Hindawi
Journal of Chemistry
Volume 2017, Article ID 9861752, 15 pages
https://doi.org/10.1155/2017/9861752

Research Article

CryoProtect: A Web Server for Classifying Antifreeze Proteins from Nonantifreeze Proteins

Reny Pratiwi,[1,2] Aijaz Ahmad Malik,[1] Nalini Schaduangrat,[1] Virapong Prachayasittikul,[3] Jarl E. S. Wikberg,[4] Chanin Nantasenamat,[1] and Watshara Shoombuatong[1]

[1]*Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand*
[2]*Department of Medical Laboratory Technology, Faculty of Health Science, Setia Budi University, Surakarta 57127, Indonesia*
[3]*Department of Clinical Microbiology and Applied Technology, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand*
[4]*Department of Pharmaceutical Biosciences, BMC, Uppsala University, SE-751 24 Uppsala, Sweden*

Correspondence should be addressed to Chanin Nantasenamat; chanin.nan@mahidol.edu
and Watshara Shoombuatong; watshara.sho@mahidol.ac.th

Received 17 October 2016; Accepted 26 December 2016; Published 9 February 2017

Academic Editor: José L. Arias Mediano

Antifreeze protein (AFP) is an ice-binding protein that protects organisms from freezing in extremely cold environments. AFPs are found across a diverse range of species and, therefore, significantly differ in their structures. As there are no consensus sequences available for determining the ice-binding domain of AFPs, thus the prediction and characterization of AFPs from their sequence is a challenging task. This study addresses this issue by predicting AFPs directly from sequence on a large set of 478 AFPs and 9,139 non-AFPs using machine learning (e.g., random forest) as a function of interpretable features (e.g., amino acid composition, dipeptide composition, and physicochemical properties). Furthermore, AFPs were characterized using propensity scores and important physicochemical properties via statistical and principal component analysis. The predictive model afforded high performance with an accuracy of 88.28% and results revealed that AFPs are likely to be composed of hydrophobic amino acids as well as amino acids with hydroxyl and sulfhydryl side chains. The predictive model is provided as a free publicly available web server called CryoProtect for classifying query protein sequence as being either AFP or non-AFP. The data set and source code are for reproducing the results which are provided on GitHub.

1. Introduction

Antifreeze protein (AFP) is an ice-binding protein produced by organisms living in extremely cold temperatures and encountering freezing environments. AFPs have been found in a wide variety of species, including bacteria, fungi, insects, plants, and animals [1]. Although AFPs have similar functions, their structures are enormously varied amongst species. The diversity of AFPs may have arisen from the fact that ice contain surfaces with different geometric arrangements of oxygen atoms [2]. Moreover, crystallography and NMR studies of AFPs revealed that there are no consensus sequences or structures for ice-binding surfaces [3–7]. As a response to climate changes, AFPs may have evolved their ice-binding abilities [8]. Despite their diversity, AFPs can be classified into two major groups: (i) antifreeze glycoproteins (AFGPs) and (ii) nonglycoproteins (types I to IV AFPs) [9]. The AFGPs constitute a major fraction of proteins in the blood serum of Antarctic Notothenioids and Arctic Cod. Each AFP consists of varying numbers of repeating $(\text{Ala-Ala-Thr})_n$ units with minor sequence variations. Furthermore, the disaccharide β-D-galactosyl-(1→3)-α-N-acetyl-D-galactosamine is joined as a glycoside to the hydroxyl oxygen of the Thr residue. Thus, these compounds allow fish to survive in subzero temperatures [10]. In this article, we use the term AFPs to refer to both the AFGP and AFP groups, emphasizing their ability to protect organisms from freezing environments.

AFPs are known to elicit protection in organisms via two mechanisms. Firstly, they act by lowering the freezing point but not the melting point. The difference between the melting

Table 1: Existing methods for predicting antifreezing protein.

Method	Classifier	Interpretable	Stand-alone program	Web server	Sequence feature	Year
AFP-Pred	RF	No	√		PCP	2011
iAFP	SVM	Yes		√[a]	*n*-peptide compositions	2011
AFP_PSSM	SVM	No		√[a]	PSSM	2012
AFP-PseAAC	SVM	No	√		PseACC	2014
TargetFreeze	SVM	No		√	AAC, PseAAC and PsePSSM	2015
AFP-ensemble	RF	No		√[a]	AAC, DPC, PCP, PSSM, disorder information, and functional domain	2015
CryoProtect	RF	Yes		√	AAC and DPC	This study

DT: decision tree, RF: random forest, SVM: support vector machine, AAC: amino composition, DPC: dipeptide acid composition, PCP: physicochemical properties, PSSM: position specific scoring matrix profiles, PseACC: pseudo amino acid composition, and PsePSSM: pseudo position-specific scoring matrix.
[a]The web server is not accessible.

and freezing temperatures, termed thermal hysteresis (TH), is used to detect and quantify antifreeze activity [11]. Secondly, upon binding with ice, AFPs modify the crystallization of ice by either developing smaller crystals or forming a different shape [1]. In addition, the key feature of AFPs activity is their ability to bind with ice surfaces. Specific amino acids are arranged in a flat pattern and bind with the ice surface on ice-binding sites. This binding is stabilized by hydrogen bonds from hydrophilic amino acids strategically arranged to match the spacing of the ice lattice [3, 8]. However, their unusual relationship with water (i.e., acting as a solvent for the protein as well as its target) render the characterization of molecular mechanisms of AFPs a challenging task [8].

Owing to its importance in the survival of cold-adapted organisms and their promising agricultural (i.e., freeze-resistant transgenic animals or plants), medical (e.g., cryopreservation and cryosurgery), and industrial (i.e., food preservation) applications, the precise identification of AFPs is vital. This can only be achieved with a better understanding of their ice-binding interaction and mechanisms [10, 12]. Therefore, an accurate computational method for the identification of AFPs is needed, particularly in the postgenomic era where protein sequence information accumulate without being functionally annotated [13]. However, the challenge in the identification of AFPs lie in their sequence diversity and different ice-binding sites amongst closely related species [8, 14]. Quantitative structure-activity/property relationship (QSAR/QSPR) is a computational paradigm that facilitates the correlation of structural feature of biological or chemical entities of interest with their respective endpoints (i.e., activity or property of interest) [15, 16].

Despite the difficulty of AFP identification, many researchers have exploited computational approaches to directly predict AFPs based on their protein sequences via the use of predictive QSAR models including AFP-Pred [12], iAFP [17], AFP_PSSM [18], AFP-PseAAC [13], AFP-Ensemble [19], and TargetFreeze [20] as summarized in Table 1. TargetFreeze provided the highest predictive performance by using support vector machine and various types of protein features, namely, amino acid composition, pseudo amino acid composition, and pseudo position-specific scoring matrix. Each of the existing methods has its own merit and did play a role in stimulating the development of this area. However, all of the mentioned studies focus mainly on increasing prediction results while possessing limitations pertaining to the characterization of important features that are essential for the identification of AFPs from non-AFPs. Furthermore, to the best of our knowledge, very few effective methods or bioinformatics tools for characterizing AFPs have been proposed.

One of the main values of bioinformatics tools should be its ability to provide insight into mechanisms of action under study. Therefore, this work attempts to develop an interpretable computational predictor for AFPs while affording a comparable accuracy. A prediction method named CryoProtect that is based on a random forest classifier and the combination of amino acid composition and dipeptide composition is proposed herein. Rigorous cross-validation suggests that the CryoProtect approach afforded the best predictive performance amongst four out of the five existing methods. Furthermore, this method also provided a comparable performance against the state-of-the-art approach, TargetFreeze. Moreover, this study also identified important features underlying the antifreeze activity based on the amino acid composition as well as physicochemical properties as derived from the AAindex database [21]. Finally, CryoProtect is provided as a free and publicly available web server at http://codes.bio/cryoprotect/ for classifying query protein sequences as AFP or non-AFP.

2. Materials and Methods

2.1. Data Set. In order to fairly compare our study with existing methods, we used the benchmark data set described by Kandaswamy et al. [12]. Briefly, the positive data set was constructed in four steps: (1) an initial positive data set consisting of 221 AFPs was taken from seed proteins in the Pfam database [22], (2) the 221 AFPs were enriched by using Position-Specific Iterated-Basic Local Alignment Search Tool (PSI-BLAST) for each sequence against a nonredundant sequence database using stringent threshold with *E*-value of 0.001, (3) the enriched data set was manually checked and all non-AFPs were removed, and (4) after the sequence identity was reduced to 40% via the use of CD-HIT program [23],

the final positive data set consisted of 481 AFPs. A negative data set consisting of 9,193 seed proteins from the Pfam database that bears no resemblance to AFPs were designated as non-AFPs [22]. After removing several protein sequences containing special characters (e.g., X and U) we obtained the final data set containing 478 AFPs and 9,139 non-AFPs.

The data set was further divided into two subsets consisting of an internal set and an external set. In consideration of the class imbalance of the data set in which the size of AFPs and non-AFPs was significantly out of proportion therefore the data set was partitioned such that the internal set now contains 300 AFPs and 300 non-AFPs. This was performed via random undersampling of the non-AFP class. The remaining 178 AFPs and 8,839 non-AFPs served as the external set for further evaluation of the extrapolation capability of the predictive model.

2.2. Descriptor Calculation. In order to develop an interpretable computational predictor for providing a comprehensive perspective of the biological and chemical properties under study, feature representation plays a pivotal role. It allows an effective predictor that can truly reflect the correlation between features and biological properties and thereby provide insights on AFPs. Previously, numerous types of protein feature representation were used to develop various sequence-based predictors. However, most of these protein descriptors were problematic and scarcely contributed any biological or chemical knowledge to users. To remedy this shortcoming, an easy and interpretable feature pertaining to the amino acid composition (AAC), dipeptide composition (DPC), and physicochemical property (PCP) were considered. The latter set of descriptors was derived from the AAindex [21] and spanned a wide range of physicochemical properties (e.g., hydrophobicity, helices, β-sheet, side chain, and buried residues). Moreover, the potential ability of these features to predict protein functions has been extensively demonstrated previously.

AAC is the proportion of each amino acid in a protein sequence that is expressed as a fixed length of 20. Given a peptide sequence with length l, the occurrence frequency of the ith amino acid (a_i) is calculated as follows:

$$a_i = \frac{AA_i}{l}, \tag{1}$$

where AA_i is the number of occurrences in the sequence for the ith amino acid.

DPC is the proportion of two consecutive amino acids or dipeptides having a fixed length of $20 \times 20 = 400$. DPC encompasses information regarding amino acid composition along the local order of amino acids. For a given peptide sequence, the occurrence frequency of the ith dipeptide (dp_i) is calculated as follows:

$$\mathrm{dp}_i = \frac{\mathrm{DP}_i}{l}, \tag{2}$$

where DP_i is occurrence of the ith amino acid in the sequence.

Physicochemical property (PCP) of amino acids is essential for the prediction and analysis of various proteins and peptides in a wide range of bioinformatics studies owing to its interpretability [24–26]. There are 544 PCPs of amino acids extracted from the amino acid index database (AAindex) [21], which is a collection of the published literature as well as different biochemical and biophysical properties of amino acids. Each physicochemical property consists of a set of 20 numerical values for amino acids. After removing 13 PCPs with the value "NA" in a value set of the amino acid index, a total of 531 PCPs were used for subsequent analysis.

2.3. Cross-Validation to Identify the Predictive Capability. In statistical predictions, three popular cross-validation (CV) methods are often used to identify the empirical predictive model for its robustness, namely, N-fold cross-validation, jackknife test, and external validation. To fairly compare the existing approaches and reduce the computational time, a 10-fold cross-validation (10-fold CV) and external validation were carried out. Additionally, to avoid the possibility of bias arising from single data split upon model training, data splitting was performed for 20 independent iterations. Particularly, each data split divides the data into two subsets consisting of an internal set and an external set. The former set was used as the training set and subjected to 10-fold CV in which the data was partitioned into 10 folds and 1 fold was left out as the testing set while the remaining were used for training the model. This process was repeated iteratively until all the folds have had the chance to be left out as the testing set. Subsequently, an external validation performed on the external set was used to assess the predictive capability of the models for inferring any unknown data not previously seen by the training model.

2.4. Multivariate Analysis. To develop an interpretable sequence-based predictor, the learning classifiers, namely, decision tree (DT) and random forest (RF), were used for the prediction and analysis of AFPs. In order to analyze the overall aspect of the informative features for classifying AFPs and non-AFPs, the principal component analysis (PCA) approach was applied. This method reduces the original feature space to a fewer dimensions while still retaining most of the variation explained by the original data set. Further details of the three learning classifiers are provided below.

PCA is a mathematical and statistical algorithm that is used for reducing the dimensionality of the data set while still retaining most of the variation [27]. In brief, PCA transforms the original variables into a set of linear combinations, namely, the principal component (PC). The PCs are linearly independent and are weighted in decreasing order of variance coverage. Thus, all original M-dimensional data patterns can be optimally transformed to the feature space with lower dimensionality [28]. PCA analysis was performed using the *FactoMineR* package [29] in R program version 3.0.1 [30].

DT is comprised of a hierarchical arrangement of nodes and branches in which nodes represent the peptide features whereas branches refer to decision rules for categorizing peptides as AFPs and non-AFPs. DT models have been successfully applied in the analysis of various types of compounds like aromatase inhibitors [31], dipeptidyl peptidase 4 inhibitors [32], influenza neuraminidase inhibitors

[33], volatile organic compounds [34], cytochrome P450-interacting compounds [35], and so forth. In this study, the DT model was constructed using the J48 algorithm from the RWeka R package using default parameters. Briefly, the J48 algorithm is a Java implementation of the C4.5 algorithm, which establishes a DT model by iteratively appending features having high information gains [36]. Finally, the algorithm automatically calculates the feature usage derived from the full decision tree or a collection of rules.

RF is an ensemble classification and regression tree (CART) classifier [37–39] whereby each tree is generated using a random vector that is sampled independently from the input vector [38]. The RF method grows many weak CART trees that enhance its prediction performance. Furthermore, the out-of-bag (OOB) approach was used for evaluating the feature importance in which two-thirds of the training data was used for constructing the predictive classifier while the remaining was used for evaluating the performance of the classifier where the decrease in the prediction performance was measured. It should be noted that the performance evaluation of the model can use either accuracy or Gini index. Herein, the RF classifier was established using the *randomFores*t R package [39, 40]. To enhance the performance of the RF model, two parameters, namely, *ntree* (i.e., the number of tree used for constructing the RF classifier) and *mtry* (i.e., the number of random candidate features), were subjected to optimization. Particularly, $ntree \in \{100, 200, 300, 400, 500\}$ was determined using 10-fold cross-validation (10-fold CV) and *mtry* was estimated using the tuneRF function in the *randomForest* R package [39, 40].

2.5. Identification of Informative Physicochemical Properties. Previously, the physicochemical properties (PCPs) of amino acids have been recognized as valuable features for providing a better understanding of protein functions from their primary sequences [24–26]. Here, the propensity scores of amino acids were utilized to provide insight into the characteristics of AFPs. The propensity scores of 20 amino acids (PS-AFP) for distinguishing AFPs from non-AFPs were calculated as follows:

$$\text{PS-AFP}_{AA(i)} = \frac{\sum \text{AFP}_{AA(i)}}{\sum \text{AFP}} - \frac{\sum \text{AFP}_{AA(i)}}{\sum \text{non-AFP}}, \quad (3)$$

where $\text{PS-AFP}_{AA(i)}$ is the propensity score for the ith amino acid and $\sum \text{AFP}_{AA(i)}$ and $\sum \text{non-AFP}_{AA(i)}$ represent the total number of ith amino acid in AFPs and non-AFPs, respectively. The $\sum$ AFP and $\sum$ non-AFP represent the total number of all amino acids in AFPs and non-AFPs, respectively. Finally, the propensity scores of all amino acids were normalized into the range of [0, 1000]. In this study, the identification of informative PCPs was performed using $\text{PS-AFP}_{AA(i)}$ in which Pearson's correlation coefficients (R) were computed between $\text{PS-AFP}_{AA(i)}$ and the 531 PCPs followed by selection of the five top-ranking PCPs affording the highest absolute R values for further analysis [26].

2.6. Performance Evaluation. One of the crucial procedures in developing a reliable and useful predictor is to objectively evaluate its performance. From the point-of-view of pattern recognition, the prediction of AFPs can be addressed as a classification problem. Herein, five standard statistical parameters, namely, accuracy (Ac), sensitivity (Sn), specificity (Sp), Matthew's correlation coefficient (MCC), and Youden's index (YI), were used to assess the predictive performance of the proposed methods. These five parameters were computed as follows:

$$\begin{aligned}
\text{Ac} &= \frac{\text{TP} + \text{TN}}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \times 100 \\
\text{Sn} &= \frac{\text{TP}}{(\text{TP} + \text{FN})} \times 100 \\
\text{Sp} &= \frac{\text{TN}}{(\text{TN} + \text{FP})} \times 100 \\
\text{MCC} &= \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})\,(\text{TP} + \text{FN})\,(\text{TN} + \text{FP})\,(\text{TN} + \text{FN})}} \\
\text{YI} &= \text{Sn} + \text{Sp} - 1,
\end{aligned} \quad (4)$$

where TP, TN, FP, and FN represent the true positive, true negative, false positive, and false negative, respectively.

2.7. Development of the CryoProtect Web Server. The CryoProtect web server was developed using the Shiny package under the R programming environment. The utilization of Shiny boasts several benefits. The first advantage is the seamless integration of the web server with the aforementioned predictive model that was also built in R. The second benefit is that there is no requirement for developers to have an extensive knowledge of web development, although it may be useful. Most importantly, Shiny facilitates rapid development and deployment of web applications, which is especially beneficial for the scientific community as predictive models can be readily deployed as a web server making it accessible to a wider group of users instead of confined to those with a background in computer science. In optimizing the loading and processing time of the web server, the data set was subjected to data balancing as to afford a balanced data set consisting of 300 AFPs and 300 non-AFPs. The CryoProtect web server is accessible at http://codes.bio/cryoprotect/ while the source code is available on GitHub at https://github.com/chaninn/cryoprotect/.

3. Results and Discussion

In this study, AFPs and non-AFPs are predicted by the proposed method CryoProtect. Firstly, analyses of PCA, propensity scores, and Gini index were performed as to characterize informative properties of antifreeze activity. Secondly, informative PCPs were used to investigate vital factors for improving the antifreeze activity of proteins. Afterwards, the proposed CryoProtect method was compared with existing methods. Finally, CryoProtect is deployed as a free prediction web server as to afford easy and rapid classification of query protein sequence as being AFP or non-AFP. Figure 2 illustrates the workflow of prediction

Table 2: Propensity score of 20 amino acids, composition difference, and summary of statistical analysis of AFPs and non-AFPs. The rank of propensity score and Gini index are shown in parenthesis.

Amino acid	Propensity score		Gini index		AFP (%)	Non-AFP (%)	Difference (%)	p value
Cys	1000	(1)	47.16	(1)	3.6	1.5	2.1	>0.05
Ala	944	(2)	18.88	(4)	9.5	8.2	1.3	>0.05
Ser	890	(3)	10.93	(12)	8.4	6.9	1.5	>0.05
Thr	867	(4)	15.49	(7)	6.6	5.3	1.3	>0.05
Gly	858	(5)	12.93	(9)	7.4	6.4	1.0	>0.05
Trp	777	(6)	30.40	(2)	2.5	1.3	1.2	>0.05
Gln	615	(7)	10.18	(13)	4.4	4.0	0.4	0.081
Asn	572	(8)	8.31	(18)	4.8	4.3	0.5	>0.05
Pro	560	(9)	9.14	(17)	4.5	4.7	−0.2	>0.05
His	559	(10)	5.52	(20)	2.4	2.2	0.2	0.066
Asp	502	(11)	7.68	(19)	4.9	5.4	−0.5	0.090
Tyr	471	(12)	11.69	(11)	2.9	3.3	−0.3	>0.05
Glu	433	(13)	12.48	(10)	5.7	6.7	−1.0	>0.05
Phe	407	(14)	9.52	(16)	3.6	4.1	−0.5	0.233
Met	404	(15)	13.48	(8)	2.2	2.5	−0.4	>0.05
Val	379	(16)	9.94	(14)	5.6	6.5	−0.9	>0.05
Arg	249	(17)	21.41	(3)	4.5	5.7	−1.2	>0.05
Ile	191	(18)	16.36	(6)	4.1	5.6	−1.5	>0.05
Lys	159	(19)	9.60	(15)	5.1	6.0	−0.8	>0.05
Leu	0	(20)	18.54	(5)	7.5	9.5	−2.0	>0.05

procedures for CryoProtect in classifying protein sequences as AFPs and non-AFPs.

3.1. Biological Space of Antifreeze Protein. In this study, the PCA analysis and propensity score analysis of 20 amino acids were used for identifying important properties governing the AFP activity as illustrated in Figure 3 and Table 2. Figure 3 shows the loadings and scores plots derived from the use of informative amino acids. These amino acids were selected by using t-test in order to compare compositions of amino acids between AFPs and non-AFPs. The result of p values and propensity scores of amino acids are shown in Table 2. As can be seen, 13 amino acids were found to be significantly different between AFPs and non-AFPs at the level of $p < 0.05$. Figure 3 shows the scores (Figure 3(a)) and loadings plot (Figure 3(b)) as derived from the informative amino acids where the red and blue circles represent AFPs and non-AFPs, respectively. Results indicated that Cys, Ser, Trp, Gly, Asn, and Thr were characteristic of AFPs while Leu, Val, Glu, Ile, and Met were characteristic of non-AFPs. Interestingly, these results are well reflected in which amino acids with the highest propensity scores were Cys, Ala, Ser, Thr, and Gly with corresponding values of 1000, 944, 890, 867, and 858, respectively, while amino acids with the lowest propensity scores were Leu, Lys, Ile, Arg, and Val with propensity scores of 0, 159, 191, 249, and 379, respectively.

Moreover, this study also made use of the Gini index from the RF model for evaluating and ranking the feature importance of amino acids as shown in Table 2. Features with the largest Gini index are deemed to be the most important owing to their contribution to the prediction performance [37–39]. Interestingly, there were seven out of ten top-ranked amino acids that were found to belong to the top five and bottom five amino acids representing highest and lowest propensity scores, respectively, such as Cys, Ala, Thr, Gly, Arg, Ile, and Leu. It was observed that results from the Gini index were complementary to the analysis of propensity scores.

The importance of Cys, Thr, Ser, Asn, and Gly in contributing to the activity of AFP is supported by several previous experimental evidences. Liou et al. [41] identified a consensus sequence consisting of repeating units (Cys-Thr-Xaa-Ser-Xaa-Xaa-Cys-Xaa-Xaa-Ala-Xaa-Thr) from the AFPs of the beetle *Tenebrio molitor* (TmAFP). This TmAFP has a 10 to 100 times lower freezing point as compared to fish AFP. Also, the X-ray crystallography and NMR studies revealed that the ice-binding surface of TmAFP is composed of Thr-Xaa-Thr motifs [42]. Moreover, Marshall et al. [43] engineered the addition and deletion of repeated TmAFP coils in order to study whether the length of AFPs and the addition of binding site enhance the antifreeze activity (Figure 4). According to their study, there was a 10- to 100-fold gain in activity for the addition of six to nine coils, depending on the concentration that was compared. The maximum freezing point depression of 6.5°C at 0.7 mg/mL was achieved by the nine-coil construct but decreased for the ten- and eleven-coil constructs. Therefore, they concluded that the antifreeze activity increases with the length of the β-helix. Although the relationship between the activity of thermal hysteresis and the concentration of AFPs are nonlinear, the differences in the activity of AFPs are not strictly proportional over a concentration range [43].

Table 3: Propensity score of amino acids and selected physicochemical properties from the AAIndex. The rank of descriptors is shown in parenthesis.

Amino acid	Propensity score		RICJ880112		SNEP660104		KOEP990101		QIAN880125	
Cys	1000	(1)	0.2	(18)	0.38	(2)	0.57	(2)	−0.02	(9)
Ala	944	(2)	0.7	(13)	−0.062	(13)	−0.04	(12)	−0.02	(10)
Ser	890	(3)	0.6	(15)	0.47	(1)	0.15	(7)	0.41	(1)
Thr	867	(4)	0.7	(14)	0.348	(3)	0.39	(3)	0.36	(3)
Gly	858	(5)	0.1	(19)	−0.017	(10)	1.24	(1)	0.38	(2)
Trp	777	(6)	0.4	(17)	0.05	(9)	0.21	(6)	−0.01	(8)
Gln	615	(7)	1.3	(6)	−0.025	(11)	−0.02	(11)	−0.17	(16)
Asn	572	(8)	0.8	(11)	0.166	(5)	0.25	(5)	0.03	(7)
Pro	560	(9)	0.0	(20)	−0.036	(12)	0.00	(9)	−0.04	(12)
His	559	(10)	1.1	(8)	0.056	(8)	−0.11	(15)	−0.09	(14)
Asp	502	(11)	0.6	(16)	−0.079	(14)	0.27	(4)	0.11	(4)
Tyr	471	(12)	1.1	(9)	0.22	(4)	0.05	(8)	−0.08	(13)
Glu	433	(13)	1.6	(4)	−0.184	(16)	−0.33	(19)	0.10	(5)
Phe	407	(14)	1.8	(3)	0.074	(7)	−0.01	(10)	−0.03	(11)
Met	404	(15)	1.0	(10)	0.077	(6)	−0.09	(14)	−0.14	(15)
Val	379	(16)	1.3	(7)	−0.212	(17)	−0.06	(13)	−0.18	(17)
Arg	249	(17)	0.8	(12)	−0.167	(15)	−0.30	(18)	0.04	(6)
Ile	191	(18)	1.4	(5)	−0.309	(19)	−0.26	(17)	−0.48	(20)
Lys	159	(19)	2.2	(1)	−0.371	(20)	−0.18	(16)	−0.39	(19)
Leu	0	(20)	1.9	(2)	−0.264	(18)	−0.38	(20)	−0.26	(18)
R	1.00		−0.741		0.736		0.695		0.683	

In order to function at subzero temperatures, AFPs rely mostly on hydrogen and disulfide bonds rather than their hydrophobic core [44]. This observation supports the analysis of PCA results and propensity scores, which revealed the importance of residues with hydroxyl and sulfhydryl side chains (e.g., Cys, Thr, Ser, Asn, and Gly) as the main amino acids responsible for the bioactivity of AFPs. Amongst these residues, only Cys is known to exhibit moderate hydrophobicity with a sulfhydryl side chain, while the rest are more hydrophilic and, thus, are prone to participate in hydrogen bond formation [45, 46]. In addition, the sulfhydryl groups of Cys form disulfide bridges in β-helix of TmAFP thereby allowing for the formation of a tight structure whereby no hydrophobic core or long sidechains are in the helix.

Meanwhile, Ser residues and Asn residues stabilize β-helix structure by forming ladder-like structure. Ser residues are lining on one side of the protein and form a ladder structure [44, 47], whereas Asn ladders were identified to be inside the β-helix structure of AFP derived from freeze-tolerant grass *Lolium perenne* (LpAFP) [48]. Two internal Asn ladders are made up of side chain amide and carbonyl groups of hydrogen bonds that bind to the main-chain atoms of neighboring coils and to the adjacent Asn side chains [48]. Moreover, the importance of hydrogen bonds can be seen in AFPs from snow flea (sfAFP) lacking hydrophobic cores, as it mainly contains Gly which is less hydrophobic. The structure of sfAFP consists of polyproline type II helix, formed by six coils whereby all the structures projected inwards are made up of Gly (Figure 5). This structure allows the coils to form carbonyl-amide hydrogen bonds with each other [44, 49].

3.2. Characterization of AFPs Using Propensity Score of Physicochemical Properties. Physicochemical properties of amino acids play an essential role in the identification and characterization of protein functions from their primary sequences. Table 3 shows selected PCPs with their corresponding R values consisting of SNEP660104 ($R = 0.736$), RICJ880112 ($R = -0.741$), KOEP990101 ($R = 0.695$), and QIAN880125 ($R = 0.683$). The analyses of four PCPs of AFPs are discussed below.

3.2.1. Contribution of Hydrophobic Residues to AFP Activity. The property of RICJ880112 was described as "amino acid preferences for specific locations at C3 ends of the α-helices." In 1988, J. S. Richardson and D. C. Richardson [50] calculated the amino acid preference at specific location such as at the end of α helices based on α-carbon positions and a sample of 215 α helices from 45 different globular protein structures. This study revealed that particular amino acids prefer to remain in certain positions at the 16 individual positions relative to the helix ends. This finding is important in order to predict a three-dimensional protein structure from amino acid sequences. According to this study, a peak preference for hydrophobic amino acids in position C3 can be observed and these peaks are especially strong for Leu [50]. As seen in Table 3, the property of RICJ880112 has the highest inverse correlation ($R = -0.741$) indicating that the AFPs tend to be composed of amino acids with low hydrophobicity. Interestingly, five amino acids with the highest propensity scores belong the group of moderate and less hydrophobic amino acids.

Although it is generally accepted that the ice-binding site of AFPs are mainly composed of hydrophilic amino acid residues, Chen and Jia [51] suggested that a larger ice-binding site might contain hydrophobic residues. By employing molecular docking simulations, they analyzed the ice-binding interaction energy of 11 different surface patches of type III AFP from fish. The simulations identified the most favorable interaction energy containing 14 residues including the highly hydrophobic amino acids, Ile, Val, and Leu [46]. Based on this analysis, the authors concluded that there is an enlargement of the ice-binding site resulting from an incorporation of surrounding hydrophobic residues.

Furthermore, Baardsnes and Davies [4] investigated the importance of hydrophobic residues of type III AFP towards protein and ice interactions by mutagenesis study. In their study, the hydrophobic residues at the ice-binding site (Leu, Ile, and Val) were mutated into the less hydrophobic Ala residue (Figure 6). It was found that single substitutions of Leu19Ala, Val20Ala, and Val41Ala decreased the activity by 20%, whereas double substitutions of Leu19Ala/Val41Ala and Leu10Ala/Ile13Ala decreased the antifreeze activity by more than 50% when compared to the wild type. Although the Ala substitutions only moderately decreased the van der Waals interactions, the overall mutations could reduce the interactions between ice and AFPs [4]. In contrast, Garnham et al. [52] reported a double mutation of less hydrophobic amino acids (Pro and Ala) to highly hydrophobic amino acid residues (Leu and Val) in AFP type III isoform SP seen in notched fin eelpout fish (SPnfe6). It was found that the double mutation of Pro19Leu/Ala20Val in the SPnfe6 mutant increased the ice-binding activity by increasing the surface coverage. Furthermore, the double mutant decreased the growth of ice crystals by greater than 30-fold when compared with the wild type SPnfe6 in the same concentration. Hence, contact surface area is important for the activity of AFPs and the enlargement of surface area will result in forming additional binding sites.

3.2.2. Hydroxythiolation of the AFP Side Chains and Its Contribution. The property of SNEP660104 is described as "relations between chemical structure and biological activity in peptides on principal component IV." Sneath [53] studied the correlation of amino acid substitution and variation of biological activity of peptides by principal component analysis. Four principal components (principal components I, II, III, and IV) were derived from calculations of 20 amino acids and interpreted as different properties. Principal component IV represents the hydroxythiolation property which can be described as the involvement of hydroxyl and sulfhydryl groups in protein activity as well as the ability of amino acids in vector IV to form hydrogen bonds. This property has the highest positive correlation ($R = 0.736$) indicating that AFPs favor amino acids that contain hydroxythiolation.

Table 3 shows 3 out of 5 amino acids having the highest propensity scores (e.g., Cys, Ser, and Thr) possessed hydroxythiolation property. Duman [54] reported the importance of this property on the structural stability of AFPs in terrestrial arthropods, *Dendroides* (DAFP) and *Tenebrio* (TmAFP, Figure 1). The sulfhydryl group in Cys residues form disulfide bridges whereby 6 out of 8 disulfide bonds are aligned in the internal loop. Although the other 2 disulfide bonds in the N-terminal do not follow this pattern, there is no distortion in the loop formation. These structures stabilize the proteins and enable polar Thr and Ser residues with a hydroxyl side chain to align in the ice-binding site and form hydrogen bonds between the AFP and ice [47, 54].

Hydrogen bonds play an important role in protein/ice interactions as they function to inhibit the growth of ice crystals by blocking surface adsorption [44]. The study of the ice-binding mechanism of AFPs from winter flounder (wfAFP) shows that the greater the number of hydrogen bonds, the higher the antifreeze activity [55]. Moreover, Wierzbicki et al. [56] employed molecular dynamics simulations and identified that the number of hydrogen bonds are determined by the type of amino acid residues that move towards the ice. In addition, they discovered that when the Thr-Ala-Ala site of the AFP is facing the ice the antifreeze activity increases as compared to when the Thr-Ala-Asx site faces the ice. This occurs because the movement of the Thr-Ala-Ala site towards the ice surface allows 13 additional sites of the AFP to come in close contact and form hydrogen bonds with the ice surface. Furthermore, the close contact of the Thr-Ala-Ala residues enables a larger surface area of the protein to associate with the ice (892 ± 4.5 Å).

3.2.3. Diversity of Secondary Structures of AFPs. We have selected two properties that describe the diversity of AFPs secondary structure based on their propensity score correlation coefficient (R value). The two properties of KOEP990101 ($R = 0.695$) and QIAN880125 ($R = 0.683$) positively correlate with our calculated propensity score of the AFP. The property of KOEP990101, obtained from Koehl and Levitt study [57], is described as "α-helix propensity derived from designed sequences," whereas the QIAN880125 property is described as "weights for β-sheet at the window position of 5" obtained from Qian and Sejnowski [58] prediction model.

A great need for an accurate 3D protein structure method has led to the development of protein secondary structure prediction in the past decades. Current methods for the secondary structure prediction of proteins are based on the algorithms adopted from simple statistical- and pattern recognition-based methods [59]. In 1988, Qian and Sejnowski [58] developed a method for predicting the secondary structure of proteins based on neural networks (NNs) model, a pattern recognition-based method. The strength of a connection from each network is called a weight while the network itself can be considered as a window. It was observed that there are certain weights for α-helix, β-sheet, and coil, in the specific window position.

Furthermore, Koehl and Levitt [57] developed a protein design method and analyzed the conformational preferences for the amino acids. From the designed sequences, the conformational preferences of amino acids were derived. In addition, a structure-based propensity scale was determined from calculations of a complete physical potential, such as van der Waals, electrostatic and hydrophobic interactions. The authors found that the values obtained from a structure-based propensity scale show significant agreement with the

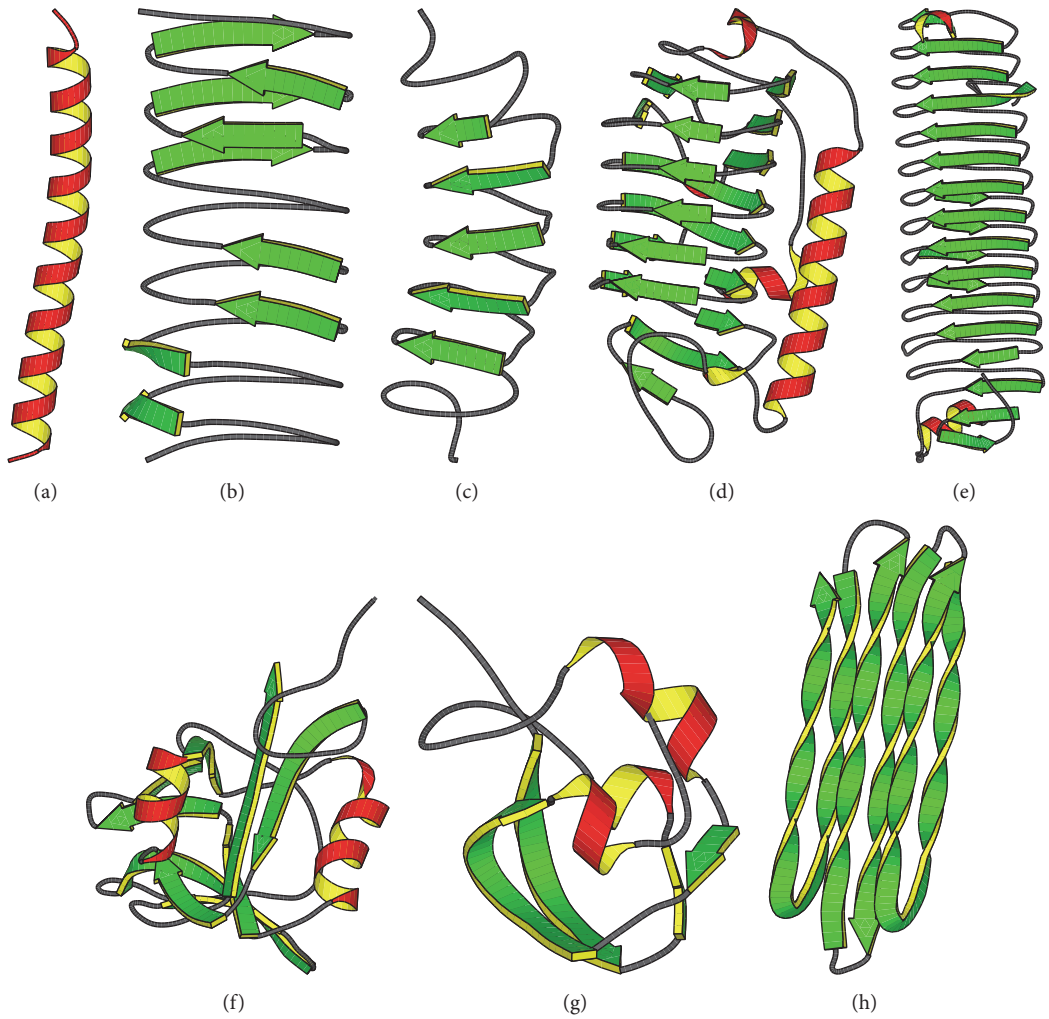


Figure 1: Diversity of protein structures of AFPs. Winter flounder *Pseudopleuronectes americanus* (PDB ID: 1WFA) (a), perennial ryegrass *Lolium perenne* (PDB ID: 3ULT) (b), insect AFP from *Tenebrio molitor* (PDB ID: 1L1I) (c), fungal AFP from *Typhula ishikariensis* (PDB ID: 3VN3) (d), β-helical AFP from Antarctic bacterium *Marinomonas primoryensis* (PDB ID: 3P4G) (e), Type II AFP from sea raven *Hemitripterus americanus* (PDB ID: 2AFP) (f), Type III AFP from ocean pout *Zoarces americanus* (PDB ID: 1KDE) (g), and snow flea AFP from *Hypogastrura harveyi* (PDB ID: 2PNE) (h). α-helix, β-sheet, and coil are shown in red, green, and grey colors, respectively, while the inner side of the α-helix and the side of the β-sheet are shown in yellow color.

experimental propensity scale values for both α-helix and β-sheet [60].

In this study, two PCPs of KOEP990101 and QIAN880125 that describe the protein secondary structure propensity score positively correlate with the propensity score of amino acid composition derived from AFPs. This result reflects the diversity of secondary structures of AFPs (Figure 1). Although AFPs have no identical amino acid sequences or structures, they were classified based on their secondary structures. According to the work of [8], fish AFPs can be classified into several subfamilies based on their secondary structures. Type I AFP is a α-helix that is mainly composed of Ala with 11 amino acid repeating units in the helical turns, whereas type II AFP is a mixture of α, β, and loop or coil structures with no observed amino acid repeats. Furthermore, type III AFP contains short β strands and, although no amino acid repeats were observed, this protein was seen to form a dimer. Likewise, the structure of insect AFPs from different species *Choristoneura fumiferana*, *Dendroides canadensis,* and *Tenebrio molitor* are β-helix.

3.3. Performance Evaluation. In this study, we investigated the predictive capability of the proposed method by considering performance comparisons between two popular interpretable machine learning algorithms (e.g., DT and RF) using protein features (e.g., AAC, DPC, and a combination of AAC and DPC). Rigorous evaluation of the predictive power of the proposed method, CryoProtect, was performed via the use of 10-fold CV and external validation. As previously mentioned, the benchmark data set described by Kandaswamy et al. [12] was used as is for comparative purposes. Table 4 lists the performance comparisons of various models using different learning methods and sequence features over 10-fold CV and external validation.

Table 4: Performance comparisons among various types of machine learning algorithms and protein sequence as evaluated on 10-fold CV and external validation tests.

Classifier	Feature(s)	10-fold CV set					External validation set				
		Ac (%)	Sn (%)	Sp (%)	MCC	YI	Ac (%)	Sn (%)	Sp (%)	MCC	YI
RF	AAC	86.33 ± 1.36	87.50 ± 1.10	85.27 ± 2.11	0.73 ± 0.03	0.73 ± 0.03	87.50 ± 0.46	78.65 ± 2.92	87.68 ± 0.51	0.27 ± 0.01	0.66 ± 0.02
	DPC	84.83 ± 0.73	86.91 ± 1.87	83.12 ± 2.40	0.70 ± 0.01	0.70 ± 0.01	84.12 ± 2.77	78.84 ± 3.74	84.22 ± 2.88	0.23 ± 0.02	0.63 ± 0.03
	AAC + DPC	89.50 ± 1.26	89.54 ± 0.14	89.50 ± 2.41	0.79 ± 0.03	0.79 ± 0.03	88.28 ± 1.00	87.27 ± 2.27	88.30 ± 1.07	0.31 ± 0.01	0.76 ± 0.01
DT	AAC	77.67 ± 0.76	78.31 ± 1.20	77.08 ± 1.09	0.55 ± 0.02	0.55 ± 0.02	81.99 ± 4.60	81.27 ± 6.26	82.00 ± 4.81	0.23 ± 0.02	0.63 ± 0.02
	DPC	73.33 ± 0.76	72.92 ± 0.48	73.76 ± 1.07	0.47 ± 0.02	0.47 ± 0.02	74.50 ± 1.47	77.34 ± 1.81	74.45 ± 1.53	0.16 ± 0.00	0.52 ± 0.00
	AAC + DPC	83.50 ± 1.92	83.33 ± 2.25	83.68 ± 1.73	0.67 ± 0.04	0.67 ± 0.04	82.26 ± 1.45	77.90 ± 1.97	82.35 ± 1.52	0.21 ± 0.01	0.60 ± 0.01

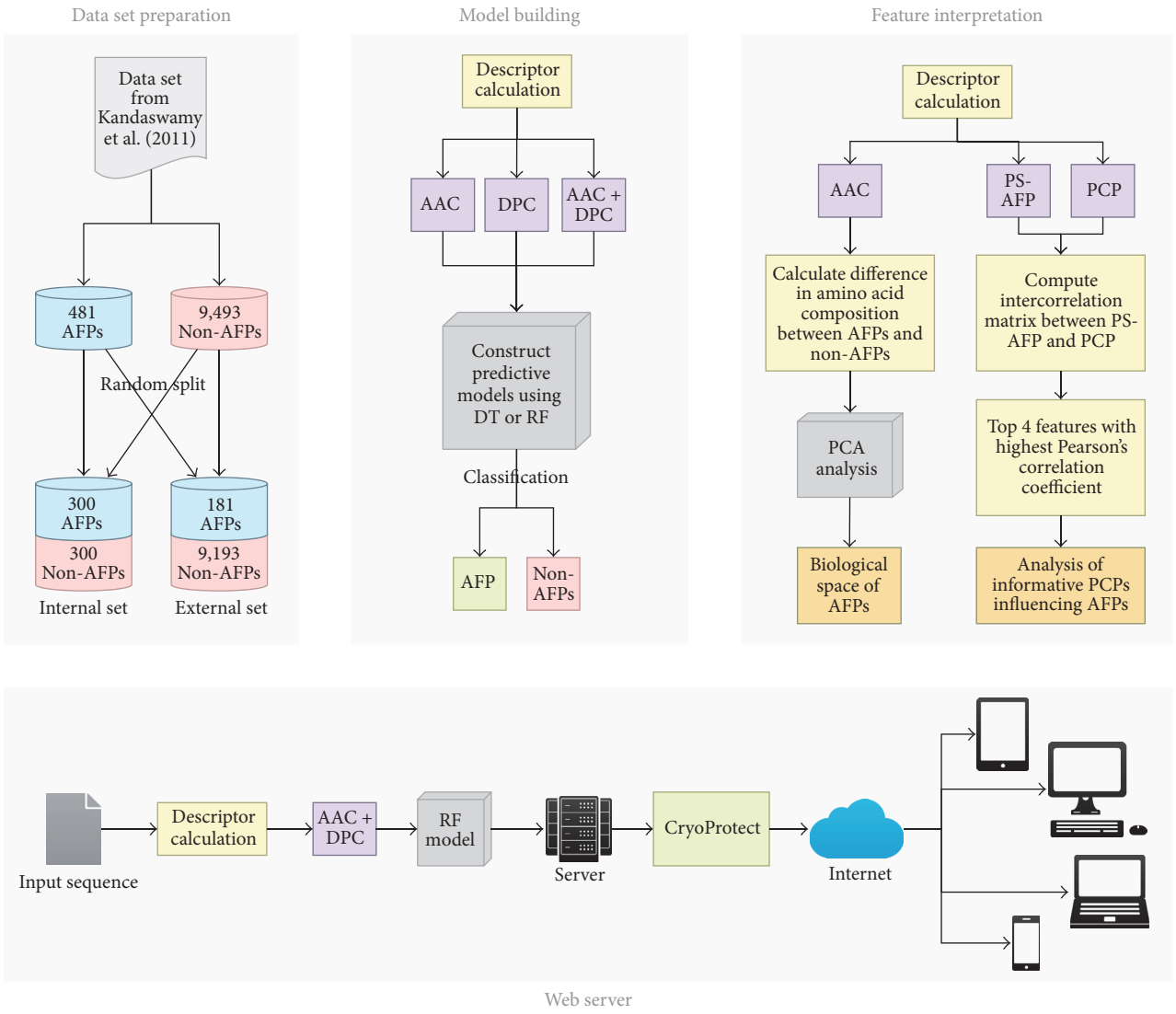


Figure 2: Flowchart of the prediction procedures of CryoProtect for classifying protein sequences as AFPs and non-AFPs.

In the case of a single feature, the RF model using AAC yielded the highest prediction results with a mean Ac, Sn, Sp, MCC, and YI of 86.33%, 87.50%, 85.27%, 0.73, and 0.73, respectively. Moreover, the 10-fold CV performed notably over external validation with a mean Ac, Sn, Sp, MCC, and YI of 87.50%, 78.65%, 87.68%, 0.27, and 0.66, respectively. Meanwhile, the RF model using DPC and the DT model using AAC performed effectively with the second and third highest mean Ac of 84.33%/84.12% and 77.67%/81.99% for 10-fold CV and external validation, respectively. As can be seen in Table 4, the prediction performances for both machine learning methods were quite consistent with those previously reported by He et al. [20]. In order to enhance the prediction performance, the combination of AAC and DPC was considered. Table 3 shows that the best Ac, Sn, Sp, MCC, and YI over 10-fold CV of 89.50%, 89.54%, 89.50%, 0.79, and 0.79, respectively, are achieved by using the RF model. Interestingly, the RF model also provided a substantial 10% improvement for both Sn and YI.

By observing the performance comparisons in Table 3, it can be briefly summarized as follows: (1) AAC plays a pivotal role in discriminating between AFP or non-AFP; (2) the RF model with the combination of AAC and DPC showed significant performance when evaluated by both 10-fold CV and external validation procedures. For convenience, from herein the best predictor for discriminating between AFP or non-AFP based on RF learning method in conjunction with the combination of AAC and DPC will be referred to as CryoProtect.

3.4. Performance Comparisons of CryoProtect and Existing Methods. In this section, we compare the proposed method CryoProtect with other popular AFP predictors, namely, iAFP [17], AFP-Pred [12], AFP_PSSM [18], AFP-PseAAC [13],

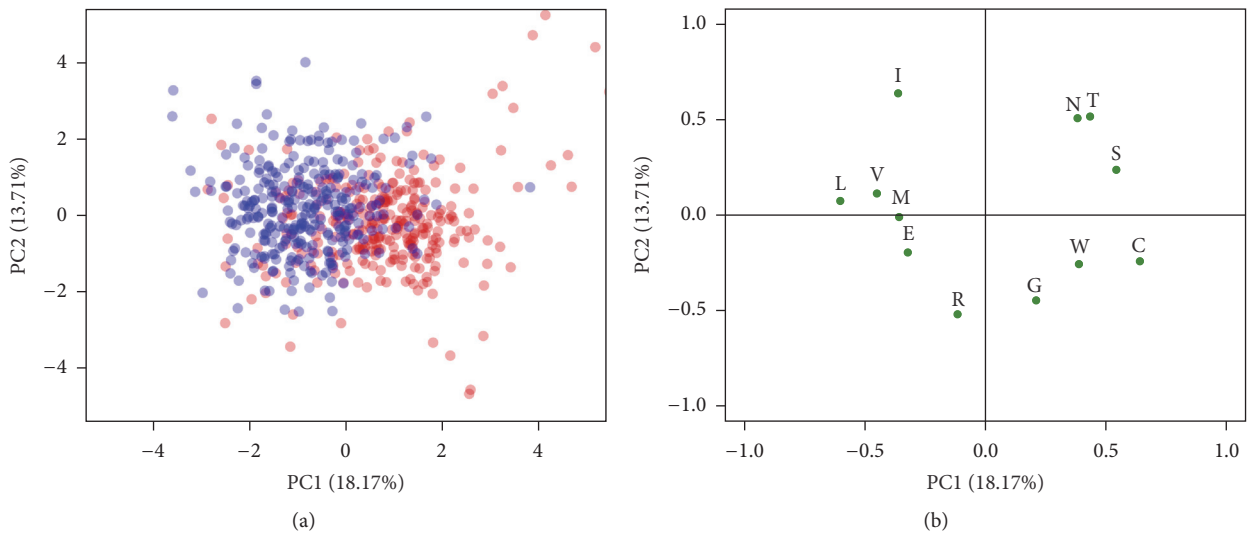


Figure 3: PCA scores (a) and loadings (b) plots of amino acid composition of AFPs. Amino acids with statistically significant difference ($p < 0.05$) in their composition were selected for PCA analysis. AFPs and non-AFPs are represented by red and blue circles.

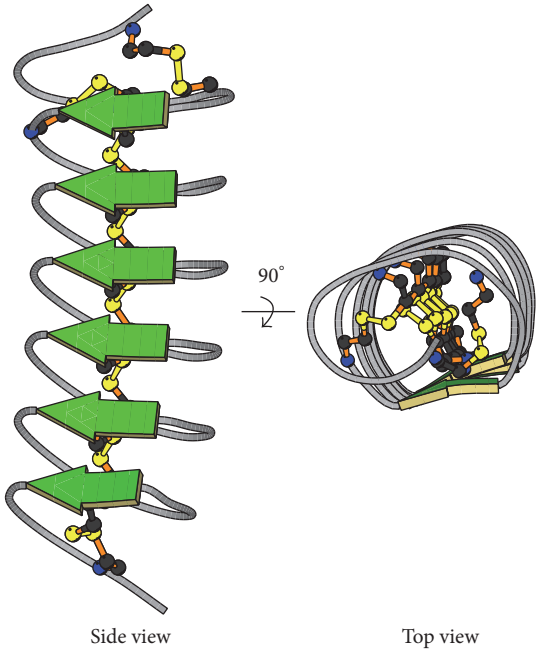


Figure 4: Crystal structure of wild type AFP from *Tenebrio molitor* (PDB ID: 1ezg) showing the β-sheet region and disulfide bonds in green and yellow colors, respectively.

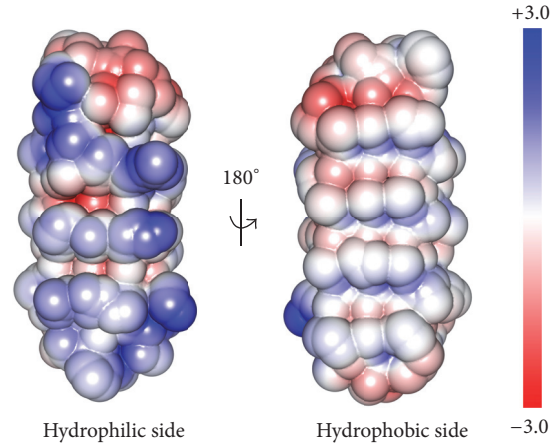


Figure 5: Gly-rich AFPs from snow flea (PDB ID: 2PNE). The structure lacks a hydrophobic core and instead one side is hydrophobic while the other side is hydrophilic. The protein surface is rendered via an APBS calculation.

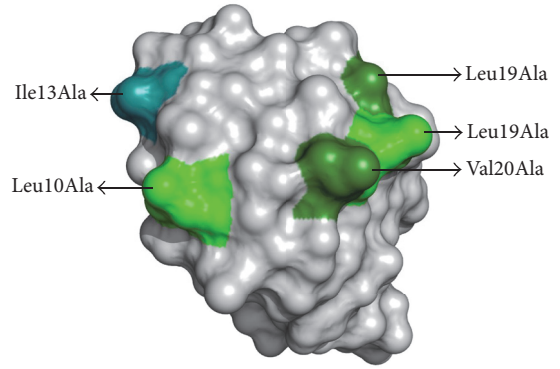


Figure 6: Crystal structure of type III AFP from fish (PDB ID: 1msi) with mutated hydrophobic residues represented in green color.

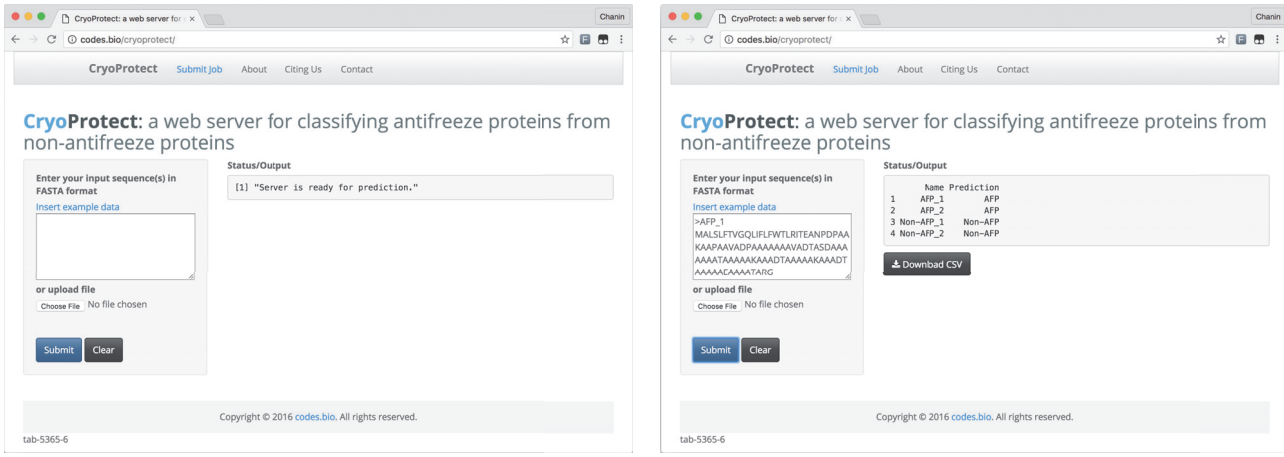
and TargetFreeze [20]. Cross-validation (e.g., 10-fold CV) provides insufficient conditions to determine which model has a higher predictive power. Thus, this study utilizes the external validation test to moderate such a problem. The reported prediction results over an external validation test of the existing predictors of AFPs shown in Table 4 are directly obtained from the work on TargetFreeze [20].

Based on the prediction results as shown in Table 5, CryoProtect achieved a greater prediction performance than

Table 5: Performance comparisons of CryoProtect with the existing methods on the external validation set(s).

Method	Ac (%)	Sn (%)	Sp (%)	MCC	YI
iAFP[a,d]	95.30	13.26	97.09	0.086	0.10
AFP-Pred[a,d]	77.34	91.16	77.04	0.23	0.68
AFP-PseAAC[a,d]	84.75	85.08	84.74	0.27	0.70
TargetFreeze[b,d]	91.30 ± 2.00	92.45 ± 1.39	91.27 ± 2.07	0.40 ± 0.04	0.84 ± 0.10
CryoProtect[c]	88.28 ± 1.00	87.27 ± 2.27	88.30 ± 1.07	0.31 ± 0.01	0.76 ± 0.01
Mean	87.39	73.84	87.69	0.29	0.61

[a]Results were obtained from 1 round of random split.
[b]Results were obtained from 3 rounds of random split.
[c]Results were obtained from 20 rounds of random split.
[d]Results reported from the work of TargetFreeze (He et al.).



(a) (b)

Figure 7: Screenshot of the CryoProtect web server before (a) and after (b) submission of the input sequence data.

iAFP [17] and AFP-Pred [12] by providing improvements of >4% and >20% on MCC and YI, respectively, while also achieving higher performance than AFP-PseAAC [13]. Thus, values for the five statistical parameters of CryoProtect were found to be superior to those of the three AFP predictors. However, TargetFreeze [20], which is considered as the best AFP predictor, uses a support vector machine (SVM) with several types of complementary protein features, namely, AAC, PseAAC, and PsePSSM as summarized in Table 1. It was observed that TargetFreeze obtained better prediction results than CryoProtect by approximately 3–5% where the former approach afforded Ac, Sn, and Sp of 91.30%, 92.45%, and 91.27%, respectively, while the latter approach afforded values of 88.28%, 87.27%, and 88.30%, respectively. However, TargetFreeze was constructed using SVM, which is regarded as a black-box approach as it is not easily interpretable. On the other hand, CryoProtect makes use of interpretable learning methods such as RF because it allows users to understand and rationalize the biological and chemical properties of AFPs. Therefore, the CryoProtect model is deemed to be a more suitable method for predicting and interpreting AFP owing to its interpretability and moderately good performance that is only a few percentage less than the best predictor.

3.5. CryoProtect Web Server. As a service to the life science community, the predictive QSAR model described herein was made publicly available as a prediction web server. A screenshot of the CryoProtect web server is shown in Figure 7. A step-by-step walkthrough of the procedures for using the CryoProtect web server is described below.

Step 1. Go to the CryoProtect web server at http://codes.bio/cryoprotect/.

Step 2. Enter the query sequence into the Input box or upload the sequence file by clicking on the *Choose file* button (i.e., found below the *Enter your input sequence(s) in FASTA format* heading). Finally, press on the *Submit* button to initiate the prediction process.

Step 3. Prediction results are automatically displayed in a grey box found below the *Status/Output* heading. Typically, it takes a few seconds for the server to process the task. Users can also download the prediction results as a CSV file by pressing on the *Download CSV button*.

Additionally, users could also run a local copy of CryoProtect on their own computer using a one-line code as follows in an R environment:

```
shiny::runGitHub('cryoprotect', 'chaninn')
```

However, prior to running the aforementioned code, it is recommended that users first install the prerequisite R

packages. This can be performed by using the following code:

```
install.packages(c('shiny', 'shinyjs',
'shinythemes', 'protr', 'seqinr', 'randomForest',
'markdown'))
```

4. Conclusion

The current study proposed a novel and interpretable RF-based CryoProtect method for prediction and analysis of AFPs from their sequences. Several machine learning approaches have been used in this study, random forest and decision tree. The performance of CryoProtect method was comparable to the SVM-based method and better than decision tree when applied in the independent set. Moreover, the propensity score analysis of informative physicochemical properties provided insight into the important features for AFPs activity. In summary, results revealed that AFPs preferred to be composed of a certain number of hydrophobic amino acids (e.g., Leu, Ile, and Val) at the end of the α-helix. Furthermore, it was also found that AFPs favor amino acids with hydroxyl and sulfhydryl side chains (e.g., Thr, Ser, and Cys). Moreover, Cys residues help to stabilize the structure of AFPs by forming disulfide bridges inside the β-helix. Finally, Thr was found to increase the activity of AFPs via the addition of hydrogen bonds on its surface area. As a service to the scientific community, the predictive model of CryoProtect was made publicly available as a prediction server to facilitate easy and rapid classification of query protein sequence as being either AFP or non-AFP.

Competing Interests

The authors declare no competing interests regarding the publication of this article.

Acknowledgments

The authors gratefully acknowledge support from the Ministry of Research, Technology and Higher Education of the Republic of Indonesia for the Ph.D. scholarship to Reny Pratiwi; the New Scholar Research Grant (no. MRG5980220) to Watshara Shoombuatong; the Goal-Oriented Research Grant (no. E09/2557) from Mahidol University to Chanin Nantasenamat; and the Swedish Research Links Program (no. C0610701) from the Swedish Research Council to Jarl E. S. Wikberg and Chanin Nantasenamat.

References

[1] R. W. R. Crevel, J. K. Fedyk, and M. J. Spurgeon, "Antifreeze proteins: characteristics, occurrence and human exposure," *Food and Chemical Toxicology*, vol. 40, no. 7, pp. 899–903, 2002.

[2] P. L. Davies, J. Baardsnes, M. J. Kuiper, and V. K. Walker, "Structure and function of antifreeze proteins," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 357, no. 1423, pp. 927–935, 2002.

[3] S. P. Graether, M. J. Kuiper, S. M. Gagné et al., "β-helix structure and ice-binding properties of a hyperactive antifreeze protein from an insect," *Nature*, vol. 406, no. 6793, pp. 325–328, 2000.

[4] J. Baardsnes and P. L. Davies, "Contribution of hydrophobic residues to ice binding by fish type III antifreeze protein," *Biochimica et Biophysica Acta—Proteins and Proteomics*, vol. 1601, no. 1, pp. 49–54, 2002.

[5] Y. Cheng, Z. Yang, H. Tan, R. Liu, G. Chen, and Z. Jia, "Analysis of ice-binding sites in fish type II antifreeze protein by quantum mechanics," *Biophysical Journal*, vol. 83, no. 4, pp. 2202–2210, 2002.

[6] C. B. Marshall, M. E. Daley, L. A. Graham, B. D. Sykes, and P. L. Davies, "Identification of the ice-binding face of antifreeze protein from *Tenebrio molitor*," *FEBS Letters*, vol. 529, no. 2-3, pp. 261–267, 2002.

[7] A. Wierzbicki, C. A. Knight, E. A. Salter, C. N. Henderson, and J. D. Madura, "Role of nonpolar amino acid functional groups in the surface orientation-dependent adsorption of natural and synthetic antifreeze peptides on ice," *Crystal Growth and Design*, vol. 8, no. 9, pp. 3420–3429, 2008.

[8] Z. Jia and P. L. Davies, "Antifreeze proteins: an unusual receptor-ligand interaction," *Trends in Biochemical Sciences*, vol. 27, no. 2, pp. 101–106, 2002.

[9] P. L. Davies and C. L. Hew, "Biochemistry of fish antifreeze proteins," *FASEB Journal*, vol. 4, no. 8, pp. 2460–2468, 1990.

[10] S. Venketesh and C. Dayananda, "Properties, potentials, and prospects of antifreeze proteins," *Critical Reviews in Biotechnology*, vol. 28, no. 1, pp. 57–82, 2008.

[11] Q. Z. Li, Y. Yeh, J. J. Liu, R. E. Feeney, and V. V. Krishnan, "A two-dimensional adsorption kinetic model for thermal hysteresis activity in antifreeze proteins," *Journal of Chemical Physics*, vol. 124, no. 20, Article ID 204702, 2006.

[12] K. K. Kandaswamy, K.-C. Chou, T. Martinetz et al., "AFP-Pred: a random forest approach for predicting antifreeze proteins from sequence-derived properties," *Journal of Theoretical Biology*, vol. 270, no. 1, pp. 56–62, 2011.

[13] S. Mondal and P. P. Pai, "Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction," *Journal of Theoretical Biology*, vol. 356, pp. 30–35, 2014.

[14] L. A. Graham, C. B. Marshall, F.-H. Lin, R. L. Campbell, and P. L. Davies, "Hyperactive antifreeze protein from fish contains multiple ice-binding sites," *Biochemistry*, vol. 47, no. 7, pp. 2051–2063, 2008.

[15] C. Nantasenamat, C. Isarankura-Na-Ayudhya, and V. Prachayasittikul, "Advances in computational methods to predict the biological activity of compounds," *Expert Opinion on Drug Discovery*, vol. 5, no. 7, pp. 633–654, 2010.

[16] C. Nantasenamat, C. Isarankura-Na-Ayudhya, T. Naenna, and V. Prachayasittikul, "A practical overview of quantitative structure-activity relationship," *EXCLI Journal*, vol. 8, pp. 74–88, 2009.

[17] C.-S. Yu and C.-H. Lu, "Identification of antifreeze proteins and their functional residues by support vector machine and genetic algorithms based on *n*-peptide compositions," *PLoS ONE*, vol. 6, no. 5, Article ID e20445, 2011.

[18] X. Zhao, Z. Ma, and M. Yin, "Using support vector machine and evolutionary profiles to predict antifreeze protein sequences," *International Journal of Molecular Sciences*, vol. 13, no. 2, pp. 2196–2207, 2012.

[19] R. Yang, C. Zhang, R. Gao, and L. Zhang, "An effective antifreeze protein predictor with ensemble classifiers and comprehensive sequence descriptors," *International Journal of Molecular Sciences*, vol. 16, no. 9, pp. 21191–21214, 2015.

[20] X. He, K. Han, J. Hu et al., "TargetFreeze: identifying antifreeze proteins via a combination of weights using sequence evolutionary information and pseudo amino acid composition," *Journal of Membrane Biology*, vol. 248, no. 6, pp. 1005–1014, 2015.

[21] S. Kawashima and M. Kanehisa, "AAindex: amino acid index database," *Nucleic Acids Research*, vol. 28, no. 1, article 374, 2000.

[22] E. L. L. Sonnhammer, S. R. Eddy, and R. Durbin, "Pfam: a comprehensive database of protein domain families based on seed alignments," *Proteins: Structure, Function and Genetics*, vol. 28, no. 3, pp. 405–420, 1997.

[23] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, 2006.

[24] W. Shoombuatong, P. Mekha, and J. Chaijaruwanich, "Sequence based human leukocyte antigen gene prediction using informative physicochemical properties," *International Journal of Data Mining and Bioinformatics*, vol. 13, no. 3, pp. 211–224, 2015.

[25] W. Shoombuatong, H.-L. Huang, J. Chaijaruwanich, P. Charoenkwan, H.-C. Lee, and S.-Y. Ho, "Predicting protein crystallization using a simple scoring card method," in *Proceedings of the 10th Annual IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB '13)—IEEE Symposium Series on Computational Intelligence (SSCI '13)*, April 2013.

[26] P. Charoenkwan, W. Shoombuatong, H.-C. Lee, J. Chaijaruwanich, H.-L. Huang, and S.-Y. Ho, "SCMCRYS: predicting protein crystallization using an ensemble scoring card method with estimating propensity scores of P-collocated amino acid pairs," *PLoS ONE*, vol. 8, no. 9, Article ID e72368, 2013.

[27] M. Ringnér, "What is principal component analysis?" *Nature Biotechnology*, vol. 26, no. 3, pp. 303–304, 2008.

[28] Z.-H. You, Y.-K. Lei, L. Zhu, J. Xia, and B. Wang, "Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis," *BMC Bioinformatics*, vol. 14, supplement 8, article S10, 2013.

[29] S. Lê, J. Josse, and F. Husson, "FactoMineR: an R package for multivariate analysis," *Journal of Statistical Software*, vol. 25, no. 1, pp. 1–18, 2008.

[30] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014.

[31] C. Nantasenamat, H. Li, P. Mandi et al., "Exploring the chemical space of aromatase inhibitors," *Molecular Diversity*, vol. 17, no. 4, pp. 661–677, 2013.

[32] W. Shoombuatong, V. Prachayasittikul, N. Anuwongcharoen et al., "Navigating the chemical space of dipeptidyl peptidase-4 inhibitors," *Drug Design, Development and Therapy*, vol. 9, pp. 4515–4549, 2015.

[33] N. Anuwongcharoen, W. Shoombuatong, T. Tantimongcolwat, V. Prachayasittikul, and C. Nantasenamat, "Exploring the chemical space of influenza neuraminidase inhibitors," *PeerJ*, vol. 2016, no. 4, Article ID e1958, 2016.

[34] D. Palomba, M. J. Martínez, I. Ponzoni, M. F. Díaz, G. E. Vazquez, and A. J. Soto, "QSPR models for predicting log Pliver values for volatile organic compounds combining statistical methods and domain knowledge," *Molecules*, vol. 17, no. 12, pp. 14937–14953, 2012.

[35] F. Hammann, H. Gutmann, U. Baumann, C. Helma, and J. Drewe, "Classification of cytochrome P450 activities using machine learning methods," *Molecular Pharmaceutics*, vol. 6, no. 6, pp. 1920–1926, 2009.

[36] J. R. Quinlan, *C4. 5:Programs for Machine Learning*, Elsevier, New York, NY, USA, 2014.

[37] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[38] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*, Wadsworth Statistics/Probability, Chapman and Hall/CRC, 1 edition, 1984.

[39] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[40] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Development Core Team, Vienna, Austria, 2016, http://www.r-project.org/.

[41] Y.-C. Liou, P. Thibault, V. K. Walker, P. L. Davies, and L. A. Graham, "A complex family of highly heterogeneous and internally repetitive hyperactive antifreeze proteins from the beetle *Tenebrio molitor*," *Biochemistry*, vol. 38, no. 35, pp. 11415–11424, 1999.

[42] Y.-C. Liou, A. Tocilj, P. L. Davies, and Z. Jia, "Mimicry of ice structure by surface hydroxyls and water of a β-helix antifreeze protein," *Nature*, vol. 406, no. 6793, pp. 322–324, 2000.

[43] C. B. Marshall, M. E. Daley, B. D. Sykes, and P. L. Davies, "Enhancing the activity of a β-helical antifreeze protein by the engineered addition of coils," *Biochemistry*, vol. 43, no. 37, pp. 11637–11646, 2004.

[44] P. L. Davies, "Ice-binding proteins: a remarkable diversity of structures for stopping and starting ice growth," *Trends in Biochemical Sciences*, vol. 39, no. 11, pp. 548–555, 2014.

[45] T. Hessa, H. Kim, K. Bihlmaier et al., "Recognition of transmembrane helices by the endoplasmic reticulum translocon," *Nature*, vol. 433, no. 7024, pp. 377–381, 2005.

[46] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *Journal of Molecular Biology*, vol. 157, no. 1, pp. 105–132, 1982.

[47] M. E. Daley, L. Spyracopoulos, Z. Jia, P. L. Davies, and B. D. Sykes, "Structure and dynamics of a β-helical antifreeze protein," *Biochemistry*, vol. 41, no. 17, pp. 5515–5525, 2002.

[48] A. J. Middleton, C. B. Marshall, F. Faucher et al., "Antifreeze protein from freeze-tolerant grass has a beta-roll fold with an irregularly structured ice-binding site," *Journal of Molecular Biology*, vol. 416, no. 5, pp. 713–724, 2012.

[49] B. L. Pentelute, Z. P. Gates, V. Tereshko et al., "X-ray structure of snow flea antifreeze protein determined by racemic crystallization of synthetic protein enantiomers," *Journal of the American Chemical Society*, vol. 130, no. 30, pp. 9695–9701, 2008.

[50] J. S. Richardson and D. C. Richardson, "Amino acid preferences for specific locations at the ends of α helices," *Science*, vol. 240, no. 4859, pp. 1648–1652, 1988.

[51] G. Chen and Z. Jia, "Ice-binding surface of fish type III antifreeze," *Biophysical Journal*, vol. 77, no. 3, pp. 1602–1608, 1999.

[52] C. P. Garnham, A. Natarajan, A. J. Middleton, M. J. Kuiper, I. Braslavsky, and P. L. Davies, "Compound ice-binding site of an antifreeze protein revealed by mutagenesis and fluorescent tagging," *Biochemistry*, vol. 49, no. 42, pp. 9063–9071, 2010.

[53] P. H. A. Sneath, "Relations between chemical structure and biological activity in peptides," *Journal of Theoretical Biology*, vol. 12, no. 2, pp. 157–195, 1966.

[54] J. G. Duman, "Antifreeze and ice nucleator proteins in terrestrial arthropods," *Annual Review of Physiology*, vol. 63, pp. 327–357, 2001.

[55] A. Cheng and K. M. Merz Jr., "Ice-binding mechanism of winter flounder antifreeze proteins," *Biophysical Journal*, vol. 73, no. 6, pp. 2851–2873, 1997.

[56] A. Wierzbicki, P. Dalal, T. E. Cheatham III, J. E. Knickelbein, A. D. J. Haymet, and J. D. Madura, "Antifreeze proteins at the ice/water interface: three calculated discriminating properties for orientation of type I proteins," *Biophysical Journal*, vol. 93, no. 5, pp. 1442–1451, 2007.

[57] P. Koehl and M. Levitt, "Structure-based conformational preferences of amino acids," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 22, pp. 12524–12529, 1999.

[58] N. Qian and T. J. Sejnowski, "Predicting the secondary structure of globular proteins using neural network models," *Journal of Molecular Biology*, vol. 202, no. 4, pp. 865–884, 1988.

[59] R. Yan, J. Song, W. Cai, and Z. Zhang, "A short review on protein secondary structure prediction methods," in *Pattern Recognition in Computational Molecular Biology: Techniques and Approaches*, M. Elloumi, C. S. Iliopoulos, J. T. L. Wang and, and A. Y. Zomaya, Eds., p. 99, Wiley, New York, NY, USA, 2015.

[60] P. Koehl and M. Levitt, "*De novo* protein design. I. In search of stability and specificity," *Journal of Molecular Biology*, vol. 293, no. 5, pp. 1161–1181, 1999.