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

J. Biomedical Science and Engineering, 2019, Vol. 12, (No. 9), pp: 427-442

# Correlating Combined Features of Amino Acid and Protein with Crystallization Propensity of Proteins from *Mycobacterium tuberculosis*

**Shaomin Yan, Guang Wu**

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

**Correspondence to:** Guang Wu, hongguanglishibahao@gxas.cn

**Keywords:** Protein Feature, *Mycobacterium tuberculosis*, Protein Crystallization

**Received:** August 17, 2019 **Accepted:** September 26, 2019 **Published:** September 29, 2019



## ABSTRACT

**Since a decade ago, both protein and amino acid features have been correlated with crystallization propensity of proteins in order to develop methods to predict whether a protein can be crystallized. In this continuing study, each of three features combining features of amino acid and protein, was correlated with the crystallization propensity of proteins from *Mycobacterium tuberculosis* using logistic and neural network models. The results showed that two combined features, amino acid distribution probability and future composition, had good predictions on whether a protein would be crystallized in comparison with the predictions obtained from each of 531 amino acid features. The results obtained from the third combined feature, amino acid pair predictability, demonstrated the trend of crystallization propensity in proteins from *Mycobacterium tuberculosis*.**

## 1. INTRODUCTION

Many features possessed by amino acid and features possessed by a protein have an influence on the process of protein crystallization. Doubtlessly, humans can find more and more features possessed by amino acids and features possessed by a protein with advance in science and technology, each feature provides us with a new insight from a viewpoint different from the rest of features, and nevertheless, every new feature may have a certain relationship with the crystallization propensity of proteins.

The notable features are the amino acid physicochemical features, which have been repeatedly correlated with propensity of protein crystallization [1]. Subsequently, these features were also correlated with propensity of protein crystallization [2], for example, protein length, protein isoelectric point, percentage

of charged residues, hydrophobicity. With the compilation of features of amino acids [3], efforts once again were made to correlate propensity of protein crystallization with amino acid features, which had not been used in previous studies [2, 4].

Apparently, all known features possessed by amino acids and a protein have been tested. However, several features, which were developed by us, have not yet been widely tested against crystallization propensity of proteins. Indeed, it is necessary to test each feature against crystallization propensity of different proteins as many as possible, and then a solid scientific conclusion can be drawn on whether a particular feature is suitable for predicting crystallization propensity of proteins.

In this context, we tested three features, which combined features possessed by both amino acids and a protein, against the crystallization propensity of proteins from *Mycobacterium tuberculosis* in this study, and compared with the results obtained from each of 530-plus features possessed by amino acids.

## 2. MATERIALS AND METHODS

### 2.1. Data

428 proteins from *Mycobacterium tuberculosis* were found in Target DB [5, 6] under the criterion of purified proteins, of which 277 were found under the criterion of crystallized protein. Those two criteria were used in previous studies [7-15]. Actually, there are many different criteria in this database as well as in other databases, but our primary interest in this study is focused on the process between purified and crystallized proteins.

### 2.2. Features Possessed by Amino Acid and Protein

The first feature is the amino acid distribution probability [16], which is based on the occupancy of subpopulations and partitions describing the distribution of elementary particles in energy states according to three assumptions with respect to whether or not to distinguish each particle and energy state, *i.e.* Maxwell-Boltzmann, Fermi-Dirac, and Bose-Einstein assumptions in statistical mechanism [17]. For its application to protein, for example, Rv1875 protein has 3 tyrosines, and the simplest question is what probability it is if 3 tyrosines are clustered together or scattered along the protein sequence. This probability can be computed according to the equation [17], $\frac{r!}{q_0!\times q_1!\times\cdots\times q_n!}\times\frac{r!}{r_1!\times r_2!\times\cdots\times r_n!}\times n^{-r}$, 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. As amino acid composition is different, each type of amino acids has its own distribution probability. Two worked examples were listed in columns 8 and 9 of **Table 2** to show the distribution probability related to each type of amino acids in proteins.

The second feature is the amino acid future composition [16], 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. For example, methionine corresponds to one RNA codon (AUG), and phenylalanine corresponds to two RNA codons (UUC and UUU) whereas leucine corresponds to 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 (**Table 1**). For instance, when a mutation occurs in alanine, it has 12/36 chances to mutate to alanine, 2/36 chances to mutate to both aspartic acid and glutamic acid, 4/36 chances to mutate to glycine, proline, serine, threonine, and valine, respectively. Two worked examples were listed in columns 10 and 11 of **Table 2** to show the characteristic of this feature.

The third feature is the amino acid pair predictability [16], which is based on permutation. For instance, there are 15 leucines (L), 17 alanines (A), and 9 isoleucines (I) in Rv1155 protein. According to the permutation, the amino acid pair LA would appear twice ($15/147 \times 17/146 \times 146 = 1.73$), and there are indeed two LAs in realty so the pair LA is predictable. However, the amino acid pair IA would appear once

Table 1. Amino acids and their translated amino acids.

| Amino acid | Mutated amino acids with their translation probability |
|---|---|
| A | 12/36A + 2/36D + 2/36E + 4/36G + 4/36P + 4/36S + 4/36T + 4/36V |
| R | 18/54R + 2/54C + 2/54Q + 6/54G + 2/54H + 1/54I + 4/54L + 2/54K + 1/54M + 4/54P + 6/54S + 2/54T + 2/54W + 2/54STOP |
| N | 2/18N + 2/18D + 2/18H + 2/18I + 4/18K + 2/18S + 2/18T + 2/18Y |
| D | 2/18A + 2/18N + 2/18D + 4/18E + 2/18G + 2/18H + 2/18Y + 2/18V |
| C | 2/18R + 2/18C + 2/18G + 2/18F + 4/18S + 2/18W + 2/18Y + 2/18STOP |
| E | 2/18A + 4/18D + 2/18E + 2/18Q + 2/18G + 2/18K + 2/18V + 2/18STOP |
| Q | 2/18R + 2/18E + 2/18Q + 4/18H + 2/18L + 2/18K + 2/18P + 2/18STOP |
| G | 4/36A + 6/36R + 2/36D + 2/36C + 2/36E + 12/36G + 2/36S + 1/36W + 4/36V + 1/36STOP |
| H | 2/18R + 2/18N + 2/18D + 4/18Q + 2/18H + 2/18L + 2/18P + 2/18Y |
| I | 1/27R + 2/27N + 6/27I + 4/27L + 1/27K + 3/27M + 2/27F + 2/27S + 3/27T + 3/27V |
| L | 4/54R + 2/54Q + 2/54H + 4/54I + 18/54L + 2/54M + 6/54F + 4/54P + 2/54S + 1/54W + 6/54V + 3/54STOP |
| K | 2/18R + 4/18N + 2/18E + 2/18Q + 1/18I + 2/18K + 1/18M + 2/18T + 2/18STOP |
| M | 1/9R + 3/9I + 2/9L + 1/9K + 1/9T + 1/9V |
| F | 2/18C + 2/18I + 6/18L + 2/18F + 2/18S + 2/18Y + 2/18V |
| P | 4/36A + 4/36R + 2/36Q + 2/36H + 4/36L + 12/36P + 4/36S + 4/36T |
| S | 4/54A + 6/54R + 2/54N + 4/54C + 2/54G + 2/54I + 2/54L + 2/54F + 4/54P + 14/54S + 6/54T + 1/54W + 2/54Y + 3/54STOP |
| T | 4/36A + 2/36R + 2/36N + 3/36I + 2/36K + 1/36M + 4/36P + 6/36S + 12/36T |
| W | 2/9R + 2/9C + 1/9G + 1/9L + 1/9S + 2/9STOP |
| Y | 2/18N + 2/18D + 2/18C + 2/18H + 2/18F + 2/18S + 2/18Y + 4/18STOP |
| V | 4/36A + 2/36D + 2/36E + 4/36G + 3/36I + 6/36L + 1/36M + 2/36F + 12/36V |
| STOP | 2/27R + 1/27C + 2/27E + 2/27Q + 1/27G + 3/27L + 2/27K + 3/27S + 2/27W + 4/27Y + 4/27STOP |

(9/147 × 17/146 × 146 = 1.04), but it appears three times in this protein, so the pair IA is unpredictable. In this way, all amino acid pairs are classified as 72.5% predictable and 27.5% unpredictable in Rv1155 protein.

Because all the three features are computed with the consideration on individual amino acids with their composition and/or distribution in a protein, so they possess characteristics of individual amino acid and a whole protein.

### 2.3. Amino Acid Features

Amino acid features are the characteristics possessed by individual amino acids, and currently a database, AAIndex, contains 540-plus amino acid features describing various aspects of amino acids [3], including physicochemical features, spatial features [18], electronic features [19], hydrophobic features [20],

**Table 2. Features for two proteins (FINA770101 is an amino acid feature that describes the helix-coil equilibrium constant).**

| Amino Acid | Number | | FINA770101 | | FINA770101 × Number | | Distribution probability | | Future composition, % | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Rv1155 | Rv1875 | Rv1155 | Rv1875 | Rv1155 | Rv1875 | Rv1155 | Rv1875 | Rv1155 | Rv1875 |
| A | 17 | 17 | 1.08 | 1.08 | 18.36 | 18.36 | 0.1098 | 0.0229 | 8.42 | 9.10 |
| R | 13 | 13 | 1.05 | 1.05 | 13.65 | 13.65 | 0.0617 | 0.0386 | 8.05 | 8.39 |
| N | 4 | 5 | 0.85 | 0.85 | 3.40 | 4.25 | 0.5625 | 0.3840 | 3.64 | 2.34 |
| D | 15 | 8 | 0.85 | 0.85 | 12.75 | 6.80 | 0.0125 | 0.0421 | 4.08 | 4.35 |
| C | 0 | 0 | 0.95 | 0.95 | 0.00 | 0.00 | 0.0000 | 0.0000 | 1.86 | 2.17 |
| E | 4 | 8 | 0.95 | 0.95 | 3.80 | 7.60 | 0.5625 | 0.1682 | 4.69 | 4.20 |
| Q | 6 | 6 | 1.15 | 1.15 | 6.90 | 6.90 | 0.1543 | 0.3472 | 2.75 | 2.57 |
| G | 8 | 14 | 0.55 | 0.55 | 4.40 | 7.70 | 0.2523 | 0.0262 | 6.70 | 8.29 |
| H | 4 | 2 | 1.00 | 1.00 | 4.00 | 2.00 | 0.5625 | 0.5000 | 4.11 | 3.33 |
| I | 9 | 2 | 1.05 | 1.05 | 9.45 | 2.10 | 0.1967 | 0.5000 | 4.79 | 4.17 |
| L | 15 | 17 | 1.25 | 1.25 | 18.75 | 21.25 | 0.1569 | 0.0366 | 8.98 | 9.15 |
| K | 4 | 2 | 1.15 | 1.15 | 4.60 | 2.30 | 0.1406 | 0.5000 | 2.71 | 2.95 |
| M | 3 | 2 | 1.15 | 1.15 | 3.45 | 2.30 | 0.6667 | 0.5000 | 1.71 | 1.35 |
| F | 2 | 3 | 1.10 | 1.10 | 2.20 | 3.30 | 0.5000 | 0.6667 | 2.73 | 2.57 |
| P | 10 | 8 | 0.71 | 0.71 | 7.10 | 5.68 | 0.1905 | 0.0280 | 6.65 | 6.37 |
| S | 8 | 5 | 0.75 | 0.75 | 6.00 | 3.75 | 0.0673 | 0.1920 | 7.34 | 7.31 |
| T | 7 | 12 | 0.75 | 0.75 | 5.25 | 9.00 | 0.1071 | 0.1241 | 6.07 | 6.15 |
| W | 2 | 4 | 1.10 | 1.10 | 2.20 | 4.40 | 0.5000 | 0.1875 | 0.77 | 0.87 |
| Y | 5 | 3 | 1.10 | 1.10 | 5.50 | 3.30 | 0.3840 | 0.6667 | 2.47 | 1.71 |
| V | 11 | 16 | 0.95 | 0.95 | 10.45 | 15.20 | 0.1616 | 0.0715 | 8.01 | 8.99 |

predictors for secondary structures [21], etc.

Amino acid features are measured through experiments and documented so that they have no need to compute for each protein, whereas the features described in previous section need to compute for each protein. Therefore an amino acid feature is a constant for an amino acid, *i.e.*, each feature has an unchanged value for a type of amino acid. In fact, only 531 amino acid features have 20 values for 20 types of amino acids. In this study, each amino acid feature served as a benchmark to compare with the results obtained from the features described in previous section.

### 2.4. Models

Logistic regression was a major tool used in previous studies [22] because it works for a relationship between yes-no event and continuously numeric values, *i.e.* the relationship between propensity of protein crystallization, which is encoded either with amino acid features or with protein features. In this study an attempt was made to correlate each of three protein features with the crystallization propensity of proteins from *Mycobacterium tuberculosis* through logistic and neural network models, whose results were compared with the results obtained from modeling each of 531 amino acid features with the crystallization

propensity of the proteins.

### 2.5. Statistics

The results were classified as true positive (TP), true negative (TN), false positive (FP) and false negative (FN), so the accuracy, sensitivity and specificity can be calculated as follows [9-15]: TP = (TP + TN)/(TP + FP + TN + FN) × 100, TN = (TP)/(TP + FN) × 100, and FP = (TN)/(TN + FP) × 100, respectively. MatLab was used to perform both logistic regression and neural network [23, 24]. The McNemar's test was used to compare the classified results. The sensitivity and specificity were compared using receiver operating characteristic (ROC) analysis [25-28]. The Mann-Whitney *U*-test was used to compare predicted accuracies at different cutoff values.

## 3. RESULTS AND DISCUSSION

**Table 2** shows differences between amino acid features and combined features. As can be seen, the amino acid feature FINA770101 that describes the helix-coil equilibrium has a constant 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 2**). By contrast, two combined features have different values for different amino acids for those two proteins (last four columns, **Table 2**). This is an important distinction between combined features and amino acid features, and a rationale to correlate with the crystallization propensity of proteins from *Mycobacterium tuberculosis*.

**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, each bar represented 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 (CHAM830108, FAUJ880111 and MITS020101) had similar accuracies (0.643 ± 0.003). Similarly, the second bar indicated that three other amino acid features (CHAM830105, GOLD730101 and MIYS990101) had similar accuracies (0.657 ± 0.004). **Figure 1** clearly showed that two combined features had a relatively good relationship with the propensity of protein crystallization. In particular, the prediction using amino acid distribution probability was the best in terms of accuracy and sensitivity.

**Figure 2** displayed the comparisons of accuracy, sensitivity and specificity obtained using neural network to correlate the propensity of protein crystallization with each of features. The presentations in this figure had similar explanations as those in **Figure 1**. Clearly, the neural network can furthermore distinguish the difference between features. Compared against amino acid features, **Figure 1** and **Figure 2** suggested that two combined features not only were involved in crystallization process, but also served better for the predictions of protein crystallization. Also, many amino acid features gave similar results, being consistent with the study that demonstrated the abundance in amino acid features [29]. In particular, **Figure 2** showed that the prediction using amino acid distribution probability was the best in terms of accuracy and specificity.

In **Figure 1** and **Figure 2**, the database was not divided, *i.e.* the model parameters obtained from the 428 *Mycobacterium tuberculosis* proteins were used for predictions. This was generally considered as the first stage in modeling, and then the database should be divided into two groups, one for the generation of model parameters while the other for the validation [30]. **Figure 3** displayed the accuracy, sensitivity and specificity obtained 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 specificity.

**Table 3** listed predictive performance with respect to each feature in terms of accuracy, sensitivity and specificity. As can be seen, the best results were obtained using amino acid distribution probability, physicochemical features and second structure features.

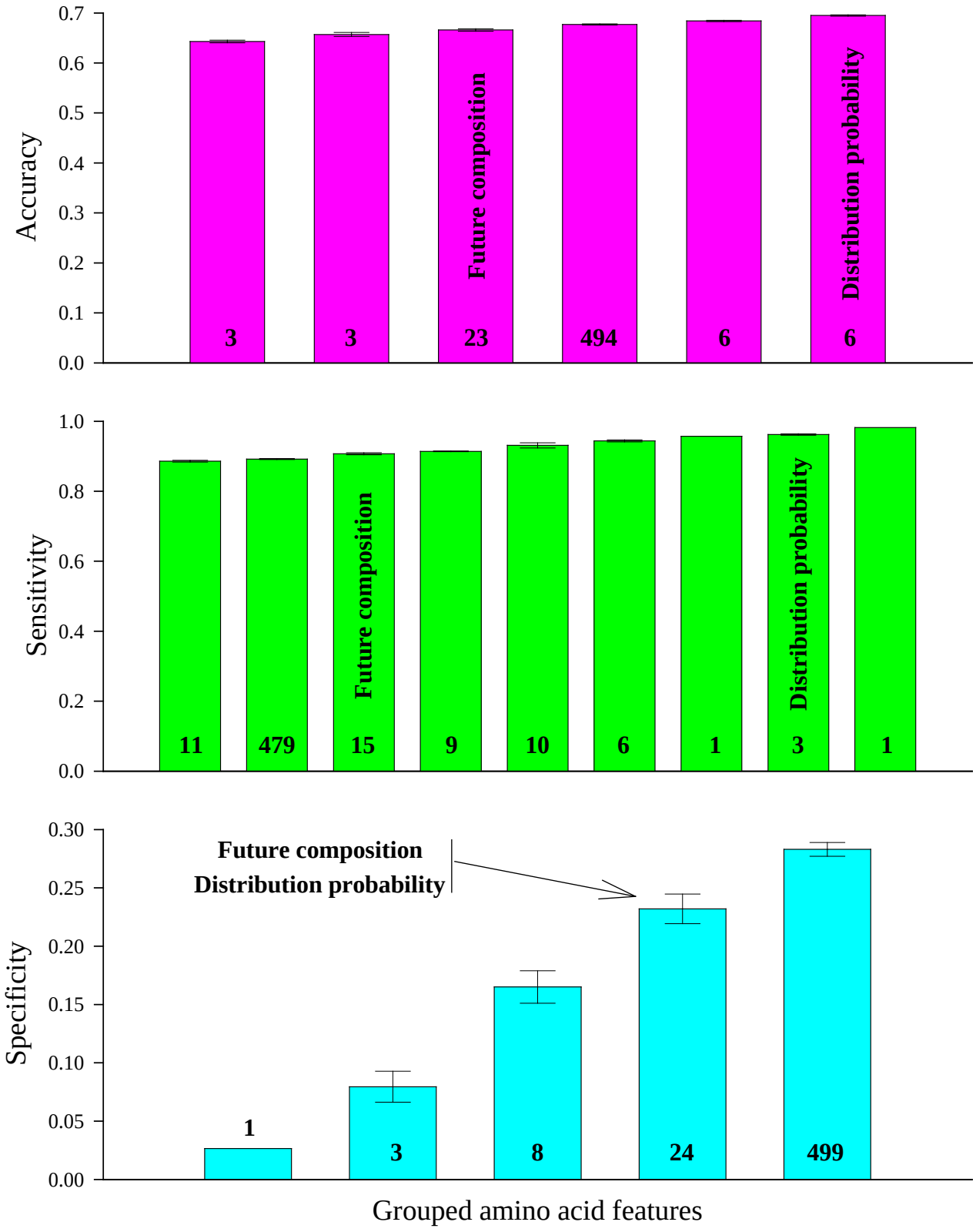


**Figure 1.** Accuracy, sensitivity and specificity obtained from logistic regression between the crystallization propensity of proteins from *Mycobacterium tuberculosis* and each of 535 features. The 535 features are grouped according to their similarity in accuracy, sensitivity and specificity.

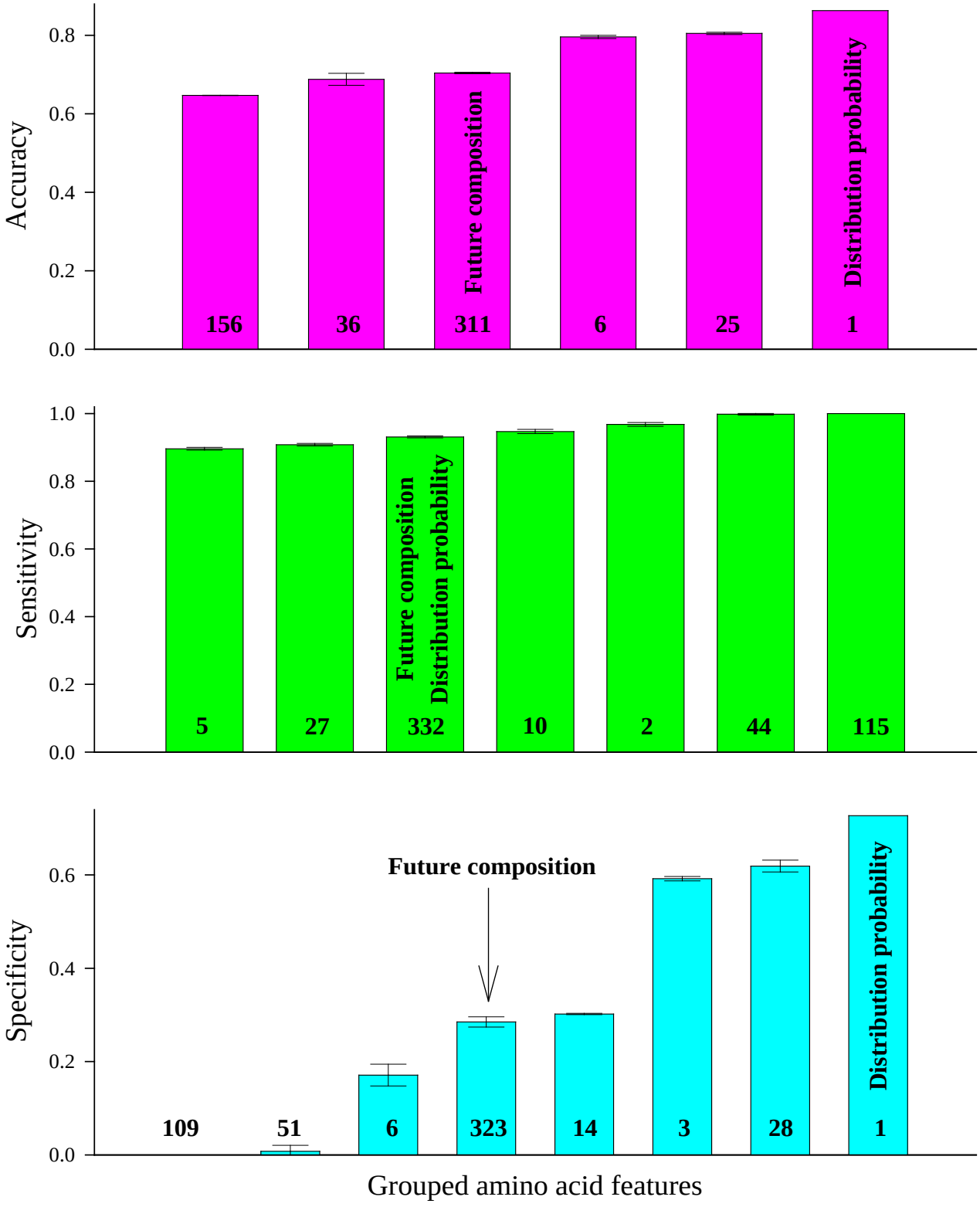


**Figure 2.** Accuracy, sensitivity and specificity obtained from fitting the relationship between the propensity of protein crystallization from *Mycobacterium tuberculosis* and each of 535 features using 20-1 feedforward backpropagation neural network. The 535 features are grouped according to their similarity in accuracy, sensitivity and specificity.

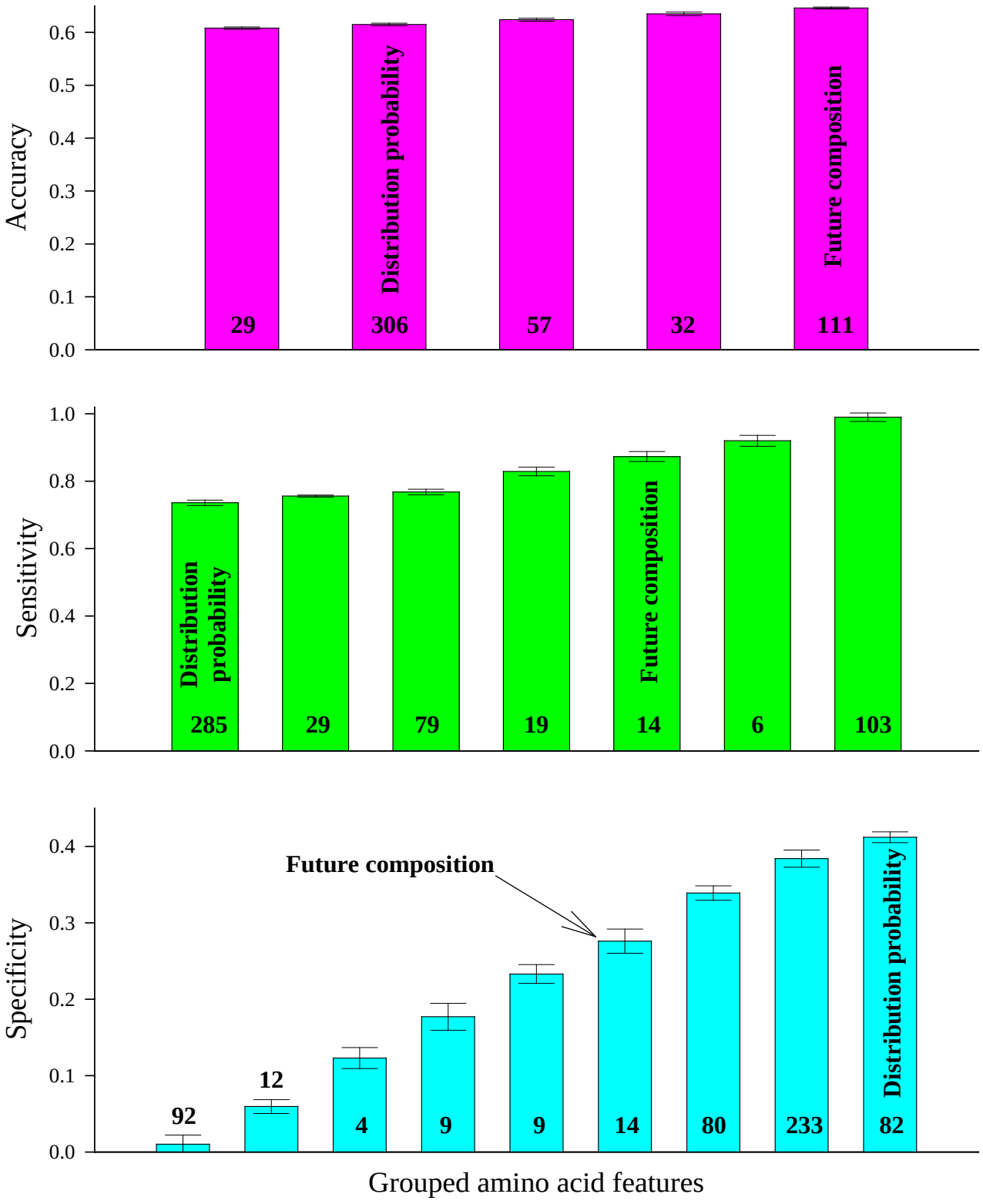


**Figure 3.** Accuracy, sensitivity and specificity of delete-1 jackknife validation obtained from modeling the relationship between crystallization propensity of proteins from *Mycobacterium tuberculosis* and each of 535 features using 20-1 feedforward backpropagation neural network. The 535 features are grouped according to their similarity in accuracy, sensitivity and specificity.

Table 3. **Predictive performance with respect to concrete features.**

| Classification | The highest value | Accession number | Description | Characteristic |
|---|---|---|---|---|
| Fitting with logistic regression | | | | |
| Accuracy | 0.6963 | | Distribution probability | Combined feature |
| | 0.6963 | TANS770107 | Normalized frequency of left-handed helix | Second structure feature |
| | 0.6963 | FAUJ880109 | Number of hydrogen bond donors | Second structure feature |
| Sensitivity | 0.9819 | FAUJ880111 | Positive charge | Physicochemical feature |
| Specificity | 0.2848 | | 40 features | Amino acid omposition |
| | 0.2848 | | 176 features | Physicochemical feature |
| | 0.2848 | | 225 features | Second structure feature |
| Fitting with neural network | | | | |
| Accuracy | 0.8631 | | Distribution probability | Combined feature |
| Sensitivity | 1 | | 24 features | Amino acid composition |
| | 1 | | 68 features | Physicochemical feature |
| | 1 | | 23 features | Second structure feature |
| Specificity | 0.7269 | | Distribution probability | Combined feature |
| Delete-1 validation with neural network | | | | |
| Accuracy | 0.6481 | NADH010101 | Hydropathy scale based on self-information values in the two-state model (5% accessibility) | Physicochemical feature |
| Sensitivity | 1 | RADA880106 | Accessible surface area | Physicochemical feature |
| | 1 | FASG760102 | Melting point | Physicochemical feature |
| | 1 | LEVM760104 | Side chain torsion angle phi (AAAR) | Second structure feature |
| Specificity | 0.4334 | HUTJ700102 | Absolute entropy | Physicochemical feature |

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. Two points could be drawn: 1) all the features gave their classifications distributing above diagonal, *i.e.* the predictions were better than random chance because the McNemar's test showed that the classified results were significantly different from those of random guess ($P < 0.01$), and 2) two combined features worked quite well in comparison with others.

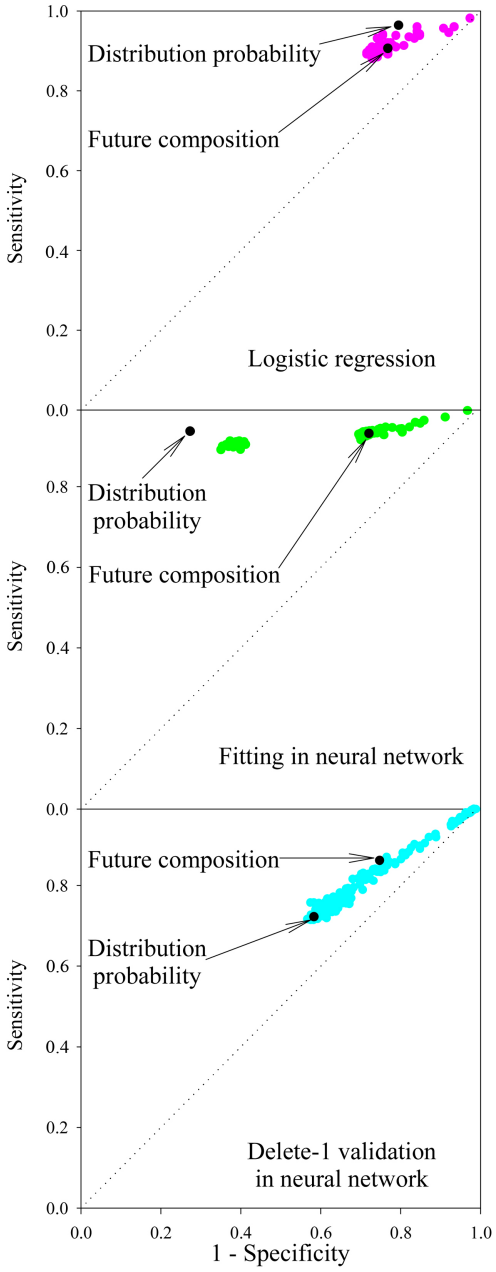


**Figure 4. Comparison of sensitivity versus specificity obtained from logistic regression and from fitting and delete-1 jackknife validation in neural network in ROC analysis. Each gray circle is a result obtained using an individual amino acid feature while each black circle is a result obtained using 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.**

Furthermore, the third combined feature that is the percentage of predictable/unpredictable amino acid pairs was used to compare the accuracy for predicting the protein crystallization. Figure 5 and Figure 6 showed such analysis in both neural network fitting and delete-1 jackknife validation. First, a cutoff value of accuracy was set at 0.75, 0.80, 0.85 and 0.90 levels; Second, 428 *Mycobacterium tuberculosis* proteins were divided into two groups according to the above-mentioned cutoff values; Third, the predictable portions of proteins were compared between two groups. Figure 5 and Figure 6 showed that the proteins, which had a large predictable portion, provided a high accuracy of predicting their crystallization propensity.

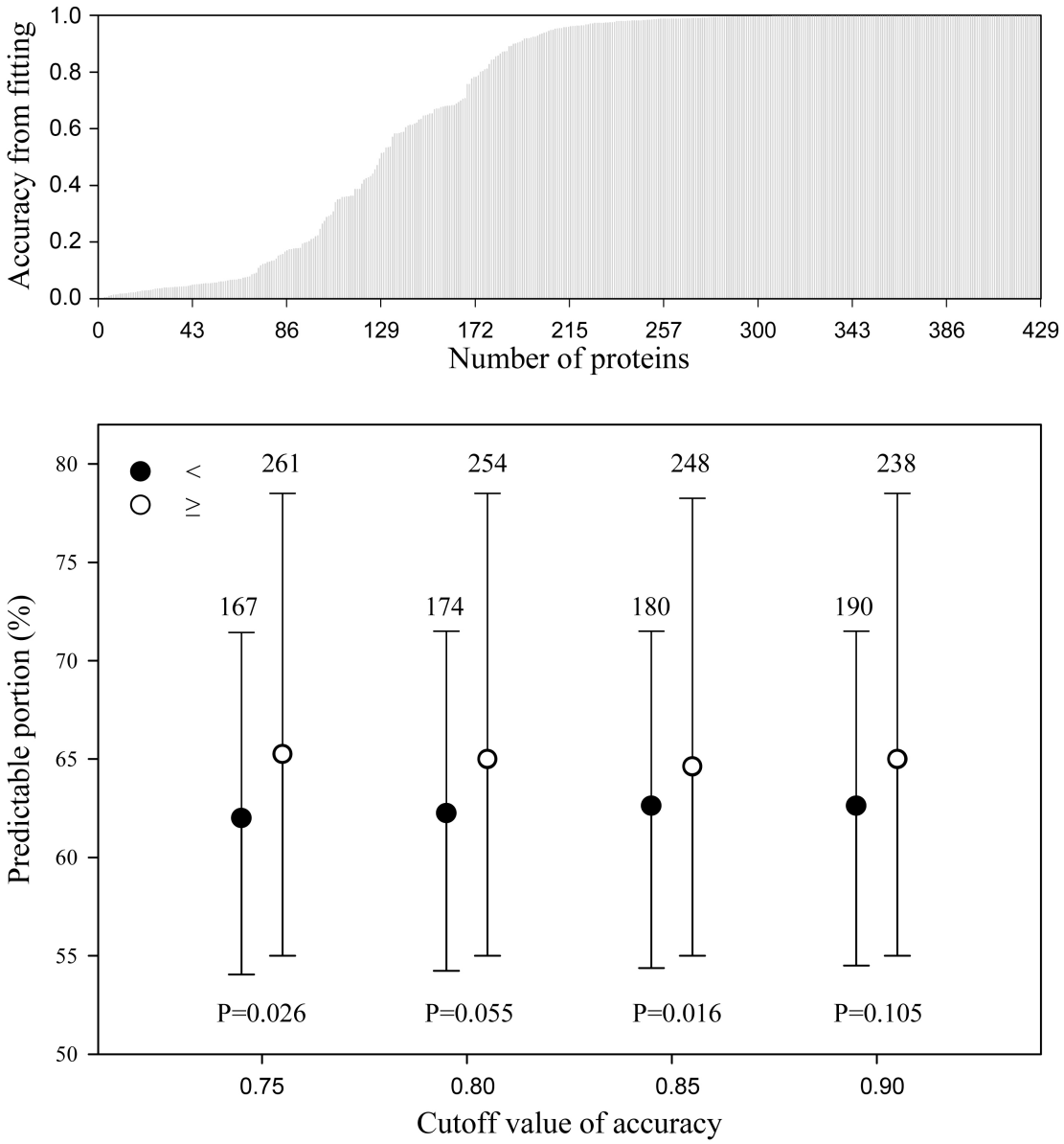


**Figure 5. Accuracy from fitting in crystallization prediction of *Mycobacterium tuberculosis* proteins (upper panel) and statistical comparison of their predictable portion of amino acid pairs at different cutoff values to separate proteins with accuracy (lower panel, the Mann-Whitney *U*-test). The data were presented as median with inter-quartiles.**

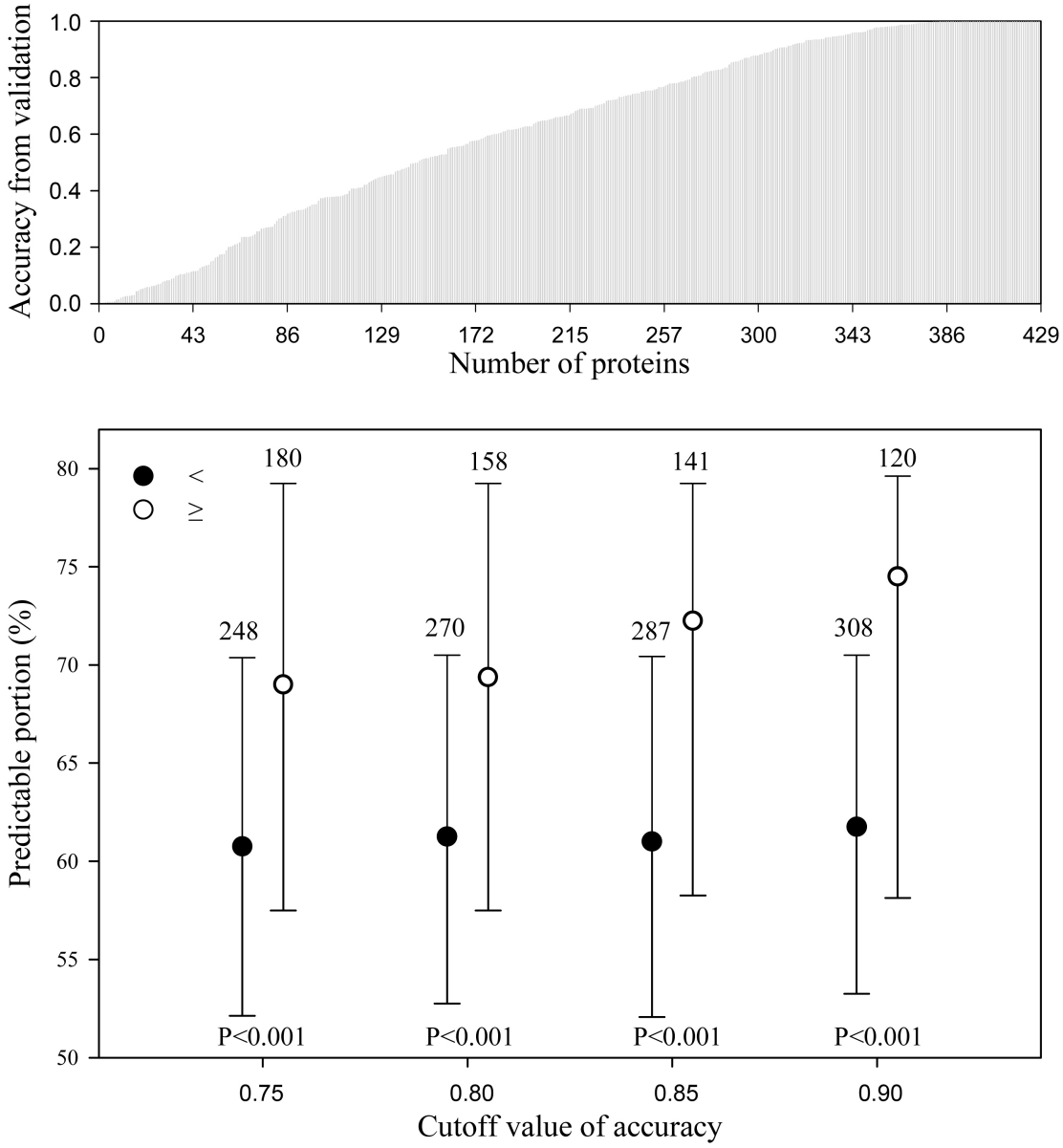


**Figure 6. Accuracy from delete-1 jackknife validation in crystallization prediction of *Mycobacterium tuberculosis* proteins (upper panel) and statistical comparison of their predictable portion of amino acid pairs at different cutoff values to separate proteins with predicted accuracy (lower panel, the Mann-Whitney *U*-test). The data were presented as median with inter-quartiles.**

Table 4 showed the third combined feature, unpredictable portion of amino acid pairs, and predictive accuracy in all, crystallized and non-crystallization proteins from *Mycobacterium tuberculosis*. As can be seen in Table 4, this feature had difference between crystallized and non-crystallized proteins from *Mycobacterium tuberculosis*, and predictive accuracy was different between crystallized and non-crystallized proteins, too. In particular, the unpredictable portion was statistically higher in crystallized proteins than in non-crystallized ones (65.25% vs. 61.50%), while the accuracy of predictions was higher in crystallized proteins than in non-crystallized ones. However, we could not find a direct correlation between unpredictable portion and prediction accuracy.

**Table 4. Unpredictable portion of amino acid pairs and accuracy of crystallization prediction in proteins from *Mycobacterium tuberculosis*. The data were presented as median with 25% - 75% interquartile range, and the Mann-Whitney *U*-test was used to determine the difference between crystallized and non-crystallized groups.**

| Characteristic | Group | Number | Median (25% - 75%) | *P* value |
|---|---|---|---|---|
| Unpredictable portion (%) | All proteins | 428 | 63.63 (54.88 - 75.25) | 0.013 |
| | Crystallized | 277 | 65.25 (55.50 - 78.25) | |
| | Non-crystallized | 151 | 61.50 (53.31 - 71.50) | |
| Accuracy in fitting | All proteins | 428 | 0.959 (0.323 - 0.998) | <0.001 |
| | Crystallized | 277 | 0.994 (0.964 - 0.999) | |
| | Non-crystallized | 151 | 0.122 (0.042 - 0.361) | |
| Accuracy in delete-1 | All proteins | 428 | 0.668 (0.377 - 0.926) | <0.001 |
| | Crystallized | 277 | 0.855 (0.659 - 0.978) | |
| | Non-crystallized | 151 | 0.268 (0.103 - 0.467) | |

The issue of whether an amino acid or protein feature can be correlated with propensity of protein crystallization has been tested through modeling [1, 4, 6, 7, 22, 31-39]. This is because it is impossible to conduct a control experiment without either amino acid or protein feature. In this study, three new features, which combined the features of individual amino acid and protein, were correlated with the crystallization propensity of proteins from *Mycobacterium tuberculosis*. The results demonstrate that these three combined features can be considered as the factors that affect the propensity of protein crystallization. Among three combined features, the amino acid pair predictability uses a single value, unpredictable portion, to represent a protein while the other two features, amino acid distribution probability and future composition, have each value for each type of amino acid. In this view, the amino acid distribution probability and future composition are somewhat similar to the 540-plus amino acid features, however, the two combined features do not have constant values as those amino acid features, therefore they more efficiently reflect certain features of amino acid in a whole protein. Clearly, more studies are needed to expend these three protein features to analyze the crystallization process in proteins from other organisms.

## FUND

This study was supported by National Natural Science Foundation of China (31560315), 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. 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

2. Canaves, J.M., Page, R., Wilson, I.A. and Stevens, R.C. (2004) Protein Biophysical Properties That Correlate with Crystallization Success in *Thermotoga Maritima*: Maximum Clustering Strategy for Structural Genomics.

*Journal Molecular Biology*, **344**, 977-991. https://doi.org/10.1016/j.jmb.2004.09.076

3. 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
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. 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
6. Berman, H.M., Gabanyi, M.J., Kouranov, A., Micallef, D.I., Westbrook, J. and Protein Structure Initiative Network of Investigators. (2017) Protein Structure Initiative—TargetTrack 2000-2017—All Data Files (Data Set). Zenodo.
7. 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
8. 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
9. 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
10. 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
11. 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
12. 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
13. 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.
14. 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
15. 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
16. Wu, G. and Yan, S. (2008) Lecture Notes on Computational Mutation. Nova Science Publishers, New York.
17. Feller, W. (1968) An Introduction to Probability Theory and Its Applications, 3rd Edition, Volume, 1, Wiley, New York.
18. 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
19. Dwyer, D.S. (2005) Electronic Properties of Amino Acid Side Chains: Quantum Mechanics Calculation of Subs-

tituent Effects. *BMC Chemical Biology*, **5**, 2. https://doi.org/10.1186/1472-6769-5-2

20. Cooper, G.M. (2004) The Cell: A Molecular Approach. ASM Press, Washington DC, 51.
21. 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
22. 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
23. Demuth, H. and Beale, M. (2001) Neural Network Toolbox for Use with MatLab. User's Guide, Version 4.
24. MathWorks Inc (1984-2001) MatLab—The Language of Technical Computing (Version 6.1.0.450, Release 12.1).
25. Shaw, P.A., Pepe, M.S., Alonzo, T.A. and Etzioni, R. (2009) Methods for Assessing Improvement in Specificity when a Biomarker is Combined with a Standard Screening Test. *Statistics in Biopharmaceutical Research*, **1**, 18-25. https://doi.org/10.1198/sbr.2009.0002
26. Pepe, M., Longton, G. and Janes, H. (2009) Estimation and Comparison of Receiver Operating Characteristic Curves. *The Stata Journal: Promoting Communications on Statistics and Stata*, **9**, 1-16. https://doi.org/10.1177/1536867X0900900101
27. 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
28. 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
29. 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
30. 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
31. Overton, I.M. and Barton, G.J. (2006) A Normalised Scale for Structural Genomics Target Ranking: the OB-Score. *FEBS Letters*, **580**, 4005-4009. https://doi.org/10.1016/j.febslet.2006.06.015
32. 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
33. 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
34. 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
35. Elbasir, A., Moovarkumudalvan, B., Kunji, K., Kolatkar, P.R., Mall, R. and Bensmail, H. (2019) DeepCrystal: 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. Derewenda, Z.S. and Godzik, A. (2017) The "Sticky Patch" Model of Crystallization and Modification of Proteins for Enhanced Crystallizability. In: Wlodawer, A., Dauter, Z. and Jaskolski, M., Eds., *Protein Crystallogra-*

*phy*. *Methods in Molecular Biology*, Humana Press, New York, 77-115.
https://doi.org/10.1007/978-1-4939-7000-1_4

38. 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

39. 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