folding efficiency. *J. Mol. Biol.* **396**, 1310–1318