Enzyme Engineering

Research Article Open Access

Purification Propensity for Proteins from *Bacillus halodurans*

Shaomin Yan and Guang Wu*

State Key Laboratory of Non-food Biomass Enzyme Technology, National Engineering Research Center for Non-food Biorefinery, Guangxi Academy of Sciences, China

Abstract

The demand for proteins with special purposes increases significantly. These proteins are generally obtained through recombinant proteins, however their purification is costly and not easy. It is necessarily important to develop a method to estimate the chance of purification beforehand in order to have a prospective on proteins in question. Purification of a protein should be related to instinct properties of a protein including its 3D structure, and so far around 540 amino acid properties are found. Thus it is possible to test each amino acid property against the successful rate of protein purification to find out which property is more suitable to estimate the purification propensity. In this study, each of 535 properties was tested against 438 purified and 429 impossible purified proteins from *Bacillus halodurans* using logistic regression and neural network model. ROC analysis was applied to the resultant sensitivity and specificity. The results show that amino acid composition properties were generally less helpful to estimate the purification propensity whereas amino acid physicochemical properties, secondary structures and dynamic properties were more useful, and dynamic properties were more promising. Therefore several types of protein properties can serve to determine purification propensity of proteins, and have the potential to reduce the cost and to speed up the production in microbiological and biotechnical fields.

Keywords: Amino acid property; *Bacillus halodurans*; Logistic regression; Neural network; Prediction; Protein; Purification

Introduction

The demand for proteins with special purposes increases significantly, for example, special proteins are in good need in development of sensitive, specific and reliable differential diagnostic assays. To meet such huge demand, proteins of interest can be expressed in either prokaryotic or eukaryotic cells, like Escherichia coli, to produce recombinant proteins. However, this is not an easy task and is often costly. Purification of recombinant proteins from plant biomass currently accounts for almost 80% of production cost [1]. A series of difficulties may be encountered in purification. For instance, purified proteins from a host may accumulate in low titers and may be mixed with infection [2,3] or form protein aggregates [4,5]. Therefore a purification scheme usually includes many steps, such as affinity chromatography, precipitation, protecting of recombinant proteins from degradation with stabilizer, centrifugation and so on.

In order to develop an efficient and cost-effective purification scheme, many efforts have been made and much has been achieved in finding the factors that affect protein purification. At codon optimization level, N-terminal rare codons increase expression [6] with some reservation [7]. At protein terminal level, hydrophilicity of histidine tag enhances the high solubility of expressed recombinant fusion proteins [8]. At protein level, polyhedrin is used as a carrier protein to facilitate antigen purification [9-13].

The proteins that require to be purified are not completely unknown in many cases, not only they are known for clinical and biotechnological applications, but also they are known for their amino acid composition, primary structure and even 3-dimensional structure. And these types of knowledge are publicly available, and are useful to estimate purification propensity. However, how to use this valuable information is a challenge. So far, more than 540 amino acid properties have been found to describe various aspects of amino acids [14]. Technically, each property is a set of 20 numeric values corresponding to 20 types of amino acids, and 535 amino acid properties are listed in Supplementary Material.

Over years, amino acid properties are constantly considered useful to correlate with various protein operations in order to minimize the operating cost. For example, amino acid properties were used to estimate whether a protein could be crystallized [15], and the whole process from cloning to expression, to purification and to crystallization with amino acid properties [16]. Actually, protein purification is different from gene expression in an organism, whose survival from generation to generation is primarily related to the evolutionary process with respect to different taxonomic groups. Indeed, purification is mainly relevant to chemical and physical processes, which are exactly described by amino acid properties. Therefore it is necessarily important to use amino acid properties to estimate the propensity of purification as an individual process.

Because each amino acid property is a set of 20 numeric values corresponding to 20 types of amino acids, the general procedure that is used to estimate any propensity is to use a set of 20 numeric values to replace their corresponding amino acids in a protein, and then use this “numeric” protein to correlate with a certain operation in process of protein, such as the outcome of protein purification. In fact, the outcome of most protein operations is either yes or no, for instance, a protein can either be purified or not. Consequently, purified protein and impossible purified protein are replaced as unity and zero, and the proteins to be predicted can be replaced by an amino acid property, and then either logistic regression or neural network can be used to estimate the relationship between “numeric” protein and unity/zero. But, it was not clear which amino acid property is useful for such estimation, so this study tested each amino acid property against 438 purified versus 429 impossible purified proteins from *Bacillus halodurans*.

B. halodurans is a Gram-positive and alkaliphilic bacterium growing above pH 9.5 and its genome was completely sequenced [17]. Alkaliphilic microorganisms have wide industrial applications on commercial enzymes [18]. Over recent years, the research interest in *B. halodurans* becomes stronger. For example, *B. halodurans* produces haloduracin, which could serve as peptide antibiotics [19]. Also, a

***Corresponding author:** Guang Wu, State Key Laboratory of Non-food Biomass Enzyme Technology, National Engineering Research Center for Non-food Biorefinery, Guangxi Biomass Industrialization Engineering Institute, Guangxi Key Laboratory of Biorefinery, Guangxi Academy of Sciences, 98 Daling Road, Nanning, Guangxi, 530007, China, Tel: +86 0771-2503940; E-mail: hongguanglishibahao@yahoo.com

Received September 01, 2016; **Accepted** September 20, 2016; **Published** September 24, 2016

Citation: Yan S, Wu G (2016) Purification Propensity for Proteins from *Bacillus halodurans*. Enz Eng 5: 151. doi:10.4172/2329-6674.1000151

considerable effort was made to crystallize proteins from *B. halodurans* [20-22]. For instance, *E. coli* has only one copy of YidC [23] while *B. halodurans* has two copies of YidC [24], which makes *B. halodurans* have more benefit for bioengineering. Nevertheless, we hope to focus our attention on the proteins that are more relevant to microbiological applications after this first stage of study.

Materials and Methods

Structural genomics initiative [25] makes experimental progresses and statuses of most target proteins become publicly available [26,27], of which 438 purified versus 429 impossible purified proteins from *B. halodurans* were documented [26]. 535 amino acid properties [14] were grouped as 40 constant composition properties, 218 physicochemical properties, 273 secondary structure properties and 4 dynamic properties (Supplementary Material). The purification outcome of a protein is either a success or a failure, which can be presented as 1 or 0. A protein is an amino acid sequence, which can be numerically replaced by an amino acid property. As a result, the purification is a relationship between 1 or 0 and a set of 20 values representing an amino acid property, where 1 or 0 event is an outcome of an amino acid property presented by 20 values.

It is true that protein purification is dependent on numerous factors, however it would be a good practice to test each factor at a time rather than to test many factors simultaneously. Unlike control experiments, amino acid properties cannot be separated from a protein during purification process. However, a stepwise regression could in principle either narrow down or extend up amino acid properties that are involved in purification, which is the rationale to estimate the purification propensity.

Technically, the first step was to determine whether logistic regression or neural network can fit the relationship between a 1 versus 0 (purified and impossible purified) and 20 weighed values (an amino acid property weighed by 20 types of amino acids). Accordingly, 438 purified and 429 impossible purified proteins could establish 857 relationships, which constructed the base to determine whether the amino acid property was useful to estimate the purification propensity of proteins from *B. halodurans*. The second step was to operate each of the 535 amino acid properties one by one, and to generate the model parameters for either logistic regression or neural network. The third step was to predict the purification propensity, that is, with numerical proteins as inputs, the obtained model parameters were used to produce the output of 0 or 1, during which a delete-1 jackknife was applied for model validation of neural network.

MatLab was used to operate both logistic regression and neural network, and the latter one was defined as 10-1 feed-forward back-propagation neural network [28]. The model output was classified into true positive, false positive, true negative and false negative. The accuracy, sensitivity and specificity were calculated as follows: Accuracy = (true positive + true negative)/ (true positive + false positive + true negative + false negative) × 100,

Sensitivity = true positive/ (true positive + false negative) × 100,

Specificity = true negative/ (true negative + false positive) × 100.

The data were presented as median with interquatile, and were analyzed by *Chi*-square test. Kruskal-Wallis one-way ANOVA on ranks and Mann-Whitney rank sum test were used to analyze the difference among and between different predictions. The receiver operating characteristic (ROC) analysis was used to compare the sensitivity and the specificity [29].

The amino acid pair predictability is dynamic value representing a protein according to permutation [30,31]. For example, a protein Q9K915 has 239 amino acids, among them there are 48 glutamic acids (E) and 9 isoleucines (I). The amino acid pair EI would appear twice in this protein $(48/239 \times 9/238 \times 238 = 1.81)$. This protein does have two EIs, so the amino acid pair EI is predictable. Taking the amino acid pair EE into account, it would appear nine times $(48/239 \times 47/238 \times 238 = 9.44)$,but it appears 6 times in Q9K915 protein, thus the amino acid pair EE is unpredictable. All amino acid pairs in this protein can be classified as predictable or unpredictable, and its predictable and unpredictable portions are 67.75% and 32.25%. This feature can be computed at the web http://www.nerc-nfb.ac.cn/calculation/pp.htm, and was used to analyze the relationship with protein purification, which was compared with Mann-Whitney Rank Sum Test. $P < 0.05$ was considered statistical significant.

Results and Discussion

It is important for biotechnological industries to make a large quantity of highly stable and purified recombinant proteins, which provide economically affordable sources for clinical and industrial applications and research. Usually, purification is laborious and unexciting although various expression systems are employed successfully, such as codon optimization in expression. This is the reason why amino acid properties were analyzed to find out which amino acid property could provide a clue on the chance of successful purification.

The upper panel of Figure 1 showed the accuracy, sensitivity and specificity resulting from logistic regression that was used to find out which of 535 amino acid properties was useful to estimate the purification propensity for 857 proteins from *B. halodurans*. In this figure, x-axis indicated each of 535 amino acid properties (Supplementary Material) while y-axis indicated the accuracy, sensitivity and specificity. At first glance, the specificity was the best followed by the accuracy and the sensitivity. Moreover, little difference appeared between 535 amino acid properties because the specificity, accuracy and sensitivity were colored similarly, but it was necessary to pick out the poorly performed amino acid properties, which were colored in blue in the upper panel of Figure 1. The amino acid properties related to electric charges were not good in estimating the chance of protein purification.

The lower panel of Figure 1 displayed the receiver operating characteristic (ROC) analysis in order to furthermore distinguish the performance of amino acid properties used in logistic regression. As can be seen, all results were located in the up-left triangle, indicating that logistic regression was effective because its outcomes surpassed a random guess. However, the results were somewhat similar, indicating that logistic regression could not effectively indicate the difference between amino acid properties.

Figure 2 illustrated the accuracy, sensitivity and specificity obtained from 10-1 neural network that was also used to find out which of 535 amino acid properties was useful to estimate the purification propensity for 857 proteins from *B. halodurans* (upper panel) and to valid this relationship with delete-1 jackknife validation (lower panel). Compared with Figures 1 and 2 included the results from both fitting and delete-1 validation, while the latter was a common procedure in development of predictive model. The results in Figure 2 varied largely because the color bar covered the whole range from zero to unity. As can be seen, the sensitivity is better than the accuracy and the specificity. However, the sensitivity and the specificity behaved oppositely, i.e., the better the sensitivity was, the worse the specificity was. Some amino acid properties provided the results with very high sensitivity but very low specificity, and they included 32/40 amino acid properties grouped as constant compositions, 76/218 amino acid properties grouped as physicochemical properties, and 18/273 amino acid properties grouped as secondary structures. Similar results could be found in other studies [31-36]. On the other hand, the rest amino acid properties provided relatively high values of both sensitivity and specificity, and they included the dynamic properties shown at right-hand in both panels of Figure 2.

When looking at estimation performance, Table 1 provided clues for the estimation of purification propensity by means of neural network. In general, the constant composition properties resulted in a lower accuracy than other properties (0.51 versus 0.71 in fitting and 0.51 versus 0.59 in validation), suggesting that the constant composition properties are less helpful to estimate purification propensity, but physicochemical properties, secondary structures and dynamic properties are more helpful in this regard.

Figure 3 demonstrated the ROC analysis in order to furthermore distinguish the performance of amino acid properties for the estimation in neural network, where the larger the distance above the diagonal was, the better the performance was. Again, some amino acid properties performed better than most amino acid properties as indicated by the cycle, because a high sensitivity was accompanied with a high specificity.

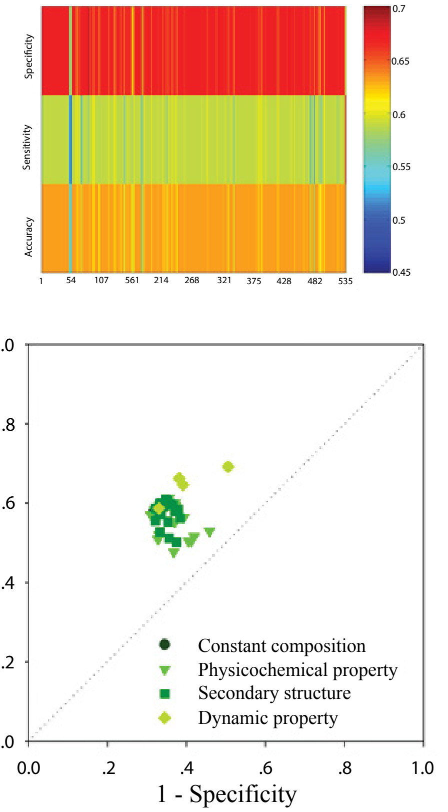

Figure 1: Accuracy, sensitivity and specificity obtained from logistic regression to determine if any of 535 amino acid properties was useful to estimate the purification propensity for 857 proteins from *B. halodurans* (upper panel), and comparison of sensitivity versus specificity in ROC analysis (lower panel). The diagonal line is the line of indiscrimination indicating a completely random guess.

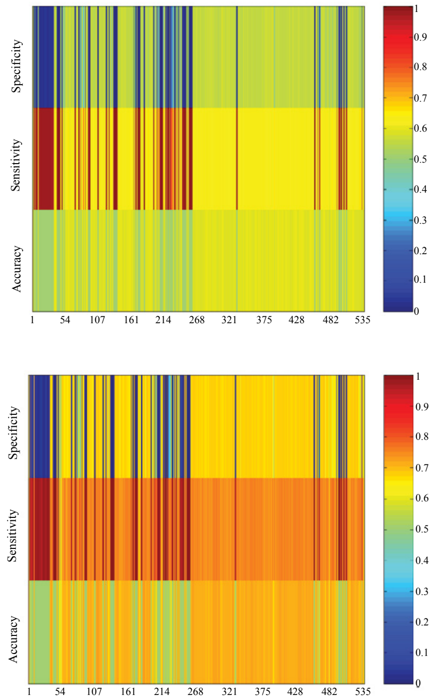

Figure 2: Accuracy, sensitivity and specificity obtained from using 10-1 neural network model to determine if any of 535 amino acid properties was useful to estimate the purification propensity for 857 proteins from *B. halodurans* (upper panel) and to valid this relationship with delete-1 jackknife validation (lower panel).

Protein purification is a long and monotonic process, which involves many factors. Although it is not clear whether an amino acid property can unambiguously reflect a certain aspect of purification process, the fact that an amino acid property cannot be separated from an amino acid renders the base for the possibility to estimate the purification propensity. At this stage, the underlined physicochemical mechanisms in protein purification are too complicated to guide such estimation. However, some reasons could be figured out for unsuitability of some amino acid properties. For example, it is difficult for some properties to reflect an overall property of a protein with different compositions, different neighboring amino acids, etc., although these properties can be weighed with their compositions (columns 6 and 7 in Table 2). On the other hand, the dynamic amino acid properties, such as amino acid distribution probability (the last 2 columns in Table 2), appear flexible because they do not have a simple weighing scheme and do change with respect to different compositions, different neighboring amino acids, and different distributions of amino acids in a protein [30,31]. In future, it is hoped to combine various amino acid properties together to estimate the purification propensity, however, various combinations of 535 amino acid properties would be tremendous, therefore this study hopefully reduced the size of such combinations, and speeds up the research to estimate the purification propensity.

For predicting protein purification, an intriguing question is what kind protein can be predicted successfully. Here we used the amino acid predictability to address this issue. In Figure 4, the upper and middle panels represented the predictive accuracy obtained from fitting (blue bars) and delete-1 jackknife validation (light blue bars), and the x-axis represents 867 *B. halodurans* proteins, which were ranked according to their predictive accuracy of purification. The acceptable accuracy was set as 75% accuracy, by which *B. halodurans* proteins were divided into two groups. Their statistical difference was showed in the lower

Amino acid properties	Accuracy	Sensitivity	Specificity
Results from fitting			
40 constant composition properties	0.51 (0.508 - 0.515)	0.969 (0.942 - 0.981)	0.041 (0.027 - 0.0696)
	[0.505 - 0.716]	[0.725 - 1]	[0 - 0.681]
218 physicochemical properties	0.7 (0.508 - 0.71) b, e	0.755 (0.741 - 0.986) c, e	0.648 (0.0196 - 0.677) b, e
	[0.505 - 0.721]	[0.551 - 1]	[0 - 0.69]
273 secondary structure properties	0.711 (0.706 - 0.714) c	0.745 (0.739 - 0.751) c	0.678 (0.672 - 0.681) c
	[0.505 - 0.722]	[0.678 - 1]	[0 - 0.693]
4 dynamic properties	0.712 (0.671 - 0.73) b	0.745 (0.712- 0.763) b	0.678 (0.629- 0.697) b
	[0.635 - 0.743]	[0.684- 0.776]	[0.586- 0.71]
Results from validation			
40 constant composition properties	0.507 (0.507 - 0.508)	0.968 (0.964 - 0.97)	0.0369 (0.0331 - 0.0408)
	[0.504 - 0.601]	[0.625 - 1]	[0.0005 - 0.569]
218 physicochemical properties	0.59 (0.506- 0.595) a, e	0.636 (0.625- 0.982) b, e	0.544 (0.021- 0.563) a, e
	[0.503 - 0.602]	[0.494- 0.999]	[0.0007- 0.578]
273 secondary structure properties	0.595 (0.592 - 0.596) c	0.626 (0.622- 0.63) c	0.563 (0.559- 0.566) c
	[0.505- 0.601]	[0.601- 0.999]	[0.0007- 0.576]
4 dynamic properties	0.593 (0.573 - 0.596) a [0.554- 0.599]	0.632 (0.607- 0.64) b [0.587- 0.644]	0.549 (0.53- 0.56) a, d [0.52- 0.561]
The data were presented as median with interquatile in parentheses and range in brackets. The letters of a, b and c indicated statistical significance at *P*<0.05, *P*<0.01 and *P*<0.001 levels, respectively, compared with constant composition properties (Mann- Whitney Rank Sum Test). The letters of d and e indicated statistical significance at *P*<0.05 and *P*<0.001 levels compared with secondary structure properties (Mann-Whitney Rank Sum Test).			

Table 1: Results obtained from fitting and delete-1 jack-knife validation by means of 10-1 feed-forward back propagation neural network.

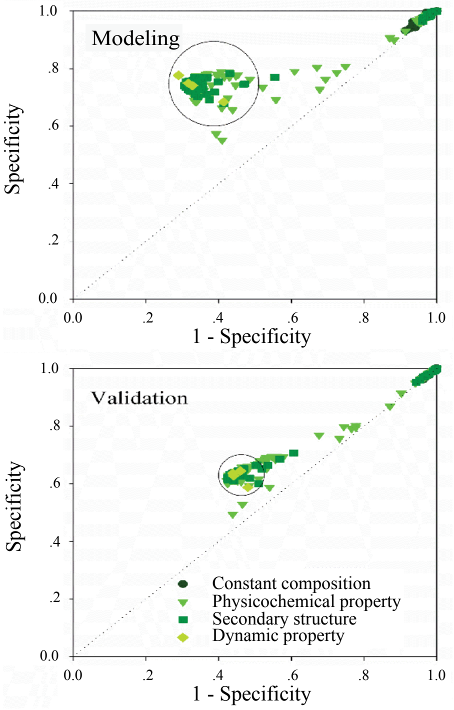

Figure 3: Comparison of predictive sensitivity versus specificity obtained from neural network in ROC analysis. The diagonal line is the line of indiscrimination indicating a completely random guess.

Amino acid	No.		RADA880107		RADA880107 × No		CC (%)		FC (%)		DP	
	P1	P2	P1	P2	P1	P2	P1	P2	P1	P2	P1	P2
A	8	8	-0.29	-0.29	-2.32	-2.32	8.70	8.70	7.25	7.81	0.1682	0.0280
R	7	1	-2.71	-2.71	-18.97	-2.71	7.61	1.09	7.19	5.80	0.0268	1.0000
N	2	1	-1.18	-1.18	-2.36	-1.18	2.17	1.09	4.09	4.57	0.5000	1.0000
D	4	7	-1.02	-1.02	-4.08	-7.14	4.35	7.61	5.86	5.31	0.5625	.2142
C	0	1	0	0	0	0	0	1.09	1.37	2.11	0	1.0000
E	14	8	-1.53	-1.53	-21.42	-12.24	15.22	8.70	5.50	5.80	0.0687	0.2243

Q	6	5	-0.9	-0.9	-5.40	-4.50	6.52	5.43	4.71	3.54	0.0386	0.1920
G	4	9	-0.34	-0.34	-1.36	-3.06	4.35	9.78	6.40	7.53	0.1875	0.0328
H	2	2	-0.94	-0.94	-1.88	-1.88	2.17	2.17	3.62	3.54	0.5000	0.5000
I	8	6	0.24	0.24	1.92	1.44	8.70	6.52	4.89	4.53	0.1682	0.2315
L	8	9	-0.12	-0.12	-0.96	-1.08	8.70	9.78	8.55	8.64	0.2243	0.1967
K	8	8	-2.05	-2.05	-16.40	-16.40	8.70	8.70	4.65	3.24	0.1682	0.2523
M	1	0	-0.24	-0.24	-0.24	0	1.09	0	2.15	1.95	1.0000	0
F	2	4	0	0	0	0	2.17	4.35	2.76	3.48	0.5000	0.1875
P	4	2	0	0	0	0	4.35	2.17	4.95	4.03	0.1875	0.5000
S	3	4	-0.75	-0.75	-2.25	-3.00	3.26	4.35	5.37	5.84	0.6667	0.5625
T	1	3	-0.71	-0.71	-0.71	-2.13	1.09	3.26	4.75	4.63	1.0000	0.6667
W	0	0	-0.59	-0.59	0	0	0	0	0.62	0.69	0	0
Y	3	5	-1.02	-1.02	-3.06	-5.10	3.26	5.43	1.69	2.58	0.6667	0.2880
V	7	9	0.09	0.09	0.63	0.81	7.61	9.78	8.45	9.42	0.1071	0.1475

RADA880107 was a physicochemical property of amino acids that described the energy transfer from out to in (95%buried). P1 and P2 were two proteins with accession number Q9KA18 and Q9K5W1. No., Number of amino acids; CC, %, the current composition of amino acids calculated by the number of a type of amino acids divided by the total number of amino acids in a protein; FC, %, the future composition of amino acids calculated according to the mutating probability (http://www.nerc-nfb.ac.cn/calculation/fc.htm); DP, the distribution probability of amino acids calculated according to the equation, r!/(q0!×q1!×...×qn!)×r!/(r1!×r2!×...×rn!)×n-r, where ! is the factorial, r is the number of a type of amino acid, q is the number of partitions with the same number of amino acids and n is the number of partitions in the protein for a type of amino acid [37].

Table 2: Comparison of weighed schemes of a physicochemical property with dynamic properties in two proteins.

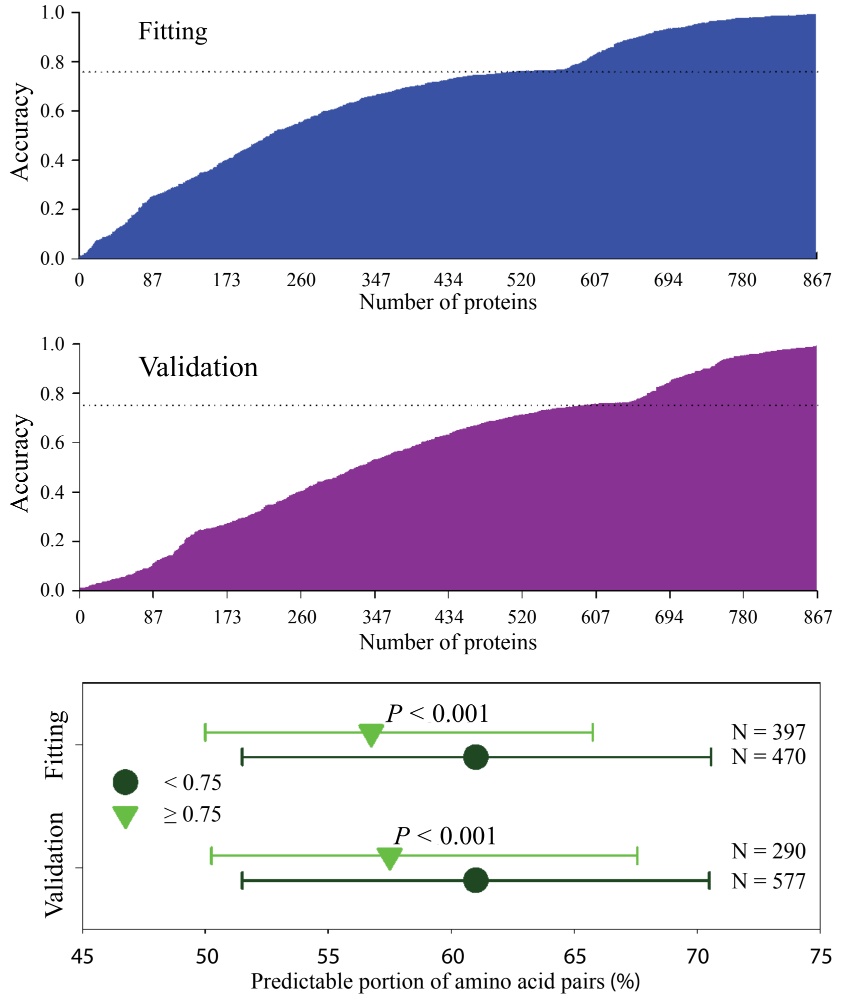

Figure 4: Purification accuracy of *B. halodurans* proteins obtained from fitting (upper panel), delete-1 jackknife validation (middle panel), and statistical comparison of their predictable portion of amino acid pairs (lower panel). The dotted lines indicate the cut-off point for separating the low accuracy from the high one. The data were presented as median with interquartile. Difference between low and high accuracy groups is statistically significant (*P* < 0.001, Mann-Whitney Rank Sum Test).

panel of Figure 4, indicating that the *B. halodurans* proteins with lower predictable portion have better predictive result for their purification state.

Acknowledgements

This study was partly supported by National Natural Science Foundation of China (31400079, 31460296, 21466007 and 31560315), Guangxi Science Foundation (2013GXNSFDA019007) and Special Funds for Building of Guangxi Talent Highland.

References

1. Kusnadi AR, Nikolov ZL, Howard JA (1997) Production of recombinant proteins in transgenic plants: Practical considerations. Biotechnol Bioeng 56: 473-484.
2. Huttinga H (1975) Purification by molecular sievin of a leak virus related to onion yellow dwarf virus. Netherland J Plant Pathol 81: 3.
3. Albrechtsen M, Heide M (1990) Purification of plant viruses and virus coat proteins by high performance liquid chromatography. J Virol Methods 28: 245-256.
4. Lin M, Trottier E, Pasick J, Sabara M (2004) Identification of antigenic regions of the Erns protein for pig antibodies elicited during classical swine fever virus. Infection J Biochem 136: 795-804.
5. Lin M, Trottier E, Pasick J (2005) Antibody responses of pigs to defined Erns fragments after infection with classical swine fever virus. Clin Diag Lab Immunol 12: 180-186.
6. Goodman DB, Church GM, Kosuri S (2013) Causes and effects of N-terminal codon bias in bacterial genes. Science 342: 475-479.
7. Pei J, Pang Q, Zhao L, Fan S, Shi H (2012) Thermoanaerobacterium thermosaccharolyticum β-glucosidase: a glucose-tolerant enzyme with high specific activity for cellobiose. Biotechnol Biofuels 5: 31.
8. Svensson J, Andersson C, Reseland JE, Lyngstadaas P, Bülow L (2006) Histidine tag fusion increases expression levels of active recombinant amelogenin in Escherichia coli. Prot Exp Purification 48: 134-141.
9. Seo JH, Yeo JS, Cha HJ (2005) Baculoviral polyhedrin-*Bacillus thuringiensis* toxin fusion protein: a protein-based bio-insecticide expressed in Escherichia coli. Biotechnol Bioeng 92: 166-172.
10. Wei Q, Kim YS, Seo JH, Jang WS, Lee IH, et al. (2005) Facilitation of expression and purification of an antimicrobial peptide by fusion with baculoviral polyhedrin in Escherichia coli. Appl Environ Microbiol 71: 5038-5043.
11. Roh AR, Nikolov ZL, Howard JA (1997) Production of recombinant proteins in transgenic plants: practical considerations. Biotechnol Bioeng 56: 473-484.
12. Lee KS, Sohn MR, Kim BY, Choo YM, Woo SD (2012) Production of classical swine fever virus envelope glycoprotein E2 as recombinant polyhedra in baculovirus-infected silkworm larvae. Mol Biotechnol 50: 211-220.
13. Bae SM, Kim HJ, Lee JB, Choi JB, Shin TY, et al. (2013) Hyper-enhanced production of foreign recombinant protein by fusion with the partial polyhedrin of nucleopolyhedrovirus. PLoS One 8: e60835.
14. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, et al. (2008) AAindex: amino acid index database, progress report 2008. Nucleic Acids Res 36: D202.
15. Kurgan L, Mizianty MJ (2009) Sequence-based protein crystallization propensity prediction for structural genomics: review and comparative analysis. Natural Sci 1: 93-106.
16. Wang H, Wang M, Tan H, Li Y, Zhang Z, et al. (2014) PredPPCrys: Accurate prediction of sequence cloning, protein production, purification and crystallization propensity from protein sequences using multi-step heterogeneous feature fusion and selection. PLoS One 9: e105902.
17. Takami H, Nakasone K, Takaki Y, Maeno G, Sasaki R, et al. (2000) Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparison with *Bacillus subtilis*. Nucleic Acids Res 28: 4317-4331.
18. Horikoshi K (1999) Alkaliphiles: some applications of their products for biotechnology. Microbiol Mol Biol Rev 63: 735-50.
19. Danesh A, Mamo G, Mattiasson B (2011) Production of haloduracin by *Bacillus halodurans* using solid-state fermentation. Biotechnol Lett 33: 1339-1344.
20. Yeo HK, Park YW, Kang J, Lee JY (2013) Crystallization and preliminary X-ray diffraction analysis of the TetR-family transcriptional repressor YhgD from *Bacillus halodurans.* Acta Crystallogr Sect F Struct Biol Cryst Commun 69: 532-534.
21. Venditto I, Santos H, Sandy J, Sanchez-Weatherby J, Ferreira LM, et al. (2014) Crystallization and preliminary X-ray diffraction analysis of a trimodular endo-β-1,4-glucanase Cel5B from *Bacillus halodurans.* Acta Crystallogr F Struct Biol Commun 70: 1628-1630.
22. Kang J, Park YW, Yeo HK, Lee JY (2015) Crystallization and preliminary X-ray diffraction analysis of the arginine repressor ArgR from *Bacillus halodurans.* Acta Crystallogr F Struct Biol Commun 71: 291-294.
23. Kumazaki K, Kishimoto T, Furukawa A, Mori H, Tanaka Y, et al. (2014) Crystal structure of Escherichia coli YidC, a membrane protein chaperone and insertase. Sci Rep 4: 7299.
24. Kumazaki K, Tsukazaki T, Nishizawa T, Tanaka Y, Kato HE, et al. (2014) Crystallization and preliminary X-ray diffraction analysis of YidC, a membrane-protein chaperone and insertase from *Bacillus halodurans*. Acta Crystallogr F Struct Biol Commun 70Pt 8: 1056-1060.
25. Burley SK (2000) An overview of structural genomics. Nat Struct Biol 7 Suppl: 932-934.
26. Chen L, Oughtred R, Berman HM, Westbrook J (2004) Target-DB: a target registration database for structural genomics projects. Bioinformatics 20: 2860-2862.
27. Joachimiak A (2009) High-throughput crystallography for structural genomics. Curr Opin Struct Biol 19: 573-584.
28. Demuth H, Beale M (2001) Neural network toolbox for use with MatLab. User’s guide, version 4.
29. Cai T, Pepe MS, Zheng Y, Lumley T, Jenny NS (2006) The sensitivity and specificity of markers for event times. Biostatistics 7: 182-197.
30. Wu G, Yan S (2008) Lecture notes on computational mutation. Nova Science Publishers, New York.
31. Yan SM, Wu G (2010) Creation and application of computational mutation. J Guangxi Academy Sci 26: 130-139.
32. Yan S, Wu G (2011) Possible random mechanism in crystallization evidenced in proteins from *Plasmodium Falciparum*. Cryst Growth Des 11: 4198-4204.
33. Yan S, Wu G (2012) Correlating dynamic amino acid properties with success rate of crystallization of proteins from *Bacteroides vulgatus*. Cryst Res Technol 47: 511-516.
34. Yan S, Wu G (2012b) Randomness in crystallization of proteins from *Staphylococcus aureus.* Protein Pept Lett 19: 784-789.
35. Yan S, Wu G (2013) Association of combined features of amino acid and protein with crystallization propensity of proteins from *Cytophaga hutchinsonii*. Zeitschrift Fur Kristallographie 228: 250-254.
36. Yan SM, Wang HJ, Wu G (2013) Correlation of combined features of amino acid and protein with crystallization propensity of proteins from Caenorhabditis elegans. Guangxi Sci 20: 234-238.
37. Feller W (1968) An Introduction to Probability Theory and Its Applications. Third ed, Vol, I. Wiley, New York.