Multiple Post-translational Modifications Affect Heterologous Protein Synthesis*  

Received for publication, March 26, 2012, and in revised form, May 29, 2012. Published, JBC Papers in Press, June 6, 2012, DOI 10.1074/jbc.M112.366351

Alexander A. Tokmakov†‡§, Atsushi Kurotani†, Tetsuo Takagi†, Mitsutoshi Toyama†, Mikako Shirouzu†, Yasuo Fukami‡¶, and Shigeyuki Yokoyama†**

From the †RIKEN Systems and Structural Biology Center and ‡RIKEN Plant Science Center, Yokohama 230-0045, the **Graduate School of Science, University of Tokyo, Bunkyo, Tokyo 113-0033, the §Graduate School of Science and the ¶Research Center for Environmental Genomics, Kobe University, Nada 657-8501, Japan

Background: Post-translational modifications (PTMs) affect protein folding.
Results: Statistically significant correlations are revealed between the yield of heterologous protein expression and the presence of multiple PTM sites bioinformatically predicted in the expressed sequences.
Conclusion: Predicting potential PTMs in polypeptide sequences can help optimize heterologous protein synthesis.
Significance: Correlations revealed provide insights into the role of specific PTMs in protein stability and solubility.

Post-translational modifications (PTMs) are required for proper folding of many proteins. The low capacity for PTMs hinders the production of heterologous proteins in the widely used prokaryotic systems of protein synthesis. Until now, a systematic and comprehensive study concerning the specific effects of individual PTMs on heterologous protein synthesis has not been presented. To address this issue, we expressed 1488 human proteins and their domains in a bacterial cell-free system, and we examined the correlation of the expression yields with the presence of multiple PTM sites bioinformatically predicted in these proteins. This approach revealed a number of previously unknown statistically significant correlations. Prediction of some PTMs, such as myristoylation, glycosylation, palmitoylation, and disulfide bond formation, was found to significantly worsen protein amenability to soluble expression. The presence of other PTMs, such as aspartyl hydroxylation, C-terminal amidation, and Tyr sulfation, did not correlate with the yield of heterologous protein expression. Surprisingly, the predicted presence of several PTMs, such as phosphorylation, ubiquitination, SUMOylation, and prenylation, was associated with the increased production of properly folded soluble proteins. The plausible rationales for the existence of the observed correlations are presented. Our findings suggest that identification of potential PTMs in polypeptide sequences can be of practical use for predicting expression success and optimizing heterologous protein synthesis. In sum, this study provides the most compelling evidence so far for the role of multiple PTMs in the stability and solubility of heterologously expressed recombinant proteins.

Heterologous protein synthesis is widely employed for production of recombinant proteins. Rather commonly, the eukaryotic proteins and their domains are expressed in the Escherichia coli bacterial cells (1–3) or cell-free extracts (3–6). However, only a minor fraction of all heterologous proteins can be successively expressed in this host system. The correct folding of eukaryotic proteins in a bacterial host remains a great challenge for their synthesis. To address this issue, eukaryotic expression systems based on the use of yeast (7), wheat germ (8), insect cells (9), rabbit reticulocytes (10), tumor HeLa cells (11), and hybridoma (12) have been developed; however, they tend to produce the proteins in relatively low yields that are often insufficient for structural and/or functional studies.

At present, the factors determining expression success of heterologous protein synthesis are poorly understood. Various physicochemical and structural features of amino acid sequences have been implicated as determining factors of soluble protein expression in a bacterial host (2, 13–15). Computational approaches to predict protein propensity for expression and solubility have been developed (13, 16, 17). Most recently, 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 have been reported (18).

Importantly, many eukaryotic proteins require multiple PTMs2 to reach a native, biologically active conformation. PTMs can significantly change the integral characteristics of proteins that affect their stability and solubility, such as charge, hydrophobicity, solvent accessibility, etc. Thus, in addition to the physicochemical and structural features of amino acid sequences, PTMs should also be considered as the major determinants of successful protein synthesis. The presence of specific sequence motifs encrypting modification sites in target proteins and the capacity of the employed expression system to carry out these modifications are the prerequisites for PTM occurrence. Notably, the bacterial expression systems have only a limited capacity for PTMs. The inability of heterologous

---

* This work was supported by the RIKEN Structural Genomics/Proteomics Initiative and National Project on Protein Structural and Functional Analyses, Ministry of Education, Culture, Sports, Science and Technology of Japan.
† This article contains supplemental Figs. S1–S34 and Tables S1–S3.
‡ To whom correspondence should be addressed: The Research Center for Environmental Genomics, Kobe University, Rokko dai 1-1, Nada, Kobe, Japan 657-8501. Tel.: 81-78-803-5953; E-mail: tokmak@phoenix.kobe-u.ac.jp.

The abbreviation used is: PTM, post-translational modification.
Role of PTMs in Heterologous Protein Expression

Protein synthesis to support all PTMs that a protein requires to fold is considered to be a major factor behind the low expression yield and pure solubility of many recombinant proteins. Thus, eukaryotic proteins produced in the bacterial expression systems are quite often misfolded or unfolded, leading to their deposition into insoluble aggregates or fast degradation. Quantifying the yields of soluble and insoluble expression provides a clue to the evaluation of folding and stability of the synthesized polypeptide products. In contrast to the numerous analyses addressing the correlations between heterologous protein synthesis and physicochemical properties of the expressed polypeptides, no systematic study concerning the correlations of heterologous protein expression with multiple PTMs has been presented.

In this study, to gain an insight into the role of multiple PTMs in the stability and solubility of heterogeneously synthesized proteins, we evaluated the expression of 1488 human proteins and their domains in a bacterial cell-free system, and we examined the correlations of the reaction yield with the presence of multiple PTM sites bioinformatically predicted in the expressed sequences. The obtained information was accumulated in the database of the structural genomics/proteomics project “Protein 3000,” launched in Japan in the year 2002 with the aim to determine the structures of 3000 proteins using NMR and x-ray analyses (19–21).

**EXPERIMENTAL PROCEDURES**

**Protein Expression and Its Evaluation**—Cell-free expression of human proteins and their domains was carried out in the *E. coli* S30 extracts as described previously (18). The main steps of the heterologous protein production are shown schematically in Fig. 1. All proteins were expressed under the same uniform set of conditions to minimize the influence of sequence-independent factors. No solubility-enhancing compounds or chaperones were included in the reaction mixture of protein synthetic reaction. All synthesized polypeptide products universally comprised the N-terminal poly-His tag. Proteins that were expressed at a lower molecular weight than expected were considered to be nonexpressed. Scores A, C, and N were 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, nonexpressed proteins (expression yield of <0.1 mg/ml). Notably, score A provided the upper estimation of soluble protein expression, and score C gave the lower estimation of insoluble expression, because the procedure of centrifugation at 10,000 × *g* used to separate soluble and insoluble proteins in this study cannot discriminate between truly soluble proteins and small protein aggregates.

**Dataset**—A complete dataset of expressed polypeptides included 1488 nonredundant human amino acid sequences. Similar sequences were filtered out at 95% identity. The length of analyzed sequences did not exceed 350 amino acids. Proteins of different functional and structural classes were represented in the dataset.

**Bioinformatics Prediction of PTMs**—The sites of PKC phosphorylation, CK2 phosphorylation, tyrosine phosphorylation, PKA phosphorylation, myristoylation, asparagine glycosylation, amidation, prenylation, sulfation, and aspartyl hydroxylation were predicted with the PROSITE scanning tool PS.Scan (22, 23). The service searches multiple protein sequences with multiple patterns from the PROSITE database. Sites of S-palmitoylation were predicted with the CSS-Palm 3.0 tool (24). Sites of N-terminal acetylation were predicted using the NetAcet 1.0 server (25). N-terminal methionine excision was analyzed using the Terminator 3 server (26, 27). Disulfide bonds were computed with the Dipro tool (28). Sites of ubiquitination were predicted using the predictor of protein ubiquitination UbPred (29). Sites of SUMOylation were predicted with the site-specific predictor SUMOsp 2.0 (30). The grand average of hydrophobicity (GRAVY) index was calculated using free software available at ExPASy server. Solvent accessibility was computed with the tool provided on line, and the contents of charged residues were calculated using Proteomix software (31).

**Homology Modeling**—Homology modeling of the dataset proteins harboring the sites of aspartyl hydroxylation was carried out using the protein structure homology-modeling server SWISS-MODEL (32–34). Visualization and rendering of the modeled structures was done with PyMOL (35).

**Statistical Analysis of Data**—In this study, all expressed proteins were categorized into three mutually exclusive classes as follows: soluble (A), insoluble (C), and nonexpressed (N) proteins. Thus, the expression data represented, in essence, a categorical dataset. To estimate statistical significance of the observed correlations, the categorical data analysis has been applied via implementation of the two-way contingency table (36). The Fisher’s exact *p* value was computed using an on-line tool. A confidence level of 95% was set up as the null hypothesis rejection threshold.

**RESULTS AND DISCUSSION**

**Overview of Protein Expression Pipeline**—The specific functions of individual PTMs can be explored by analyzing heterologous protein production in the expression system that does not support these PTMs. In this study, human proteins and their domains were expressed in the cell-free extracts of *E. coli*. The workflow of the heterologous protein production is presented in Fig. 1. Linear DNA templates were amplified by the two-step PCR from the nonredundant source human cDNA clones selected from the RIKEN cDNA collection. The linear PCR products thus generated were used to program batch-mode coupled transcription/translation protein synthesis in the bacterial S30 extracts under the same uniform set of conditions. Soluble and insoluble reaction products were separated by centrifugation and subjected to SDS-PAGE and protein staining to evaluate the yields of soluble and insoluble protein expression. Each experimentally expressed protein was categorized into one of the following three groups: soluble (A), insoluble (C), and nonexpressed proteins (N), as described under “Experimental Procedures.” Overall estimation of protein expression showed that the proteins of group A represented 30.9% (460); the proteins of group C represented 52.6% (782), and the proteins of group N represented 16.5% (246) of all proteins in the dataset. Similar rates of soluble expression have been reported previously for human proteins expressed in *E. coli* (37) and in cell-free bacterial extracts (18).
Role of PTMs in Heterologous Protein Expression

PTMs Whose Occurrence Negatively Correlates with Soluble Protein Expression—Predictive bioinformatics analysis carried out on the dataset of 1488 human polypeptide sequences expressed in the cell-free bacterial system revealed several PTMs associated with significantly worsened protein amenability to heterologous expression. They included Asn glycosylation, S–S bond formation, myristoylation, and palmitoylation (Fig. 2, Table 1, and supplemental Table S1). All of them were highly abundant PTMs that could be predicted repeatedly in many proteins of the studied dataset (Fig. 2, B, D, F, and H).

As expected, the predicted presence of N-glycosylation sites in the dataset sequences was found to be associated with a lower rate of soluble expression and a higher rate of insoluble expression (Fig. 2A, curves A and C). These correlations were statistically confirmed (Table 1 and supplemental Table S1). The observed tendencies are most likely related to the fact that the N-linked protein glycosylation pathway is absent from the E. coli-based system of protein synthesis used in this study. N-Linked protein glycosylation was shown to be important for the folding, stability, trafficking, and pharmacokinetics of many proteins (38). The increased folding efficiency of glycosylated proteins is attributed to the chaperone-like activity of glycans, which can be observed even when glycans are not covalently linked to proteins (39). In addition, glycosylation increases the overall hydrophilicity of nascent polypeptide chains, preventing their potential aggregation due to intermolecular hydrophobic interactions. Importantly, although the bacterial N-linked protein glycosylation pathway (Pgl) has been discovered in the e-proteobacterium Campylobacter jejuni and the characteristic enzyme of the pathway, PgL, was found in at least 49 bacterial species (40, 41), N-glycosylation does not occur in E. coli, and only the transfer of N-linked glycosylation systems to this bacterium enables the production of recombinant glycoproteins.

In accordance with our previous results (18), the predicted presence of disulfide bonds in proteins was found to be negatively correlated with soluble protein expression and positively correlated with the insoluble expression (Fig. 2C, curves A and C). It comes as no surprise, considering that the disulfide bridges effectively stabilize protein molecules by linking distant parts of polypeptide chains. However, the formation of S–S bonds in human proteins is greatly compromised in bacterial expression systems, as the reducing conditions and oxidative folding in bacteria differ from those in eukaryotes (42). It has been established that the formation of intra- and intermolecular disulfides is not possible in the reducing cytoplasm of wild-type E. coli, resulting in the aggregation of some disulfide bond-rich proteins, e.g. Fab antibody fragments (43). Practically, it is often possible to optimize the conditions for the S–S bond formation in a given polypeptide by adjusting reducing conditions of the protein synthetic reaction in the cell-free systems of heterologous protein synthesis. The pretreatment of bacterial extracts with iodoacetamide was shown to abolish the disulfide reducing activity of bacterial extracts, resulting in the recovery of oxidizing redox environment (44).

Similarly to N-glycosylation and disulfide bond formation, the predicted presence of the lipid modification sites, such as myristoylation and palmitoylation, was found to be associated with worsened soluble expression and increased rates of non-expressed and insoluble-expressed proteins (Fig. 2, E and G, Table 1, and supplemental Table S1). The most plausible explanation for the observed tendencies may be related to the fact that the amino acid sequences with the predicted lipid modification sites have far higher-than-average overall hydrophobicity, typical of the membrane-localized and membrane-associated proteins. The average values of the GRAVY parameter in the subsets of modified and unmodified proteins included −0.398 and −0.603 for myristoylation subsets and −0.355 and −0.493 for palmitoylation subsets. The highest GRAVY value in the C-ranked targets (data not shown) suggests that this parameter is primarily related to protein solubility. Previously, high surface hydrophobicity has been shown to increase protein aggregation due to intermolecular hydrophobic interactions and to worsen soluble protein expression yield in bacterial expression systems (16, 18, 45, 46). Therefore, the amino acid sequences with the predicted lipid modification sites constitute a subset of very hydrophobic proteins, which have an intrinsically low potential for soluble expression.

PTMs Whose Occurrence Does Not Correlate with Protein Expression—In addition to the abovementioned modifications, the predicted presence of aspartyl hydroxylation sites seemingly worsened expression amenability of amino acid
sequences. Indeed, prediction of this PTM decreased the rate of soluble expression and increased the proportion of nonexpressed proteins (Fig. 3A, columns A and N, respectively). Aspartyl hydroxylation was found to be a relatively low abundant PTM; only 12 proteins in the analyzed dataset displayed this feature. The categorical data analysis showed that the two subsets of data, ASX(+) and ASX(−), were not significantly different at the 95% confidence level (Table 1 and supplemental...
Table S1); therefore, this modification was categorized into the group of PTMs whose presence does not correlate with the protein expression amenability in this expression system.

Notably, the PROSITE scanning tool employed in this study for prediction of multiple PTMs, including aspartyl hydroxylation, is based on the identification of predefined PTM consensus signature patterns in linear amino acid sequences. The tool does not take into account spatial accessibility of the potential PTM sites, suggesting that at least a fraction of these sites cannot be modified as they may be located in the inaccessible parts of protein molecules. This should result in a number of false-positive predictions, obscuring the correlations between the protein expression yield and the presence of the PROSITE-predicted PTM sites. Importantly, misidentification of even a single modification site may be crucial in the case of low abundant PTMs, such as aspartyl hydroxylation, because it can bring about a statistically significant difference in the analyzed data subsets.

To address this issue, solvent accessibility of the predicted sites of aspartyl hydroxylation has been investigated. Three-dimensional structures of the 12 dataset proteins harboring this modification were built using structural homology modeling, and the locations of the modification sites were mapped in the modeled structures. Importantly, all of the predicted aspartyl hydroxylation sites were found to be located in the solvent-accessible regions of protein molecules (supplemental Fig. S1), suggesting their validity and functional relevance. This result rules out the presence of false-positive PROSITE-predicted aspartyl hydroxylation sites in the analyzed dataset, thereby reinforcing the conclusions of the categorical data analysis.

Still, the existence of negative correlation between aspartyl hydroxylation and heterologous protein synthesis cannot be completely ruled out at this stage, and the study of a more extended dataset is necessary to further clarify the effect of this PTM. The expression analysis of the ankyrin repeat domain-containing proteins, the recently identified substrates of hydroxylation by FIH hydroxylase (47), may help address this point.

About 18% of human polypeptide sequences in the analyzed dataset have been predicted to contain the amidated C terminus (Fig. 3B). Amidation prevents ionization of the C terminus, rendering it more hydrophobic, with implications for protein solubility. The exact biological role of this PTM is largely unknown. It has been suggested that C-terminal amidation may contribute to protein stability (38). The bioinformatics analysis performed in this study could not reveal any statistically significant differences in the ratios of soluble, insoluble, and nondetectable expression in the subset of proteins predicted to harbor this PTM in comparison with the subset of unmodified proteins (Fig. 3B, Table 1, and supplemental Fig. S1). This result implies
that, rather than stability and solubility, amidation may affect protein function. Indeed, the studies of several amidated peptides indicate that the amide moiety may be a key determinant of ligand-receptor interactions (48–50).

Similarly, we could not reveal any statistically significant correlations between the predicted presence of tyrosine sulfation sites and heterologous protein synthesis (Fig. 3C, Table 1, and supplemental Table S1). In total, about 29% of human polypeptide sequences in the analyzed dataset have been predicted to contain sulfated Tyr residues (Fig. 3D), which largely agrees with the previous estimation that the mouse genome may encode over 2000 Tyr-sulfated proteins (51). Tyr sulfation is often observed in bioactive peptides, such as neuropeptides, peptide hormones, conotoxins, etc., isolated from different vertebrate and invertebrate organisms (52). However, Tyr sulfation does not occur in prokaryotes and unicellular eukaryotes (51). In most cases, this modification is required for full biological activity of the peptides. Structural data indicate that sulfotyrosine is involved in hydrogen bonding networks and salt bridge interactions (53). Similarly to amidation, the lack of correlation between the predicted presence of sulfation sites and heterologous protein synthesis suggests that Tyr sulfation has little to do with protein stability and solubility in the employed expression system.

Quite expectedly, we failed to observe any statistically significant correlations between protein synthesis efficiency and the predicted presence of the N-terminal modifications, such as N-terminal methionine excision, myristoylation, and acetylation (supplemental Fig. S2 and supplemental Table S2), because all synthesized polypeptide products in this study universally composed the N-terminal poly-His tag to allow their purification. Evidently, the expression of untagged or C-terminal-tagged polypeptides should be analyzed to deduce the correlations between the presence of N-terminal modifications and the yield of heterologous protein synthesis.

PTMs Whose Occurrence Positively Correlates with Soluble Protein Expression—The bacterial cell-free expression systems do not support multiple PTMs that eukaryotic proteins require to properly fold. This is considered to be a major factor behind the low expression yield and pure solubility of many recombinant proteins. Unexpectedly, the predicted presence of several PTMs, such as prenylation, ubiquitination, SUMOylation, and phosphorylation, was found to be associated with the increased production of properly folded soluble protein. As a rule, this tendency was accompanied by a reciprocal decrease in the rate of insoluble expression.

Among these PTMs, prenylation was found to be a quite low abundant modification, and only 15 proteins in the analyzed dataset have been predicted to contain the potential prenylation sites (Fig. 4A). Nevertheless, categorical data analysis confirmed a statistically significant difference in the ratios of soluble proteins in the two expression subsets, Pre(+) and Pre(−)
Role of PTMs in Heterologous Protein Expression

(Table 1 and supplemental Table S1). The low abundance of prenylated proteins in the analyzed dataset is consistent with the previous estimate, which put the number of possible prenylated proteins in the mammalian proteome to less than 2% (54, 55). Prenyl groups were shown to be involved in protein-protein interactions through specialized prenyl-binding domains (56, 57). In addition, the long chain hydrophobic prenyl groups increase net hydrophobicity of proteins and facilitate their attachment to cell membranes. Therefore, eukaryotic proteins synthesized in a eukaryotic system, and they should be preferentially retained in the cytosolic (i.e. soluble) fraction. This may explain the increased rate of soluble expression observed in the subset of human proteins predicted to contain potential prenylation sites (Fig. 4A).

Other PTMs, whose predicted presence was found to be associated with the increased yield of soluble protein product, included such highly abundant protein modifications, as ubiquitination and SUMOylation. No SUMOylation and ubiquitination machineries are known to exist in bacteria, suggesting that the presence of ubiquitination and SUMOylation sites in amino acid sequences may be associated with the intrinsically better solubility of these sequences even in the physical absence of the modifications. We hypothesized that, most probably, it should be related to the physicochemical and/or structural characteristics of the modification sites themselves. In this connection, ubiquitination has been established to occur on lysine side chains of the acceptor substrate. Although ubiquitin attachment sites were resolved for some subsets of proteins, little sequence consensus was detected, implying that specific residues surrounding the modified lysine are not important determinants for ubiquitination (29, 58). Nevertheless, it has been reported that the pronounced feature of ubiquitination sites is the abundance of charged and polar amino acids, especially negatively charged Asp and Glu, and the depletion of hydrophobic residues, such as Leu, Ile, Phe, and Pro around these sites. In addition, ubiquitination sites display the increased solvent accessibility and high propensity for intrinsic disorder (29). Importantly, the high content of charged residues, low hydrophobicity, increased solvent accessibility, and high content of intrinsically disordered sequences have all been implicated as the factors that augment soluble protein synthesis in the expression system used (18). These factors may account for the increased yields of soluble expression observed for the sequences predicted to contain multiple ubiquitination sites.

The physicochemical and structural features of SUMOylation sites have not been investigated in depth. Although the catalytic mechanisms of ubiquitination and SUMOylation are similar, the sites of ubiquitination and SUMOylation are not the same. At the functional level, SUMOylation often acts antagonistically to ubiquitination and serves to stabilize proteins rather than to target them for degradation. Most SUMO-modified proteins have been shown to contain the tetrapeptide consensus motif $ΨKX(D/E)$ (where $Ψ$ is Ala, Ile, Leu, Met, Pro, Phe, or Val, and X is any amino acid residue), although the existence of some nonconsensus sites has also been reported (30, 59). The above consensus sequence suggests that the proteins with multiple sites of SUMOylation may have an increased percentage of charged residues. Indeed, the average content of charged residues in the analyzed subsets of modified and unmodified proteins were significantly different, composing 29 and 24%, respectively. This observation may provide a plausible explanation for the better solubility of the polypeptide sequences containing multiple predicted sites of SUMOylation. Notably, SUMO and ubiquitin fusion tags have recently been employed to express recombinant proteins in E. coli. These fusions often improve solubility and stability of the expressed proteins (60, 61).

Correlations of Protein Expression with the Presence of Phosphorylation Sites—One of the most interesting findings of this study was the revelation of positive correlation between the predicted presence of phosphorylation sites and protein amenability to heterologous cell-free expression. Phosphorylation has been established to greatly influence protein structure and function; thus, it could be expected that the reduced capacity of the bacterial expression system for phosphorylation would complicate the correct folding of heterologously synthesized proteins, resulting in the decreased yields of their soluble expression. Quite unexpectedly, the opposite tendency has been observed (Fig. 5A, curve A). It was accompanied by a decrease in the rates of insoluble and nondetectable protein expression (Fig. 5A, curves C and N).

Similarly to the case of ubiquitination and SUMOylation, we reasoned out that the predicted presence of phosphorylated sites in proteins should confer the intrinsically better solubility to these proteins, even if they remain unmodified, and, most probably, this property is related to the physicochemical and/or structural characteristics of the phosphorylation sites themselves. Notably, phosphorylation can occur on multiple sites in a given protein. Early estimates suggested that the majority of human proteins may be phosphorylated at multiple sites (in total $>100000$ sites) (62). In agreement with this, the total number of phospho-sites predicted by the PROSITE PS_SCAN scanning tool in the studied dataset was 12,308. The average number of phospho-sites per protein in the dataset was 8.27, and only 8 proteins in the dataset have been found not to harbor this modification (Fig. 5B).

Evidently, due to the nature of the PROSITE prediction algorithms, which do not take into account spatial accessibility of modification sites, not all of the predicted sites can be accessible to phosphorylation. This may bring about a number of false-positive predictions, as discussed above for the case of aspartyl hydroxylation. Interestingly, the solvent accessibility of the predicted phosphorylatable residues calculated for a subset of phosphorylated proteins composed 81.1%. This value is significantly higher than the average solvent accessible surface of the proteins in the expression dataset ($\sim$51.8%). Still, it is difficult to correctly evaluate the functional accessibility of the predicted phosphorylation sites as some of them may be rendered accessible after interactions with their effector protein kinases.

Phosphorylation of eukaryotic proteins on multiple sites is carried out by numerous divergent members of the eukaryotic protein kinase superfamily, including Ser/Thr- and Tyr-specific protein kinases (63). No single substrate consensus sequence can be provided for these enzymes. More than 500
putative protein kinase genes have been identified in the human genome (64). Together, the most processive Ser/Thr-specific protein kinases, such as PKA, PKC, PKG, CK2, CamII, and Cdk1, account for phosphorylation of more than 90% of all phosphorylatable residues in proteins. Their corresponding PROSITE consensus patterns are as follows: $RX_1\_2(S/T)X$, $RX_3\_4(S/T)X$, and $RX_5\_6(S/T)X$. 

FIGURE 5. Correlations of heterologous protein expression with phosphorylation. Relative rates of soluble (curve A), insoluble (curve C), and nonexpressed (curve N) proteins with different numbers of predicted phosphorylation sites are presented in A. Distribution of the dataset proteins according to the number of total phosphorylation sites is shown in B. The relative expression rates of proteins with different numbers of predicted PKC-phosphorylated sites, PKA-phosphorylated sites, and Tyr-phosphorylated sites with their distributions in the dataset are shown in C–H, respectively.
Role of PTMs in Heterologous Protein Expression

X(S/T)X(R/K), (R/K)_2–3X(S/T)X, X(S/T)XX(D/R), RXX(S/T)X, and X(S/T)PX(R/K) (65). Evidently, the specificity of these kinases is directed by basic Lys and Arg residues in a close proximity to the acceptor residue. This fact suggests that the proteins with multiple phosphorylation sites may have an increased ratio of charged residues and a higher solvent accessibility associated with a larger solvent exposure of charged residues. Indeed, a direct correlation has been observed between the predicted number of phosphorylation sites and the content of charged residues/solvent accessibility (supplemental Fig. S3, A and B). Previously, these parameters were found to be positively correlated with the soluble protein expression (18). Considering the similarity of the substrate consensus sequences, it comes as no surprise that protein amenability to soluble expression correlates positively with the predicted presence of the specific phosphorylation sites for PKC, PKA, and CK2. The amenability to insoluble expression correlates negatively with the presence of these sites (Fig. 5, C and E, supplemental Fig. S4, and supplemental Table S3).

Interestingly, these correlations have not been observed for the proteins with the predicted sites of Tyr phosphorylation (Fig. 5G; and supplemental Table S3). One explanation for this may be the drastic difference in the amino acid composition of Tyr and Ser/Thr phosphorylation sites. The specificity of Tyr kinases is dominated by acidic, basic, and hydrophobic residues adjacent to the acceptor residue; however, a large variation makes it difficult to deduce a generalized consensus sequence for these sites (66, 67). Moreover, in contrast to the polypeptide sequences with the predicted sites of Ser/Thr phosphorylation, no tendency to the increased content of charged residues and solvent accessibility has been observed for the proteins with predicted sites of Tyr phosphorylation. The values of the average solvent accessible surface for the proteins in the analyzed subsets of Tyr-phosphorylated and unphosphorylated proteins were very close (51.5 and 51.9%, respectively).

In addition, Tyr phosphorylation is a relatively rare PTM. The total number of the predicted Tyr phospho-sites in the expression dataset was estimated to be 568, which represents 4.61% of the total number of phospho-sites in the dataset. This figure agrees well the estimate (~4%) provided by the “Kinexus” protein phosphorylation resource. No more than two sites of Tyr phosphorylation could be predicted in any dataset protein (Fig. 5H). Plausibly, in general, a single event of phosphorylation cannot change significantly the integral characteristics of a protein molecule, which define its solubility. Thus, the low abundance of Tyr phosphorylation may be related to the negligible effect of this PTM on the yield of heterologous protein expression.

The existence of correlations presented in supplemental Fig. S3, A and B raises a question about the extent to which the algorithms for PTM prediction may reflect physicochemical properties of proteins. Evidently, some PROSITE consensus patterns are dominated by the amino acids of a certain type, charged, polar, hydrophobic, etc. The presence of multiple predicted PTM sites in a protein molecule should confer the properties associated with these sites to the whole molecule. For example, a direct correlation between the number of phosphorylation sites and the content of charged residues in proteins (supplemental Fig. S3A) is based on the presence of charged residues in the multiple consensus sequences of phosphorylation sites. However, the consensus patterns of some other PTMs cannot confer the unique features to proteins because they are low abundant and lack the prominent sequence determinants associated with specific physicochemical properties. For instance, the PROSITE consensus sequence for asparagine glycosylation is NX(S/T) (65). It is short, hydrophilic neutral, and relatively low abundant; most of the N-glycosylated proteins contain only a single site for this modification (Fig. 2B). Therefore, its presence should have only a minor impact on the total amino acid composition of a protein. Nevertheless, the polypeptides with predicted sites of N-glycosylation have a significantly lower propensity for soluble expression and a higher propensity for insoluble expression (Fig. 2A). Importantly, protein folding in the region of the consensus site was shown to play an important role in the regulation of glycosylation. This region adopts a specific conformation during the glycosyl transfer reaction, although a defined secondary structure seems not to be essential for recognition of the substrate by the oligosaccharyltransferase (68). Thus, it is plausible that the PROSITE algorithm for prediction of N-glycosylation sites may inadvertently cover some important structural determinants associated with these sites and linked to overall protein stability and solubility. Similarly, certain conformational properties may also be associated with the consensus sequences of other PTMs. At present, the structural prerequisites of most PTMs remain largely unknown.

In conclusion, our study demonstrates that the amenability of human polypeptide sequences to heterologous cell-free expression correlates with the presence of multiple PTM sites bioinformatically predicted in these sequences (Table 2). For each analyzed PTM, a credible explanation for the existence of observed correlations is presented. These findings provide a plethora of important insights into the role of multiple PTMs in the stability and solubility of heterogeneously expressed proteins. In a practical sense, identification of potential PTM sites in the polypeptide sequences can be used for predicting expression success and optimizing heterologous protein synthesis.
Role of PTMs in Heterologous Protein Expression

combination of multiple PTMs, whose occurrence correlates most robustly with protein amenability to heterologous cell-free expression, may be used to set up the computational algorithms for predicting the outcome of protein synthetic reactions. In addition, a number of calculated physicochemical and structural characteristics of polypeptide sequences can serve as the parameters for predicting expression success (18). Currently, we are developing a discriminant-based machine-learning algorithm that utilizes multiple features of amino acid sequences to predict the success rate of heterologous protein expression.

REFERENCES

1. Yokoyama, S. (2003) Protein expression systems for structural genomics and proteomics. Curr. Opin. Chem. Biol. 7, 39–43
2. Marsden, R. L., and Orengo, C. A. (2008) Target selection for structural genomics. An overview. Methods Mol. Biol. 426, 3–25
3. Farrokhzad, N., Hrmova, M., Burton, R. A., and Fincher, G. B. (2009) Heterologous and cell-free protein expression systems. Methods Mol. Biol. 513, 175–198
4. Spirin, A. S. (2004) High throughput cell-free systems for synthesis of functionally active proteins. Trends Biotechnol. 22, 538–545
5. Katzen, F., Chang, G., and 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. 25, 126–132
7. Eckart, M. R., and Bussinse, C. M. (1996) Quality and authenticity of heterologous proteins synthesized in yeast. Curr. Opin. Biotechnol. 7, 525–530
8. Madin, K., Sawasaki, T., Ogasawara, T., and Endo, Y. (2000) A highly efficient and robust cell-free protein synthesis system prepared from wheat germ embryos. Plants apparently contain a suicide system directed at ribosomes. Proc. Natl. Acad. Sci. U.S.A. 97, 559–564
9. Ezure, T., Suzuki, T., Higashide, S., Shintani, E., Endo, K., Kobayashi, S., Shikata, M., Ito, M., Tanimizu, K., and Nishimura, O. (2006) Cell-free protein synthesis system prepared from insect cells by freeze-thawing. Biotechnol. Prog. 22, 1570–1577
10. Merrick, W. C., and Bausch, D. (2007) Use of reticulocyte lysates for high throughput experimental analysis. J. Mol. Biol. 371, 538–545
11. Mikami, S., Masutani, M., Sonenberg, N., Yokoyama, S., and Imataka, H. (2006) An efficient mammalian cell-free translation system supplemented from ribosomes. Proc. Natl. Acad. Sci. U.S.A. 103, 1241–1246
12. Mikami, S., Kobayashi, T., Yokoyama, S., and Imataka, H. (2006) A hybrid-based in vitro translation system that efficiently synthesizes glycoproteins. J. Biotechnol. 127, 65–78
13. Obata, Y., Hasegawa, K., Nagao, Y., Kato, M., Kusumoto, M., and Tomita, M. (2002) Identification of protein features that correlate with successful expression. J. Mol. Biol. 220, 115–130
14. Bertone, P., Kluger, Y., Lan, 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
15. Dyson, M. R., Shadbolt, S. P., Vincent, K. J., Pereira, R. L., and McCafferty, J. (2004) Production of soluble mammalian proteins in Escherichia coli. Identification of protein features that correlate with successful expression. BMC Biotechnol. 4, 32
16. Iidculla-Thomas, S., and 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
17. Iidculla-Thomas, S., Kulkarni, A. J., Kulkarni, B. D., Jayaraman, V. K., and Balaji, P. V. (2006) A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on overexpression in Escherichia coli. Bioinformatics 22, 278–284
18. Kurotani, A., Takagi, T., Toyama, M., Shirouzu, M., Yokoyama, S., Fukami, Y., and Tokmakov, A. A. (2010) Comprehensive bioinformatics analysis of cell-free protein synthesis. Identification of multiple protein properties that correlate with successful expression. FASEB J. 24, 1095–1104
19. Yokoyama, S., Hirota, H., Kigawa, T., Yabuki, T., Shirouzu, M., Terada, T., Ito, Y., Matsu, Y., Kuroda, Y., Nishimura, Y., Kiyogoku, Y., Miki, K., Masui, R., and Kuramitsu S. (2000) Structural genomics projects in Japan. Nat. Struct. Biol. 7, suppl. 943–945
20. Yokoyama, S. (2005) Large scale structural proteomics project at RIKEN. Present and future. Tanpakushitsu Kakusan Koso 50, 836–845
21. Yokoyama, S., Kigawa, T., Shirouzu, M., Miyano, M., and Kuramitsu, S. (2008) RIKEN structural genomics/proteomics initiative. Tanpakushitsu Kakusan Koso 53, 632–637
22. Gattiker, A., Gasteiger, E., and Bairoch, A. (2002) ScanProsite. A reference implementation of a PROSITE scanning tool. Appl. Bioinformatics 1, 107–108
23. Sigrist, C. J., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A., and Bucher, P. (2002) PROSITE. A documented database using patterns and profiles as motif descriptors. Brief. Bioinform. 3, 265–274
24. Ren, J., Wen, L., Gao, X., Jin, C., Xue, Y., and Yao, X. (2008) CSS-Palm 2.0: An updated software for palmitoylation sites prediction. Protein Eng. Des. Sel. 21, 639–644
25. Kiemer, L., Bengtsen, J. D., and Blom, N. (2005) NetAcet. Prediction of N-terminal acetylation sites. Bioinformatics 21, 1269–1270
26. Frottin, F., Martinez, A., Peynot, P., Mitra, S., Holz, R. C., Giglione, C., and Meinnel, T. (2006) The proteomics of N-terminal methionine cleavage. Mol. Cell. Proteomics 5, 2336–2349
27. Martinez, A., Traverso, J. A., Valot, B., Ferro, M., Espagne, C., Ephritikhine, G., Zivy, M., Giglione, C., and Meinnel, T. (2008) Extent of N-terminal modifications in cytotoxic proteins from eukaryotes. Proteomics 8, 289–2431
28. Cheng, J., Saigo, H., and 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
29. Radivojac, P., Vacic, V., Haynes, C., Cocklin, R. R., Mohan, A., Heyen, J. W., Goebl, M. G., and Parsegov, G. K. (2010) Identification, analysis, and prediction of protein ubiquitination sites. Proteins 78, 365–380
30. Ren, J., Gao, X., Jin, C., Zhu, M., Wang, X., Shaw, A., Wen, L., Yao, X., and Xue, Y. (2009) Systematic study of protein sumoylation. Development of a site-specific predictor of SUMOpro 2.0. Proteomics 9, 3409–3412
31. Iikeyama, E., Kurotani, A., Kuroda, Y., and Yokoyama, S. (2004) ProteomeX. An integrated and flexible system for interactively analyzing large numbers of protein sequences. Bioinformatics 20, 2836–2838
32. Peitsch, M. C., Wells, T. N., Stampf, D. R., and Sussman, J. L. (1995) The Swiss-3DImage collection and PDB-Browser on the World-Wide Web. Trends Biochem. Sci. 20, 82–84
33. Schwede, T., Kopp, J., Guex, N., and Peitsch, M. C. (2003) SWISS-MODEL. An automated protein homology-modeling server. Nucleic Acids Res. 31, 3381–3385
34. Arnold, K., Bordoli, L., Kopp, J., and Schwede, T. (2006) The SWISS-MODEL workspace. A web-based environment for protein structure homology modeling. Bioinformatics 22, 195–201
35. Delano, W. L. (2002) The PyMOL Molecular Graphics System, Delano Scientific LLC, San Carlos, CA
36. Xu, B., Feng, X., and Burdine, R. D. (2010) Categorical data analysis in experimental biology. Dev. Biol. 348, 3–11
37. Ding, H. T., Ren, H., Chen, Q., Fang, G., Li, L., Li, R., Wang, Z., Jia, X. Y., Liang, Y. H., Hu, M. H., Li, Y., Luo, J. C., Gu, X. C., Su, X. D., Luo, M., and Lu, S. Y. (2002) Parallel cloning, expression, purification, and crystallization of human proteins for structural genomics. Acta Crystallogr. D Biol. Crystallogr. 58, 2102–2108
38. Walsh, G., and Jefferis, R. (2006) Post-translational modifications in the context of therapeutic proteins. Nat. Biotechnol. 24, 1241–1252
39. Mitra, N., Sinha, S., Ramya, T. N., and Surolia, A. (2006) N-Linked oligosaccharides as outfitters for glycoprotein folding, form, and function. Trends Biochem. Sci. 31, 156–163
Role of PTMs in Heterologous Protein Expression

42. Wacker, M., Feldman, M. F., Callewaert, N., Kowari, M., Clarke, B. R., Pohl, N. L., Hernandez, M., Vines, E. D., Valvano, M. A., Whitfield, C., and Aebi, M. (2006) Substrate specificity of bacterial oligosaccharyltransferase suggests a common transfer mechanism for the bacterial and eukaryotic systems. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 7088–7093

41. Nothaft, H., and Szymanski, C. M. (2010) Protein glycosylation in bacteria. Sweeter than ever. *Rev. Microbiol.* **8**, 765–778

40. Tu, B. P., and Weissman, J. S. (2004) Oxidative protein folding in eukaryotes. *J. Cell Biol.* **164**, 341–346

39. Baneyx, F., and Mujacic, M. (2004) Recombinant protein folding and misfolding in *Escherichia coli*. *Nat. Biotechnol.* **22**, 1399–1408

38. Braun, P., Hu, Y., Shen, B., Halleck, A., Koundinya, M., Harlow, E., and LaBae, J. (2002) Proteome scale purification of human proteins from bacteria. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 2654–2659

37. Kim, D. M., and Swartz, J. R. (2004) Efficient production of a bioactive, multiple disulfide-bonded protein using modified extracts of *Escherichia coli*. *Bioeng**. **85**, 122–129

36. Moore, K. L. (2003) The biology and enzymology of protein tyrosine phosphorylation. *Science* **299**, 1912–1934

35. Somers, W. S., Tang, J., Shaw, G. D., and Camphausen, R. T. (2000) Insights into the molecular basis of leukocyte tethering and rolling revealed by structures of P- and E-selectin bound to SL(X) and PSGL-1. *Cell* **103**, 467–479

34. Gao, J., Liao, J., and Yang, G. Y. (2009) CAAX-box protein, prenylation process, and carcinogenesis. *Annu. Rev. Transl. Res.* **1**, 312–325

33. Catanzariti, A. M., Soboleva, T. A., Jans, D. A., Board, P. G., and Baker, R. T. (2004) An efficient system for high level expression and easy purification of authentic recombinant proteins. *Protein Sci.* **13**, 1331–1339

32. Zhang, H., Zha, X., Tan, Y., Hornbeck, P. V., Mastrangelo, A. J., Alessi, D. R., Polakiewicz, R. D., and Comb, M. J. (2002) Phosphoprotein analysis using antibodies broadly reactive against phosphorylated motifs. *J. Biol. Chem.* **277**, 39379–39387

31. Morris, K. S. (2003) Genomic analysis of the eukaryotic protein kinase superfamily. A perspective. *Genome Biol.* **4**, 111

30. Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002) The protein kinase complement of the human genome. *Science* **298**, 1912–1934

29. Blom, N., Sicheritz-Pontén, T., Gupta, R., Gammeltoft, S., and Brunak, S. (2004) Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics* **4**, 1633–1649

28. Amanchy, R., Periaswamy, B., Mathivanan, S., Reddy, R., Tattikota, S. G., and Pandey, A. (2007) A curated compendium of phosphorylation motifs. *Nat. Biotechnol.* **25**, 285–286

27. Blom, N., Sicheritz-Pontén, T., Gupta, R., Gammeltoft, S., and Brunak, S. (2004) Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics* **4**, 1633–1649

26. Hunter, T. (2009) Tyrosine phosphorylation. Thirty years and counting. *Curr. Opin. Cell Biol.* **21**, 140–146

25. Schwarz, F., and Aebi, M. (2011) Mechanisms and principles of N-linked protein glycosylation. *Curr. Opin. Struct. Biol.* **21**, 576–582