CHAPTER V

PREDICTING PROTEIN CRYSTALLIZATION USING A SIMPLE SCORING CARD METHOD

5.1 Introduction

Knowledge of three-dimensional protein structures is crucial when investigating the functions of proteins. Three-dimensional structural knowledge is considered to be important when designing drugs involving protein function [88]. X-ray crystallography and nuclear magnetic resonance spectroscopy are commonly used for determining the structures of proteins. Approximately 80% of the protein structures in the Protein Data Bank (PDB) were obtained using the X-ray crystallography method [89]. In fact, these two approaches involve very complex, time-consuming, laborious and expensive processes. Because of the difficulties in determining the crystal structures, the current protocol has only a 30% success rate [90]. Thus, many researchers take advantage of computational approaches for directly predicting protein crystallizability.

Canaves et al. [91] and Goh et al. [92] have proposed methods for extracting informative features to predict protein crystallization. Many sequence-based computational methods, including OB-Score [93], SECRET [94], CRYSTALP [95], XtalPred [96], ParCrys [97], CRYSTALP2 [98], SVMCRYS [99], PPCpred [100] and RECRYS [101], predict protein crystallization, which are summarized in Table 5.1.

Both support vector machine (SVM) [94, 99, 100] and the ensemble mechanism [100, 101] are well-known techniques to enhance prediction accuracy. Because of the

different design aims and benchmarks used, it is not easy to assess which method and features are the most effective. From the study in [101] and Table 5.1, we can see that the SVM_POLY method (see the work [100]) using SVM has the highest accuracy among the non-ensemble methods. This method is one of the four SVM predictors that are integrated into PPCpred [100]. The state-of-the-art ensemble methods PPCpred and RFCRYS have high prediction accuracies using the SVM and Random Forest classifiers, respectively. PPCpred utilizes a comprehensive set of inputs that are based on energy and hydrophobicity indices, the composition of certain amino acid types, predicted disorder, secondary structure, solvent accessibility, and the content of certain buried and exposed residues [100]. RFCRYS predicts the protein crystallization by utilizing the mono-, di- and tri-peptide compositions; the frequencies of amino acids in different physicochemical groups; the isoelectric point; the molecular weight; and the length of the protein sequences [101]. However, the mechanism of these two ensemble classifiers suffers from low interpretability for biologists. It is not clear which sequence features provide the essential contribution to the high prediction accuracy.

Rather than increasing both the complexity of prediction methods and the number of feature types while pursuing high accuracy, the motivation of this study is to provide a simple and highly interpretable method with a comparable accuracy from the viewpoint of biologists. The *p*-collocated AA pairs (p=0 for a dipeptide) are shown to be significant in influencing or enhancing protein crystallization because of the impact of folding corresponding to the interaction between local AA pairs [8,11]. The *p*-collocated AA

pairs provide the additional information on which the interaction between local AA pairs reflects besides the simple AA composition.

Table 5.1 Some existing methods for predicting protein crystallization from sequences.

Method	Classifier	Single/ensemble
OB-Score [93]	Single Threshold	Single
SECERT [94]	SVM	Single
CRYSTALP [95]	NN	Single
XtalPred [96]	Logarithm Method	Single
ParCrys [97]	Parzen Window Density Estimator	Single
CRYSTAIP2 [98]	Normalized Gaussian radial basis function network	Single
SVMCRYS [99]	SVM	Single
PCCpred [100]	SVM	Ensemble
RFCRYS [101]	Random Forest	Ensemble
Our work [3]	**SCM**	**Single**

The propensity scores of dipeptides and amino acids to be crystallizable are highly correlated with the crystallization ability of sequences and can provide insights into protein crystallization. This study also proposes a mutagenesis analysis method for illustrating the additional advantage of SCM [3]. The analysis result reveals the

hypothesis that the mutagenesis of surface residues Ala and Cys has large and small probabilities of enhancing protein crystallizability in applying protein engineering approaches [102].

5.2 Dataset

We obtained training and testing sets containing 3587 and 3585 protein sequences, respectively, which were provided by Mizianty and Kurgan [100]. Two sequences with lengths of 9 and 11 were removed to make the dataset suitable for *p*-collocated AA pair analysis (p=0 to 9). We also removed several protein sequences containing special characters, such as X and U, to facilitate the computation using the SCM method. In our experiment, we considered both the training set (CRYS3575) and the testing set (CRYS3572) with both positive (crystallizable) and negative (non-crystallizable) classes, as summarized in Table 5.2.

Table 5.2 - Summary of the dataset for evaluating the performance of protein crystallizability predictors.

Data	Total	Positive	Negative
CRYS3575	3575	1197	2378
CRYS3572	3572	1198	2374

5.3 Results and Discussion

We first describe the data set developed by Mizianty and Kurgan [100] and the most frequently used SVM for predicting protein crystallization [99,100] is briefly

introduced. The amino acid and dipeptide compositions which are informative features for predicting protein crystallization are described. The implemented SCM with dipeptide composition is presented. Finally, the SCM method using the collocated dipeptides is given.

5.3.1 Prediction performance of the scoring card method based on collocated dipeptides

The scoring card method based on dipeptide composition was performed to predict protein crystallization on the TEST3572 data set. Additional detail of the SCM method can be found in Chapter 2 and [51]. For this work, the CSM was utilized in three ways: the initial, optimized, and combined CSMs. One important characteristic of the scoring card method is to optimize the dipeptide propensity scores to conserve the AAC information encoded in the protein sequence by considering R values between the initial and optimized CSMs. Many studies have been developed using this kind of dipeptide composition [98].

Due to the non-deterministic characteristics of the GA, we needed to perform 10 independent runs to obtain the optimized CSM. For using IGA to optimize the parameters, we obtained a small variance as low as 0.00, shown in Table 5.4. Our approach is able to predict the protein crystallization using only the *S(P)* based on a threshold cutoff. Table 5.3 shows the values for the full training, MCC, accuracy, sensitivity, and specificity of the initial CSM which are 72.14%, 0.30, 71.47%, 0.33, and 0.91, respectively. Using the optimized CSM yielded average results of 77.95±0.20%, 0.38±0.02, 73.9±0.57%, 0.45±0.03, and 0.88±0.01 for the full train, MCC, accuracy, sensitivity and specificity

respectively, shown in Table 5.4. It can be noted that the highest score of the fitness function was 0.84, which was found in the optimized CSM of the 10th experiment. Therefore, in this work, the 10th experiment was used as the best case for the optimized CSM. Finally, using the combined CSM, we obtained values of the full train, MCC, accuracy, sensitivity, and specificity of 78.99%, 0.39, 73.66%, 0.55, and 0.83, respectively, as is shown in Table 5.5. The combined CSM was calculated by averaging the 10 optimized CSMs.

Tables 5.3-5.4 show the results for the full train, MCC, accuracy, sensitivity, and specificity performed with a zero gap. Using the optimized CSM gave higher results, especially for specificity as compared to the initial CSM. Using the combined CSM we obtained higher sensitivity compared with the optimized (from 0.39 to 0.55) shown in Table 5.5.

Table 5.3 Performances of the scoring card method based on initail CSM with various gaps.

#GAP(*p*)	Full Train (%)	MCC	Accu(%)	Sens	Spec	Correlation
0	72.14	0.30	71.47	0.33	0.91	1.00
1	71.69	0.30	71.72	0.32	0.92	1.00
2	71.78	0.29	71.05	0.32	0.91	1.00
3	71.55	0.30	71.42	0.37	0.89	1.00
4	71.30	0.29	71.02	0.33	0.90	1.00

Moreover, we further examined the data for between one to four gaps which are represented by using the ROC curve, shown in Figure 5.1. This figure shows that using the

zero gap yields higher results, especially for the value of the AUC, as compared to the other gap sizes.

Table 5.4 Performances of the scoring card method based on optimized CSM with zero gaps.

#Expe	Fitness	Full Train(%)	MCC	Acc(%)	Sens	Spec
1	0.83	78.07	0.36	73.43	0.44	0.88
2	0.82	77.96	0.38	74.13	0.42	0.90
3	0.83	78.24	0.35	72.84	0.44	0.88
4	0.83	77.54	0.37	73.54	0.43	0.89
5	0.83	77.90	0.40	74.52	0.48	0.88
6	0.83	77.96	0.40	74.50	0.48	0.88
7	0.83	77.76	0.38	74.22	0.44	0.90
8	0.83	77.82	0.40	74.55	0.48	0.88
9	0.83	78.10	0.37	73.60	0.42	0.90
10	0.84	78.13	0.38	73.66	0.50	0.86
Avg	0.83±0.00	77.95±0.20	0.38±0.02	73.9±0.57	0.45±0.03	0.88±0.01

Table 5.5 Performances of the scoring card method based on combined CSM with various gaps.

#GAP(*p*)	Full Train (%)	MCC	Acc(%)	Sens	Spec	Correlation
0	78.99	0.39	73.66	0.55	0.83	0.98
1	78.60	0.37	73.88	0.43	0.90	0.97
2	78.46	0.34	71.36	0.50	0.82	0.97
3	78.07	0.38	72.65	0.56	0.81	0.98
4	78.66	0.38	73.88	0.48	0.87	0.97

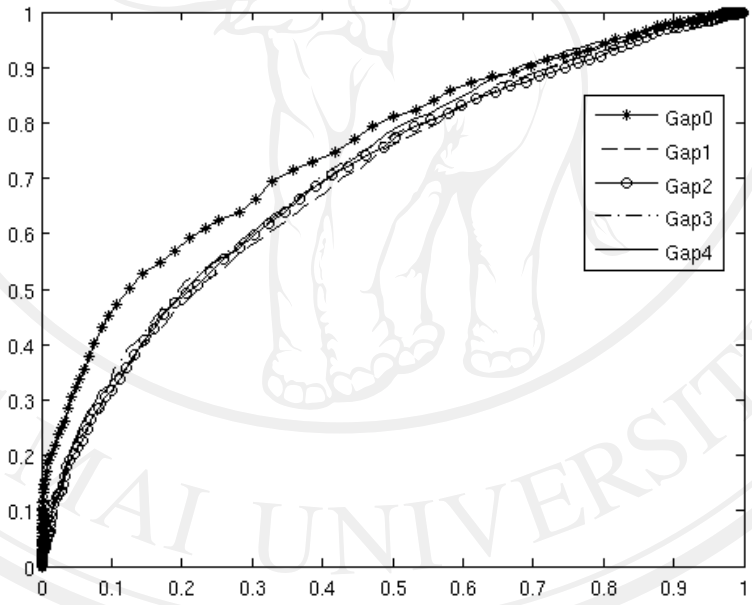

Figure 5.1. The ROC curve of the scoring card method by using various gaps [3]

5.3.2 Performance of SVM using amino acid and collocated dipeptide compositions

In this study, the SVMs with a radial basis function [103] were used to predict the protein crystallization using only the AAC and collocated dipeptide with zero to four gaps. The simplest feature (AAC) obtained full train, MCC, accuracy, sensitivity, and

specificity of 75.64%, 0.35, 73.12%, 0.38, and 0.91, respectively. Table 5.6 shows the results for the SVM of the full train, MCC, accuracy, sensitivity, and specificity as 92.56%, 0.46, 77.02%, 0.49, and 0.91, respectively, using one gap. Using the zero gap, the SVM yielded higher specificity than using other gaps. It should be noted that using the basic dipeptide composition had the strongest ability to predict protein crystallization.

Although Tables 5.4 and 5.6 show that the SVM classifier obtains higher accuracies than the scoring card method, the results of the training and test sets are not fully comparable, for example in the evaluation results, the training set has accurate results ranging in [88.14%, 93.31%] whereas the test set has the range [75.08, 77.02%]. This result could be a case of overfitting the training data. Moreover, SVM suffers from gaining insight into the mechanisms affecting protein crystallization. More detail about this application can be seen in the analysis section.

Table 5.6 Performances SVM using various compositions.

Types of feature	**Full Train(%)**	**MCC**	**Acc(%)**	**Sens**	**Spec**	**Cost**	**Gamma**
AAC	75.64	0.35	73.12	0.38	0.91	64.00	0.06
p=0	88.14	0.45	76.93	0.41	0.95	1.00	0.03
p=1	92.56	0.46	77.02	0.49	0.91	2.00	0.03
p=2	93.20	0.42	75.08	0.55	0.85	2.00	0.03
p=3	93.26	0.41	75.22	0.47	0.89	1.00	0.06
p=4	93.31	0.41	75.42	0.41	0.93	1.00	0.06

5.3.3 Comparison of the scoring card method with existing classifiers

Using the optimized CSM (CSM_best fit) from the 10^{th} experiment and combined CSM (CSM_com), our experimental results show that using dipeptide composition had better results than other collocated dipeptides. Table 5.7 shows the comparison of the scoring card method, i.e. CSM_best fit and CSM_com, with existing methods for predicting crystallizable proteins in the TEST3572 data set. Previously, the SVM-based classifier was successfully used in many predictions of protein functions [104,105]. Therefore, we used the SVM classifier to compare with our method. The scoring card method based on CSM_com yielded a higher sensitivity value of 0.55 than SVM_POLY and RFCRYS. The value of the specificity was not significantly different compared with SVM_POLY and PPCpred. In the case of ensemble methods, i.e. PPCpred and RFCRYS, these approaches had reasonably high results compared with other classifier methods.

In the case of classifier-based conventional methods, the SVM_POLY method is generally considered to provide the best results, since this tool uses both an SVM and additional features not used in our approach [100]. In contrast, the scoring card method used with only the dipeptide composition provides a clear interpretation to the predictive results. The experimental results show that our approach is comparable to the other methods and significantly simpler from a theoretical standpoint. Our approach uses only the *S(P)* which is a single decision boundary. For the SVM, the kernel uses a mapping function which either has a polynomial [100] or radial basis function [99]. Obviously, the radial basis function is dependent on the cost and gamma values which lead to a more complex decision boundary.

Table 5.7 Comparison of the scoring card method with existing methods.

Classifier	MCC	Accuracy(%)	Sensitivity	Specificity
CRYSTALP2[a]	0.19	55.3	0.74	0.46
SVMCRYS[a]	0.21	56.3	0.75	0.47
SVM_POLY[a]	0.40	74.6	0.48	0.88
PPCpred[a]	0.47	76.8	0.61	0.85
RFCRYS[a]	0.53	80.0	0.51	0.95
CSM_best fit	0.38	73.7	0.50	0.86
CSM_com	0.39	73.7	0.55	0.83

[a]Result reported from Jahandideh and Mahdavi [101].

5.3.4 Propensity scores of amino acids and dipeptides

Previously, we mentioned that there is one advantage of the scoring card method which is interpretability which allows us to gain insight into the mechanisms of protein crystallization. In this part, we consider the score of the dipeptide compositions derived from the scoring card method. Figure 5.2 shows the mean score of the dipeptide compositions represented using a 20×20 matrix. The scores of the dipeptides in the CSM are highly correlated to the relative contributions of the dipeptides to protein crystallizations shown in Figure 5.2 and Tables 5.8-5.9. The 20 propensity scores of amino acids to be crystallizable derived from the scores of dipeptides (Figure 5.2) are shown in Table 5.8. Glu, Gly, Ala, His and Val are the five top-ranked amino acids to be

crystallizable, and Ser, Asn, Cys, Gln and Pro are the five top-ranked amino acids to be non-crystallizable.

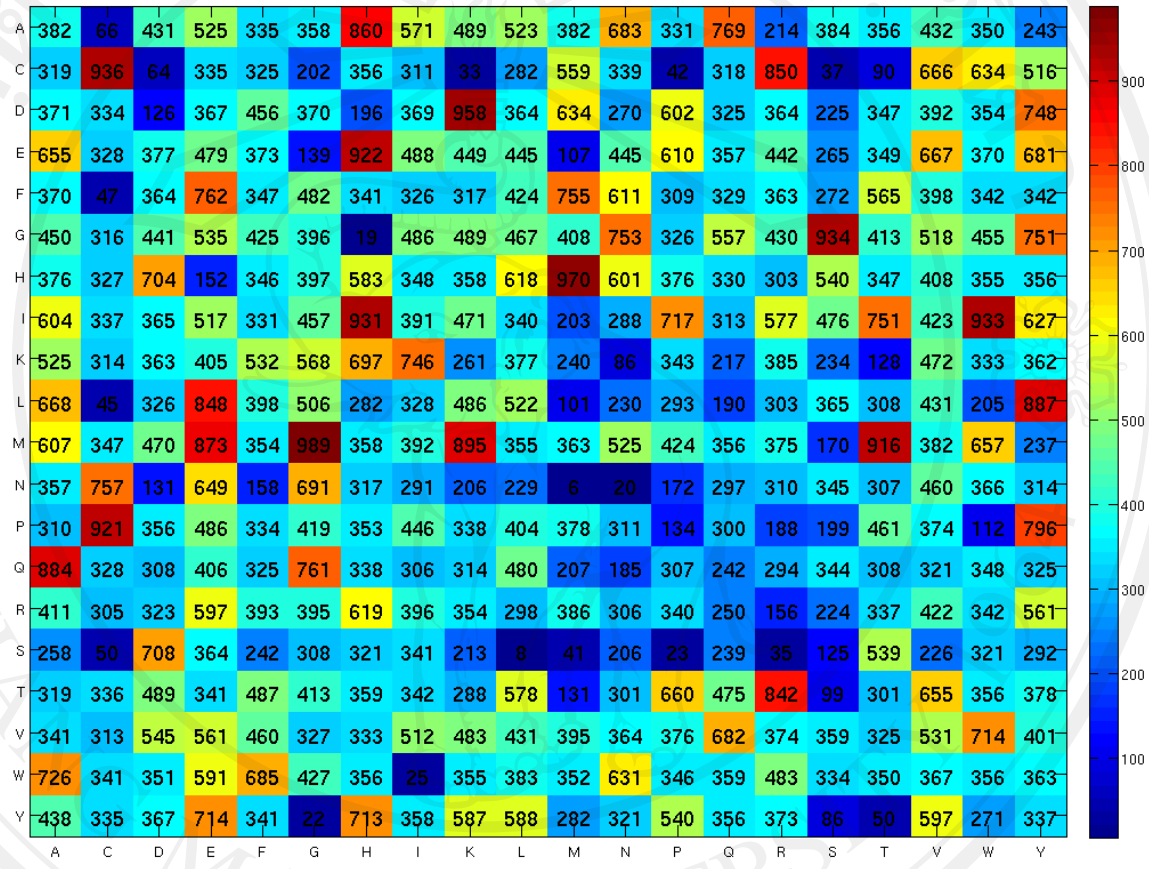

Figure 5.2. The propensity scores of dipeptides from averaging results of 10 independent runs [3].

Table 5.8 The propensity scores of amino acids to be crystallizable

Amino	Name	Score	Amino	Name	Score
E	Glu	486.38	L	Leu	395.95
G	Gly	454.90	D	Asp	394.53
A	Ala	451.38	F	Phe	392.83
H	His	451.23	T	Thr	392.45
V	Val	449.23	R	Arg	376.90
I	Ile	445.63	P	Pro	372.28
Y	Tyr	429.83	Q	Gln	364.80
M	Met	423.63	C	Cys	357.43
W	Trp	408.88	N	Asn	346.48
K	Lys	398.30	S	Ser	271.93

Table 5.9 The 10-top of highest and lowest propensity scores of dipeptides to be crystallizable

Dipeptides	Score	Dipeptides	Score
MG	989	CS	37
HM	970	SR	35
DK	958	CK	33
CC	936	WI	25
GS	934	SP	23
IW	933	YG	22
IH	931	NN	20
EH	922	GH	19
PC	921	SL	8
MT	916	NM	6

5.3.5 Application of propensity score of amino acids for the crystallization of protein

Several approaches have been developed to enhance protein crystallizability. One of the most popular approaches is protein engineering used for increasing the success rate in crystallization. The substitution of single-site amino acids can dramatically affect the crystallization of proteins. However, it is reported that the question of which substituting residue would perform better than others is more difficult to answer [106]. Many studies further presented advantages of single-site mutations for increasing the solubility of proteins and obtained higher quality of crystals [103]. In this part, we are interested of surface mutagenesis of Ala, Cys, and Ser for enhancing the crystallizability of protein.

From the propensity scores of amino acids in Table 5.8, Ala is the most frequenctly used amino acid for substituting in the mutagenesis of surface residue. The most frequently used mutation of X→Ala (replacing amino acid X by Ala) is Glu→Ala and Lys→Ala from the literature survey. It is not surprising that the mutation of Ala→X in enhancing crystallizability in literature is rare and ineffective. We found one mutation of Ala→Cys which is not effective in enhancing crystallizability [109]. This result of Ala→Cys can be well recognized from the analysis of Table 5.8.

We would examine the mutations Cys→X and X→Cys from literature survey where Cys has the ranks of crytallizabilitty equal to 18. Hence, Cys is possibly the important obstacle for protein crystallization. Most mutations of Cys→X enhanced crystallizability according to the reasons of enhancing protein solubility and decreasing aggregation and molecular size [120-125]. Ser is a well-known substituted mutant of Cys

because Ser can conserve a similar protein function of Cys [120, 122-125]. The mutation Cys→Ala could be the perfect mutation of enhancing crystalizability according to our hypothesis in this study, which is reported as successful enhancement in the study [121]. The mutations of X→Cys are believed to enhance crystallizability for obtaining useful heavy-atom derivation [109, 126-128]. However, all mutations of X→Cys in these studies [109, 126-128] could not successfully improve crystallizability. All these experimental results are reflected in our hypothesis.

Ser has the lowest crystallizability scores. Therefore, the mutations of Ser→X should increase the probability of enhancing protein crystallization for the same reason with Cys. We found the mutations of Ser→Cys for obtaining useful heavy-atom derivatives to enhance crystallizability [126-128]. However, all these mutations of Ser→Cys in the studies [126-128] fail to increase the crystallizability. It might be reasonable that the mutation Cys→X demonstrated the high probability of enhancing protein crystallization [120-125]. Relatively few mutations of X→Ser were conducted to enhance crystallizability. However, some results of the mutations X→Ser demonstrated the successful enhancement of protein crystallization [106]. And, the results reported by Cooper et al. [106] showed that both Ser and His residues performed less well, but these two amino acids were better than the wild type in promoting protein crystallization. Further studies are needed to evaluate effectiveness of the mutation X→Ser in increasing the probability of enhancing protein crystallization for some specific proteins and conditions. More details about the application of the SCM method can be found in [102].

5.4 Conclusion

We have applied the scoring card method to predict protein crystallization based on dipeptide compositions. Many studies have shown that dipeptide composition provides an essential source of information about protein crystallization. Table 5.5 shows the results of the scoring card method using a combined CSM to obtain full train, accuracy, sensitivity, and specificity of 78.99%, 73.66%, 0.55, and 0.83, respectively. The proposed method using the combined CSM performed on a test set obtained higher accuracies than the other two CSMs. The scoring card method was then compared to existing methods. The experimental results showed that our approach obtained higher MCC, accuracy, and specificity than SVMCRYS, and are comparable to SVM_POLY. However, in fact, the ensemble-based approach, i.e., PPCpred, and RFCRYS, obtained results slightly better than the proposed method which is not an ensemble approach. Unfortunately, it is difficult to interpret these methods to further realize what mechanisms could affect protein crystallization. In conclusion, the SCM method gives a simple approach which allows a direct interpretation of dipeptides in the crystallization prediction based on the propensity scores of the dipeptides only. The propensity scores of dipeptides and amino acids to be crystallizable are highly correlated with the crystallization ability of sequences and can provide insights into protein crystallization. Nevertheless, future studies are needed to develop the scoring card method by combining our approach with other techniques such as the ensemble approach for improving the accuracy of protein crystallization prediction.