Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Prediction and analysis of antifreeze proteins

Ryosuke Miyata, Yoshitaka Moriwaki, Tohru Terada, Kentaro Shimizu *

Department of Biotechnology, The University of Tokyo, 1-1-1 yayoi, Bunkyo-ku, Tokyo, 113-8657, Japan

HIGHLIGHTS

- Construct reliable datasets of antifreeze proteins (AFPs).
- Develop a bioinformatics tool to predict AFPs from their sequences by using random forest.
- Achieved accuracies of 0.961 and 0.947 for non-redundant sequences with less than 90% and 30% identities.
- Using the ability of random forest, dentified the sequence features that contributed to the prediction.

ARTICLE INFO

Keywords:
Antifreeze proteins
Prediction
Protein sequences
Amino acids
Random forest

ABSTRACT

Antifreeze proteins (AFPs) are proteins that protect cellular fluids and body fluids from freezing by inhibiting the nucleation and growth of ice crystals and preventing ice recrystallization, thereby contributing to the maintenance of life in living organisms. They exist in fish, insects, microorganisms, and fungi. However, the number of known AFPs is currently limited, and it is essential to construct a reliable dataset of AFPs and develop a bioinformatics tool to predict AFPs. In this work, we first collected AFPs sequences from UniProtKB considering the reliability of annotations and, based on these datasets, developed a prediction system using random forest. We achieved accuracies of 0.961 and 0.947 for non-redundant sequences with less than 90% and 30% identities and achieved the accuracy of 0.953 for representative sequences for each species. Using the ability of random forest, we identified the sequence features that contributed to the prediction. Some sequence features were common to AFPs from different species. These features include the Cys content, Ala-Ala content, Trp-Gly content, and the amino acids' distribution related to the disorder propensity. The computer program and the dataset developed in this work are available from the GitHub site: https://github.com/ryomiya/Prediction-and-analysis-of-antifreeze-proteins.

1. Introduction

Antifreeze proteins (AFPs) are proteins that protect cellular fluids and body fluids from freezing by inhibiting the nucleation and growth of ice crystals and preventing ice recrystallization, thereby contributing to the maintenance of life in living organisms. They exist in fish, insects, microorganisms, and fungi [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]. Because of their functions, AFPs are expected to be utilized in food processing, cryopreservation of food, tissues, and organs, medicine, and the development of anti-icing materials. Thus, it is imperative to research antifreeze proteins, which are expected to be applied to a vast range of fields.

Previously known AFPs are antifreeze glycoproteins (AFGPs), Type-1 AFP, Type-2 AFP, Type-3 AFP, and Type-4 AFP, which were discovered early in the history of AFP research. AFGPs are proteins with the repeating sequence Ala-Ala-Thr, which is found in polar fishes, and the Thr is modified with β-d-galactosyl-(1,3)-α-n-acetyl-d-galactosamine disaccharide, which is involved in ice crystal binding [11]. Some low-molecular-weight AFGPs have been found to have alanine residues replaced by proline and threonine residues replaced by arginine [12]. Type-1 AFP is an Ala-rich, helix-rich protein found in fish such as flatfish, flounder, and sculpin and is the most widely documented of all AFP types due to its historical background as the first protein structure to be determined [13]. Type-2 AFP is a 14–24 kDa Ca^{2+}-dependent, a cysteine-rich globular protein that has been found in fishes such as herring (*Clupea harengus*), cucumber fish (*Osmerus mordax)*, wagtail (*Hypomesus nipponensis), sculpin* (*Hemitripterus americanus*), and whitefish (*Brachyopsis rostratus*) [14, 15, 16]. Type-2 AFP is the largest globular AFP to date homologous to the Ca^{2+}-dependent C-type lectin-like domain. Type-3 AFP is a small globular protein with an average molecular weight of about 6.5 kDa with a one-turn alpha-helix and beta-sheet

[*] Corresponding author.
E-mail address: shimizu@bi.a.u-tokyo.ac.jp (K. Shimizu).

https://doi.org/10.1016/j.heliyon.2021.e07953
Received 19 May 2021; Received in revised form 28 July 2021; Accepted 3 September 2021

and can be found in Antarctic eelgrass (*Macrozoarces americanus*), Antarctic *eelpout*, and *wolffish* [17]. Type-4 AFP is a new AFP species found in the horned sculpin's plasma, *Myoxocephalus octodecemspinosus* [18]. The protein has been named LS-12 and registered under Uniprot ID:P80961. This protein is approximately 12 kDa and 128 amino acids in length. The N-terminal 20 residues comprise a signal peptide for extracellular presentation, and the amino acid sequence of the AFP active domain from 21-128 is Gln/Glu-rich (~17%).

In this way, four types have been described to date; however, many antifreeze proteins that are not classified in the above types have been discovered, and the above classifications are not sufficient to classify all antifreeze proteins that have been discovered. A great variety of AFPs have been found in insects, microorganisms, and fungi. Despite their functional similarities, AFPs show great diversity in their structural and sequence characteristics among species. Therefore, it is difficult to predict AFPs only by homology search.

Machine learning has been used to identify AFPs from amino acid sequences. Kandaswamy et al. [19] developed an AFP prediction system called AFP-Pred, which used random forest (RF) as a classifier. They extracted AFP sequences using the Pfam database [20] and collected their homologous sequences, eliminating redundant sequences with more than 40% identity by CD-HIT [21]. As a result, 481 sequences were obtained and were used as the positive dataset for prediction. Amino acid frequency, dipeptide frequency, secondary structure information, and physicochemical properties in the protein sequences were used as features. The AFP prediction tools AFP_PSSM [22], AFP_pseAAC [23], TargetFreeze [24] used support vector machine (SVM) as a classifier. AFP_PSSM used the position-specific scoring matrix (PSSM) as a sequence feature; AFP_pseAAC improved prediction performance using pseudo amino acid composition (pseAAC), and TargetFreeze further improved prediction performance by using pseudo PSSM in addition to pseAAC. Yang et al. [25] developed a tool called AFP-ensemble which used RF as a classifier and used many sequence features, including amino acid composition (AAC), dipeptide composition (DC), physicochemical properties, PSSM, disorder information, and functional domain. They compared performance between different features with RFs and SVM as classifiers. One unique contribution of their work is that it incorporates the features related to protein disorder. They considered that the disorder regions may relate to ice-binding because the disorder regions are always rich in binding sites and carry important roles in regulating protein functions [25]. Their work showed that the disorder's contribution is the smallest among their physicochemical properties, whereas our work suggests that the disorder is one of the important features as described later. Khan et al. developed another AFP prediction system called RAFP-Pred by using RF [26]. RAFP-Pred divides a protein sequence into two sub-sequences, calculates AAC and DC of each sub-sequence, and combines the results to obtain features. AFP-LSE, developed by Usman et al., uses an autoencoder and deep neural network with the feature “composition of *k*-spaced amino acid pairs.” The performance results when using AFP-LSE were as follows: balanced accuracy = 0.903; Youden's Index = 0.81; and Matthews correlation coefficient (MCC) = 0.52; these results indicate that this method performs better than the existing methods [27]. We discuss the performance of our method and that of AFP-LSE in the Results section. Usman et al. also developed another prediction method, AFP-SRC [28], which uses amino acid composition and dipeptide composition as features and also applied principal component analysis (PCA) to reduce the dimensions of features. discussed the effect of PCA comprehensively. Alim et al. developed PoGB-pred [29]; this prediction model was gradient boosting and the author applied PCA to reduce the dimensions of features, AAC, DC, and pseudo AAC. Alim et al. also discussed the effect of dimension reduction via a comparison with the performance of RAFP-Pred [26].

Recent studies also focus on the characterization of essential features for identifying AFPs. Pratiwi et al. [30] developed an AFP prediction tool called CryoProtect, which uses RF as a classifier and uses AAC and DC as sequence features. They also analyzed the sequences of AFPs and showed that Cys, Ser, Trp, Gly, Asn, and Thr were characteristic residues of AFPs while Leu, Val, Glu, Ile, and Met were characteristic of non-AFPs. These results reflected the amino acid propensities of AFP and non-AFPs, and the Gini index from the RF model for evaluating and ranking the feature importance of amino acids was complementary to the above results. Eslami et al. [31] developed an AFP prediction tool called afpCOOL, which uses SVM as a classifier and uses four types of descriptors: hydropathy, physicochemical properties, AAC, and evolutionary profile (400 descriptors). They showed physicochemical descriptors are the most informative features to discriminate between AFPs and non-AFPs. Sun et al. [32] developed a tool for identifying antifreeze proteins which used SVM as a classifier and evolutional information derived from PSSM as sequence features. They showed interesting findings that Cys, Trp, and Gly are conservative, and their replacements by Ala, Met, and Ala, respectively, are rare in AFPs. In this work, we applied AAC, DC, the composition of two amino acids with one arbitrary amino acid in between, and composition, transition, and distribution (CTD) based on various amino acid properties as described in the “Materials and Methods” section. Thus a wide range of features that are essential for the identification of AFPs can be found.

As for the dataset of AFP used for the machine learning benchmark, the above AFP prediction tools used the AFP-Pred dataset [19] as the basis. However, this dataset is limited in number, and it contains some sequences that we could not find any evidence or annotation related to antifreeze function, such as many C-type lectins and MADS-box domain-containing protein. UniProtKB contains a large number of sequences annotated as AFP and related keywords. The collection of AFPs from UniProtKB should be reconsidered. For this reason, RAFP-Pred [26] used 41 sequences from Protein Data Bank (PDB) and 369 sequences from UniProtKB, and the sequences in the AFP-Pred dataset. Eslami et al. also used 517 AFP sequences from UniProtKB and the same number of non-AFP sequences using PDB structures. They showed the results of performance analysis for the AFP-Pred dataset and the UniProtKB-based dataset. In this work, we constructed datasets of AFPs considering UniProtKB's reliability and removed the redundancy based on sequence identities and organic species. The current annotated sequences of UniProtKB are biased and we tried to identify common features that are characteristic to AFPs regardless of sequence families and species.

The current study does not predict the ice-binding residues of AFP directly, as in the study of Yu et al. [33] who used the PDB structures as a positive dataset for the purpose of predicting ice-binding sites. Although the dataset was limited and the accuracy of AFP identification was evaluated using this limited dataset, the identification of ice-binding sites was unique. In the present study, the amino acid motifs that ranked high in terms of feature score (see the Results) can be regarded as candidates for functional sites (not limited to ice-binding sites) of AFP.

Figure 1 illustrates the outline of this work. We first constructed datasets of AFPs considering reliability, developed a binary classification prediction system based on these datasets, and discussed the features of AFPs that contributed to the prediction.

2. Materials and methods

2.1. Datasets

We first searched UniProtKB for AFP sequences using “antifreeze” and the synonyms of “antifreeze proteins” registered in NCBI MeSH as keywords. InterPro and PRINTS, linked to UniProtKB, were also specified as search targets. The total number of AFP sequences collected was 4,194. The details of the collection method (search keywords) are described in Table S1. We call the dataset of these sequences the primary dataset.

Using the primary dataset, we constructed the non-redundant datasets: Dataset 30 and Dataset 90 created by clustering the primary dataset with 30% and 90% sequence identity cutoffs, respectively, using PSI-CD-HIT (version 4.5.3) [21]. In Dataset 30, sequences from a Pfam family were grouped into a single and several clusters for some families. For

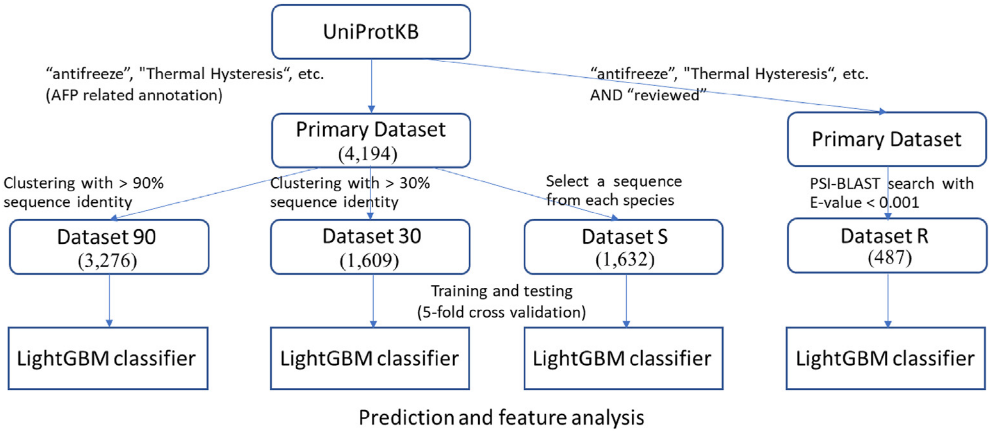


Figure 1. Outline of the present research. The numbers in parentheses represent the number of sequences in each dataset.

example, 279 sequences with the SAF domain (PF08666) were divided into six clusters; 69 sequences with the AFP domain (PF02420) were grouped into one cluster; 32 sequences with the CfAFP domain (PF05264) were divided into four clusters; 244 sequences with NeuB domain (PF03102) were divided into six clusters. Dataset 30 and Dataset 90 consists of 1,609 and 3,276 sequences, respectively.

The sequences in the primary dataset were derived from 1,632 species. Figure 2 shows the species with the ten largest number of sequences along with their dataset. There is a certain amount of bias by species. To eliminate the bias's possible influence, we also constructed another dataset, Dataset S, by selecting a representative sequence from each species. Dataset S consists of 1,632 sequences.

Among the 4,194 sequences in the primary dataset, there were 57 sequences with "reviewed" annotations of UniProtKB, which means that the sequence is manually annotated and considered to be highly reliable. Other sequences are "unreviewed," which contains computationally generated annotations and protein sequences from large-scale functional analysis experiments. We performed the PSI-BLAST search (version 2.9.0+) with E-value $<$ 0.001 for the 57 sequence. The number of hit sequences was 29,255. Out of 4,137 "unreviewed" sequences in the primary dataset, 430 sequences were included in the hit sequences, and they were added to the "reviewed" 57 sequences to construct Dataset R. The number of sequences in the Dataset R is 487.

Besides these datasets, we also prepared the dataset used in AFP-Pred. This dataset was collected based on the Pfam database; 221 sequences were collected from the Pfam database, and homologous sequences were collected with a strict threshold (E-value 0.001) by PSI-BLAST search. The final dataset contained only protein sequences with less than 40% identity after redundancy elimination using CD-HIT. We call this dataset Dataset AP. Dataset AP consists of 481 sequences.

Figure 3 (a) shows the comparison of Dataset AP and the primary dataset. As shown in Figure 3 (a), 121 sequences are included in common. To compare the sequences between Dataset AP and the primary dataset at the same redundancy level, we created Dataset 40 by removing redundancy in the primary dataset with a 40% identity cutoff using PSI-CD-HIT. The number of sequences in Dataset 40 is 1,782, and the number of sequences common with Dataset AP is 47, as shown in Figure 3 (b). For the other redundancy levels, Dataset 90 and Dataset 30 have 109 and 38 sequences common to Dataset AP, respectively. Figure 3 (c) compares Dataset AP and Dataset R. The number of common sequences is 28. These figures show that the Dataset AP includes a tiny part of AFP-annotated sequences in UniProtKB.

The above datasets, Dataset 90, Dataset 30, Dataset S, and Dataset R, were used as positive datasets of machine learning. We constructed a non-AFP dataset (negative dataset) by randomly selecting the sequences tagged as "reviewed" from UniProtKB, excluding those hit by the search

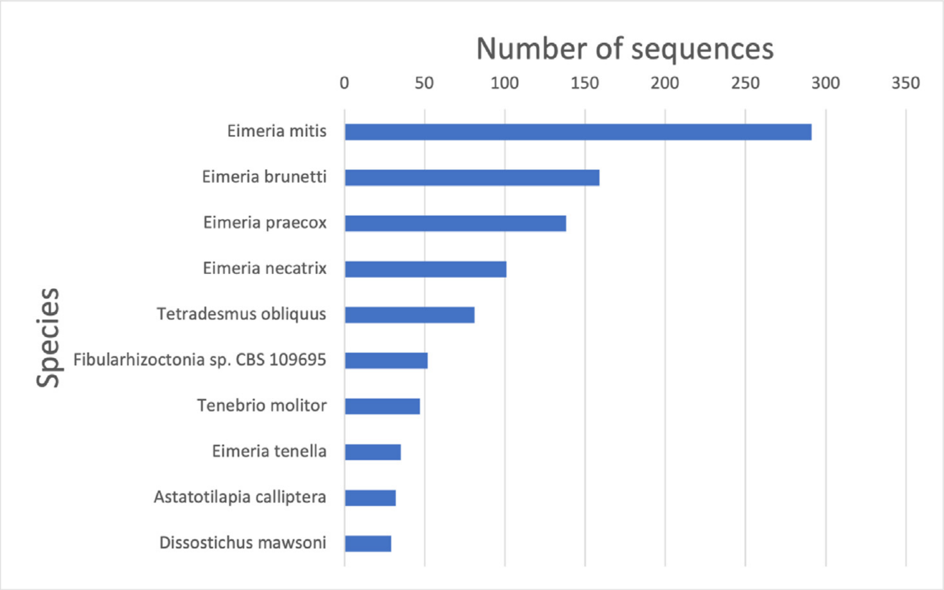


Figure 2. Top 10 species of the primary dataset.

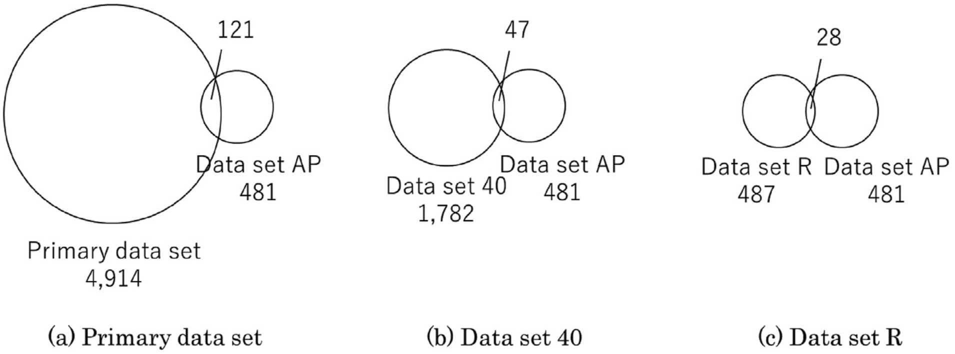


Figure 3. Comparison between Dataset AP and other datasets.

under the keywords "antifreeze" and "afp." The resulting number of sequences in the negative dataset was 560,147; however, for training and testing, the same number of sequences was selected from the negative dataset as were selected from the positive dataset. Therefore, all training was performed on balanced positive and negative datasets. Table 1 summarizes the number of sequences of each dataset.

2.2. Features

In this work, we used the following features for machine learning:

1. Amino acid (AA) composition (AAC).
2. Dipeptide composition (DC).
3. Composition of two amino acids with one arbitrary amino acid in between (AA-x-AA composition).
4. CTD.

The AAC can be expressed as a 20-dimensional vector. The DC is a composition of two consecutive amino acids, expressed as a vector of 20 × 20 = 400 dimensions. The AA-x-AA composition is a composition of the pattern AA-x-AA where x is an arbitrary amino acid and can also de expressed as a vector of 20 × 20 = 400 dimensions. This feature was used because some AFPs have this sequence pattern, especially in the ice-binding regions [34, 35]. CTD is a feature that indicates the distribution of amino acid patterns along the primary sequence of a protein, based on physicochemical and structural properties [36, 37]. C (Composition) indicates the composition of amino acids of a particular property; T (Transition) indicates the percent frequency with which amino acids with a different property follow amino acids with a particular property; *D* (Distribution) indicates the distribution of the initial, 25%, 50%, 75%, and 100% positions of the amino acids with a particular property.

Table 1. Number of sequences in the datasets.

Datasets		Number of sequences
Primary dataset	Sequences annotated with AFP-related keywords	4,194
Dataset 90	Non-redundant sequences <90% sequence identity	3,276
Dataset 40	Non-redundant sequences <40% sequence identity	1,782
Dataset 30	Non-redundant sequences <30% sequence identity	1,609
Dataset S	Dataset of the representative sequence for each species	1,632
Dataset R	Reviewed sequences plus their 90% similar sequences	487
Dataset AP	Dataset used in AFP-Pred	481
Negative dataset+	UniProtKB sequences excluding AFP-related keywords. For training, the same number of negative sequences was selected from this dataset as were selected from the positive datasets.	560,147

Dataset 40 is only used for comparison with Dataset AP (same redundancy level).

In this work, eight properties: hydrophobicity, normalized van der Waals volume, polarity, charge, secondary structure, solvent accessibility, polarizability, and disorder propensity [38], were considered as shown in Table S2. Each amino acid was classified into three classes Group 1, 2, and 3 for each property. For each property, Composition has a three-dimensional feature, Transition has a three-dimensional feature, and Distribution has a 15-dimensional feature (three groups times five positions). Therefore, the dimension of the CTD features of this work is 8 × (3 + 3 + 15) = 168. As an example, consider the sequence "MAGGDLVYAGSIAEHRKL." When we consider the polarizability of this sequence, it can be encoded as "311112231112123332" according to the grouping described in Table S2. Figure S1 shows the detailed calculation of CTD for this example. The compositions of group 1, group 2, and group 3 are 8/18, 5/18, and 5/18, respectively. The compositions of transition 1/2, transition 1/3, and transition 2/3 are 4/17, 2/17, and 3/17, respectively. Transition a/b represents transition a to b or transition b to a. The distributions of the initial (first), 25% (second), 50% (fourth), 75% (sixth), and 100% (eight) of group 1 were calculated as 2/18, 3/18, 5/18, 10/18, and 13/18, respectively.

2.3. Classification method

We used LightGBM [39] as a machine learning framework. LightGBM is a type of Gradient Boosting Decision Tree (GBDT) and uses an ensemble learning method that combines several weak learners (decision trees) to build a single strong learner. LightGBM is particularly unique among GBDTs in that it employs the Leaf-wise method, Gradient-based One-Side Sampling (GOSS), and Exclusive Feature Bundling (EFB). Figure S2 shows the flowchart of LigthtGBM.

The Leaf-wise method is used in the decision tree construction process of the GBDT algorithm [40]. In LightGBM, Leaf-wise is used because it has an advantage in short training time. GOSS is a learning method that excludes data with small gradients and samples only data with large gradients [41]. GOSS's advantage is that the computational cost can be significantly reduced because the instance split point is determined using the estimated value of variance gains. On the other hand, EFB is a method to reduce the number of features by bundling mutually exclusive features. This is based on an idea that the computational cost of searching for branch points is so heavy that it is not necessary to search for the entire features. Especially in large datasets, there are many sparse features, and it is not uncommon for the features of non-zero elements to have no overlap at all, such as one-hot encoding features.

2.4. Evaluation method

We performed five-fold cross-validation on each dataset. Five-fold cross-validation divides the dataset into five blocks and treats one data

Table 2. Prediction performance.

	Accuracy	Precision	Recall	MCC	AUC	Youden's index
Dataset 90	0.961	0.978	0.944	0.923	0.994	0.919
Dataset 30	0.947	0.979	0.914	0.896	0.989	0.882
Dataset S	0.953	0.973	0.934	0.908	0.994	0.990
Dataset R	0.986	0.992	0.979	0.972	1	0.945

block as the test data and the remaining data blocks as the training data. By repeating this process five times, each divided data block is treated as the test data.

2.5. Performance measure

The precision, specificity, recall, accuracy, and MCC were used to evaluate the prediction performance in this work. These are expressed as follows:

$$Precision = \frac{TP}{TP + FP}$$

$$Specifity = \frac{TN}{TN + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TP, FP, TN, and FN are: true positive, false positive, true negative, and false negative, respectively. Precision is the percentage of true positive sequences against the sequences that are predicted to be positive, specificity is the percentage of true negative sequences against the truly negative sequences, recall is the percentage of true positive sequences against sequences that are truly positive, and accuracy is the percentage of the correctly predicted sequences against the total sequences. MCC takes account of the accuracy of prediction for both positive and negative sequences.

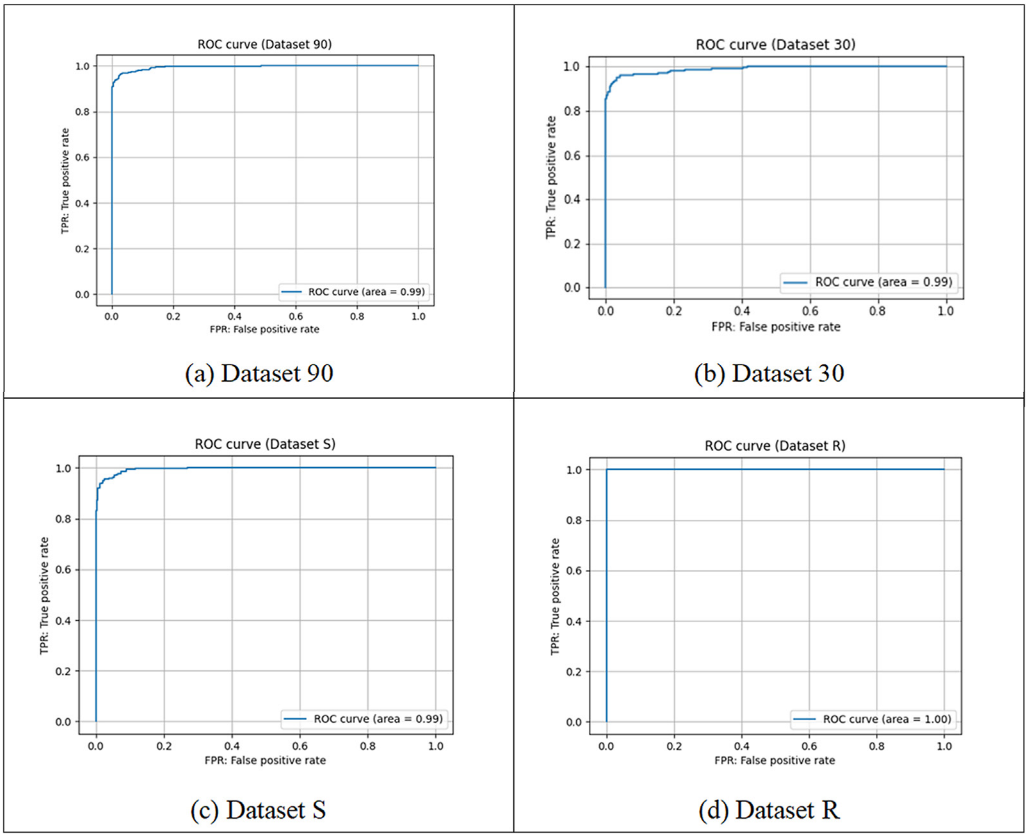
The Receiver Operating Characteristic curve (ROC curve) was also used to visualize the prediction performance. The ROC curve is a plot of the false positive rate ($= FP/(TN + FP)$) against the true positive rate (= reproducibility; recall).

3. Results and discussions

3.1. Prediction performance

The prediction performances for the four datasets are shown in Table 2. The performance was evaluated based on the accuracy (ACC), specificity (SPE), recall (REC), and MCC using five-fold cross-validation. The average value of each performance measure in the five-fold cross-validation is shown in Table 2. LightGBM created a classification predictor for each dataset. The LightGBM Tuner was used to find optimal hyperparameter values of the model. Figure 4 (a)-(d) shows the ROC curves for Dataset 90, Dataset 30, Dataset S, and Dataset R, respectively.

Dataset 90 has a larger number of sequences and facilitates learning compared with Dataset 30. Many sequences of specific species (*e.g., Eimeria*) are included in the primary dataset (Dataset 90 and Dataset 30), which may influence the precision of AFP of the specific species to some extent. Accuracy, recall, and MCC of Dataset S are higher than Dataset 30. Although Dataset S only contains one sequence for each species,



(a) Dataset 90

(b) Dataset 30

(c) Dataset S

(d) Dataset R

Figure 4. ROC curves.

Table 3. Prediction performances of independent tests.

	Accuracy	Precision	Recall	MCC	AUC	Youden's index
Dataset S	0.946	0.958	0.930	0.893	0.986	0.900
Dataset R	0.990	0.981	1	0.980	0.999	1

sequence redundancy was not removed. If some common sequence properties exist in AFPs among species, the learning effect seems higher than Dataset 30. Since Dataset R consists of 57 "reviewed" sequences and their close homologs, there may be a possibility that learning and testing were performed within a limited variety of sequences.

We performed independent tests on Dataset S and Dataset R. In both datasets, 64% of the whole dataset was treated as the training set, 16% as the validation set, and 20% as the test set. The prediction performances are shown in Table 3. Good performance was also obtained via an independent test.

We also performed three cross-data tests by training the model on Dataset 90, Dataset 30, and Dataset S (as well as each of the negative datasets) and by testing it on Dataset R. Table 4 shows the performance of these cross tests.

Results from cross-testing Dataset 90 and Dataset R showed high performance. In this case, the features of AFPs seemed to be trained and used for predicting AFPs. However, the performance from cross-testing Dataset 30 and Dataset R was relatively low. This suggests that there are some features of AFPs that failed to be trained by removing the sequence redundancy. Cross-testing Dataset S and Dataset R also resulted in a high performance, although the number of sequences in Dataset 30 and Dataset S are almost the same. Since Dataset S consists of representative sequences for each species, the features common to the AFPs of each species were effectively trained. In other words, the features with high scores in Dataset S are considered to be included in features common to the AFPs of each species.

For comparison with other methods, we evaluated the performance of a recent method, AFP-LSE [27], relative to our method. Independence tests on Dataset S and Dataset R were performed under the same conditions: 64% of the whole dataset was treated as the training set, 16% as the validation set, and 20% as the test set. The prediction performances are shown in Table 5. Comparing the results from the two methods, our method outperformed AFP-LSE for both Dataset S and Dataset R (Table 3). Since AFP-LSE was based on the AFP-Pred dataset [19], it could not train the features of our datasets that were collected from a broader set of AFP sequences.

Table 4. Prediction performance of cross testing.

	Accuracy	Precision	Recall	MCC	AUC	Youden's index
Training on Dataset 90 and Testing on Dataset R	0.990	0.990	0.990	0.979	1	0.986
Training on Dataset 30 and Testing on Dataset R	0.729	0.991	0.462	0.542	0.912	0.688
Training on Dataset S and Testing on Dataset R	0.982	0.990	0.973	0.963	0.999	0.971

Table 5. Performance comparison of our method with AFP-LSE.

	Accuracy	Precision	Recall	MCC	AUC	Youden's index
Data set S	0.896	0.897	0.886	0.791	0.950	0.795
Data set R	0.841	0.929	0.760	0.698	0.950	0.812

Table 6. Top 10 features contributing to classification prediction.

	Feature	Score
a) Dataset 90		
1	Polarity Composition; Group 2	1.57×10^{-1}
2	Cys	7.81×10^{-2}
3	Ala-Ala	6.69×10^{-2}
4	Ala-x-Ala	4.50×10^{-2}
5	Disorder Propensity Distribution; Group 1 (50%)	3.27×10^{-2}
6	Met-Gly	3.24×10^{-2}
7	Trp-x-Asp	2.57×10^{-2}
8	Normalized VDWV Composition; Group 1	2.42×10^{-2}
9	Trp-Gly	1.89×10^{-2}
10	Hydrophobicity Composition; Group 2	1.47×10^{-2}
b) Dataset 30		
1	Ala-x-Ala	2.73×10^{-1}
2	Ala-Ala	2.65×10^{-1}
3	Cys	4.33×10^{-2}
4	Polarity Composition; Group 2	3.20×10^{-2}
5	Ala	2.57×10^{-2}
6	Polarity Composition; Group 1	1.56×10^{-2}
7	Normalized VDWV Composition; Group 1	1.41×10^{-2}
8	Leu	1.13×10^{-2}
9	Polarizability Composition; Group 1	7.72×10^{-3}
10	Hydrophobicity Composition; Group 2	6.68×10^{-3}
c) Dataset S		
1	Disorder Propensity Distribution; Group 2 (50%)	8.68×10^{-2}
2	Disorder Propensity Distribution; Group 1 (50%)	5.24×10^{-2}
3	Ala	4.39×10^{-2}
4	Polarity Composition; Group 2	3.86×10^{-2}
5	Cys	3.73×10^{-2}
6	Trp-Gly	2.93×10^{-2}
7	Normalized VDWV Composition; Group 1	2.90×10^{-2}
8	Hydrophobicity Composition; Group 2	1.97×10^{-2}
9	Ala-Ala	1.96×10^{-2}
10	Arg	1.92×10^{-2}
d) Dataset R		
1	Disorder Propensity Distribution; Group 3 (50%)	6.18×10^{-2}
2	Polarity Composition; Group 2	5.59×10^{-2}
3	Trp-Gly	4.89×10^{-2}
4	Normalized VDWV Composition; Group 1	4.57×10^{-2}
5	Cys	3.82×10^{-2}
6	Arg	3.19×10^{-2}
7	Ala-Ala	2.24×10^{-2}
8	Ala	2.15×10^{-2}
9	Disorder Propensity Distribution; Group 1 (50%)	1.63×10^{-2}
10	Leu-Leu	1.59×10^{-2}

Score refers to a normalized decrease (Gini impurity) in the objective function from the decision branch that uses the feature. Groups 1–3 are classifications of AA properties shown in Table S1.

3.2. Contributing features to the prediction

From the features that contributed to the classification prediction, it is possible to investigate the antifreeze function's essential properties. Table 6 (a)-(d) shows the top ten features with a considerable contribution to classification. In particular, the result of Dataset 90 (a) shows the overall tendency of AFPs, the result of Dataset 30 (b) shows some properties common to domains and families of AFPs, the result of Dataset S (c) shows the properties common across different species in biological taxonomy, and the result of Dataset R (d) shows the tendency of the current, reliable AFP sequences.

Many of the features that contributed to prediction in Dataset 30 are similar for prediction in Dataset 90; this shows that these features are shared among the various families and AFP domains. Focusing on the AA characteristics, Cys content, Ala-Ala content, Ala-x-Ala content, Met-Gly content, Trp-x-Asp content, and Trp-Gly content of AFP are much larger than those of other proteins. The Cys content is higher in Type-2 AFP, and the Ala-Ala and Ala-x-Ala contents are higher in Type-1 AFP. In particular, Ala-Ala and Ala-x-Ala are the most abundant peptides on the ice-binding surface of Type-1 AFP [42]. Met-Gly, Trp-Xx-Asp, and Trp-Gly are also considered essential peptides on the ice-binding surface of AFP physicochemical properties; AFP contains more AA with moderate polarity, small van der Waals volume (VDWV), and moderate hydrophobicity than other proteins. As for the VDWV, it is thought that the AA with a large VDWV can sterically inhibit other AAs from binding to ice through van der Waals interactions and hydrogen bonds.

As shown in Table 2 (c), the contribution of high (Group 1) and moderate (Group 2) disorder propensities is high compared with the above datasets. This shows disorder seems to have some relationship with antifreeze function regardless of species, although such a relationship has not yet been reported.

The features such as the Cys content, Ala-Ala content, Trp-Gly, polarity composition, small van der Waals, moderate hydrophobicity also appear as common features. On the other hand, Met-Gly content and Trp-x-Asp content, which contributed to Dataset 90 but did not contribute significantly to the prediction for the other datasets, might be characteristic of the specific group of sequences.

4. Conclusion

In this work, we created a new dataset of AFPs, made binary classification predictions for AFPs and non-AFPs, and discussed the properties of AFPs based on the features that contributed to the classification predictions.

As described in the “Datasets” section, the reliability of the sequences in the dataset used in AFP-Pred (Dataset AP) is not high, and therefore, in this work, we carefully collected the AFP sequences. The AFP sequences in the primary dataset (selected from UniProtKB) contain many sequences from specific species, as shown in Table 1. Dataset S was constructed by eliminating the influence of any bias of the species.

The classification predictors developed in this work showed high accuracy, precision, and recall. They can be used as tools for predicting whether or not a given protein is an AFP in proteome-wide analyses. As shown in the results for non-redundant datasets, AFP prediction by sequence homology (similarity) seems to be difficult. However, machine learning was a useful technique for learning the common features of AFPs.

Regarding the features that contributed to the prediction of AFP and non-AFP classification, several features contributed commonly. A typical example is short AA sequence motifs, including Ala, Cys, Trp, and Gly. Our results also suggest that the disorder propensity was related to the function of AFPs. For further research, it is necessary to confirm the antifreeze property by experimental methods.

A small number of reliable annotated AFP sequences (Dataset R) exist in the current database. In the present study, we obtained a high AUC value of 0.99 in the independent test of Dataset R. However, a lower AUC value of 0.767 was obtained by training with Dataset R and performing a cross test with Dataset S. This may be due to the small number of sequences in Dataset R and the characteristics of AFP not being fully learned, despite the potential inclusion of sequences other than AFP in Dataset S. To develop a predictor that is more reliable, it will be desirable to achieve reliable annotations of AFP sequences.

As future research to extend the present study, it will be necessary to confirm thermal hysteresis activity via experimental methods. For example, by mutating amino acid residues with high disorder tendencies using those with low disorder tendencies, and then by comparing the thermal hysteresis activity, the relationship between the disorder tendencies of amino acids and the function of AFP could be further elucidated.

Declarations

Author contribution statement

Ryosuke Miyata: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Yoshitaka Moriwaki: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

Tohru Terada: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

Kentaro Shimizu: Conceived and designed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This work was supported by Platform Project for Supporting Drug Discovery and Life Science Research (Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS)) from AMED under Grant Number JP21am0101107, and The Uehara Memorial Foundation under Grant Number 202080071.

Data availability statement

Data associated with this study has been deposited at https://github.com/ryomiya/Prediction-and-analysis-of-antifreeze-proteins.

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2021.e07953.

References

[1] J.A. Gilbert, P.L. Davies, J. Laybourn-Parry, A hyperactive, Ca2+-dependent antifreeze protein in an Antarctic bacterium, FEMS Microbiol. Lett. 245 (1) (2005) 67–72.

[2] H. Do, et al., Structure-based characterization and antifreeze properties of a hyperactive ice-binding protein from the Antarctic bacterium Flavobacterium frigoris PS1, Acta Crystallogr. Sect. D-Struct. Biol. 70 (2014) 1061–1073.

[3] J.A. Raymond, B.C. Christner, S.C. Schuster, A bacterial ice-binding protein from the Vostok ice core, Extremophiles 12 (5) (2008) 713–717.

[4] P. Singh, et al., Antifreeze protein activity in Arctic cryoconite bacteria, FEMS Microbiol. Lett. 351 (1) (2014) 14–22.

[5] N.H.F. Hashim, et al., Characterization of Afp1, an antifreeze protein from the psychrophilic yeast Glaciozyma Antarctica PI12, Extremophiles 17 (1) (2013) 63–73.

[6] R. Kiko, Acquisition of freeze protection in a sea-ice crustacean through horizontal gene transfer? Polar Biol. 33 (4) (2010) 543–556.

[7] W. Jung, et al., Isolation and characterization of antifreeze proteins from the antarctic marine microalga pyramimonas gelidicola, Mar. Biotechnol. 16 (5) (2014) 502–512.

[8] I.G. Gwak, et al., Antifreeze protein in antarctic marine diatom, chaetoceros neogracile, Mar. Biotechnol. 12 (6) (2010) 630–639.

[9] Z.C. Jia, P.L. Davies, Antifreeze proteins: an unusual receptor-ligand interaction, Trends Biochem. Sci. 27 (2) (2002) 101–106.

[10] L. Ramya, Physicochemical properties of insect and plant antifreeze proteins: a computational study, Curr. Sci. 112 (7) (2017) 1512–1520.

[11] D. Slaughter, et al., Antifreeze proteins from the sea raven, Hemitripterus-americanus - further evidence for diversity among fish polypeptide antifreezes, J. Biol. Chem. 256 (4) (1981) 2022–2026.

[12] H.R. Morris, et al., Antifreeze glycoproteins from the blood of an antarctic fish. The structure of the proline-containing glycopeptides, J. Biol. Chem. 253 (14) (1978) 5155–5162.

[13] F. Sicheri, D.S.C. Yang, Ice-binding structure and mechanism of an antifreeze protein from winter flounder, Nature 375 (6530) (1995) 427–431.

[14] K.V. Ewart, et al., Ca2+-dependent antifreeze proteins - modulation of conformation and activity by divalent metal ions, J. Biol. Chem. 271 (28) (1996) 16627–16632.

[15] W. Gronwald, et al., The solution structure of type II antifreeze protein reveals a new member of the lectin family, Biochemistry 37 (14) (1998) 4712–4721.

[16] Y. Nishimiya, et al., Crystal structure and mutational analysis of Ca2+-independent type II antifreeze protein from longsnout poacher, Brachyopsis rostratus, J. Mol. Biol. 382 (3) (2008) 734–746.

[17] C.L. Hew, et al., Multiple genes provide the basis for antifreeze protein diversity and dosage in the ocean pout, Macrozoarces americanus, J. Biol. Chem. 263 (24) (1988) 12049–12055.

[18] G.J. Deng, D.W. Andrews, R.A. Laursen, Amino acid sequence of a new type of antifreeze protein: from the longhorn sculpin Myoxocephalus octodecimspinosis, FEBS Lett. 402 (1) (1997) 17–20.

[19] K.K. Kandaswamy, et al., AFP-Pred: a random forest approach for predicting antifreeze proteins from sequence-derived properties, J. Theor. Biol. 270 (1) (2011) 56–62.

[20] e.a. Sara El-Gebali, The Pfam protein damilies database in 2019, Nucleic Acids Res. 47 (D1) (2019) D427–D432.

[21] L. Fu, , et al.CD-HIT, Accelerated for clustering the next generation sequencing data, Bioinformatics 28 (2012) 3150–3152.

[22] X. Zhao, Z. Ma, M. Yin, Using support vector machine and evolutionary profiles to predict antifreeze protein sequences, Int. J. Mol. Sci. 13 (2) (2012) 2196–2207.

[23] S. Mondal, P.P. Pai, Chou′s pseudo amino acid composition improves sequence-based antifreeze protein prediction, J. Theor. Biol. 356 (2014) 30–35.

[24] X. He, et al., TargetFreeze: identifying antifreeze proteins via a combination of weights using sequence evolutionary information and pseudo amino acid composition, J. Membr. Biol. 248 (2015) 1005–1014.

[25] R. Yang, et al., An effective antifreeze protein predictor with ensemble classifiers and comprehensive sequence descriptors, Int. J. Mol. Sci. 16 (9) (2015) 21191–21214.

[26] S. Khan, et al., RAFP-pred: robust prediction of antifreeze proteins using localized analysis of n-peptide compositions, IEEE ACM Trans. Comput. Biol. Bioinf. 15 (1) (2018) 244–250.

[27] M. Usman, S. Khan, J.A. Lee, AFP-LSE: antifreeze proteins prediction using latent space encoding of composition of k-spaced amino acid pairs, Sci. Rep. 10 (1) (2020) 7197.

[28] M. Usman, S. Khan, S. Park, A. Wahab, AFP-SRC: Identification of Antifreeze ProteinsUsing Sparse Representation Classifier, arXiv, 2021.

[29] A. Alim, A. Rafay, I. Naseem, PoGB-pred: prediction of antifreeze proteins sequences using amino acid composition with feature selection followed by a sequential-based ensemble approach, Curr. Bioinf. (2021) 446–456.

[30] R. Pratiwi, et al., CryoProtect: a web server for classifying antifreeze proteins from nonantifreeze proteins, J. Chem. 2017 (2017) 9861752.

[31] M. Eslami, et al., afpCOOL: a tool for antifreeze protein prediction, Heliyon 4 (7) (2018).

[32] S. Sun, et al., Identifying antifreeze proteins based on key evolutionary information, Front. Bioeng. Biotechnol. 8 (2020) 244.

[33] C.S. Yu, 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 6 (5) (2011), e20445.

[34] M. Mangiagalli, et al., Cryo-protective effect of an ice-binding protein derived from Antarctic bacteria, FEBS J. 284 (1) (2017) 163–177.

[35] T.D.R. Vance, et al., Ice-binding proteins and the 'domain of unknown function' 3494 family, FEBS J. 286 (5) (2019) 855–873.

[36] G. Govindan, A.S. Nair, Composition, transition and distribution (CTD) - a dynamic feature for predictions based on hierarchical structure of cellular sorting, in: Annual IEEE India Conference - Engineering Sustainable Solutions, BITS Pilani, Hyderabad Campus, Hyderabad, INDIA, 2011.

[37] I. Dubchak, et al., Prediction of protein folding class using global description of amino acid sequence, Proc. Natl. Acad. Sci. U. S. A 92 (19) (1995) 8700–8704.

[38] V. Vacic, et al., Composition Profiler: a tool for discovery and visualization of amino acid composition differences, BMC Bioinf. 8 (2007).

[39] G.L. Ke, et al., LightGBM: a highly efficient gradient boosting decision tree, Adv. Neural Inf. Process. Syst. 30 (2017) 30 (Nips 2017).

[40] T.Q. Chen, C. Guestrin, M. Assoc comp, XGBoost: a scalable tree boosting system, in: Kdd'16: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[41] R. Wang, et al., Power system transient stability assessment based on bayesian optimized LightGBM, in: 2019 IEEE 3rd Conference on Energy Internet and Energy System Integration (EI2), IEEE, 2019.

[42] J. Baardsnes, et al., New ice-binding face for type I antifreeze protein, FEBS Lett. 463 (1-2) (1999) 87–91.