http://www.scirp.org/journal/jbise

J. Biomedical Science and Engineering, 2019, Vol. 12, (No. 4), pp: 245-256

Correlation of Combined Characters of Amino Acid and Whole Protein with Success Rate of Crystallization of *Lactobacillus* Proteins

Shaomin Yan, Guang Wu

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

Correspondence to: Guang Wu, hongguanglishibahao@gxas.cn

Keywords: Amino Acid Character, Distribution Probability, Future Composition, *Lactobacillus*, Protein Crystallization

Received: February 18, 2019 **Accepted:** April 25, 2019 **Published:** April 28, 2019

ABSTRACT

Crystallization of proteins is a very delicate process, which is influenced by many known and unknown factors. Of tested factors, many factors are exclusively related to individual amino-acid characters such as molecular weight or protein characters such as protein length. It is considered necessary to test factors that combine both individual amino-acid characters and protein characters with respect to success rate of crystallization. In this study, two combined characters characterizing individual amino-acid character and protein character, amino acid distribution probability and future composition, were used to correlate the success rate of crystallization of proteins from *Lactobacillus* via modeling. The results obtained from logistic regression and neural network were compared against the results obtained from each of 533 individual amino-acid characters. This study demonstrated that the combined characters are involved in crystallization process and should be taken into account for predicting the success rate of crystallization process.

1. INTRODUCTION

Crystallization of proteins is a very delicate process and costs time, because many known and unknown factors influence the process of crystallization. Therefore, it is hoped that a law between affecting factors and crystallization can be found to facilitate this process, *i.e.* to predict whether its protein is likely to be crystallized. Over last years, intensive efforts are made to search various factors, and then correlate these factors with the success rate of crystallization of proteins [1-8]. Technically, these factors should be

numeric in order to correlate with success rate of crystallization. Efforts have become less impressed recently because almost all known factors have been tested without much improvement on predictions. Of tested factors, many factors are exclusively related to individual amino-acid characters, for example, molecular weight of amino acid, whereas a small number of tested factors are related to the whole protein characters, for example, the length of a protein.

Really, it is necessary to correlate the factors that combine both individual amino-acid characters and whole protein characters with the rate of protein crystallization. This is because 1) an individual amino-acid character is a fixed numerically number, for example, molecular weight, no matter whether an amino acid is in a protein or exists individually, and 2) protein characters appear simple in the previous studies. In fact, there is a combined character, *i.e.* the amino acid composition that represents very basic character of proteins and has been widely used in various analyses. However, new combined characters are needed in order to understand the nature of protein from different angles.

Over the last decade, we have developed three combined characters characterizing individual amino acid and protein together, and we have applied them to many different studies, for example, protein evolution, drug target designing, determination of mutation patterns, analysis of genetic disorder, protein structure and function, and prediction of mutation of influenza A viruses [9-12]. The results demonstrate the applicability and advantage of the combined characters, thus it is our desire to correlate these combined characters with the success rate of crystallization of proteins.

Technically, the relationship between various factors and success rate of crystallization of proteins was established via modeling, because it is impossible to run a control experiment without individual amino-acid characters and protein characters. So far, logistic regression was a major tool to model the relationship, because whether a protein can be crystallized is a yes-no event while protein sequences were encoded using individual amino-acid characters [4-6]. In this study, an attempt was made to test the role of combined characters in crystallization of *Lactobacillus* proteins via logistic regression and neural network model, whose results were compared with the results obtained from each of 531 individual amino-acid characters.

We chose *Lactobacillus*, not only because it is important for human health with food industrial perspective [13-15], but also because big efforts were made to crystallize its proteins. The sample of data is relatively larger than proteins from other species of interests [16].

2. MATERIALS AND METHODS

2.1. Data

314 proteins from *Lactobacillus* were found in TargetDB [16] under the criterion of purified proteins before 2011, of which 141 were found under the criterion of crystallized protein. Those two criteria were used in previous studies [17-22].

2.2. Combined Characters

The combined characters means that a character that combines a character of an individual amino acid and a character of a protein in terms of numerical value. For example, the molecular weight of an amino acid is a character of an individual amino acid and is unchangeable no matter where the given amino acid is located at any position in a protein. Although it is true that the molecular weight is unchangeable, the amino acid should affect the crystallization of a protein differently when it is located at different position. Similarly, the length of protein is a character associated with a whole protein, however it losses the individuality of composed amino acids, because the proteins with same length do not grantee the same crystallization propensity because they can have different amino acid compositions. So it is important to have a combined character forming from both the character of an individual amino and the character of a whole protein.

The first combined character is the amino acid distribution probability, which is based on the occu-

pancy of subpopulations and partitions [23] with its online computation [24]. Two worked examples were listed in columns 8 and 9 of Table 1 to show how this combined character is different from protein to protein.

Table 1. Comparison of characters of individual amino acid and combined character of individual amino acid and of a whole protein.

Amino Acid	Number		OOBM850101		OOBM850101 × Number		Distribution probability		Future composition, %	
	Protein 1	Protein 2	Protein 1	Protein 2	Protein 1	Protein 2	Protein 1	Protein 2	Protein 1	Protein 2
A	29	23	2.01	2.01	58.29	46.23	0.0069	0.0153	9.00	8.12
R	10	8	0.84	0.84	8.4	6.72	0.0008	0.2243	7.10	6.82
N	9	7	0.03	0.03	0.27	0.21	0.1475	0.3213	3.60	3.78
D	6	21	−2.05	−2.05	−12.3	−43.05	0.1543	0.0270	4.64	4.06
C	1	0	1.98	1.98	1.98	0	1.0000	0.0000	1.71	2.04
E	15	6	1.02	1.02	15.3	6.12	0.0196	0.3472	4.44	4.92
Q	13	12	0.93	0.93	12.09	11.16	0.0221	0.1241	3.09	2.76
G	22	15	0.12	0.12	2.64	1.8	0.0878	0.0981	7.69	6.49
H	5	5	−0.14	−0.14	−0.7	−0.7	0.1920	0.3840	3.20	4.13
I	14	12	3.7	3.7	51.8	44.4	0.0550	0.1241	5.28	5.27
L	17	23	2.73	2.73	46.41	62.79	0.0183	0.0791	7.71	9.86
K	12	9	2.55	2.55	30.6	22.95	0.0621	0.1475	3.96	3.00
M	5	6	1.75	1.75	8.75	10.5	0.1920	0.2315	1.82	1.66
F	3	9	2.68	2.68	8.04	24.12	0.6667	0.1770	2.59	2.99
P	4	14	0.41	0.41	1.64	5.74	0.0938	0.0550	5.09	6.07
S	14	9	1.47	1.47	20.58	13.23	0.0550	0.0492	7.24	6.73
T	14	11	2.39	2.39	33.46	26.29	0.1649	0.1616	6.71	5.87
W	0	4	2.49	2.49	0	9.96	0.0000	0.5625	0.76	0.60
Y	6	7	2.23	2.23	13.38	15.61	0.3472	0.2142	1.76	2.64
V	20	18	3.5	3.5	70	63	0.0338	0.0831	8.68	8.57

OOBM850101 is a character of individual amino acid that describes the optimized beta-structure-coil equilibrium constant. P1 and P2 are two proteins with accession number LdR34 and LpR114. The amino acid distribution probability was computed according to the equation,

$r!/(q_0!\times q_1!\times\cdots\times q_n!)\times r!/(r_1!\times r_2!\times\cdots\times r_n!)\times n^{-r}$, 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.

The second combined character is the amino acid future composition, which is based on the relationship between RNA codons and their translated amino acids [25-27] with its online computation [28]. Two worked examples were listed in columns 10 and 11 of Table 1 to show how this combined character is different from protein to protein.

2.3. Characters for Comparison

A database, called AAIndex, contains more than 530 different individual amino-acid characters [29]. Some are quite familiar to us, for examples, physicochemical properties, spatial properties [29], electronic properties [30], hydrophobic properties [31], predictors for secondary structures [32], and so on. These individual amino-acid characters are constants, *i.e.*, each character generally has an unchangeable value for an amino acid, for example, molecular weight for alanine is 89.09. Each individual amino-acid character is put into model to predict the success rate of crystallization of *Lactobacillus* proteins each time for comparison with the results obtained from combined characters.

2.4. Models

Logistic regression and 18-1 neural network were used, because the success rate of protein crystallization was a yes-no event while any character is a number for a type of amino acid, *i.e.* the model outcome is defined as unity when a protein can be crystallized and the model outcome is defined as zero when a protein cannot be crystallized.

2.5. Statistics

MatLab was used to perform both logistic regression and neural network [33, 34]. The results obtained from each predictor were classified as true positive (TP), true negative (TN), false positive (FP) and false negative (FN). The accuracy, sensitivity and specificity were calculated as follows: Accuracy = (TP + TN)/(TP + FP + TN + FN) × 100, Sensitivity = (TP)/(TP + FN) × 100, and Specificity = (TN)/(TN + FP) × 100. The McNemar's test was used to compare the classified results. Sensitivity and specificity were compared using receiver operating characteristic (ROC) analysis [35-37].

3. RESULTS AND DISCUSSION

Table 1 compares the difference between an individual amino-acid character, OOBM850101, and combined characters. No matter what an amino-acid character describes, its value for each type of amino acid is unchangeable (columns 4 and 5). This appears counter-intuitive when we use it to describe an amino acid in a protein because intuitively an amino acid should have different values in terms of different position, different neighboring amino acid and different composition. On the other hand, we can weigh an individual amino-acid character with amino acid composition (columns 6 and 7). As a result, the combined characters do have different values for the same type of amino acids when they are located at different positions, when their neighboring amino acids are different and when their number in a protein is different (last four columns). Therefore, the combined characters are more meaningful but their values have to be computed for each type of amino acid in each protein.

Figure 1 showed the results of accuracy, sensitivity and specificity obtained using logistic regression to correlate the success rate of protein crystallization with each of two combined characters and each of 533 individual amino-acid characters. In this figure: each bar represented how many characters used in predictions resulted in a similar accuracy, sensitivity and specificity. For example, the most right bars in upper, middle and lower indicated that the predictions using each of 483 individual amino-acid characters produced a similar accuracy of 0.6, the predictions using each of 488 individual amino-acid characters produced a similar sensitivity of 0.6, and the prediction using an individual amino-acid character produces the highest specificity. For another example, VENT840101 and FAUJ880112 had the accuracy of 0.53 and 0.55 in the first and second bars from left-hand in the upper panel, while the third bar indicated that three

individual amino-acid characters, FAUJ880111, CHAM830107 and NOZY710101, had similar accuracies (0.58 ± 0.01). Figure 1 clearly showed that two combined characters, distribution probability and future composition, had a relative good relationship with the success rate of crystallization of *Lactobacillus* proteins.

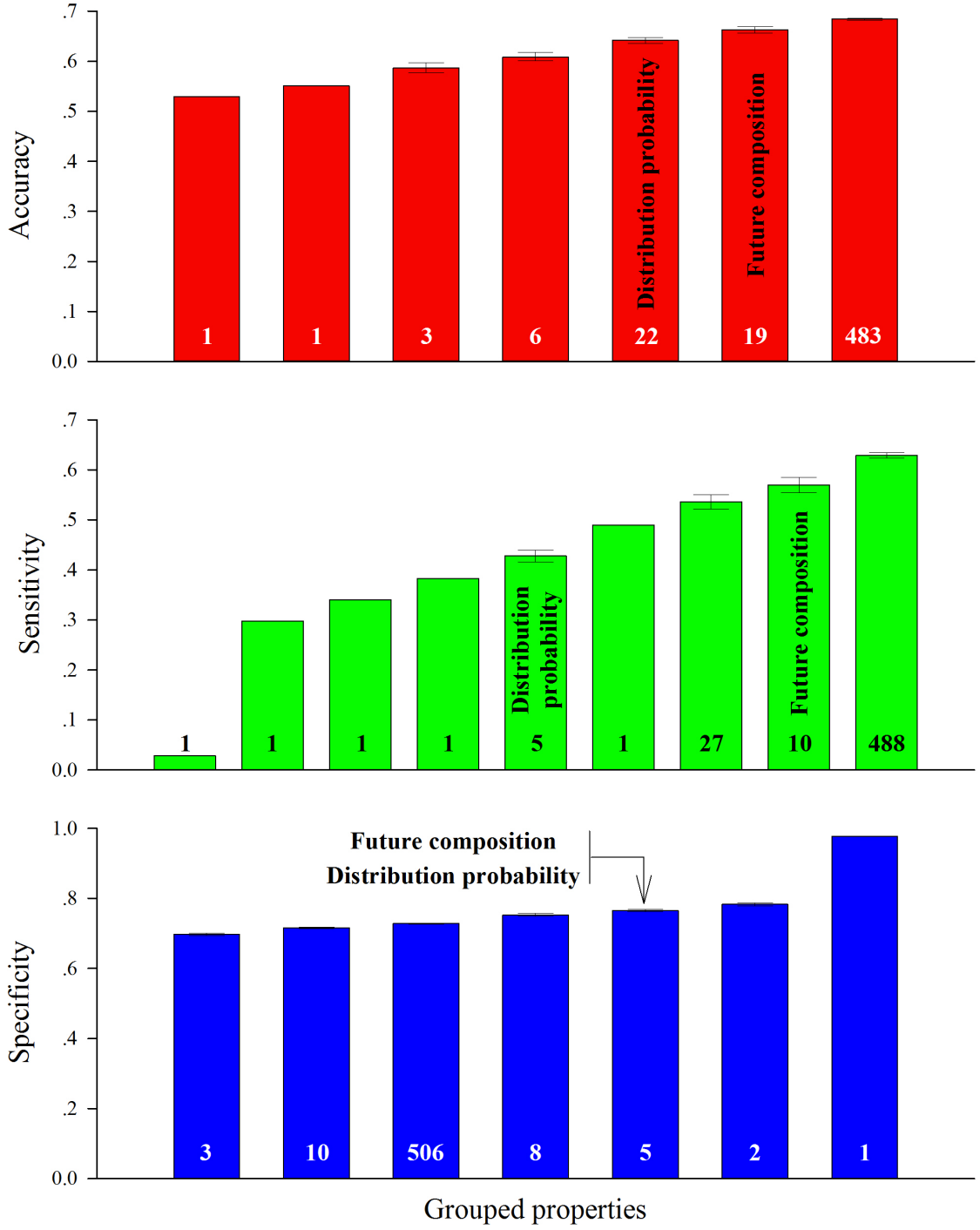

Figure 1. Accuracy, sensitivity and specificity of predictions using logistic regression to model the success rate of crystallization of proteins from *Lactobacillus* and each of 535 characters. The text labels are the combined characters introduced in this study.

A frequent question in modeling is whether predictors result in a random prediction, which especially is the case for yes-no event prediction because yes-no event can easily connect with random tossing a coin. As good performance includes high true positive rate and low false positive rate, these render the ROC (receiver operating characteristic) analysis, where *x*-axis represented the false positive rate and *y*-axis represented the true positive rate. **Figure 2** demonstrated the comparison of sensitivity versus 1-specificity obtained from logistic regression, where *x*-axis represented 1-specificity and *y*-axis represented the sensitivity. As can be seen, the ratios of sensitivity versus 1-specificity appear on upper-left area above the diagonal, indicating these characters give a good prediction. The McNemar's test shows that such classified results are significantly different from those of random guess ($P < 0.05$). However, only one circle is located near the lower left corner, which resulted from an individual amino-acid character, FAUJ880112, reflecting negative charge. Thus, this individual amino-acid character, FAUJ880112, is not suitable to predict the success rate of crystallization of *Lactobacillus* proteins.

Figure 3 showed the results of accuracy, sensitivity and specificity obtained using 18-1 feedforward backpropagation neural network to correlate the success rate of protein crystallization with each of two combined characters and each of 533 individual amino-acid characters. **Figure 3** had similar explanations and implications as those in **Figure 1**. Clearly, the neural network can furthermore distinguish the difference between characters for prediction of the success rate of protein crystallization. Compared against individual amino-acid characters, **Figures 1-3** suggested that the two combined characters are sensitive to the crystallization process of *Lactobacillus* proteins. Not surprisingly, many individual amino-acid characters generated similar results, being consistent with the study showing the abundance in individual amino-acid characters [38].

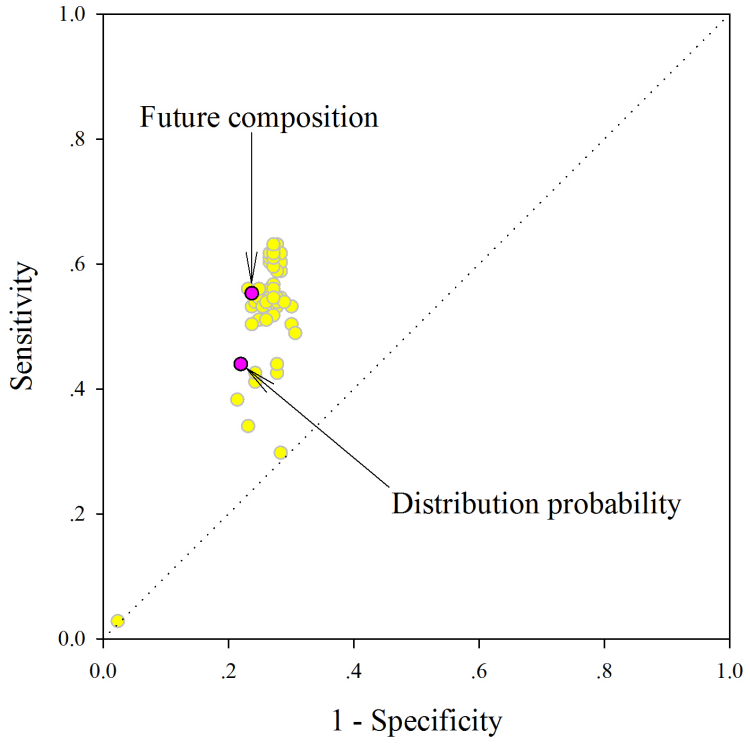

Figure 2. Comparison of sensitivity versus specificity obtained from logistic regression in ROC analysis. Each yellow circle is a result obtained using an individual amino-acid character while each pink circle is a result obtained using one of two combined characters. The diagonal line is the line of indiscrimination indicating a completely random guess. The text labels are the combined characters introduced in this study.

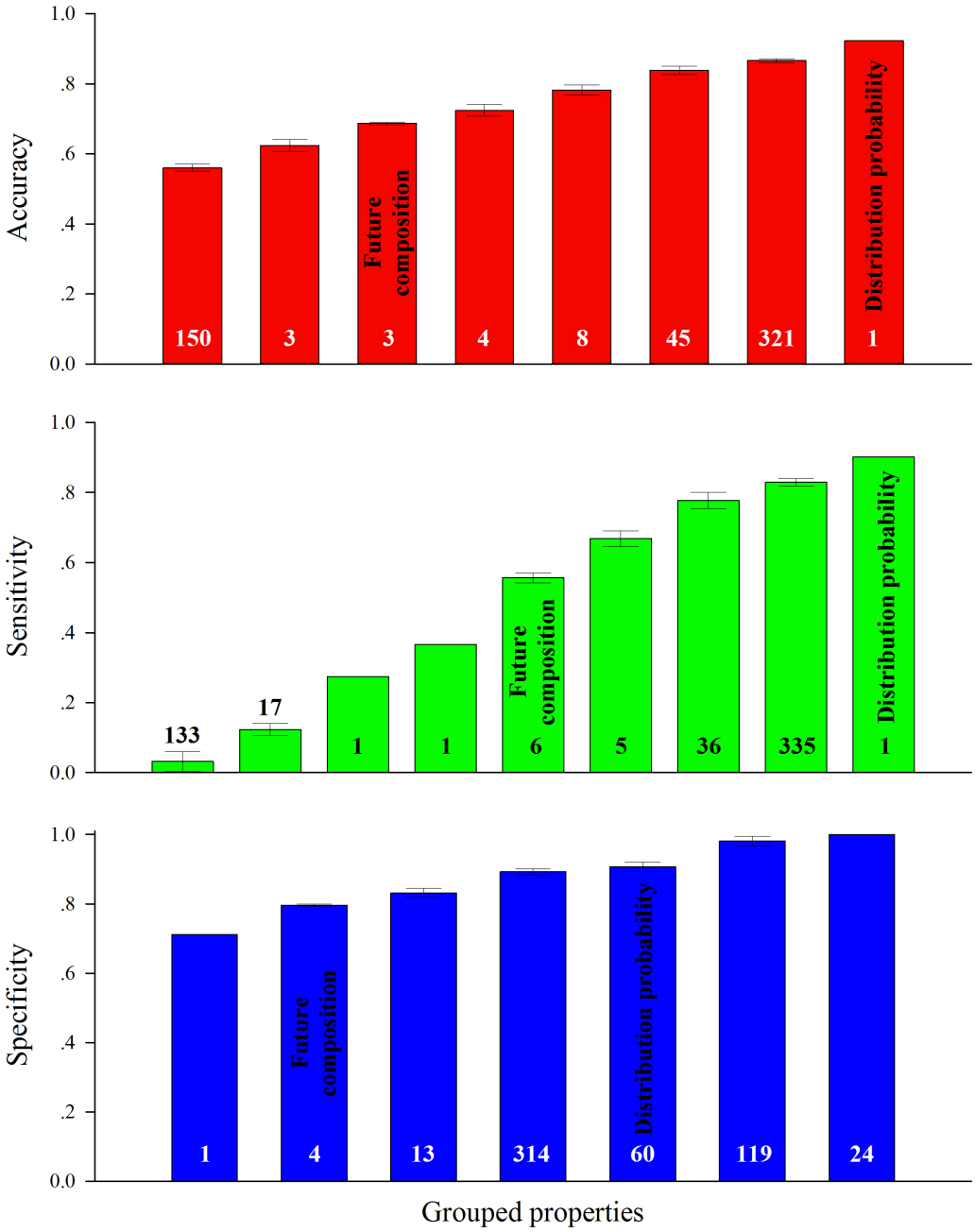

Figure 3. Accuracy, sensitivity and specificity obtained using neural network to model the success rate of protein crystallization from *Lactobacillus* and each of 535 characters. The text labels are the combined characters introduced in this study.

For the results in Figures 1-3, the database was not divided, *i.e.*, the model parameters obtained from the 314 *Lactobacillus* proteins were used for predictions. This was generally considered as the first stage in modeling, and then the database should be divided as two groups, one for the generation of model parameters while the other for the validation [39]. Figure 4 displayed the accuracy, sensitivity and specificity obtained using delet-1 jackknife validation, which further demonstrated that the predictions using combined characters were not worse than those using individual amino-acid characters.

Figure 5 displayed the results of ROC analysis with respect to fitting and delete-1 jackknife validation

using 18-1 feed forward back propagation neural network. Although the McNemar's test shows that such classified results are significantly different from those of random guess ($P < 0.05$), a cluster of circles appear at the lower left corner and near the diagonal indicating that 152 individual amino-acid characters result in the sensitivity smaller than 0.5 in the fitting (upper panel of Figure 5) therefore these characters cannot be used as predictors. On the contrary, the two combined characters and other individual amino-acid characters can be used to predict the success rate of crystallization of *Lactobacillus* proteins.

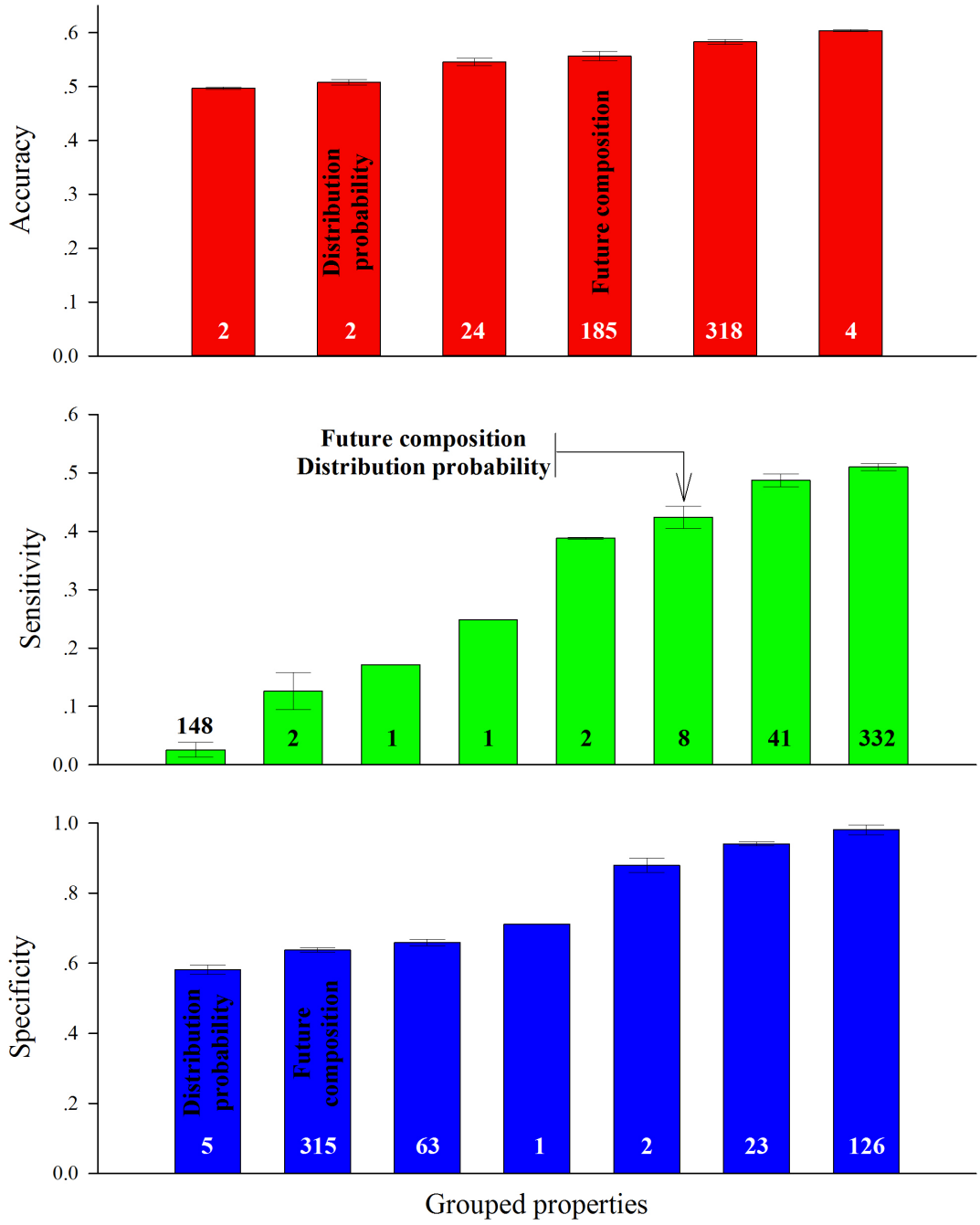

Figure 4. Accuracy, sensitivity and specificity of delete-1 jackknife validation obtained using neural network to model the success rate of crystallization of proteins from *Lactobacillus* and each of 535 characters. The text labels are the combined characters introduced in this study.

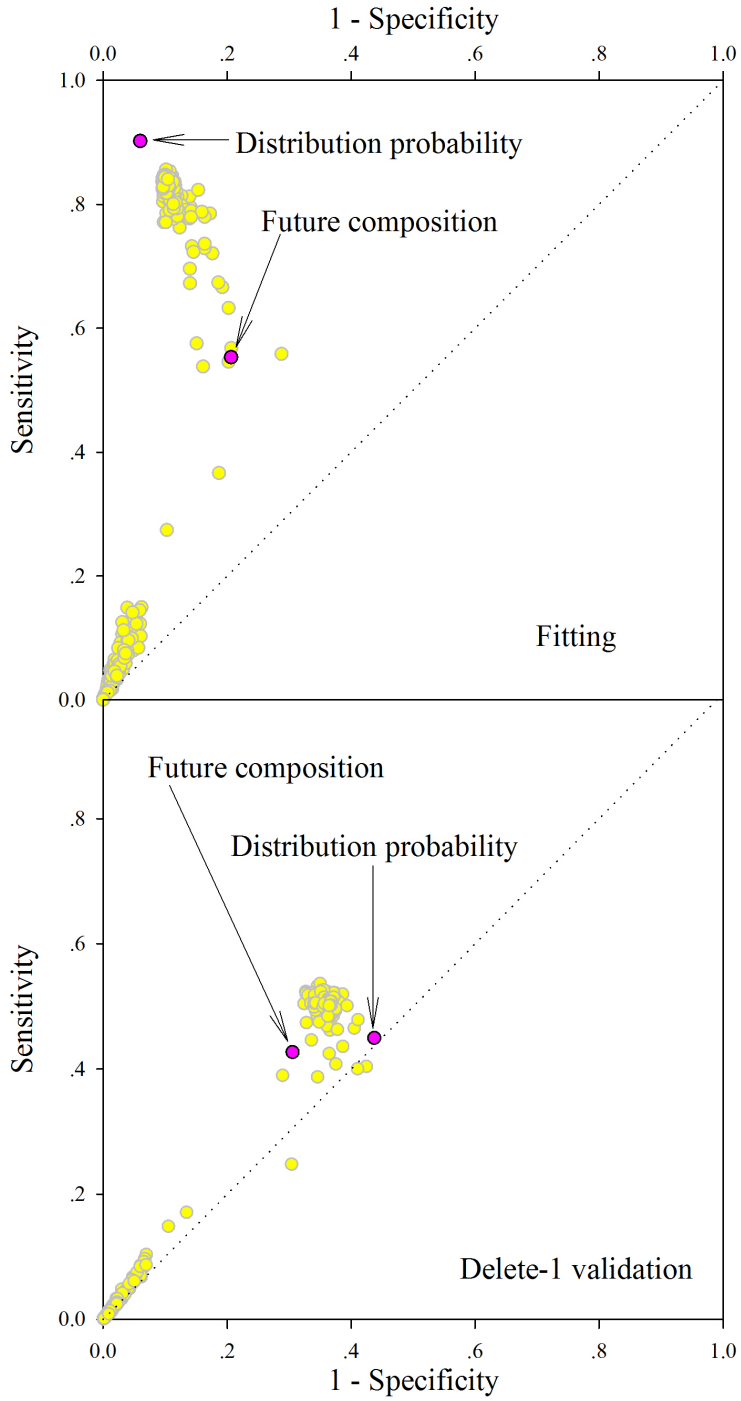

Figure 5. Comparison of sensitivity versus specificity obtained from neural network in ROC analysis. Each yellow circle is a result obtained using an individual amino-acid character while each pink circle is a result obtained using one of two combined characters. The diagonal line is the line of indiscrimination indicating a completely random guess. The text labels are the combined characters introduced in this study.

Actually, the workload in this study is not small at all because the proposed combined characters has been checked against each of 532 individual amino-acid characters in order to get a solid conclusion.

The current practice on prediction of success rate of crystallization employs as many characters as possible, such as hybrid crystal growth predictive model [7], "sticky patch" model [40], theoretical underpinning using a solubility phase diagram [41]. Therefore, we would expect that our proposed combined characters would be included in the factors, which influence the success rate of crystallization of *Lactobacillus* proteins.

At present, to build a predictable relationship between individual protein and its crystallization propensity is still difficult when using either logistical model or neural network model. This suggests that the more sophisticated model could be more suitable for such studies in future, for example, deep learning model. On the other hand, the introduction of cryo-electron microscopy to determine the protein 3-demensional structure reduces the demand for crystallization of proteins for X-ray crystallography [42], however the relationship between individual protein and its crystallization propensity is still important.

FUND

This study was supported by National Natural Science Foundation of China (31460296 and 31560315), Key Project of Guangxi Scientific Research and Technology Development Plan (AB17190534) and Special Funds for Building of Guangxi Talent Highland.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest regarding the publication of this paper.

REFERENCES

1. Smialowski, P., Schmidt, T., Cox, J., Kirschner, A. and Frishman, D. (2006) Will My Protein Crystallize? A Sequence-Based Predictor. *Proteins*, **62**, 343-355. https://doi.org/10.1002/prot.20789
2. Slabinski, L., Jaroszewski, L., Rychlewski, L., Wilson, I.A., Lesley, S.A. and Godzik, A. (2007) XtalPred: A Web Server for Prediction of Protein Crystallizability. *Bioinformatics*, **23**, 3403-3405. https://doi.org/10.1093/bioinformatics/btm477
3. Chen, K., Kurgan, L. and Rahbari, M. (2007) Prediction of Protein Crystallization Using Collocation of Amino Acid Pairs. *Biochemical and Biophysical Research Communications*, **355**, 764-759. https://doi.org/10.1016/j.bbrc.2007.02.040
4. Overton, I.M., Padovani, G., Girolami, M.A. and Barton, G.J. (2008) ParCrys: A Parzen Window Density Estimation Approach to Protein Crystallization Propensity Prediction. *Bioinformatics*, **24**, 901-907. https://doi.org/10.1093/bioinformatics/btn055
5. Kurgan, L. and Mizianty, M.J. (2009) Sequence-Based Protein Crystallization Propensity Prediction for Structural Genomics: Review and Comparative Analysis. *Natural Science*, **1**, 93-106. https://doi.org/10.4236/ns.2009.12012
6. Kurgan, L., Razib, A.A., Aghakhani, S., Dick, S., Mizianty, M.J. and Jahandideh, S. (2009) CRYSTALP2: Sequence-Based Protein Crystallization Propensity Prediction. *BMC Structural Biology*, **9**, 50. https://doi.org/10.1186/1472-6807-9-50
7. Zucker, F.H., Stewart, C., dela Rosa, J., Kim, J., Zhang, L., Xiao, L., Ross, J., Napuli, A.J., Mueller, N., Castaneda, L.J., Nakazawa Hewitt, S.R., Arakaki, T.L., Larson, E.T., Subramanian, E., Verlinde, C.L., Fan, E., Buckner, F.S., Van Voorhis, W.C., Merritt, E.A. and Hol, W.G. (2010) Prediction of Protein Crystallization Outcome Using a Hybrid Method. *Journal of Structural Biology*, **171**, 64-73. https://doi.org/10.1186/1472-6807-9-50
8. Wang, H., Feng, L., Webb, G.I., Kurgan, L., Song, J. and Lin, D. (2018) Critical Evaluation of Bioinformatics Tools for the Prediction of Protein Crystallization Propensity. *Briefings in Bioinformatics*, **19**, 838-852. https://doi.org/10.1093/bib/bbx018

9. Wu, G. and Yan, S.M. (2002) Randomness in the Primary Structure of Protein: Methods and Implications. *Molecular Biology Today*, **3**, 55-69.
10. Wu, G. and Yan, S. (2006) Mutation Trend of Hemagglutinin of Influenza A Virus: A Review from Computational Mutation Viewpoint. *Acta Pharmacologica Sinica*, **27**, 513-526. https://doi.org/10.1111/j.1745-7254.2006.00329.x
11. Wu, G. and Yan, S. (2008) Lecture Notes on Computational Mutation. Nova Science Publishers, New York.
12. Wu, G. and Yan, S. (2010) Creation and Application of Computational Mutation. *Journal of Guangxi Academy of Sciences*, **17**, 145-150.
13. Seddik, H.A., Bendali, F., Gancel, F., Fliss, I., Spano, G. and Drider, D. (2019) *Lactobacillus plantarum* and Its Probiotic and Food Potentialities. *Probiotics and Antimicrobial Proteins*, **9**, 111-122. https://doi.org/10.1007/s12602-017-9264-z
14. Salas-Jara, M.J., Ilabaca, A., Vega, M. and García, A. (2016) Biofilm Forming *Lactobacillus*: New Challenges for the Development of Probiotics. *Microorganisms*, **4**, pii: E35. https://doi.org/10.3390/microorganisms4030035
15. Martín, R., Miquel, S., Ulmer, J., Kechaou, N., Langella, P. and Bermúdez-Humarán, L.G. (2013) Role of Commensal and Probiotic Bacteria in Human Health: A Focus on Inflammatory Bowel Disease. *Microbial Cell Factories*, **12**, 71. https://doi.org/10.1186/1475-2859-12-71
16. Chen, L., Oughtred, R., Berman, H.M. and Westbrook, J. (2004) TargetDB: A Target Registration Database for Structural Genomics Projects. *Bioinformatics*, **20**, 2860-2862. https://doi.org/10.1093/bioinformatics/bth300
17. Yan, S. and Wu, G. (2011) Possible Random Mechanism in Crystallization Evidenced in Proteins from *Plasmodium falciparum*. *Crystal Growth & Design*, **11**, 4198-4204. https://doi.org/10.1021/cg200814k
18. Yan, S. and Wu, G. (2012) Correlating Dynamic Amino Acid Properties with Success Rate of Crystallization of Proteins from *Bacteroides vulgatus*. *Crystal Research and Technology*, **47**, 511-516. https://doi.org/10.1002/crat.201200007
19. Yan, S. and Wu, G. (2012) Randomness in Crystallization of Proteins from *Staphylococcus aureus*. *Protein & Peptide Letters*, **19**, 784-789. https://doi.org/10.2174/092986612800793190
20. Yan, S. and 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. https://doi.org/10.1524/zkri.2013.1570
21. Yan, S., Wang, H. and Wu, G. (2013) Correlation of Combined Features of Amino Acid and Protein with Crystallization Propensity of Proteins from *Caenorhabditis elegans*. *Guangxi Sciences*, **20**, 234-238.
22. Yan, S. and Wu, G. (2015) Predicting Crystallization Propensity of Proteins from *Arabidopsis thaliana*. *Biological Procedures Online*, **17**, 16. https://doi.org/10.1186/s12575-015-0029-3
23. Feller, W. (1968) An Introduction to Probability Theory and Its Applications. Third Edition, Wiley, New York, Vol. 1.
24. http://www.gxas.cn/dp.htm
25. Wu, G. and Yan, S. (2005) Determination of Mutation Trend in Proteins by Means of Translation Probability between RNA Codes and Mutated Amino Acids. *Biochemical and Biophysical Research Communications*, **337**, 692-700. https://doi.org/10.1016/j.bbrc.2005.09.106
26. Wu, G. and Yan, S. (2006) Determination of Mutation Trend in Hemagglutinins by Means of Translation Probability between RNA Codons and Mutated Amino Acids. *Protein & Peptide Letters*, **13**, 601-609. https://doi.org/10.2174/092986606777145779
27. Wu, G. and Yan, S. (2007) Translation Probability between RNA Codons and Translated Amino Acids, and Its

Applications to Protein Mutations. In: Ostrovskiy M.H., Ed., *Leading-Edge Messenger RNA Research Communications*, Nova Science Publishers, New York, Chapter 3, 47-65.

28. http://www.gxas.cn/fc.htm
29. Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T. and Kanehisa, M. (2008) AAindex: Amino Acid Index Database, Progress Report 2008. *Nucleic Acids Research*, **36**, D202-D205. https://doi.org/10.1093/nar/gkm998
30. Darby, N.J. and Creighton, T.E. (1993) Dissecting the Disulphide-Coupled Folding Pathway of Bovine Pancreatic Trypsin Inhibitor. Forming the First Disulphide Bonds in Analogues of the Reduced Protein. *Journal of Molecular Biology*, **232**, 873-896. https://doi.org/10.1006/jmbi.1993.1437
31. Dwyer, D.S. (2005) Electronic Properties of Amino Acid Side Chains: Quantum Mechanics Calculation of Substituent Effects. *BMC Chemical Biology*, **5**, 2. https://doi.org/10.1186/1472-6769-5-2
32. Chou, P.Y. and Fasman, G.D. (1978) Prediction of Secondary Structure of Proteins from Amino Acid Sequence. *Advances in Enzymology and Related Subjects of Biochemistry*, **47**, 45-48.
33. Demuth, H. and Beale, M. (2001) Neural Network Toolbox for Use with MatLab. User's Guide, Version 4.
34. MathWorks Inc. (1984-2001) MatLab—The Language of Technical Computing. Version 6.1.0.450, Release 12.1.
35. Alonzo, T. and Pepe, M.S. (2002) Distribution-Free ROC Analysis Using Binary Regression Techniques. *Biostatistics*, **3**, 421-432. https://doi.org/10.1093/biostatistics/3.3.421
36. Cai, T.X., Pepe, M.S., Zheng, Y.Y., Lumley, T. and Jenny, N.S. (2006) The Sensitivity and Specificity of Markers for Event Times. *Biostatistics*, **7**, 182-197. https://doi.org/10.1093/biostatistics/kxi047
37. Pepe, M., Longton, G. and Janes, H. (2009) Estimation and Comparison of Receiver Operating Characteristic Curves. *Stata Journal*, **9**, 1. https://doi.org/10.1177/1536867X0900900101
38. Atchley, W.R., Zhao, J., Fernandes, A.D. and Druke, T. (2005) Solving the Protein Sequence Metric Problem. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 6395-6400. https://doi.org/10.1073/pnas.0408677102
39. Chou, K.C. (2011) Some Remarks on Protein Attribute Prediction and Pseudo Amino Acid Composition (50th Anniversary Year Review). *Journal of Theoretical Biology*, **273**, 236-247. https://doi.org/10.1016/j.jtbi.2010.12.024
40. Derewenda, Z.S. and Godzik, A. (2017) The "Sticky Patch" Model of Crystallization and Modification of Proteins for Enhanced Crystallizability. *Methods in Molecular Biology*, **1607**, 77-115. https://doi.org/10.1007/978-1-4939-7000-1_4
41. Altan, I., Charbonneau, P. and Snell, E.H. (2016) Computational Crystallization. *Archives of Biochemistry and Biophysics*, **602**, 12-20. https://doi.org/10.1016/j.abb.2016.01.004
42. Cressey, D. and Callaway, E. (2017) Cryo-Electron Microscopy Wins Chemistry Nobel. *Nature*, **550**, 167. https://doi.org/10.1038/nature.2017.22738