Identification of multiple physicochemical and structural properties associated with soluble expression of eukaryotic proteins in cell-free bacterial extracts

Alexander A. Tokmakov*
Research Center for Environmental Genomics, Kobe University, Kobe, Japan

Bacterial extracts are widely used to synthesize recombinant proteins. Vast data volumes have been accumulated in cell-free expression databases, covering a whole range of existing proteins. It makes possible comprehensive bioinformatics analysis and identification of multiple features associated with protein solubility and aggregation. In the present paper, an approach to identify the multiple physicochemical and structural properties of amino acid sequences associated with soluble expression of eukaryotic proteins in cell-free bacterial extracts is presented. The method includes: (1) categorical assessment of expression data; (2) calculation and prediction of multiple properties of expressed sequences; (3) correlation of the individual properties with the expression scores; and (4) evaluation of statistical significance of the observed correlations. Using this method, a number of significant correlations between calculated and predicted properties of amino acid sequences and their propensity for soluble cell-free expression have been revealed.

Keywords: cell-free protein synthesis, protein solubility, physicochemical and structural protein properties, categorical data analysis, correlation analysis

METHOD OVERVIEW

The developed method is intended for analysis of output from an existing cell-free protein production pipeline. Thus, this paper does not cover the experimental workflow of protein production. It is described in detail in the previous publications (Yabuki et al., 2007; Kigawa et al., 2008; Kurotani et al., 2010; Tokmakov et al., 2012). Here, the focus is set on the processing of experimental data with the purpose of identification of multiple physicochemical and structural properties associated with soluble expression of eukaryotic proteins in bacterial cell-free extracts, and highlight major correlations obtained using this approach.
acids sequences using computational and predictive bioinformatics tools; (3) correlation of the individual protein properties with the experimental expression scores; and (4) evaluation of statistical significance of the observed correlations. The developed approach has been extensively used to analyze experimental expression of human proteins and their domains in *Escherichia coli* bacterial extracts (Kurotani et al., 2010; Tokmakov et al., 2012; see Results and Discussion section). Notably, this setting, the score A provides the upper estimation of soluble protein expression, because the procedure of centrifugation at 10,000 × g cannot discriminate between small protein aggregates and truly soluble proteins. Often, expressed proteins can be found in both soluble and insoluble fractions of the bacterial extract. Lane-to-lane comparison of total and supernatant fractions of the extract in PAGE gels is usually sufficient to establish the preferential pattern of protein expression.

**CALCULATION AND PREDICTION OF MULTIPLE PROPERTIES OF EXPRESSED SEQUENCES**

In this step, multiple features of the amino acid sequences in the expression dataset are calculated or predicted using existing bioinformatics tools. Various protein properties can be classified into the four major types, including physicochemical parameters, structural properties, the presence of specific sequence motifs, and the presence of PTM sites (Figure 3). Many of the physicochemical parameters, such as protein length, molecular weight, amino acid composition, number of charged residues, pI, hydrophobicity, etc., can be calculated using the free ProtParam tool available at the Expasy server1. On the other hand, it is difficult to precisely calculate high-dimensional protein properties, because the 3D structures of expressed protein targets are usually unknown. Still, it is possible to deduce some structural features of the proteins in the expression dataset using existing prediction algorithms. Admittedly, some of these algorithms have quite low prediction accuracy, not exceeding 80%. The low accuracy of prediction thwarts the following correlation analysis, making impossible detection of weak correlations.

---

1http://www.expasy.org/tools/
Solvant accessibility can be assessed with the ACCpro 4.0 software downloaded from the SCRATCH Protein Predictor server (Cheng et al., 2005) and content of secondary structure is evaluated with the PREDATOR 2.1.2 tool (Frishman and Argos, 1997) provided online. Coiled coil structures are predicted with the pepcoil tool provided online (Lupas et al., 1991) and content of disordered structure is predicted with the RONN software (Yang et al., 2005). The specific sequence motifs in proteins can also be predicted using available bioinformatics tools. PEST regions, signal sequences, and transmembrane domains are predicted with the tools provided online (Hu et al., 1998; Lan et al., 2009). The sites of ubiquitination and SUMOylation are predicted using the site-specific predictors UbPred (Radivojac et al., 2010) and SUMOsp 2.0 (Ren et al., 2009) freely downloadable for academic research from http://ubpred.org/ and http://sumosp.biocuckoo.org/, respectively. The sites of S-palmitoylation are predicted with the CSS-Palm tool (Ren et al., 2008) and S–S bonds can be predicted using the DIPtro tool (Cheng et al., 2006) downloadable free from http://download.igb.uci.edu/intro.html.

**CORRELATION OF THE INDIVIDUAL PROPERTIES WITH EXPRESSION SCORES**

The multiple protein properties calculated and predicted using the above bioinformatics tools can be categorized into the three types, including yes/no, discrete, and continuous variables (Figure 4). Data processing and presentation differs for the three types of variables. The yes/no type variables, such as single-event PTMs, are the features that can be either present in or absent from proteins. To present the expression data associated with these variables, the bar graphs can be built, which show the ratio of proteins in the expression categories A, C, and N. The graphs should represent two subsets of proteins, excluding and including the analyzed feature. Total number of sequences in the two subsets should be defined. Using these graphs, it is easy to make a side-by-side comparison of the data for the two subsets and deduce the tendencies in protein expression amenability associated with the analyzed feature. To present the expression correlations associated with the discrete variables related to the protein features repeatedly observed in the analyzed sequences, such as abundant multi-site PTMs, another type of data presentation is more convenient. In this case, the percentage of proteins in the expression categories A, C, and N is plotted at different values of analyzed parameter, covering the entire parameter range in the dataset. In addition, the distribution of dataset proteins according to parameter values should be presented. The distribution graphs provide important information concerning the abundance of studied protein features in the analyzed dataset. The processing of data associated with continuous variables, such as sequence hydrophobicity, solvent accessibility, content of intrinsic disorder, etc., is similar to that described for discrete variables. The graphs of A, C, and N scores, as well as the distribution graphs should be provided in the full range of continuous feature values. Curve smoothing is recommended to straighten the graphs obtained with continuous variables. It can be performed using the Excel chart smoothing algorithm. The examples of data presentation for the three types of variables associated with different protein properties are provided in our recent publication (Tokmakov et al., 2014).
STATISTICAL SIGNIFICANCE OF THE OBSERVED CORRELATIONS

The expression data processed by the proposed method represent categorical datasets, where all expressed sequences are classified into three categories – soluble (A), insoluble (C), and non-expressed targets (Figure 2). Thus, to evaluate the statistical significance of the observed correlations between the multiple protein features and protein amenability to cell-free expression, the categorical data analysis should be applied (Xu et al., 2010). The estimation of statistical significance should be provided for each expression category (A, C, and N). In addition, multiple protein properties are also categorized into the three types, such as yes/no, discrete, and continuous variables (Figure 4). Evaluation of statistical significance differs for the three types of variables. To deduce the statistical differences associated with yes/no type variables, the two-way contingency table test can be applied (Figure 5). The Fisher’s exact p-values can be computed using the tool provided on line at http://statpages.org/ctab2x2.html. Usually, a confidence level of 95% is set up as the null hypothesis rejection threshold. To evaluate the statistical significance of expression correlations associated with the discrete variables, which have a finite number of possible values, as well as the continuous variables, Pearson’s pairwise correlation coefficients should be calculated (Figure 5). The percentage of proteins in the expression categories A, C, and N should be paired with the values of the analyzed variable in the full range of variable values observed in the dataset. Statistical significance of the correlation coefficients is validated by calculating one-tailed probability values, given the value of correlation coefficient (r) and the sample size (n), with the significance level set to 0.05. Calculations of both correlation coefficients and p-values can be performed using the online statistics calculators available at http://www.danielsoper.com/statcalc3/. As a general comment, it should be noted that the confidence level of categorical data analysis increases greatly with the number of sequences in the expression datasets (Norman and Streiner, 2000).

RESULTS AND DISCUSSION

Using the developed method, expression of 3066 human proteins and their domains in a cell-free bacterial system has been analyzed. It was found that the rate of soluble expression (score A) in the investigated dataset constituted 25.7% (Kurotani et al., 2010). This value should be considered as a benchmark, as the similar success rate has been reported for a different subset of human proteins expressed in E. coli (Ding et al., 2002). Furthermore, a number of statistically significant correlations between calculated and predicted properties of amino acid sequences and their amenability to bacterial cell-free expression have been identified using the developed approach. The most influential features that affect protein amenability to cell-free expression are listed in Table 1.

Notably, some of these features, such as protein pI, hydrophobicity, presence of localization signals, etc., are mostly related to protein solubility, whereas the others, such as protein length, Table 1 | Correlations of cell-free protein expression with calculated and predicted properties of amino acid sequences.

| Expression property | Soluble | Insoluble | Undetectable |
|--------------------|---------|-----------|--------------|
| Length             | −       | ND        | +            |
| pI                 | ±       | ±         | ND           |
| Charge             | +       | ±         | −            |
| Hydrophobicity     | −       | +         | ND           |
| Solvent accessibility | +     | ±         | −            |
| Secondary structure | +     | ±         | −            |
| Intrinsic disorder | +       | −         | +            |
| Protein domains    | −       | −         | +            |
| S–S bonds          | −       | +         | +            |
| Coiled coil        | +       | −         | −            |
| Transmembrane seqs | −      | −         | +            |
| Localization signals | −    | +         | ND           |
| PEST regions       | +       | −         | +            |
| Prenylation        | +       | ND        | ND           |
| Phosphorylation    | +       | −         | −            |
| Asn glycosylation  | −       | +         | ND           |
| Palmitoylation     | −       | ±         | +            |
| Ubiquitination     | +       | −         | ND           |
| SUMOylation        | +       | −         | ±            |
| Amidation          | ND      | ND        | ND           |
| Asx hydroxylation  | ND      | ND        | ND           |
| Sulfation          | ND      | ND        | ND           |

The signs (+) and (−) indicate positive and negative correlations, respectively; (±) refers to the opposite tendencies of expression estimates at different values of calculated parameters; and ND denotes the lack of correlation.
charge, solvent accessibility, presence of S–S bonds, transmembrane sequences, PEST regions, etc., also affect the overall expression propensity. The presence of some specific sequence motifs was found to be one of the most discriminative parameters for expression propensity. The correlations revealed can be of practical use for protein engineering with the aim of increasing expression success. The rationales for these correlations are discussed in detail in the published paper (Kurotani et al., 2010).

In addition, it was found that amenability of human polypeptide sequences to bacterial cell-free expression correlates with the presence of multiple PTM sites bioinformatically predicted in these sequences (Tokmakov et al., 2012; Table 1). Surprisingly, the presence of predicted sites for several PTMs, such as ubiquitination, SUMOylation, etc. (Table 1), was associated with increased production of properly folded soluble protein. However, no SUMOylation and ubiquitination machineries are known to exist in bacteria, suggesting that the presence of these PTM sites in amino acid sequences is related to intrinsically better protein solubility even in the absence of the modifications. It was hypothesized that physicochemical and/or structural characteristics of the modification sites themselves convey the better solubility (Tokmakov et al., 2012). Altogether, these findings indicate that identification of potential PTM sites in polypeptide sequences can be of practical use for predicting expression success and optimizing heterologous protein synthesis. Currently, a discriminant-based machine-learning algorithm that utilizes multiple features of amino acid sequences to predict the success rate of heterologous protein synthesis is being developed based on the reported findings. The algorithm will provide a basis for the internet-based tool for predicting amenability of eukaryotic proteins to cell-free expression in a prokaryotic system.

ACKNOWLEDGMENTS

This work was supported by the research fund for Foreign Visiting Professor from Kobe University and the Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan (no. 25440023).

REFERENCES

Bertone, P., Kluger, Y., Lan, N., Zheng, D., Christendat, D., Yee, A., et al. (2001). SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. Nucleic Acids Res. 29, 2884–2898. doi: 10.1093/nar/29.13.2884

Chen, R. (2012). Bacterial expression systems for recombinant protein production: E. coli and beyond. Biotechnol. Adv. 30, 1102–1107. doi: 10.1016/j.biotechadv.2011.09.013

Cheng, J., Randall, A. Z., Sweredoski, M. J., and Baldi, P. (2005). SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res. 33, W72–W76. doi: 10.1093/nar/gki396

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. doi: 10.1002/prot.20787

Ding, H. T., Ren, H., Chen, Q., Fang, G., Li, L. F., Li, R., et al. (2002). Parallel cloning, expression, purification, and crystallization of human proteins for structural genomics. Acta Crystallogr. D Biol. Crystallogr. 58, 2102–2108. doi: 10.1107/S0907444902016359

Dyson, M. R., Shadbolt, S. P., Vincent, K. J., Perera, 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. doi: 10.1186/1427-6750-4-32

Frisman, D., and Argos, P. (1997). Seventy-five percent accuracy in protein secondary structure prediction. Proteins 27, 329–335. doi: 10.1002/(SICI)1097-0134(199703)27:3<329:AID-PROT1>3.0.CO;2-8

Goh, C. S., Lan, N., Douglas, S. M., Wu, B., Echols, N., Smith, A., et al. (2004). Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis. J. Mol. Biol. 336, 115–130. doi: 10.1016/j.jmb.2003.11.053

He, M. (2008). Cell-free protein synthesis: applications in proteomics and biotechnology. Nat. Biotechnol. 25, 126–132. doi: 10.1016/j.nbt.2008.08.004

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. doi: 10.1110/s04109005

Katzen, F., Chang, G., and Kudlicki, W. (2005). The past, present and future of cell-free protein synthesis. Trends Biotechnol. 23, 150–156. doi: 10.1016/tibitech.2005.01.003

Kigawa, T., Matsuda, S., Yabuki, T., and Yokoyama, S. (2008). "Bacterial cell-free system for highly efficient protein synthesis," in Cell-Free Protein Synthesis, eds A. S. Sspin and J. R. Swartz (Wiley VCH: Weinheim), 83–97.

Kigawa, T., Yabuki, T., Yoshiida, Y., Tsutsui, M., Ito, Y., Shabata, T., et al. (1999). Cell-free production and stable-isotope labeling of milligram quantities of proteins. FEBS Lett. 442, 15–19. doi: 10.1016/S0014-5793(99)01620-2

Kurotani, A., Takagi, T., Toyama, M., Shirozu, M., Yokoyama, S., Fukami, Y., et al. (2010). Comprehensive bioinformatic analysis of cell-free protein synthesis: identification of multiple protein properties that correlate with successful expression. FASEB J. 24, 1095–1104. doi: 10.1096/fj.09-139527

Lupas, A., Van Dyke, M., and Stock, J. (1991). Predicting coiled coils from protein sequences. Science 252, 1162–1164. doi: 10.1126/science.252.5009.1162

Norman, G. R., and Streiner, D. L. (2000). Biostatistics: The Bare Essentials. B. C. Decker, Hamilton.

Radivojac, P., Vacic, V., Haynes, C., Cocklin, R. R., Mohan, A., Heyen, J. W., et al. (2010). Identification, analysis, and prediction of protein ubiquitination sites. Proteins 78, 365–380. doi: 10.1002/pro.2255

Ren, J., Gao, X., Jin, C., Zhu, M., Wang, X., Shaw, A., et al. (2009). Systematic study of protein sumoylation: development of a site-specific predictor of SUMPs 2.0. Proteomics 9, 3409–3412. doi: 10.1002/pmic.200800646

Ren, J., Wen, L., Gao, X., Jin, C., Xue, Y., and Yao, X. (2008). CSS-Palm 2.0: an updated software for palmoylation sites prediction. Protein Eng. Des. Sel. 21, 639–644. doi: 10.1093/protein/gzn059

Sivasankarumugam, A., Murray, V., Cui, C., Zhang, Y., Wang, J., and Li, Q. (2009). Practical protocols for production of very high yields of recombinant proteins using Escherichia coli. Protein Sci. 18, 936–948. doi: 10.1002/pro.102

Sorensen, H. P., and Mortensen, K. K. (2005). Advanced genetic strategies for recombinant protein expression in Escherichia coli. J. Biotechnol. 115, 113–128. doi: 10.1016/j.jbiotec.2004.08.004

Spirin, A. S. (2004). High-throughput cell-free systems for synthesis of functionally active proteins. Trends Biotechnol. 22, 538–545. doi: 10.1016/j.tibtech.2004.08.012

Tokmakov, A. A., Kurotani, M., Shirozu, M., Fukami, Y., and Yokoyama, S. (2014). Bioinformatics analysis and optimization of cell-free protein synthesis. Methods Mol. Biol. 1118, 17–33. doi: 10.1007/978-1-62703-782-2_2

Tokmakov, A. A., Kurotani, M., Takagi, T., Toyama, M., Shirozu, M., Fukami, Y., et al. (2012). Multiple post-translational modifications affect heterologous protein synthesis. J. Biol. Chem. 287, 27106–27116. doi: 10.1074/jbc.M112.366351

Xu, B., Feng, X., and Burdine, R. D. (2010). Categorical data analysis in experimental biology. Dev. Biol. 348, 3–11. doi: 10.1016/j.ydbio.2010.08.018

Yabuki, T., Motoda, Y., Hanada, K., Nunokawa, E., Saito, M., Seki, E., et al. (2007). A robust two-step PCR method of template DNA production for high-throughput cell-free protein synthesis. J. Struct. Funct. Genomics 8, 173–191. doi: 10.1007/s11480-007-9038-z

Yang, Z. R., Thomson, R., McMeil, P., and Ensouf, R. M. (2005). RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics 21, 3369–3376. doi: 10.1093/bioinformatics/bti534

Yokoyama, S. (2003). Protein expression systems for structural genomics and proteomics. Curr. Opin. Chem. Biol. 7, 39–43. doi: 10.1016/S1367-5931(02)00019-4
Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 10 April 2014; accepted: 29 May 2014; published online: 20 June 2014.

Citation: Tokmakov AA (2014) Identification of multiple physicochemical and structural properties associated with soluble expression of eukaryotic proteins in cell-free bacterial extracts. Front. Microbiol. 5:295. doi: 10.3389/fmicb.2014.00295

This article was submitted to Microbial Physiology and Metabolism, a section of the journal Frontiers in Microbiology.

Copyright © 2014 Tokmakov. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.