J Membrane Biol (2015) 248:1005–1014
DOI 10.1007/s00232-015-9811-z

TargetFreeze: Identifying Antifreeze Proteins via a Combination of Weights using Sequence Evolutionary Information and Pseudo Amino Acid Composition

Xue He[1] · Ke Han[1] · Jun Hu[1] · Hui Yan[1] · Jing-Yu Yang[1] · Hong-Bin Shen[2] · Dong-Jun Yu[1,3]

Received: 26 March 2015 / Accepted: 19 May 2015 / Published online: 10 June 2015

Abstract Antifreeze proteins (AFPs) are indispensable for living organisms to survive in an extremely cold environment and have a variety of potential biotechnological applications. The accurate prediction of antifreeze proteins has become an important issue and is urgently needed. Although considerable progress has been made, AFP prediction is still a challenging problem due to the diversity of species. In this study, we proposed a new sequence-based AFP predictor, called TargetFreeze. TargetFreeze utilizes an enhanced feature representation method that weightedly combines multiple protein features and takes the powerful support vector machine as the prediction engine. Computer experiments on benchmark datasets demonstrate the superiority of the proposed TargetFreeze over most recently released AFP predictors. We also implemented a user-friendly web server, which is openly accessible for academic use and is available at http://csbio.njust.edu.cn/bioinf/TargetFreeze. TargetFreeze supplements existing AFP predictors and will have potential applications in AFP-related biotechnology fields.

Keywords Antifreeze protein prediction · Multi-view protein features · Support vector machine · Machine learning

Introduction

Antifreeze proteins (AFPs) are ubiquitous and indispensable in various species, such as animals, plants, microbes, fishes, etc., that live in cold environments (Kim 2013). According to a structure and function study on the antifreeze protein of the winter flounder, the antifreeze mechanism was shown to be the result of the polypeptide or carbohydrate binding to an ice nucleation structure in ice lattice, subsequently stopping, or retarding the growth of ice crystals to depress the freeze point (Chou 1992). AFPs play a vital role in the survival of cold-adapted organisms in low and subzero temperatures by adsorbing onto and interacting with the surface of ice (Davies and Hew 1990; Fletcher et al. 2001). AFPs also have potential applications in industrial, medical, biotechnological and agricultural fields, such as food technology, preservation of cell lines, organs, cryosurgery and freeze-resistant transgenic plants and animals (Griffith and Ewart 1995; Feeney and Yeh 1998; Breton et al. 2000). Hence, the accurate identification of AFPs is of significant importance both for understanding the mechanism of protein-ice interactions and creating new ice-binding domains in other proteins (Kandaswamy et al. 2011). Designing accurate and automated computational methods for AFP identification is highly desired, especially in the post-genome era in which large volumes of sequenced proteins quickly accumulate without being functionally annotated (Mondal and Pai 2014).

Unfortunately, although AFPs in different species have a common anti-cold functionality that is obtained by

✉ Dong-Jun Yu
njyudj@njust.edu.cn

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, Xiaolingwei 200, Nanjing 210094, China

[2] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Dongchuan Road 800, Shanghai 200240, China

[3] Changshu Institute, Nanjing University of Science and Technology, Changshu 215513, China

contributing to freeze avoidance or freeze tolerance or both (Levitt 1980; Sformo et al. 2009), the sequences and structures of ice-binding domains differ significantly among AFPs although they are from closely related species (Jia and Davies 2002; Griffith and Yaish 2004; Graham et al. 2008). Consequently, it is challenging to correlate the relationships between sequences, structures, and functions (Yu and Lu 2011), which leads to difficulty in developing effective computational models for the identification and annotation of AFPs (Zhao et al. 2012). Nevertheless, researchers in this field have made efforts to develop computational models for identifying AFPs directly from protein sequences by utilizing machine-learning methods, and several promising predictors have emerged. Kandaswamy et al. (2011) performed pioneering work in antifreeze protein prediction from sequences using a machine-learning technique. In their method, called AFP-Pred, multiple physicochemical properties were used to encode the feature of a protein sequence, and the random forest is used as classifier. However, they do not provide an online web server, which limits the applicability of AFP-Pred; subsequently, Yu and Lu (2011) developed a web-based predictor, called iAFP, which, for the first time, utilizes multiple sets of *n*-peptide compositions as features for antifreeze protein prediction; Zhao et al. (2012) released AFP_PSSM for antifreeze protein prediction using a support vector machine (SVM) and position-specific scoring matrix (PSSM) profiles. Recently, Mondal and Pai (2014) reported the state-of-the-art sequence-based antifreeze protein predictor, called AFP-PseAAC, which utilizes Chou's pseudo amino acid composition (PseAAC) and SVM. All these works demonstrate the feasibility of identifying an antifreeze protein from the shared sequential features among the various structural types of AFPs with machine-learning methods.

All in all, much progress has been achieved in machine-learning based antifreeze protein prediction. Nevertheless, the prediction performances of the aforementioned methods are still not fully satisfactory, and there is still room for further improvement. Motivated by these considerations, we thus developed a new sequence-based antifreeze protein predictor, called TargetFreeze, which utilizes the weighted combination feature extracted from multi-view protein feature sources including amino acid composition (AAC), Chou's PseAAC, and the pseudo PSSM. Our computer experimental results on the benchmark dataset showed that the proposed TargetFreeze outperformed existing sequence-based antifreeze protein predictors, demonstrating the feasibility and efficacy of the proposed method.

As demonstrated by a series of recent publications (Chou 2011; Chen et al. 2014; Ding et al. 2014; Lin et al. 2014; Xu et al. 2014; Liu et al. 2015) in response to the call (Chou 2011) to establish a really useful sequence-based statistical predictor for a biological system, we need to consider the following procedures: (a) construct or select a valid benchmark dataset to train and test the predictor; (b) formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (c) introduce or develop a powerful algorithm (or engine) to operate the prediction; (d) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (e) establish a user-friendly web server for the predictor that is accessible to the public. Below, let us describe how to address these steps one by one.

Materials and Methods

Benchmark Datasets

To objectively evaluate the performance of the proposed method and fairly compare it with existing antifreeze predictors, the same dataset that has been used in previous studies (Kandaswamy et al. 2011; Zhao et al. 2012) was taken as the benchmark dataset in this study. The positive subset of the benchmark dataset consists of 481 AFPs constructed with stringent steps as follows: First, an initial dataset consisting of 221 antifreeze protein sequences was extracted from the Pfam database (Sonnhammer et al. 1997); then, the initial dataset was enriched by aligning each sequence in it against a non-redundant sequence database using PSI-BLAST with *E*-value (0.001); the enriched dataset was further manually inspected, and all those non-AFPs were removed; next, the maximal pairwise sequence identity of the proteins in the manually inspected dataset was culled to $\leq$40 % using CD-HIT (Li and Godzik 2006); the remaining 481 antifreeze protein sequences constitute the positive subset. The negative subset was constructed from 9193 seed proteins of Pfam protein families, which are unrelated to AFPs (Sonnhammer et al. 1997; Kandaswamy et al. 2011). For further details about the construction of the benchmark dataset, refer to (Kandaswamy et al. 2011).

To make a fair comparison with previously developed predictors, the benchmark dataset was further divided into a training set (for cross-validation test purposes) and a corresponding validation set (for independent validation test purposes) as in AFP-Pred (Kandaswamy et al. 2011) and AFP_PSSM (Zhao et al. 2012). More specifically, we constructed training and validation sets as follows: The training set was obtained by randomly selecting 300 antifreeze and 300 non-AFPs for three times from the benchmark dataset; the remaining 181 antifreeze and $8293 = 9193 - 3 * 300$ non-AFPs constituted the independent validation set.

Feature Representation

To develop a machine-learning-based AFP predictor, the critical step is to represent a protein sequence by a fixed-length feature vector. In this study, three types of protein sequential features, i.e., AAC, PseAAC, and pseudo position-specific scoring matrix (PsePSSM) were explored for AFP prediction.

Amino Acid Composition Feature

AAC (Block and Bolling 1951; Ahmad et al. 2004) is a classic protein feature representation method describing the composition information of a protein sequence. The AAC feature of a protein sequence is a 20-dimensional vector that can be formulated as follows:

$$\mathrm{AAC} = (f_1, f_2, \ldots, f_{20})^T, \tag{1}$$

where $f_i (i = 1, 2, \ldots, 20)$ represents the occurrence frequency of the 20 native amino acids in a protein sequence.

Pseudo Amino Acid Composition Feature

As shown in Eq. 1, the AAC feature only provides the composition information of a protein sequence, and all the sequence-order information is lost. To remedy this shortcoming of AAC, Chou (Chou 2001a, b; Shen and Chou 2008) proposed a PseAAC feature by incorporating protein sequential information into the traditional AAC feature. PseAAC has been widely used in protein attribute prediction problems (Huang et al. 2009; Roy et al. 2009; Jahandideh and Mahdavi 2012; Huang and Yuan 2013; Zou 2014) including AFP prediction (Kandaswamy et al. 2011). In this study, we generate the PseAAC feature of a protein sequence by feeding it to our previously developed web server, available at: http://www.csbio.sjtu.edu.cn/bioinf/PseAAC/ (Shen and Chou 2008).

When generating the PseAAC feature of a protein, the following parameter set was applied: (1) The PseAAC mode was set to Type 2; (2) all of the six amino acid characters, i.e., hydrophobicity, hydrophilicity, side chain mass, pK of the α-COOH group, pK of the $\alpha\text{-NH}_3^+$ group and pI at 25 °C, were selected ($\xi = 6$); (3) The weight factor was set to 0.05; (4) the counted rank (or tier) of the correlation was set to 20 ($\lambda = 20$). Accordingly, the dimensionality of the PseAAC feature is $20 + \xi \times \lambda = 20 + 6 \times 20 = 140$. For details regarding PseAAC and the PseAAC web server, please refer to Chou (2001a, b) and Shen and Chou (2008).

Pseudo Position-Specific Scoring Matrix Feature

We generate the PSSM for a given protein sequence by applying the PSI-BLAST (Schäffer et al. 2001) to search the Swiss-Prot database through three iterations with $E = 0.001$ as the cutoff against the query sequence. Then, we normalize each element in the obtained PSSM with the logistic function $f(x) = 1/(1 + e^{-x})$. Let $\mathbf{S}_{\text{pssm}} = (s_{i,j})_{L\times 20}$ be the normalized PSSM of a protein sequence with L residues; we can calculate its PsePSSM feature as follows (Wold et al. 1993; Liu et al. 2012; Yu et al. 2012):

First, we compute the 20-dimentional PSSM composition feature, denoted as $\mathbf{P}_{\overline{\text{pssm}}}$, by averaging the scores along each of the 20 columns of $\mathbf{S}_{\text{pssm}}$ as follows:

$$\mathbf{P}_{\overline{\text{pssm}}} = (\bar{s}_1, \bar{s}_2, \ldots, \bar{s}_j, \ldots, \bar{s}_{20})^T, \tag{2}$$

where

$$\bar{s}_j = \frac{1}{L}\sum_{i=1}^{L} s_{i,j}. \tag{3}$$

Second, we compute the PSSM sequence-order feature buried in the $\mathbf{S}_{\text{pssm}}$ by calculating the correlation factor of each column of $\mathbf{S}_{\text{pssm}}$ as follows:

$$\boldsymbol{\mu}^d = \left(\mu_1^d, \mu_2^d, \ldots, \mu_j^d, \ldots, \mu_{20}^d\right)^T, \tag{4}$$

where

$$\mu_j^d = \frac{1}{L-d}\sum\nolimits_{i=1}^{L-d} (s_{i,j} - \bar{s}_j) \cdot (s_{i+d,j} - \bar{s}_j), \tag{5}$$

where d $(1 \le d \le D)$ is the rank of correlation along the protein sequence and D is the maximum rank of correlation.

Finally, the PsePSSM feature of the protein sequence is obtained by combining its PSSM composition feature and PSSM sequence-order features as follows:

$$\mathrm{PsePSSM} = \begin{pmatrix} \mathbf{P}_{\overline{pssm}} \\ \boldsymbol{\mu}^1 \\ \boldsymbol{\mu}^2 \\ \vdots \\ \boldsymbol{\mu}^D \end{pmatrix} \tag{6}$$

In this study, the maximum rank of correlation D was set to 6, which was optimized on the benchmark dataset over ten-fold cross-validation. Accordingly, the dimensionality of the PsePSSM feature is $20 + D \times 20 = 20 + 6 \times 20 = 140$.

Weightedly Combining Multiple Features

AAC and PseAAC represent protein sequential information, while PsePSSM encodes the protein evolutionary information. These features from different views may potentially complement each other. The most straightforward and simplest method for utilizing the potential complementary information is to serially combine these multi-view features. However, studies have previously demonstrated that directly combining different features will not definitely lead to the improvement in prediction accuracy compared with a single-view feature (Kohavi and John 1997; Yu et al. 2013). One important reason for this phenomenon is that the scale of different features may differ significantly. In view of this, we will try to weightedly combine different features to improve the discriminative capability of the combined feature. How to set the weights of different features will be further discussed in the subsequent "Experimental Results and Analysis" section.

Classifier Selection

For a machine-learning-based predictor, its prediction performance will depend not only on the feature representation method but also on the classifier used. In this study, two popular classifiers, i.e., SVM (Vapnik 1998) and random forests (RFs) (Breiman 2001), were considered because they have been demonstrated to be extremely powerful in many protein attribute prediction problems including AFP prediction (Kandaswamy et al. 2011). Here, we briefly introduce the basic concepts and the parameter optimization procedures of the two classifiers.

SVM: Vapnik (1998) proposed the SVM based on the Vapnik–Chervonenkis theory of statistical learning and the principle of structural risk minimization. SVM seeks the best compromise between the computational complexity and the learning capability to obtain the best generalization ability (Vapnik 1998; Kecman 2001; Chang and Lin 2011). SVM can be categorized into two groups, i.e., support vector classification (SVC) and support vector regression (SVR). In this study, SVC was applied and the LIBSVM software package (Fan et al. 2005; Chang and Lin 2011), which is freely available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/, was used. Note that to apply SVC, the two parameters, i.e., the regularization parameter (ξ) and the kernel parameter (γ), were optimized based on tenfold cross-validation using a grid search strategy provided in the LIBSVM software package.

RFs: Breiman (2001) proposed the random forest learning algorithm, which is an ensemble method that adds an additional layer of randomness to bagging. In a standard forest, each node is split using the best split among all variables, while in a random forest, each node is split using the best among a subset of predictors randomly chosen at that node (Liaw and Wiener 2002). Numerous studies have demonstrated the success of the randomness strategy in RFs (Breiman 2001; Liaw and Wiener 2002). RFs can be used to perform both classification and regression. In this study, a random forest regression algorithm was applied. The RF code, which is freely accessible at http://scikit-learn.org/stable/modules/ensemble.html#random-forests, is used to evaluate and implement the proposed TargetFreeze.

Two parameters should be seriously considered when applying RFs. The first parameter (*mTry*) is the number of dimensions randomly sampled as candidates at each split. Previous studies have shown that the $sqrt(N)$ is a preferred strategy for setting the value of *mTry*, where N is the dimensionality of the feature vector (Breiman 2001; Liaw and Wiener 2002). Because the dimensionality of the combined feature is 300 ($N = 20 + 140 + 140$), the parameter *mTry* was set to be 17 $\approx$ sqrt(300). The second parameter is the number of trees to grow (*nTree*). In this study, *nTree* was optimized (under *mTry* = 17) and set to 500 with a tenfold cross-validation procedure on the training set.

Workflow of the Proposed TargetFreeze

Figure 1 illustrates the workflow of the proposed TargetFreeze. For a query protein sequence, TargetFreeze first extracts its AAC, PseAAC, and PsePSSM features and weightedly combines the extracted features; then, the combined feature will be fed to the trained SVM model, which outputs the probability of being an antifreeze protein of the query protein, denoted as T_{p}; the query protein will be predicted as an antifreeze protein if its predicted probability, i.e., T_{p}, is larger than the prescribed threshold T; otherwise, it will be predicted as a non-antifreeze protein. How to choose the threshold T will be further discussed in the subsequent section.

Evaluation Indices

In this study, we evaluated the performance of the proposed method with five routine evaluation indices in this field, i.e., Sensitivity (Sn), Specificity (Sp), Accuracy (Acc), Matthew's Correlation Coefficient (MCC), and Youden's Index. The definitions of the five evaluation indices are as follows:

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{7}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \tag{8}$$

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}}, \tag{9}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FN} \times \mathrm{FP}}{\sqrt{(\mathrm{TP}+\mathrm{FN}) \times (\mathrm{TP}+\mathrm{FP}) \times (\mathrm{TN}+\mathrm{FN}) \times (\mathrm{TN}+\mathrm{FP})}}, \tag{10}$$

$$\text{Youden's Index} = \mathrm{Sn} + \mathrm{Sp} - 1, \tag{11}$$

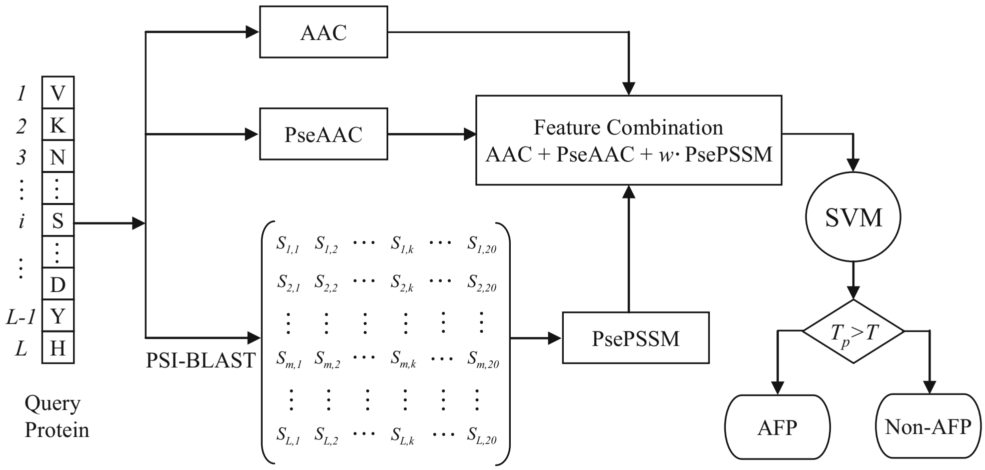
1
2
3
i
L-1
L
V
K
N
S
D
Y
H
Query Protein
AAC
PseAAC
PSI-BLAST
$S_{1,1}$ $S_{1,2}$ ⋯ $S_{1,k}$ ⋯ $S_{1,20}$
$S_{2,1}$ $S_{2,2}$ ⋯ $S_{2,k}$ ⋯ $S_{2,20}$
$S_{m,1}$ $S_{m,2}$ ⋯ $S_{m,k}$ ⋯ $S_{m,20}$
$S_{L,1}$ $S_{L,2}$ ⋯ $S_{L,k}$ ⋯ $S_{L,20}$
Feature Combination
AAC + PseAAC + w· PsePSSM
PsePSSM
SVM
$T_p > T$
AFP
Non-AFP

Fig. 1 Workflow of the proposed TargetFreeze

where *TP*, *FN*, *TN*, and *FP* are abbreviations of true positive, false negative, true negative, and false positive, respectively.

Although the four metrics (Eqs. 7–10) above were often used in the literature to measure the prediction quality of a prediction method, they are no longer the best ones because they lack intuitiveness and are not easy to understand for most biologists, particularly the *MCC* (the Matthews correlation coefficient). To make it easy to read, we adopt an additional four metrics proposed by Chou (Chou 2001a, b; Chen et al. 2013; Lin et al. 2014; Liu et al. 2014; Guo et al. 2014):

$$\mathrm{Sn} = 1 - \frac{N_-^+}{N^+}, \tag{12}$$

$$\mathrm{Sp} = 1 - \frac{N_+^-}{N^-}, \tag{13}$$

$$\mathrm{Acc} = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-}, \tag{14}$$

$$\mathrm{MCC} = \frac{1 - \left(\frac{N_-^+}{N^+} + \frac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \frac{N_+^- - N_-^+}{N^+}\right) \cdot \left(1 + \frac{N_-^+ - N_+^-}{N^-}\right)}}, \tag{15}$$

where N^+ represents the total number of the antifreeze protein sequences investigated, while N_-^+ is the number of antifreeze protein sequences incorrectly predicted as non-antifreeze protein sequences; N^- represents the total number of non-antifreeze protein sequences investigated, while N_+^- is the number of non-antifreeze protein sequences incorrectly predicted as antifreeze protein sequences. We can find the relations between Eqs. 7–10 and 12–15 as follows (Chou 2001a, b):

$$\begin{cases} \mathrm{TP} = N^+ - N_-^+ \\ \mathrm{TN} = N^- - N_+^- \\ \mathrm{FP} = N_+^- \\ \mathrm{FN} = N_-^+ \end{cases} \tag{16}$$

Obviously, when $N_-^+ = 0$, it means that none of the AFPs are incorrectly predicted, and we have the sensitivity Sn = 1; $N_-^+ = N^+$ means that all of AFPs are predicted incorrectly, and the sensitivity is 0 (Sn = 0). Similarly, we will have the specificity Sp = 1 if $N_+^- = 0$, and specificity will be 0 (Sp = 0) if $N_+^- = N^-$. When $N_-^+ = 0$ and $N_+^- = 0$, meaning that all of the proteins are predicted correctly, we have the accuracy Acc = 1, whereas $N_-^+ = N^+$ and $N_+^- = N^-$ indicate that none of the proteins in the datasets are correctly predicted, and the accuracy is zero (Acc = 0). The Matthews correlation coefficient (MCC) is usually used for measuring the quality of binary (two-class) classifications. When $N_+^- = N_-^+ = 0$, showing that the predictions of all proteins are correct, we have MCC = 1. If $N_+^- = N^-/2$ and $N_-^+ = N^+/2$, we have MCC = 0, indicating no better than random prediction; and when $N_+^- = N^-$ and $N_-^+ = N^+$, meaning the prediction for all the proteins is wrong, we have MCC = −1. Therefore, as described above, it is much more intuitive and easier-to-understand when using Eqs. 12–15 to examine a predictor with the four metrics, particularly for the Matthews correlation coefficient (Chou 2001a, b). Please note that the set of metrics defined in Eqs. 12–15 is valid only for single-label systems. For multi-label systems, which have become more frequent in systems biology (Lin et al. 2013) and systems medicine (Xiao et al. 2013), a completely different set of metrics as defined in Chou (2013) is needed.

On the other hand, the above-mentioned evaluation indices are threshold-dependent, i.e., their values will be affected by the threshold used. To fairly compare with previously developed predictors, we choose the threshold with the following strategy: we first identify the threshold, denoted as *T*, that maximizes the value of MCC of the predictions on the training set using cross-validation, and then, the identified threshold *T*, rather than another optimized one, was used to evaluate the performance of the proposed method on the corresponding independent validation set.

Experimental Results and Analysis

In this section, we will perform both a cross-validation test and independent validation test on the benchmark dataset to demonstrate the efficacy of the proposed method. As described in the Benchmark Datasets section, the training and validation sets were constructed with a random sampling procedure. To objectively evaluate the impact of the random sampling procedure, we repeated this construction procedure three times. On each of the constructed training and validation sets, we performed a cross-validation test and the corresponding independent validation test. The final prediction performances of the cross-validation test and independent validation test of the proposed method were obtained by averaging the three corresponding performances on training sets and validation sets, respectively.

Which Features Are Beneficial to Antifreeze Protein Prediction?

First, we will try to determine the features that are beneficial to antifreeze protein prediction by performing performance comparisons between different features with RFs and SVM as classifiers on the training set over both tenfold cross-validation and leave-one-out cross-validation. In this experiment, AAC, PseAAC, and PsePSSM were considered as three basic features. Two weighted combinations of the three basic features, i.e., PseAAC + $w\cdot$ PsePSSM and AAC + PseAAC + $w\cdot$ PsePSSM, were also considered to investigate the complementarity of the three basic features, where w is the combination weight. In this study, the combination weight w is 0.8, which was optimized by varying its value from 0 to 1 with a step size of 0.1 on the training set over tenfold cross-validation.

Tables 1 and 2 list the results of performance comparisons between different features on the training set over tenfold cross-validation and leave-one-out cross-validation, respectively. Note that in Tables 1 and 2, for each feature and classifier, the average performance of each evaluation index is reported, followed by a standard deviation. Figure 2a, b plot the receiver operating characteristic (ROC) curves of the five considered features with classifier RFs and SVM, respectively.

From Tables 1 and 2, several observations can be made as follows:

1. Each of the three basic features (i.e., AAC, PseAAC, and PsePSSM) can be effectively used to predict AFPs with MCC >0.689 and Youden's Index >0.686, indicating that both the protein sequential composition information and the protein evolutionary information are important clues for determining whether a protein is antifreeze. This observation is quite consistent with previous works (Kandaswamy et al. 2011; Yu and Lu 2011; Zhao et al. 2012; Mondal and Pai 2014).
2. As shown in Tables 1 and 2, for both RFs and SVM, we can find that the prediction performances were improved by weightedly combining the two or three basic features. Improvements of more than 2 % on both MCC and Youden's Index were observed. This observation demonstrates that the three basic features may contain complementary information for each other.
3. The best prediction performances on the training set over both tenfold cross-validation and independent validation tests were achieved by the method that takes AAC + PseAAC + $w\cdot$ PsePSSM as the input feature and SVM as the prediction engine. For the convenience of the subsequent description, we will refer to this method as TargetFreeze.

Comparisons with Existing Antifreeze Protein Predictors

In this section, we will compare the proposed TargetFreeze with popular sequence-based antifreeze protein predictions by performing both cross-validation and independent validation tests.

As described in the Introduction section, researchers in this field have made efforts to develop computational models for identifying AFPs and several promising predictors have emerged. In this study, we compared the proposed TargetFreeze with three sequence-based antifreeze protein predictors, i.e., AFP-Pred (Kandaswamy et al. 2011), iAFP (Yu and Lu 2011), and AFP-PseAAC (Mondal and Pai 2014), because they are the three most recent ones.

Comparisons with the State-of-the-Art Predictor Over the Cross-Validation Test

First, we will compare TargetFreeze with other predictors over the cross-validation test. On the one hand, among the three methods [i.e., AFP-Pred (Kandaswamy et al. 2011), iAFP (Yu and Lu 2011), and AFP-PseAAC (Mondal and Pai 2014)] used for comparison, AFP-PseAAC (Mondal and Pai 2014) is the most recently released predictor; on the other hand, only AFP-PseAAC (Mondal and Pai 2014) provides both tenfold cross-validation and leave-one-out cross-validation performances on the same benchmark dataset used in this study. In view of this, here, we only

Table 1 Performance comparisons between different features with RFs and SVM as classifiers on the training set over tenfold cross-validation

Feature	Classifier	Sn (%)	Sp (%)	Acc (%)	MCC	Youden's Index
AAC	RFs	80.55 (±1.35)	89.89 (±3.79)	85.22 (±1.93)	0.708 (±0.041)	0.704 (±0.039)
	SVM	84.22 (±0.69)	88.67 (±3.48)	86.45 (±1.49)	0.730 (±0.032)	0.729 (±0.030)
PseAAC	RFs	78.33 (±4.93)	90.33 (±3.06)	84.33 (±1.04)	0.693 (±0.016)	0.687 (±0.021)
	SVM	86.56 (±1.26)	86.89 (±3.67)	86.72 (±1.41)	0.735 (±0.029)	0.734 (±0.028)
PsePSSM	RFs	78.45 (±6.26)	91.78 (±6.24)	85.11 (±1.65)	0.712 (±0.032)	0.702 (±0.033)
	SVM	85.78 (±6.77)	91.00 (±5.17)	88.39 (±2.12)	0.772 (±0.039)	0.768 (±0.042)
PseAAC + *w*· PsePSSM[a]	RFs	83.67 (±2.91)	90.89 (±2.11)	87.28 (±1.18)	0.748 (±0.023)	0.746 (±0.024)
	SVM	88.89 (±4.50)	91.44 (±1.83)	90.16 (±1.89)	0.804 (±0.037)	0.803 (±0.038)
AAC + PseAAC + *w*· PsePSSM[a]	RFs	84.67 (±4.51)	89.89 (±4.44)	87.28 (±1.18)	0.748 (±0.025)	0.746 (±0.024)
	SVM	90.11 (±2.50)	91.78 (±0.51)	90.95 (±1.07)	0.819 (±0.021)	0.819 (±0.021)

For each feature and classifier, the average performance of each evaluation index is reported, followed by a standard deviation

[a] The weight ($w = 0.8$) is optimized by varying its value from 0 to 1 with a step size of 0.1 on the training set over tenfold cross-validation

Table 2 Performance comparisons between different features with RFs and SVM as classifiers on the training set over leave-one-out cross-validation

Feature	Classifier	Sn (%)	Sp (%)	Acc (%)	MCC	Youden's Index
AAC	RFs	76.22 (±7.93)	93.22 (±3.56)	84.72 (±2.37)	0.707 (±0.034)	0.694 (±0.047)
	SVM	84.11 (±1.35)	88.67 (±3.18)	86.39 (±1.11)	0.729 (±0.024)	0.728 (±0.022)
PseAAC	RFs	80.34 (±5.13)	88.22 (±2.99)	84.28 (±1.27)	0.689 (±0.021)	0.686 (±0.025)
	SVM	86.33 (±2.61)	87.44 (±3.91)	86.89 (±1.86)	0.739 (±0.037)	0.738 (±0.037)
PsePSSM	RFs	79.78 (±6.55)	91.45 (±3.97)	85.61 (±1.50)	0.720 (±0.022)	0.712 (±0.030)
	SVM	82.44 (±7.72)	94.89 (±3.56)	88.67 (±2.08)	0.783 (±0.033)	0.773 (±0.042)
PseAAC + *w*· PsePSSM[a]	RFs	82.00 (±2.73)	92.00 (±3.84)	87.00 (±1.00)	0.745 (±0.022)	0.740 (±0.020)
	SVM	87.67 (±5.51)	92.67 (±2.33)	90.17 (±1.69)	0.806 (±0.031)	0.803 (±0.034)
AAC + PseAAC + *w*· PsePSSM[a]	RFs	83.00 (±2.00)	92.33 (±2.00)	87.67 (±1.00)	0.757 (±0.020)	0.753 (±0.020)
	SVM	88.22 (±1.84)	93.44 (±1.02)	90.83 (1.17)	0.818 (±0.023)	0.817 (±0.023)

For each feature and classifier, the average performance of each evaluation index is reported, followed by a standard deviation

[a] The weight ($w = 0.8$) is optimized by varying its value from 0 to 1 with a step size of 0.1 on the training set over tenfold cross-validation

perform comparisons between TargetFreeze and AFP-PseAAC (Mondal and Pai 2014).

Table 3 lists the performance comparisons between TargetFreeze and AFP-PseAAC over tenfold cross-validation and leave-one-out cross-validation.

From Table 3, we can find that TargetFreeze achieves very comparable or even slightly better performances than AFP-PseAAC concerning all the five evaluation indices under tenfold cross-validation. On the other hand, TargetFreeze outperforms AFP-PseAAC with improvements of >4 % on both MCC and Youden's Index under leave-one-out cross-validation. Considering that leave-one-out cross-validation is the most rigorous cross-validation method, we thus can claim that TargetFreeze is superior to AFP-PseAAC, at least for the considered benchmark dataset.

Comparisons with Existing Predictors Over the Independent Validation Test

In statistical prediction, the sub-sampling or *K*-fold cross-validation test, jackknife test and independent dataset test are three often used cross-validation methods to examine a predictor for its effectiveness in practical application. Of the three test methods, the jackknife test is considered as the least arbitrary, because it can always yield a unique result for a given benchmark dataset as described in Chou (2011) and demonstrated by Eqs. 28–30 in (Chou 2011). Accordingly, the jackknife test has been widely recognized and increasingly used by investigators to examine the quality of various predictors (Xu et al. 2013; Chen et al. 2014; Ding et al. 2014; Lin et al. 2014; Dehzangi et al.

2015; Khan et al. 2015; Mandal et al. 2015). However, to reduce the computational time, we adopted the independent dataset test in this study as performed by many investigators, with SVM as the prediction engine.

More specifically, we will demonstrate the better generalization capability of TargetFreeze by comparing it with other predictors on an independent validation set. Performance comparisons between TargetFreeze, AFP-PseAAC (Mondal and Pai 2014), AFP-Pred (Kandaswamy et al. 2011), and iAFP (Yu and Lu 2011) on the independent validation set are listed in Table 4. Note that for TargetFreeze, the results were obtained by testing the protein sequences in the independent validation set on the model trained with the training set; and the results for AFP-PseAAC, AFP-Pred, and iAFP were obtained by feeding the protein sequences in the independent validation set to their web servers.

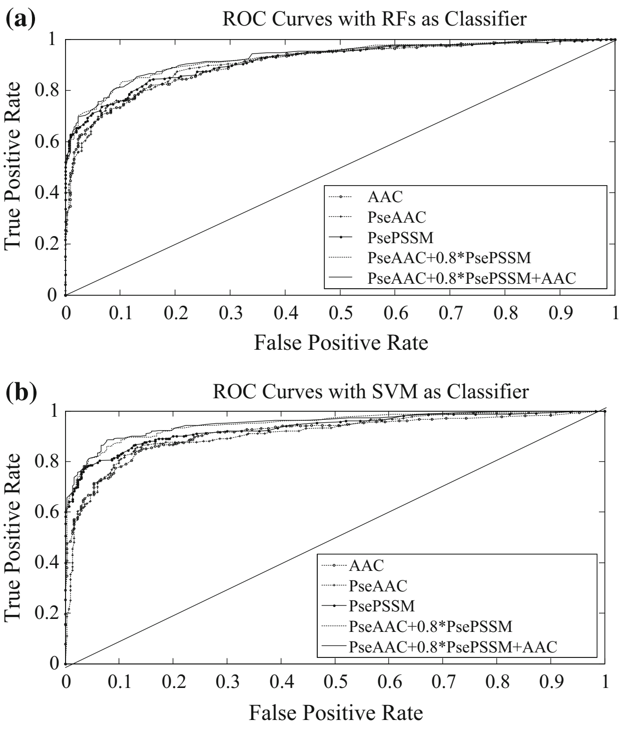

Fig. 2 Receiver operating characteristic (ROC) curves for the five considered features with **a** RFs and **b** SVM as classifier

By observing the results listed in Table 4, we can clearly find that TargetFreeze outperforms all the three other predictors with the highest MCC (0.398) and Youden's Index (0.837) and acts as the best performer. Improvements of >12 and >13 % for MCC and Youden's Index, respectively, were observed when compared with the second-best performer, i.e., AFP-PseAAC. It has not escaped our notice that iAFP achieves the best performances for Sp (97.09 %) and Acc (95.30 %); however, the Sn of iAFP is only 13.26 %, which is significantly lower than that of other predictors. In other words, iAFP predicts too many false negatives, leading to extremely lower MCC (0.086) and Youden's Index (0.10), which are two overall measurements of the quality of predictions. In summary, TargetFreeze has better generalization capability than iAFP, AFP-Pred, and AFP-PseAAC. We believe the better generalization capability of TargetFreeze benefits from the improved discriminative performance obtained by weightedly combining multiple protein sequential and evolutionary features.

Table 3 Performance comparisons between TargetFreeze and AFP-PseAAC over tenfold cross-validation and leave-one-out cross-validation

Method	Sn (%)	Sp (%)	Acc (%)	MCC	Youden's Index
TargetFreeze[a]	90.11 (±2.50)	91.78 (±0.51)	90.95 (±1.07)	0.819 (±0.021)	0.819 (±0.021)
AFP-PseAAC[a,c]	88.89 (±1.835)	91.00 (±0.330)	89.69 (±0.706)	0.800 (±0.0095)	0.799 (N/A)
TargetFreeze[b]	88.22 (±1.84)	93.44 (±1.02)	90.83 (±1.17)	0.818 (±0.023)	0.817 (±0.023)
AFP-PseAAC[b,c]	87.89 (±1.261)	89.22 (±0.840)	88.56 (±1.019)	0.771 (±0.0203)	0.771 (N/A)

[a] Results over tenfold cross-validation

[b] Results over leave-one-out cross-validation

[c] Results excerpted from reference Mondal and Pai (2014)

Table 4 Performance comparisons between TargetFreeze, AFP-PseAAC, AFP-Pred, and iAFP on the independent validation set

Method	Sn (%)	Sp (%)	Acc (%)	MCC	Youden's Index
TargetFreeze	92.45 (±1.39)	91.27 (±2.07)	91.30 (±2.00)	0.398 (±0.036)	0.837 (±0.010)
AFP-PseAAC[a]	85.08	84.74	84.75	0.270	0.70
AFP-Pred[a]	91.16	77.04	77.34	0.230	0.68
iAFP[a]	13.26	97.09	95.30	0.086	0.10

[a] Results were obtained by feeding the protein sequences in the independent validation set to the web servers of AFP-PseAAC, AFP-Pred, and iAFP

Conclusions

Accurate prediction of antifreeze protein is a challenging problem because of the biological diversity of different species. In this study, we developed a new sequence-based antifreeze protein predictor, called TargetFreeze, with high prediction performance obtained by effectively utilizing multiple protein features and a powerful SVM classifier. Computer experiments for both cross-validation and independent validation tests on the benchmark dataset demonstrate the superiority of TargetFreeze over existing predictors. To help potential users of TargetFreeze, we have implemented a user-friendly web server, which is freely available at http://csbio.njust.edu.cn/bioinf/TargetFreeze. We believe that the proposed TargetFreeze will supplement the existing AFPs predictors and can facilitate faster and broader applications of AFPs in biotechnology.

Acknowledgments This work was supported by the National Natural Science Foundation of China (No. 61373062, 61222306, 61202134, and 61233011), the Natural Science Foundation of Jiangsu (No. BK20141403), Jiangsu Postdoctoral Science Foundation (No. 1201027C), the Jiangsu University Graduate Research and Innovation Project (No. KYZZ_0123), the China Postdoctoral Science Foundation (No. 2013M530260, 2014T70526), "The Six Top Talents" of Jiangsu Province (No. 2013-XXRJ-022), the Fundamental Research Funds for the Central Universities (No. 30920130111010).

References

Ahmad S, Gromiha MM et al (2004) Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics 20:477–486

Block RJ, Bolling D (1951) The amino acid composition of proteins and foods. Analytical methods and results. Charles C Thomas, Springfield

Breiman L (2001) Random forests. Mach Learn 45:5–32

Breton G, Danyluk J et al (2000) Biotechnological applications of plant freezing associated proteins. Biotechnol Annu Rev 6:59–101

Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 27:1–27

Chen W, Feng PM et al (2013) iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res 41:e68

Chen W, Feng P-M et al (2014) iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. Anal Biochem 462:76–83

Chou K-C (1992) Energy-optimized structure of antifreeze protein and its binding mechanism. J Mol Biol 223:509–517

Chou K (2001a) Using subsite coupling to predict signal peptides. Protein Eng 14:75–79

Chou KC (2001b) Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 43:246–255

Chou K-C (2011) Some remarks on protein attribute prediction and pseudo amino acid composition. J Theor Biol 273:236–247

Chou KC (2013) Some remarks on predicting multi-label attributes in molecular biosystems. Mol Biosyst 9:1092–1100

Davies PL, Hew CL (1990) Biochemistry of fish antifreeze proteins. FASEB J 4:2460–2468

Dehzangi A, Heffernan R et al (2015) Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC. J Theor Biol 364:284–294

Ding H, Deng E-Z et al (2014) iCTX-type: A sequence-based predictor for identifying the types of conotoxins in targeting ion channels. BioMed Res Int. doi:10.1155/2014/286419

Fan RE, Chen PH et al (2005) Working set selection using second order information for training SVM. J Mach Learn Res 6:1889–1918

Feeney RE, Yeh Y (1998) Antifreeze proteins: current status and possible food uses. Trends Food Sci Technol 9:102–106

Fletcher GL, Hew CL et al (2001) Antifreeze proteins of teleost fishes. Annu Rev Physiol 63:359–390

Graham LA, Lougheed SC et al (2008) Lateral transfer of a lectin-like antifreeze protein gene in fishes. PLoS One 3:e2616

Griffith M, Ewart KV (1995) Antifreeze proteins and their potential use in frozen foods. Biotechnol Adv 13:375–402

Griffith M, Yaish MW (2004) Antifreeze proteins in overwintering plants: a tale of two activities. Trends Plant Sci 9:399–405

Guo SH, Deng EZ et al (2014) iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. Bioinformatics 30:1522–1529

Huang C, Yuan J-Q (2013) A multilabel model based on Chou's pseudo-amino acid composition for identifying membrane proteins with both single and multiple functional types. J Membr Biol 246:327–334

Huang W-L, Tung C-W et al (2009) Predicting protein subnuclear localization using GO-amino-acid composition features. Biosystems 98:73–79

Jahandideh S, Mahdavi A (2012) RFCRYS: sequence-based protein crystallization propensity prediction by means of random forest. J Theor Biol 306:115–119

Jia Z, Davies PL (2002) Antifreeze proteins: an unusual receptor–ligand interaction. Trends Biochem Sci 27:101–106

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

Kecman V (2001) Learning and soft computing: support vector machines, neural networks, and fuzzy logic models. MIT press, Cambridge

Khan ZU, Hayat M et al (2015) Discrimination of acidic and alkaline enzyme using Chou's pseudo amino acid composition in conjunction with probabilistic neural network model. J Theor Biol 365:197–203

Kim S-K (2013) Marine proteins and peptides: biological activities and applications. Wiley, Chichester

Kohavi R, John GH (1997) Wrappers for feature subset selection. Artif Intell 97:273–324

Levitt J (1980) Responses of plants to environmental stresses, vol II., Water, radiation, salt, and other stressesAcademic Press, New York

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659

Liaw A, Wiener M (2002) Classification and regression by random forest. R News 2:18–22

Lin WZ, Fang JA et al (2013) iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins. Mol Biosyst 4:634–644

Lin H, Deng E-Z et al (2014) iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Res 42:12961–12972

Liu T, Geng X et al (2012) Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles. Amino Acids 42:2243–2249

Liu B, Xu J et al (2014) iDNA-Prot|dis: Identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition. PLoS One 9:e106691

Liu B, Fang L et al (2015) Identification of real microRNA precursors with a pseudo structure status composition approach. PLoS One 10:e0121501

Mandal M, Mukhopadhyay A et al (2015) Prediction of protein subcellular localization by incorporating multiobjective PSO-based feature subset selection into the general form of Chou's PseAAC. Med Biol Eng Compu 53:331–344

Mondal S, Pai PP (2014) Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction. J Theor Biol 356:30–35

Roy S, Martinez D et al (2009) Exploiting amino acid composition for predicting protein-protein interactions. PLoS One 4:e7813

Schäffer AA, Aravind L et al (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 29:2994–3005

Sformo T, Kohl F et al (2009) Simultaneous freeze tolerance and avoidance in individual fungus gnats, Exechia nugatoria. J Comp Physiol B 179:897–902

Shen HB, Chou KC (2008) PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. Anal Biochem 373:386–388

Sonnhammer EL, Eddy SR et al (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 28:405–420

Vapnik VN (1998) Statistical learning theory. Wiley, New York

Wold S, Jonsson J et al (1993) DNA and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures. Anal Chim Acta 277:239–253

Xiao X, Wang P et al (2013) iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. Anal Biochem 436:168–177

Xu Y, Shao XJ, Wu LY, Deng NY, Chou KC (2013) iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. PeerJ 1:e171

Xu Y, Wen X et al (2014) iNitro-Tyr: prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. PLoS One 9:e105018

Yu C-S, Lu C-H (2011) Identification of antifreeze proteins and their functional residues by support vector machine and genetic algorithms based on n-peptide compositions. PLoS One 6:e20445

Yu D, Wu X et al (2012) Enhancing membrane protein subcellular localization prediction by parallel fusion of multi-view features. IEEE Trans Nanobioscience 11:375–385

Yu D-J, Hu J et al (2013) Learning protein multi-view features in complex space. Amino Acids 44:1365–1379

Zhao X, Ma Z et al (2012) Using support vector machine and evolutionary profiles to predict antifreeze protein sequences. Int J Mol Sci 13:2196–2207

Zou H-L (2014) A multi-label classifier for prediction membrane protein functional types in animal. J Membr Biol 247:1141–1148