Prediction of Crystallization Propensity of Proteins from *Bacillus haloduran* Using Various Amino Acid and Protein Features

Shaomin Yan, Guang Wu

State Key Laboratory of Non-Food Biomass and Enzyme Technology, Guangxi Key Laboratory of Bio-Refinery, Guangxi Biomass Engineering Technology Research Center, National Engineering Research Center for Non-Food Bio-Refinery, Guangxi Academy of Sciences, Guangxi, China

Correspondence to: Guang Wu, hongguanglishibahao@gxas.cn

Keywords: Protein Feature, *Bacillus haloduran*, Protein Crystallization

Received: October 29, 2019 **Accepted:** December 7, 2019 **Published:** December 10, 2019

ABSTRACT

Correct prediction of propensity of crystallization of proteins is important for cost- and time-saving in determination of 3-demensional structures because one can focus to crystallize the proteins whose propensity is high through predictions instead of choosing proteins randomly. However, so far this job has yet to accomplish although huge efforts have been made over years, because it is still extremely hard to find an intrinsic feature in a protein to directly relate to the propensity of crystallization of the given protein. Despite of this difficulty, efforts are never stopped in testing of known features in amino acids and proteins versus the propensity of crystallization of proteins from various sources. In this study, the comparison of the features, which were developed by us, with the features from well-known resource for the prediction of propensity of crystallization of proteins from *Bacillus haloduran* was conducted. In particular, the propensity of crystallization of proteins is considered as a yes-no event, so 185 crystallized proteins and 270 uncrystallized proteins from *B. haloduran* were classified as yes-no events. Each of 540 amino-acid features including the features developed by us was coupled with these yes-no events using logistic regression and neural network. The results once again demonstrated that the predictions using the features developed by us are relatively better than the predictions using any of 540 amino-acid features.

1. INTRODUCTION

The prediction of propensity of crystallization of proteins from various bacteria is an important aspect of our studies [1-8] because this research direction is still active [9-15], though, after years of investigation. Statistically, the predictions are better than random chance throughout studies, even with very high successful rate. However, this is still a phenomenon based approach because it still cannot figure out the deeply-uncovered factors, which determine the propensity of crystallization.

Of predictors, an important group of predictors is physicochemical features of amino acids. However, no solid and general conclusion could be easily reached on which physicochemical feature is better to predict the crystallization propensity [16]. Yet, the protein crystallization more and more becomes a routine work in many laboratories, which require simple and reliable methods to predict the propensity of crystallization of proteins of interests.

Clearly, much effort and many studies are still in need to approach this problem because the number of proteins is still increasing rapidly although the crystallization already is no longer the only technology to find the 3-dimensional structure of proteins.

Accordingly, it necessarily tests each physicochemical feature against the propensity of crystallization for as many different proteins as possible although all known physicochemical features have been tested in different occasions under different circumstances.

In this study, the three features, which combined features from amino acid and protein, were tested against the propensity of crystallization of proteins from *B. haloduran*, and compared with the results obtained from each of 540-plus features possessed by amino acid. The results of this study once again demonstrate the wide-ranged applicability of three features developed by us because they catch the intrinsic random characteristic from protein sequences.

2. MATERIALS AND METHODS

2.1. Data

Four hundred fifty five proteins of *B. haloduran* were obtained from Target DB [17, 18] under the criterion of purified proteins including 185 under the criterion of crystallized protein as used in previous studies [1-8, 19, 20].

2.2. Features Possessed by both Amino Acid and Protein

The amino acid distribution probability is the first feature possessed by both amino acid and protein. This feature comes from the occupancy of subpopulations and partitions describing the distribution of elementary particles in energy states according to three assumptions of whether or not to distinguish each particle and energy state, *i.e.* Maxwell-Boltzmann, Fermi-Dirac, and Bose-Einstein assumptions in statistical mechanism [21]. This feature has been used in many occasions, whose probability can be computed with the following 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. For a type of amino acids, it has only one distribution probability in a protein (Columns 8 and 9, **Table 1**).

The amino acid future composition is the second feature possessed by both amino acid and protein, which comes from the observation that there are 64 RNA codons but only 20 types of amino acids, so each type of amino acids corresponds to different number of RNA codons, e.g., methionine has one RNA codon (AUG), phenylalanine has two RNA codons (UUC and UUU) but leucine has six RNA codons (CUA, CUC, CUG, CUU, UUA and UUG). These naturally lead to different translation probabilities when a single RNA code mutates, and consequently the probability that an amino acid mutates to another amino acid is different (Columns 10 and 11 in **Table 1**). And this feature has been used in many occasions.

The amino acid pair predictability is the third feature possessed by both amino acid and protein, which is based on permutation. And this feature has been used in many occasions.

2.3. Amino Acid Features

By contrast, a physicochemical feature is only related to a single aspect of individual amino acids,

Table 1. **Comparison of BEGF750101 feature, which is an amino acid feature that describes the helix-coil equilibrium, with features of amino-acid distribution probability and amino-acid future composition for two proteins, 359060 and 367736.**

Amino Acid	Number		BEGF750101		BEGF750101 × Number		Distribution Probability		Future Composition, %	
	359,060	367,736	359,060	367,736	359,060	367,736	359,060	367,736	359,060	367,736
A	6	9	1	1	6.00	9.00	0.2315	0.1967	5.48	8.02
R	9	4	0.52	0.52	4.68	2.08	0.1770	0.1875	8.12	6.05
N	3	7	0.35	0.35	1.05	2.45	0.6667	0.1071	3.78	4.23
D	9	8	0.44	0.44	3.96	3.52	0.1967	0.2243	5.30	5.89
C	2	0	0.06	0.06	0.12	0.00	0.5000	0.0000	2.96	2.16
E	17	20	0.44	0.44	7.48	8.80	0.1098	0.0422	4.26	4.63
Q	7	2	0.73	0.73	5.11	1.46	0.1071	0.5000	4.21	2.98
G	8	13	0.35	0.35	2.80	4.55	0.2243	0.1158	6.22	7.09
H	9	2	0.6	0.6	5.40	1.20	0.1475	0.5000	3.77	2.46
I	12	9	0.73	0.73	8.76	6.57	0.1241	0.1967	5.29	5.67
L	16	7	1	1	16.00	7.00	0.1192	0.3213	9.81	6.80
K	5	8	0.6	0.6	3.00	4.80	0.3840	0.0421	3.63	4.43
M	4	4	1	1	4.00	4.00	0.5625	0.1406	1.82	1.67
F	5	6	0.6	0.6	3.00	3.60	0.1920	0.2315	3.22	2.60
P	3	5	0.06	0.06	0.18	0.30	0.6667	0.1920	4.42	4.42
S	9	14	0.35	0.35	3.15	4.90	0.1770	0.0087	6.69	7.93
T	6	15	0.44	0.44	2.64	6.60	0.3472	0.0981	4.67	7.58
W	7	0	0.73	0.73	5.11	0.00	0.2142	0.0000	0.83	0.60
Y	6	6	0.44	0.44	2.64	2.64	0.2315	0.1543	2.74	2.49
V	7	11	0.82	0.82	5.74	9.02	0.1285	0.2020	7.26	8.07

therefore there are more than 540 amino acid features documented in AA Index database [22], for example, spatial features [23], electronic features [24], hydrophobic features [25], predictors for secondary structures [26].

2.4. Models

Logistic regression was a major tool used to model the relationship between crystallization propensity of proteins and amino-acid/protein features for proteins from *B. haloduran* because whether a protein can be crystallized can be defined as a yes-no event as the output of logistic regression, whereas various amino-acid/protein features can serve as the input of logistic regression. Similarly, the 10-1 feedforward back-propagation neural network was also used to model the relationship between crystallization propensity of proteins and amino-acid/protein features for proteins from *B. haloduran*. MatLab was used to perform both logistic regression and neural network [27, 28].

2.5. Statistics

The results were grouped into true positive (TP), true negative (TN), false positive (FP) and false negative (FN), so the accuracy, sensitivity and specificity can be calculated as follows: (TP + TN)/(TP + FP + TN + FN) × 100, (TP)/(TP + FN) × 100, and (TN)/(TN + FP) × 100, respectively. The McNemar's test was used to compare the classified results. The sensitivity and specificity were compared using receiver operating characteristic (ROC) analysis [29-31]. The Mann-Whitney *U*-test was used to compare predicted accuracies at different cutoff values.

3. RESULTS AND DISCUSSION

Table 1 shows differences between amino acid features and combined features. As seen, the amino acid feature BEGF750101 that describes the helix-coil equilibrium has a invariable value for each type of amino acid (Columns 4 and 5) regardless of amino acid's location, composition (Columns 2 and 3), and neighboring amino acids. A simple remedy is to multiply this amino acid feature by its corresponding composition (Columns 6 and 7, **Table 1**). In contrast, two combined features have different values for different amino acids for those two proteins (last four columns, **Table 1**). As can be seen, there are differences among these features, which can be used to correlate with the propensity of crystallization of proteins from *B. haloduran*, as well as for the comparison of their predictability.

Figure 1 showed the comparisons of accuracy, sensitivity and specificity obtained using logistic regression to correlate the propensity of protein crystallization with each of features. In this figure, every bar indicated how many features resulted in a similar accuracy, sensitivity or specificity. For example, the first bar from left-hand in the upper panel indicated that three amino acid features (CHAM830107, MITS020101 and FAUJ880112) had similar accuracies (0.588 ± 0.001). Interestingly, similar features (CHAM830108, FAUJ880111 and MITS020101) also have the worst performance in prediction of propensity of crystallization of proteins from *Mycobacterium tuberculosis* [8] and from *Lactobacillus* [7]. Similarly, the second bar indicated that two amino acid features (FAUJ880111 and KLEP840101) had the same accuracy (0.593), so the features, FAUJ880111 and FAUJ880112, should be completely eliminated for any prediction in this regard in future. **Figure 1** strongly displayed that two combined features had relatively good relationship with the propensity of crystallization of protein. In particular, the prediction using amino acid distribution probability was the best with respect to accuracy and sensitivity.

Figure 2 displayed the comparisons of accuracy, sensitivity and specificity obtained using neural network to correlate the propensity of crystallization of protein with each of features. The presentations in this figure had similar explanations as those in **Figure 1**. As shown in previous studies [1-8] and this study, the neural network can more accurately perceive difference between features. Compared against amino acid features, **Figure 1** and **Figure 2** suggested that two combined features not only are actively involved in crystallization process, but also worked better for the predictions of propensity of protein crystallization. Again, many amino acid features render similar results, being identical to the argument of abundance in amino acid features [32]. Indeed, the prediction using amino acid distribution probability was the best with respect to accuracy and specificity in **Figure 2**.

The database in the computation for both **Figure 1** and **Figure 2** was not regrouped, that is, the model parameters got from the 428 *B. haloduran* proteins were employed for predictions. This procedure is usually regarded as the initial stage of modeling, and then the database should be regrouped into two groups; one produces the model parameters whereas the other serves for the validation [33]. **Figure 3** illustrated the accuracy, sensitivity and specificity got from delete-1 jackknife validation, which further demonstrated the predictions using combined features were not worse than those using amino acid features. In fact, **Figure 3** showed that the prediction using amino acid distribution probability and future composition had the best predictions in terms of accuracy and sensitivity.

Table 2 listed predictive performance with respect to each feature in terms of accuracy, sensitivity

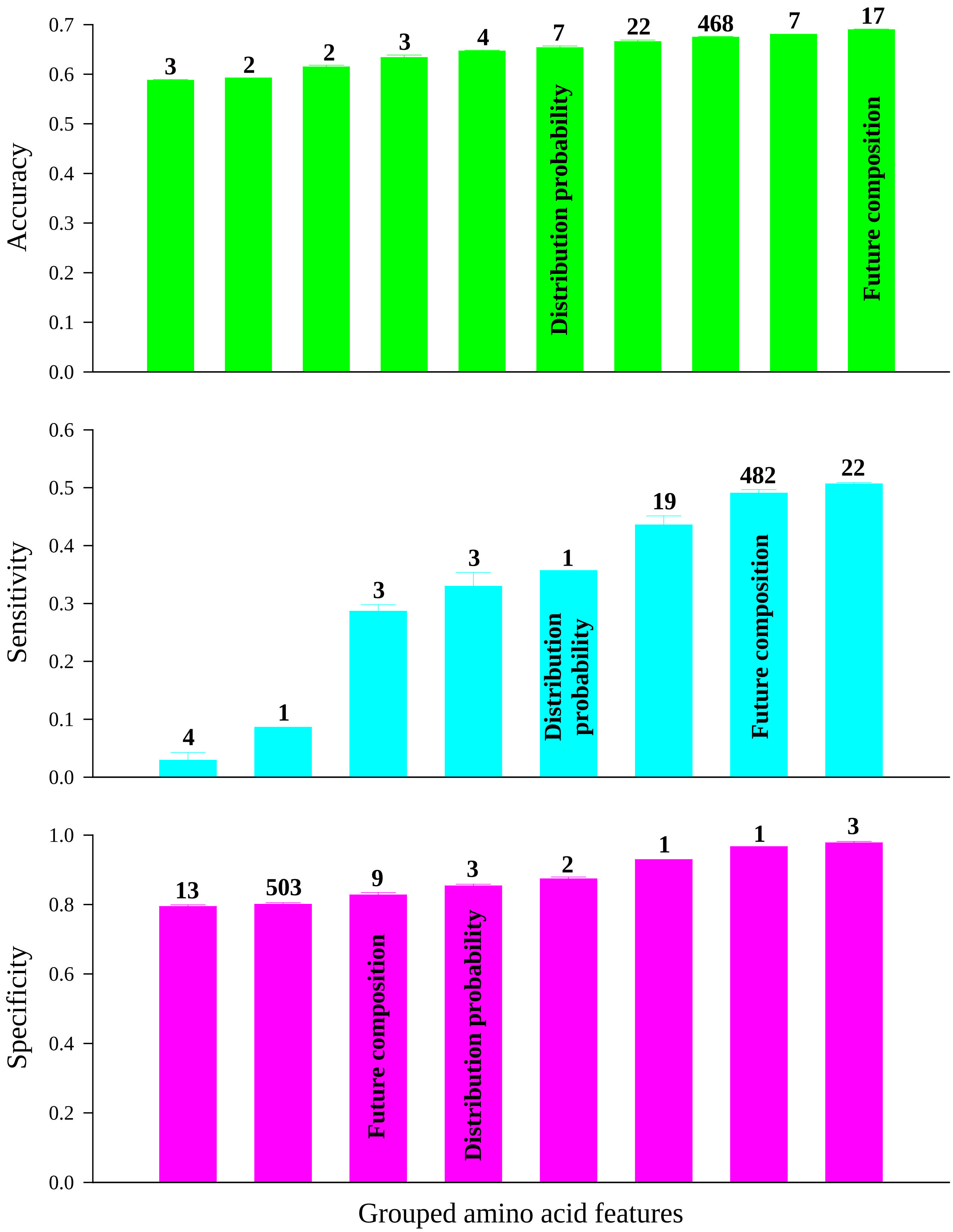

Figure 1. Accuracy, sensitivity and specificity obtained from logistic regression between the crystallization propensity of proteins from *B. haloduran* and each of 535 features.

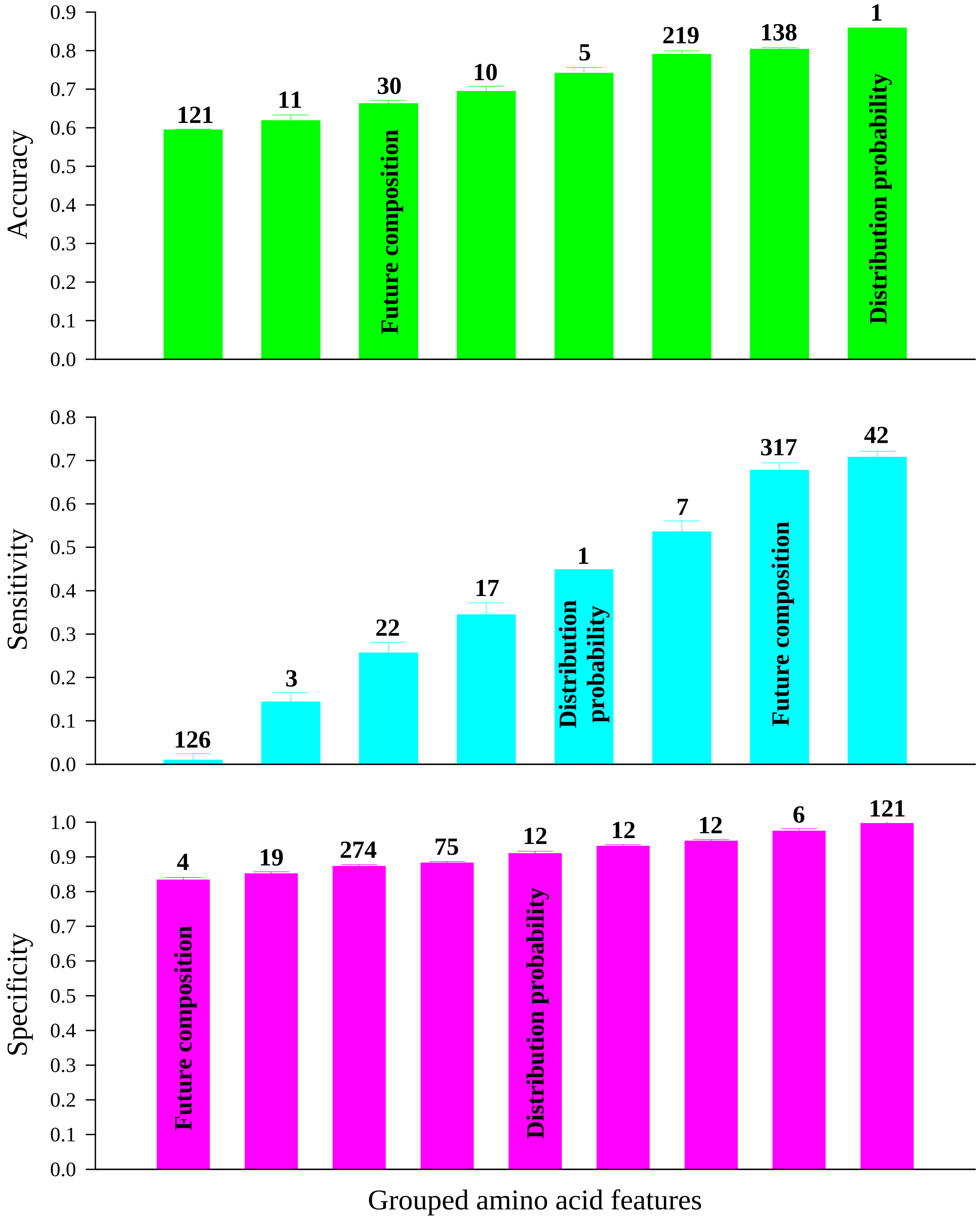

Figure 2. Accuracy, sensitivity and specificity obtained from fitting the relationship between the propensity of protein crystallization from *B. haloduran* and each of 535 features using 10-1 feed-forward backpropagation neural network.

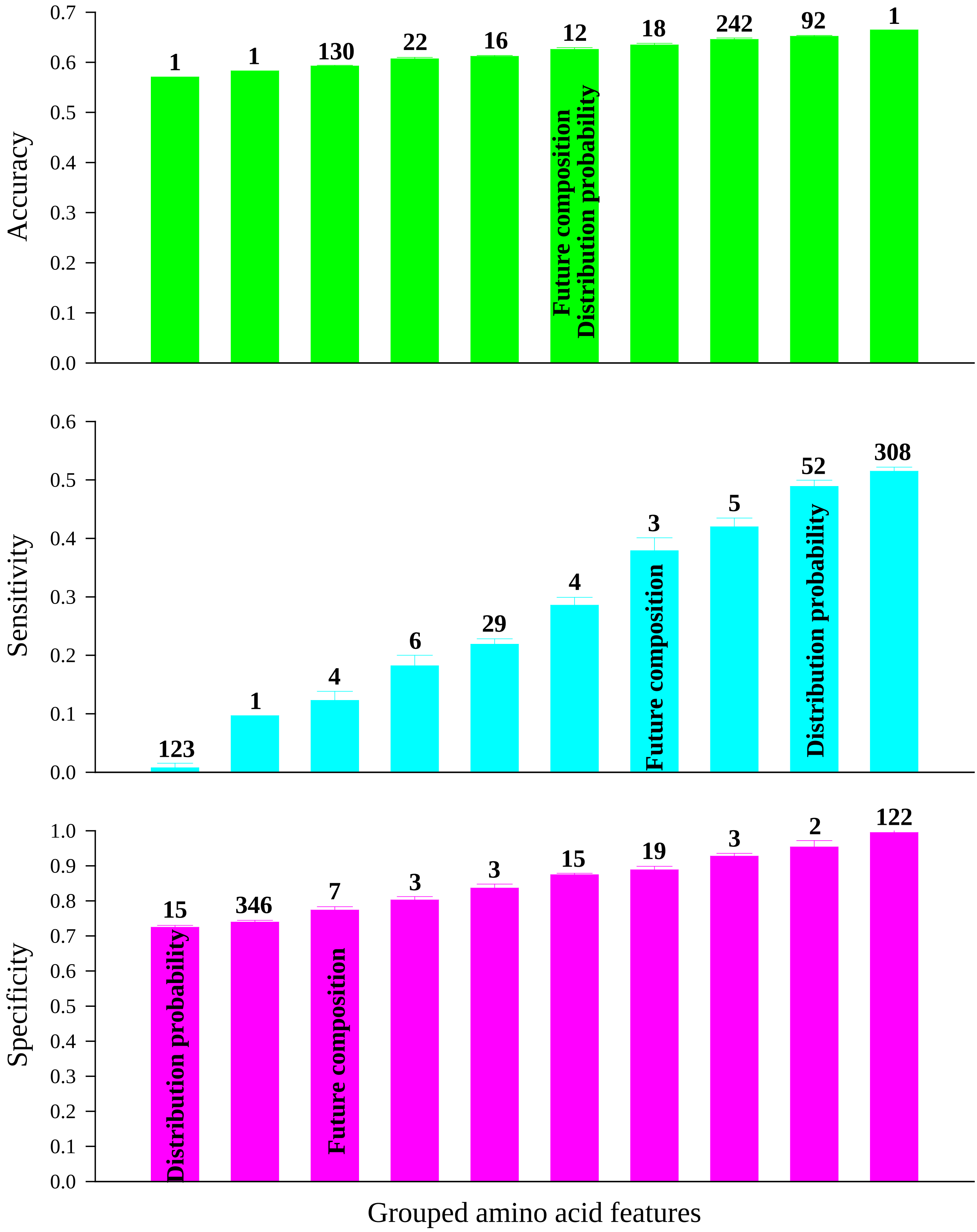

Figure 3. Accuracy, sensitivity and specificity of delete-1 jackknife validation obtained from modeling the relationship between crystallization propensity of proteins from *B. haloduran* and each of 535 features using 10-1 feedforward backpropagation neural network.

Table 2. **Predictive performance with respect to concrete features.**

Classification	The highest value	Accession number	Description	Characteristic
Fitting with logistic regression				
Accuracy	0.6945		Current composition	Combined feature
Sensitivity	0.5081		19 features	Second structure feature
Specificity	0.9815	KLEP840101	Net charge	Amino acid composition
Fitting with neural network				
Accuracy	0.8591		Distribution probability	Combined feature
Sensitivity	0.7827		Distribution probability	Amino acid composition
Specificity	1		45 features	Combined feature
Delete-1 validation with neural network				
Accuracy	0.9997	BEGF750101	Conformational parameter of inner helix	Physicochemical feature
Sensitivity	0.7225	WOLS870102	Principal property value z2	Physicochemical feature
	0.7225	MIYS990105	Optimized relative partition energies—method D	Physicochemical feature
	0.7225	KARP850103	Flexibility parameter for two rigid neighbors	Second structure feature
Specificity	0.9997	BEGF750101	Conformational parameter of inner helix	Physicochemical feature

and specificity. As shown, the delete-1 validation with neural network produces different features sensitive to predictions. This difference between delete-1 validation and other methods of validation is still unclear, suggesting more studies in need.

Figure 4 displayed the results of ROC analysis with respect to logistic regression, fitting and delete-1 jackknife validation using 20-1 feedforward backpropagation neural network. As expected: all the prediction features generate their classifications distributing above diagonal, so the predictions are not a random event because the McNemar's test showed that the classified results were significantly different from those of random guess ($P < 0.01$). Still, the combined features worked quite well in comparison with others.

Table 3 showed the third combined feature, unpredictable portion of amino acid pairs, and predictive accuracy in all, crystallized and non-crystallization proteins from *B. haloduran*. In Table 3, this feature had difference between crystallized and non-crystallized proteins from *B. haloduran*, and predictive accuracy was different between crystallized and non-crystallized proteins, too. In particular, the predictable portion is statistically higher in crystallized proteins than in non-crystallized ones (40.07% vs. 38.37%), which suggests the difference between crystallized and uncrystallized proteins in terms of the predictable portion, while other physicochemical features cannot show such difference. This difference perhaps explains the reason in the accuracy of predictions.

In conclusion, the present study once again demonstrated that the predictions using the features developed by us are relatively better than the predictions using any of 540 amino-acid features because they catch the intrinsic random characteristic from protein sequences so they have a wide-ranged applicability.

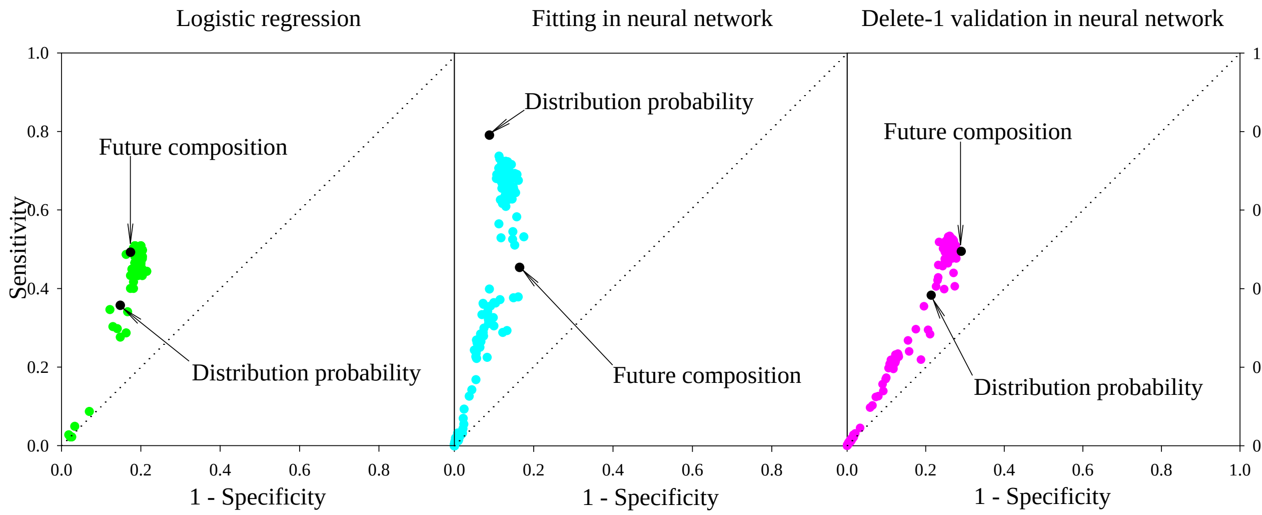

Figure 4. Comparison of sensitivity versus specificity obtained from logistic regression and from fitting and delete-1 jackknife validation in neural network in ROC analysis. Each gray circle is a result obtained using an individual amino acid feature while each black circle is a result obtained using one of two combined features. The diagonal line is the line of indiscrimination indicating a completely random guess. The text labels are the combined features.

Table 3. Predictable portion of amino acid pairs and accuracy of crystallization prediction in proteins from *B. haloduran* (The data were presented as median with 25% - 75% interquartile range, and the Mann-Whitney Rank Sum test was used to determine the difference between non-crystallized and crystallized groups).

Characteristic	Group	Number	Median (25% - 75%)	*P* value
Predictable portion (%)	Non-crystallized	270	40.07 (33.33 - 44.21)	0.032
	Crystallized	185	38.37 (33.79 - 42.78)	
	All proteins	455	39.39 (33.61 - 43.32)	
Accuracy in fitting	Non-crystallized	270	0.97 (0.89 - 0.99)	<0.001
	Crystallized	185	0.54 (0.31 - 0.70)	
	All proteins	455	0.79 (0.59 - 0.98)	
Accuracy in delete-1	Non-crystallized	270	0.90 (0.70 - 0.96)	<0.001
	Crystallized	185	0.36 (0.11 - 0.69)	
	All proteins	455	0.71 (0.39 - 0.92)	

Although many studies have been carried with respect to the prediction of propensity of crystallization of various proteins [1-15, 19, 20, 34-50], this issue is definitely unsolved. Therefore, effects are needed. In particular, how to find the features, which really represent the propensity of crystallization of various proteins is still unsolved.

FUND

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

CONFLICTS OF INTEREST

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

REFERENCES

1. 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
2. Yan, S. and Wu, G. (2012) Randomness in Crystallization of Proteins from *Staphylococcus aureus*. *Protein & Peptides Letters*, **19**, 784-789. https://doi.org/10.2174/092986612800793190
3. 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
4. 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
5. Yan, S.M., Wang, H.J. and Wu, G. (2013) Correlation of Combined Features of Amino Acid and Protein with Crystallization Propensity of Proteins from *Caenorhabditis elegans*. *Guangxi Sciences*, **20**, 234-243.
6. 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
7. Yan, S. and Wu, G. (2019) Correlation of Combined Characters of Amino Acid and Whole Protein with Success Rate of Crystallization of *Lactobacillus* proteins. *Journal of Biomedical Science and Engineering*, **12**, 245-256. https://doi.org/10.4236/jbise.2019.124017
8. Yan, S. and Wu, G. (2019) Correlating Combined Features of Amino Acid and Protein with Crystallization Propensity of Proteins from *Mycobacterium tuberculosis*. *Journal of Biomedical Science and Engineering*, **12**, 427-442. https://doi.org/10.4236/jbise.2019.129034
9. 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
10. Elbasir, A., Mall, R., Kunji, K., Rawi, R., Islam, Z., Chuang, G.Y., Kolatkar, P.R. and Bensmail, H. (2019) BCrystal: An Interpretable Sequence-Based Protein Crystallization Predictor. *Bioinformatics*, btz762. https://doi.org/10.1093/bioinformatics/btz762
11. Varga, J.K. and Tusnády, G.E. (2018) TMCrys: Predict Propensity of Success for Transmembrane Protein Crystallization. *Bioinformatics*, **34**, 3126-3130. https://doi.org/10.1093/bioinformatics/bty342
12. Gao, J., Wu, Z., Hu, G., Wang, K., Song, J., Joachimiak, A. and Kurgan L. (2018) Survey of Predictors of Propensity for Protein Production and Crystallization with Application to Predict Resolution of Crystal Structures. *Current Protein & Peptide Science*, **19**, 200-210. https://doi.org/10.2174/1389203718666170921114437
13. Wang, H., Feng, L., Webb, G.I., Kurgan, L., Song, J. and Lin, D. (2017) Critical Evaluation of Bioinformatics Tools for the Prediction of Protein Crystallization Propensity. *Briefings in Bioinformatics*, **18**, 1092. https://doi.org/10.1093/bib/bbx076
14. 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
15. 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

16. Fusco, D., Barnum, T.J., Bruno, A.E., Luft, J.R., Snell, E.H., Mukherjee, S. and Charbonneau, P. (2014) Statistical Analysis of Crystallization Database Links Protein Physico-Chemical Features with Crystallization Mechanisms. *PLoS ONE*, **9**, e101123. https://doi.org/10.1371/journal.pone.0101123

17. 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

18. Berman, H.M., Gabanyi, M.J., Kouranov, A., Micallef, D.I., Westbrook, J. and Protein Structure Initiative Network of Investigators (2017) Protein Structure Initiative-Target Track 2000-2017-All Data Files [Data Set]. Zenodo.

19. Slabinski, L., Jaroszewski, L., Rodrigues, A.P.C., Rychlewski, L., Wilson, I.A., Lesley, S.A. and Godzik, A. (2007) The Challenge of Protein Structure Determination—Lessons from Structural Genomics. *Protein Science*, **16**, 2472-2482. https://doi.org/10.1110/ps.073037907

20. 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

21. Feller, W. (1968) An Introduction to Probability Theory and Its Applications, 3rd Edition, Volume, I, Wiley, New York.

22. 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

23. 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 Molecular Biology*, **232**, 873-896. https://doi.org/10.1006/jmbi.1993.1437

24. 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

25. Cooper, G.M. (2004) The Cell: A Molecular Approach. ASM Press, Washington DC, 51.

26. 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**, 45-148. https://doi.org/10.1002/9780470122921.ch2

27. Demuth, H. and Beale, M. (2001) Neural Network Toolbox for Use with MatLab. User's Guide, Version 4.

28. MathWorks Inc. (1984-2001) MatLab-The Language of Technical Computing (Version 6.1.0.450, Release 12.1).

29. Pepe, M., Longton, G. and Janes, H. (2009) Estimation and Comparison of Receiver Operating Characteristic Curves. *Stata Journal*, **9**, 1-16. https://doi.org/10.1177/1536867X0900900101

30. 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

31. 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

32. 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

33. Chou, K.C. (2011) Some Remarks on Protein Attribute Prediction and Pseudo Amino Acid Composition. *Journal of Theoretical Biology*, **273**, 236-247. https://doi.org/10.1016/j.jtbi.2010.12.024

34. Chen, K., Kurgan, L. and Rahbari, M. (2007) Prediction of Protein Crystallization Using Collocation of Amino

Acid Pairs. *Biochemical and Biophysical Research Communications*, **355**, 764-769. https://doi.org/10.1016/j.bbrc.2007.02.040

35. Elbasir, A., Moovarkumudalvan, B., Kunji, K., Kolatkar, P.R., Mall, R. and Bensmail, H. (2019) Deep Crystal: A Deep Learning Framework for Sequence-Based Protein Crystallization Prediction. *Bioinformatics*, **35**, 2216-2225. https://doi.org/10.1093/bioinformatics/bty953

36. Meng, F., Wang, C. and Kurgan, L. (2018) fDETECT Webserver: Fast Predictor of Propensity for Protein Production, Purification, and Crystallization. *BMC Bioinformatics*, **18**, 580. https://doi.org/10.1186/s12859-017-1995-z

37. Wang, H., Feng, L., Zhang, Z., Webb, G.I., Lin, D. and Song, J. (2016) Crysalis: An Integrated Server for Computational Analysis and Design of Protein Crystallization. *Scientific Reports*, **6**, 21383. https://doi.org/10.1038/srep21383

38. Wang, H., Wang, M., Tan, H., Li, Y., Zhang, Z. and Song, J. (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. https://doi.org/10.1371/journal.pone.0105902

39. Charoenkwan, P., Shoombuatong, W., Lee, H.C., Chaijaruwanich, J., Huang, H.L. and Ho, S.Y. (2013) SCMCRYS: Predicting Protein Crystallization Using an Ensemble Scoring Card Method with Estimating Propensity Scores of P-Collocated Amino Acid Pairs. *PLoS ONE*, **8**, e72368. https://doi.org/10.1371/journal.pone.0072368

40. Jahandideh, S. and Mahdavi, A. (2012) RFCRYS: Sequence-Based Protein Crystallization Propensity Prediction by Means of Random Forest. *Journal of Theoretical Biology*, **306**, 115-119. https://doi.org/10.1016/j.jtbi.2012.04.028

41. Mizianty, M.J. and Kurgan, L.A. (2012) CRYSpred: Accurate Sequence-Based Protein Crystallization Propensity Prediction Using Sequence-Derived Structural Characteristics. *Protein & Peptide Letters*, **19**, 40-49. https://doi.org/10.2174/092986612798472910

42. Mizianty, M.J. and Kurgan, L. (2011) Sequence-Based Prediction of Protein Crystallization, Purification and Production Propensity. *Bioinformatics*, **27**, i24-i33. https://doi.org/10.1093/bioinformatics/btr229

43. Overton, I.M., van Niekerk, C.A. and Barton, G.J. (2011) XANNpred: Neural Nets That Predict the Propensity of a Protein to Yield Diffraction-Quality Crystals. *Proteins*, **79**, 1027-1033. https://doi.org/10.1002/prot.22914

44. Kandaswamy, K.K., Pugalenthi, G., Suganthan, P.N. and Gangal, R. (2010) SVMCRYS: An SVM Approach for the Prediction of Protein Crystallization Propensity from Protein Sequence. *Protein & Peptide Letters*, **17**, 423-430. https://doi.org/10.2174/092986610790963726

45. Babnigg, G. and Joachimiak, A. (2010) Predicting Protein Crystallization Propensity from Protein Sequence. *Journal of Structural and Functional Genomics*, **11**, 71-80.

46. Mizianty, M.J. and Kurgan, L. (2009) Meta Prediction of Protein Crystallization Propensity. *Biochemical and Biophysical Research Communications*, **390**, 10-15. https://doi.org/10.1016/j.bbrc.2009.09.036

47. Kurgan, L., Razib, A.A., Aghakhani, S., Dick, S., Mizianty, M. 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

48. Price, W.N., Chen, Y., Handelman, S.K., Neely, H., Manor, P., Karlin, R., Nair, R., Liu, J., Baran, M., Everett, J., Tong, S.N., Forouhar, F., Swaminathan, S.S., Acton, T., Xiao, R., Luft, J.R., Lauricella, A., DeTitta, G.T., Rost, B., Montelione, G.T. and Hunt J.F. (2009) Understanding the Physical Properties That Control Protein Crystallization by Analysis of Large-Scale Experimental Data. *Nature Biotechnology*, **27**, 51-57.

49. 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

50. 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