J Membrane Biol (2016) 249:845–854
DOI 10.1007/s00232-016-9935-9

iAFP-Ense: An Ensemble Classifier for Identifying Antifreeze Protein by Incorporating Grey Model and PSSM into PseAAC

Xuan Xiao[1,3] · Mengjuan Hui[1] · Zi Liu[2]

Received: 26 July 2016 / Accepted: 24 October 2016 / Published online: 3 November 2016

Abstract Antifreeze proteins (AFPs), known as thermal hysteresis proteins, are ice-binding proteins. AFPs have been found in many fields such as in vertebrates, invertebrates, plants, bacteria, and fungi. Although the function of AFPs is common, the sequences and structures of them show a high degree of diversity. AFPs can be adsorbed in ice crystal surface and inhibit the growth of ice crystals in solution. However, the interaction between AFPs and ice crystal is not completely known for human beings. It is vitally significant to propose an automated means as a high-throughput tool to timely identify the AFPs. Analyzing physicochemical characteristics of AFPs sequences is very significant to understand the ice-protein interaction. In this manuscript, a predictor called "iAFP-Ense" was developed. The operation engine to run the AFPs prediction is an ensemble classifier formed by a voting system to fuse eleven different random forest classifiers based on feature extraction. We also compare our predictor with the AFP-PseAAC via the tenfold cross-validation on the same benchmark dataset. The comparison with the existing methods indicates the new predictor is very promising, meaning that many important key features which are deeply hidden in complicated protein sequences. The predictor used in this article is freely available at http://www.jci-bioinfo.cn/iAFP-Ense.

Keywords Antifreeze proteins · iAFP-Ense · Voting system · Tenfold cross-validation · Ensemble classifier

✉ Xuan Xiao
jdzxiaoxuan@163.com

Mengjuan Hui
huimengjuan@163.com

Zi Liu
liuzi189836@163.com

[1] Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403, China

[2] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

[3] School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China

Introduction

Antifreeze proteins (AFPs) which can unite with ice crystals under low concentration are the diverse group of polypeptides or glycoproteins. The material can inhibit the growth of ice crystals to improve the frost resistance. Accounting for these beneficial properties, AFP has been incorporated into the freeze-resistance or freeze tolerance strategies of many organisms such as animals, plants, microbes, especially in fish inhabitants of ice-laden sea water, which provide protection from freezing in extremely cold environments. In 1957, the professors (Scholander et al. 1957) noticed that many a fish species can be capable of surviving in the conditions where the temperature is lower than the freezing point of their body fluids. Later, it was reported that some overwintering plants can survive at temperatures of less than –50 °C (Sakai and Larcher 1987; Yoshida et al. 1997; Moriyama et al. 1995), suggesting that these organisms and plants have special antifreeze mechanisms to protect themselves against freezing stress. This antifreeze activity makes the organisms less sensitive to cold temperatures. Previous studies reported that the antifreeze effect is due to a group of proteins called ''antifreeze proteins'' (Levitt 1980; Sformo et al. 2009; Chou 1992; Mondal and Pai 2014). Depending upon the surrounding, organisms adopt two strategies

namely freeze tolerance and freeze avoidance to survive at low and subzero temperatures (Levitt 1980; Sformo et al. 2009; Chou 1992) which may account for the diversity observed among various species.

Analyses of AFPs from fish, insects, and plants have shown that there is no consensus sequence or structure for an ice-binding domain. Such research in sequence or structural level is important for understanding protein–ice interactions and freeze tolerance mechanisms of AFPs. These diverse achievements provide many clues to explain ice-binding affinity and specificity, but no one single mechanism has emerged to date. Since these proteins hold a promising scope for wide range of biotechnological applications in industry, medicine, food technology, cell lines and organ preservation, cryosurgery, and transgenic, gaining knowledge into their functional mechanisms has become increasingly essential (Mondal and Pai 2014; Kandaswamy et al. 2011).

AFPs have now been isolated from fishes, insects, plants, and microorganisms. AFPs are structurally diverse; they typically present a large proportion of their surface area for binding to ice. For example, in fish, AFPs are divided into five basically types (Xu et al. 2016; Jia and Davies 2002): AFP I, AFP II, AFP III, AFP IV, and AFGP. And AFP I which can be as alanine-rich α helix. AFP II are made up *C*-type lectin fold of mixed α, β, and loop structure. AFP III are Globular protein contains short β strands. Type IV AFPs are helix-bundle protein. AFGP antifreeze glycoprotein (Kandaswamy et al. 2011; Xu et al. 2016; Jia and Davies 2002). Besides, insect have two different β helixes that are right-handed and left-handed. Unfortunately, so far, no plant AFP structures have yet been solved (Jia and Davies 2002).

With the avalanche of newly generated protein sequences in the postgenomic age, a rapid, specific, and highly precise automated approach is desirable for identification and annotations of AFPs. Researchers encouraged by the overwhelming success of machine learning methods in protein classification and function prediction (Kandaswamy et al. 2011; Anand et al. 2008; Cai et al. 2004; Chou 2001, 2005; Chou and Cai 2005; Chou and Shen 2009; Zhao et al. 2012; Huang et al. 2009; Yu and Lu 2011). A large amount of analyses display that there is low sequence or structure similarity for an ice-binding domain and lack of common features among different AFPs (Ewart et al. 1999; Davies et al. 2002; Cheng 1998). These make it difficult to establish powerful prediction methods to identify AFPs. What is more, AFPs play vital roles in various fields, such as freeze-resistant transgenic plants and animals, food technology, and preservation of cell lines, organs, and cryosurgery (Griffith and Ewart 1995; Breton et al. 2000). The prediction approach has been designed on the basis of previously reported successful studies (Mondal and Pai 2014; Chen et al. 2013; Min et al. 2013; Qiu et al. 2014). It is highly desirable to develop an efficient means to determine the antifreeze proteins, which is also consequential to advance the understanding of protein–ice interactions and creating new ice-binding domains in other proteins.

In biology, it is virtually axiomatic that the sequence specifies conformation, which implies an intriguing hypothesis: the amino acid sequence alone might be sufficient to determine the preserving the sequence order information. So the concept of pseudo amino acid composition (pseAAC) was proposed (Chou 2009). In this concept, protein sequences are represented as discrete models yet without completely losing the sequence order information. Many novel models for addressing various kinds of problems in proteins and protein-related systems have been put forth since the idea of pseAAC was proposed. Further, different modes of optimal pseAAC composition are known to correspond to different protein attributes. Subsequently, the process of key components selection from the trivial ones for obtaining its pseAAC has become challenging. However, as pseAAC gives important direction for further improvement of the quality of protein attribute prediction, it has captured the interest of biologists.

AFPs have potential application in different fields including medical, biotechnological, agricultural, preservation of cell lines, organs, cryosurgery, and freeze-resistant transgenic plants and animals (Griffith and Ewart 1995; Breton et al. 2000). Identification of novel AFPs is important in understanding protein–ice interactions and also in creating novel ice-binding domains in other proteins.

Actually, in a pioneer work, Zhao et al. (2012) developed a method called "AFP_PSSM", which by incorporating SVM into evolutionary information to predict antifreeze protein. Yu and Lu (2011) developed a method based on the *n*-peptide composition coding schemes to identify different structural types of AFPs. Mondal and Pai (2014) using Chou's pseAAC features and SVMs proposed a novel method AFP-PseAAC to predict antifreeze protein and achieved better performance.

The aim of this work is to propose a new predictor for determination of the AFPs based on the features composed of sequence information by incorporating its sequential evolution information into the general form of pseudo amino acid composition, and the prediction engine was operated by the RF (random forest) algorithm. It anticipated that the iAFP-Ense protector can become a useful high-throughput tool for both basic research and drug development, and that the current approach may be easily extended to study the interactions of drug with other targets as well.

Materials and Methods

Benchmark Datasets

To develop a iAFP-Ense prediction model, we need construct or select a valid benchmark dataset to train and test the predictor. The data used in this paper were collected from the publicly available article (Kandaswamy et al. 2011); the antifreeze protein sequences were obtained from seed proteins of the Pfamdatabase (Sonnhammer et al. 1997). By performing Position Specific Iteration-Basic Local Alignment Search Tool (PSI-BLAST) search for each sequence against non-redundant sequence database with stringent threshold (E value 0.001) (Sonnhammer et al. 1997), the datasets were enlarged. The final positive dataset contained 481 (positive examples) AFPs and 9193 (negative examples) non-antifreeze proteins which were unrelated to AFPs. Here, the benchmark dataset $\mathbb{S}$ can be formulated as

$$\mathbb{S} = \mathbb{S}^+ \cup \mathbb{S}^-, \tag{1}$$

where $\mathbb{S}^+$ is the positive subset that consists of the AFPs only, while the $\mathbb{S}^-$ is negative subset that contains of the non-antifreeze proteins only, and the symbol $\cup$ represents the union in the set theory and S represents Online Supporting Information.

Sample Representation

In the traditional prediction, the data are mostly sacrificed and the quantity of dataset is inferior. Some sequence order information of the samples is inevitably missing in the process of prediction. That is to say, in the field of machine learning, which has a relatively mature development, single classifiers have gradually begun to show drawbacks. One of the challenging problems in computational biology today is to formulate a biological sequence with a discrete model or a vector, yet still keep considerable sequence order information. One important step to predict AFPs using sequence information is to find a suitable encoding of the protein sequence. It means to convert the protein sequence to a vector space. In this paper, the amino acid composition (AAC) of protein and Grey position specific scoring matrix (PSSM) were selected to translate a protein sequence to vector space.

Given a protein sample with L residues as expressed by

$$\mathrm{P} = R_1 R_2 R_3 R_4 R_5 R_6 R_7 \cdots R_L, \tag{2}$$

where R_1 represents the 1st amino acid residue of the protein **P**, R_2 the 2nd residue, and so forth. Now the problem is how to effectively represent the sequence of Eq. (2) with a or non-sequential discrete model (Chou and Shen 2007). This is because, the number of biological sequences with different sequence orders is extremely high and their lengths vary widely, meanwhile all the existing operation engines, such as covariance discriminant (Chen et al. 2012; Chou 2005; Wang et al. 2005), neural network (Feng et al. 2005), support vector machine (SVM) (Chen et al. 2013; Liu et al. 2014), random forest (RF) (Kandaswamy et al. 2011; Lin et al. 2011), conditional random field (Xu et al. 2013), nearest neighbor (Chou and Cai 2006), K-nearest neighbor (KNN) (Chou and Shen 2006), OET-KNN (Chou and Shen 2007; Shen and Chou 2009), Fuzzy KNN (Shen et al. 2006; Xiao et al. 2013) ML-KNN algorithm (Chou 2013), and SLLE algorithm (Wang et al. 2005), can only handle vector but not length-different sequences. However, a vector defined in a discrete model may completely lose all the sequence order information and hence limit the quality of prediction. Facing such a dilemma, can we find an approach of incorporating the sequence order effects? Ever since the concept of PseAAC was proposed in 2001 (Chou 2001), it has penetrated into almost all the areas of computational proteomics. Moreover, the concept of PseAAC was further extended to represent the feature vectors of nucleotides (Chen et al. 2012), as well as other biological samples (see, e.g., (Li et al. 2012; Huang et al. 2012; Jiang et al. 2013). Because it has been widely and increasingly used, recently two powerful soft-wares, called "PseAAC-Builder"(Du et al. 2012) and "propy" (Cao et al. 2013), were established for generating various special Chou's pseudo-AACs, in addition to the web-server "PseAAC" (Shen and Chou 2008; Chou 2011) built in 2008. According to a comprehensive review (Chou 2011), the general form of PseAAC for a protein sequence P is formulated by

$$\mathrm{P} = [\psi_1 \quad \psi_2 \quad \cdots \quad \psi_u \quad \cdots \quad \psi_\Omega]^T, \tag{3}$$

where **T** is the transpose operator, while Ω an integer to reflect the vector's dimension. The value of Ω as well as the components P in Eq. (3) will depend on how to extract the desired information from a peptide sequence. Below, we are to describe how to extract the useful information from the aforementioned benchmark datasets (cf. Eq. 1) to define the working peptides via Eq. (3).

First, many earlier studies (see, e.g., [136–141]) have indicated that the AAC of a protein plays an important role in determining its attributes. The AAC contains 20 components with each representing the occurrence frequency of one of the 20 native amino acids in the protein concerned. Thus, such 20 AAC components were used here to define the first 20 elements in Eq. (3); i.e.,

$$\psi_i = f_i^{(1)} \quad (i = 1, 2, \cdots, 20), \tag{4}$$

where $f_i^{(1)}$ is the normalized occurrence frequency of the i-th type native amino acid in the antifreeze proteins. Since AAC did not contain any sequence order information, the following steps were taken to make up for this shortcoming.

To avoid completely losing the local or short-range sequence order information, we incorporate the global or long-range sequence order information, let us consider the following approach. According to molecular evolution, all biological sequences have developed starting out from a very limited number of ancestral samples. To extract the evolutionary information, the profile of each protein sequence is generated by running Position Specific Iterated BLAST (PSI-BLAST) program (Schäffer et al. 2001; Altschul et al. 1997). Then this information can be represented as a two-dimensional matrix which is known as the PSSM of the protein. According to Schäffer et al. (2001), the sequence evolution information of a antifreeze protein P with L amino acid residues can be expressed by a 20 × Lmatrix, as given by

$$P_{PSSM}^{(0)} = \begin{bmatrix} D_{1\to1}^0 & D_{1\to2}^0 & \cdots & D_{1\to20}^0 \\ D_{2\to1}^0 & D_{2\to2}^0 & \cdots & D_{2\to20}^0 \\ \vdots & \vdots & \vdots & \vdots \\ D_{L\to1}^0 & D_{L\to2}^0 & \cdots & D_{L\to20}^0 \end{bmatrix}, \tag{5}$$

where $D_{i\to j}^0$ represents the original score of the i-th amino acid residue ($i = 1, 2,\ldots, L$) in the antifreeze sequence changed to amino acid type j ($j = 1, 2,\ldots, 20$) in the process of evolution. Here, the numerical codes 1, 2,…, 20 are used to, respectively, represent A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, the 20 single-letter codes for the 20 native amino acids. The L × 20 scores in Eq. (5) were generated using PSI-BLAST (Altschul et al. 1997) to search the UniProtKB/Swiss-Prot database (The Universal Protein Resource (UniProt); http://www.uniprot.org/) through three iterations with 0.001 as the D value cutoff for multiple sequence alignment against the sequence of the antifreeze concerned. In order to make every element in Eq. (5) be scaled from their original score ranges into the region of [0, 1], we performed a conversion through (Min et al. 2013) the standard sigmoid function to make it become

$$P_{PSSM}^{(1)} = \begin{bmatrix} D_{1\to1}^1 & D_{1\to2}^1 & \cdots & D_{1\to20}^1 \\ D_{2\to1}^1 & D_{2\to2}^1 & \cdots & D_{2\to20}^1 \\ \vdots & \vdots & \vdots & \vdots \\ D_{L\to1}^1 & D_{L\to2}^1 & \cdots & D_{L\to20}^1 \end{bmatrix}, \tag{6}$$

where

$$D_{i\to j}^1 = \frac{1}{1+e^{-D_{i\to j}^0}} \quad (1 \le i \le L, 1 \le j \le 20). \tag{7}$$

Now we extract the useful information from Eq. (6) to define the next 20 components of Eq. (3) via the following equation

$$\psi_{j+20} = \alpha_j \quad (j = 1, 2, \cdots, 20), \tag{8}$$

where

$$\alpha_j = \frac{1}{L} \times \sum_{K=1}^{L} D_{K\to j}^1 \quad (j = 1, 2, \cdots, 20). \tag{9}$$

Moreover, we used the grey system model approach as elaborated in (Min et al. 2013) to further define the next 60 components of Eq. (3)

$$\psi_{j+40} = \delta_j \quad (j = 1, 2, \cdots, 60), \tag{10}$$

where

$$\begin{cases} \delta_{3j-2} = w_1 f_j^{(1)} a_1^j \\ \delta_{3j-1} = w_2 f_j^{(1)} a_2^j \\ \delta_{3j} = w_3 f_j^{(1)} b^j \end{cases} \quad (j = 1, 2, \cdots, 20). \tag{11}$$

In the above equation, w_1, w_2, and w_3 are weight factors, which were all set to 1 in the current study; $f_j^{(1)}$ has the same meaning as in Eq. (4); a_1^j, a_2^j, and b^j are given by

$$\begin{bmatrix} a_1^j \\ a_2^j \\ b^j \end{bmatrix} = (B_j^T B_j)^{-1} B_j^T U_j \quad (j = 1, 2, \cdots, 20), \tag{12}$$

where

$$B_j = \begin{bmatrix} -D_{2\to j}^1 & -(D_{1\to j}^1 + 0.5D_{2\to j}^1) & 1 \\ -D_{3\to j}^1 & -(\sum_{i=1}^{2} D_{i\to j}^1 + 0.5D_{3\to j}^1) & 1 \\ \vdots & \vdots & \vdots \\ -D_{L\to j}^1 & -(\sum_{i=1}^{L-1} D_{i\to j}^1 + 0.5D_{L\to j}^1) & 1 \end{bmatrix} \tag{13}$$

and

$$U_j = \begin{bmatrix} D_{2\to j}^1 - D_{1\to j}^1 \\ D_{3\to j}^1 - D_{2\to j}^1 \\ \vdots \\ D_{L\to j}^1 - D_{L-1\to j}^1 \end{bmatrix}. \tag{14}$$

Incorporating the Eqs. (4), (8), and (10), we found that the total number of the components obtained via the current approach for the PseAAC of Eq. (3) is

$$\Omega = 20 + 20 + 60 = 100. \tag{15}$$

Each of the components is given by

$$\psi_j = \begin{cases} f_j^{(1)} & (1 \le j \le 20) \\ \alpha_j & (21 \le j \le 40) \\ \delta_j & (41 \le j \le 100). \end{cases} \tag{16}$$

Ensemble Learning Algorithm

The random forest algorithm, introduced by Breiman (2001), has been applied successfully in various biological

problems (Kandaswamy et al. 2011; Lin et al. 2011; Jia et al. 2015). The random forests algorithm is an ensemble of unpruned classification and regression trees that operates by constructing a multitude of decision trees at the training time and outputting the final class that is the majority vote of the classes output by individual trees. These trees are generated by bootstrap samples of the training data and using random feature selection in the tree generation process. Random forests algorithm usually exhibits a remarkable improvement of performance over the single-decision tree classifier. In our experiments, we found that the random forests algorithm is not sensitive to the number of trees. 100 trees are used in our model when the computational cost and overfitting problem are considered. The final output of the ensemble learning framework was the fusion of the results engendered by random forests algorithm.

It should be pointed out, however, that the number of negative samples in the current case is much larger than that of positive ones, but most classifiers (including RF) are usually working properly for the benchmark datasets consisting of balanced subsets. To deal with situation, an asymmetric boostrap approach was adopted as elaborated in (Lin et al. 2011) and illustrated in Fig. 1.

The predictor obtained via the aforementioned procedure is called iAFP-Ense, where "i" means identify, and "iAFP-Ense" means the ensemble classifier for identifying antifreeze proteins. To provide an intuitive overall picture, a flowchart is provided in Fig. 2 to show the process of how the predictor works in identifying the antifreeze proteins.

Metrics for Measuring Prediction Quality

To provide a more intuitive and easier to understand method to measure the prediction quality, the following set of metrics based on the formulation used by Chou (2001a, b, c, d) in predicting signal peptides was adopted. According to Chou's formulation, the sensitivity, specificity, overall accuracy, and Matthew's correlation coefficient can be, respectively, expressed as expressed as (Chen et al. 2012, 2013; Xu et al. 2013a, b).

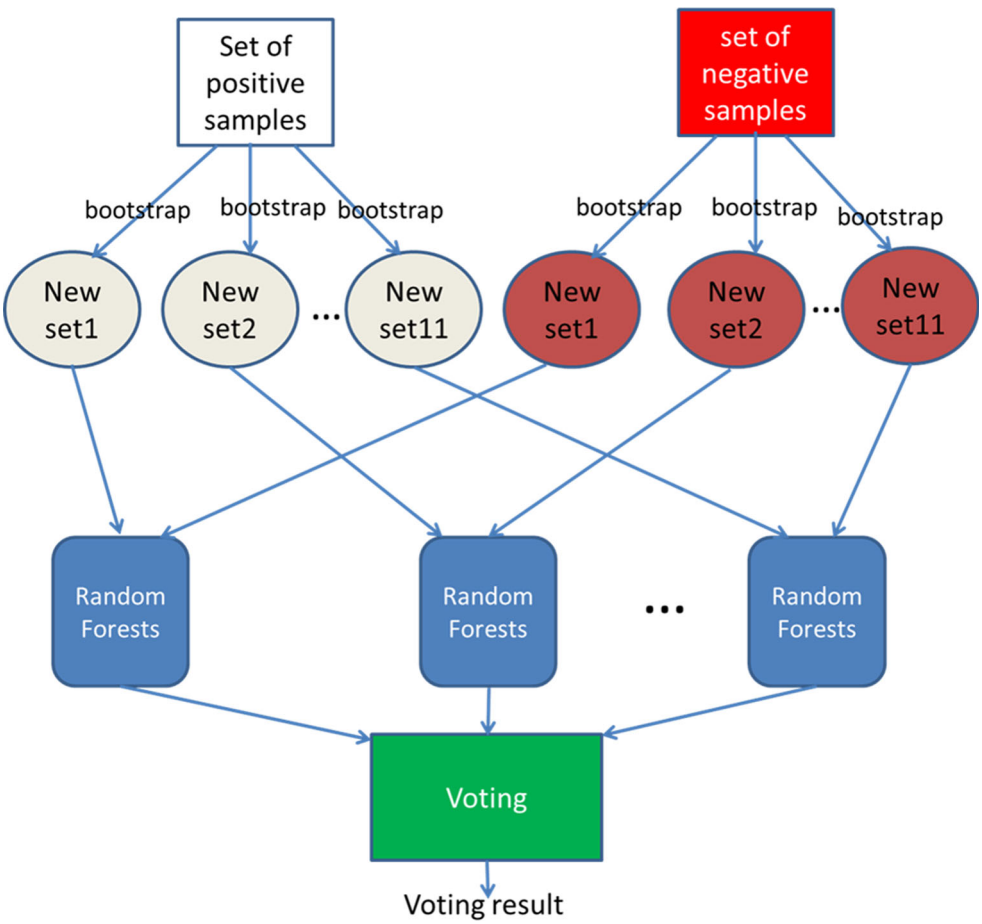

Fig. 1 A flowchart to show how an ensemble classifier is formed via a voting system

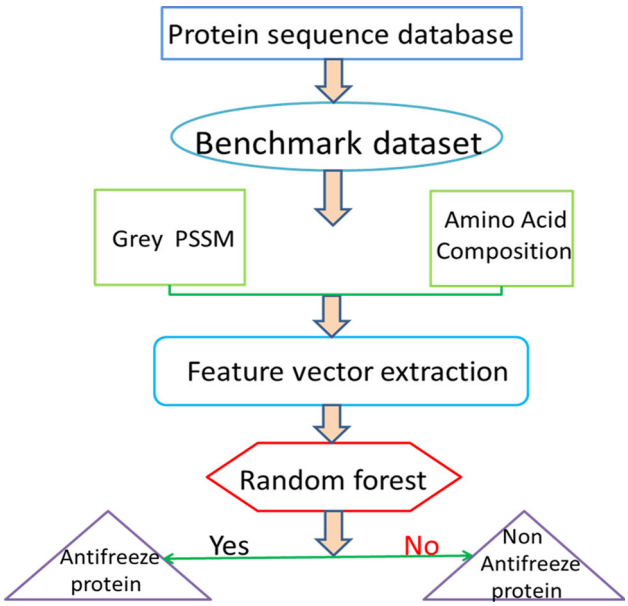

Fig. 2 The flowchart to show the process of how the predictor works in identifying the antifreeze proteins

$$\begin{cases} \mathrm{Sn} = 1 - \dfrac{N_-^+}{N^+} & 0 \le \mathrm{Sn} \le 1 \\ \mathrm{Sp} = 1 - \dfrac{N_+^-}{N^-} & 0 \le \mathrm{Sp} \le 1 \\ \mathrm{Acc} = \Lambda = 1 - \dfrac{N_-^+ + N_+^-}{N^+ + N^-} & 0 \le \mathrm{Acc} \le 1 \\ \mathrm{MCC} = \dfrac{1 - \left(\dfrac{N_-^+}{N^+} + \dfrac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \dfrac{N_+^- - N_-^+}{N^+}\right)\left(1 + \dfrac{N_-^+ - N_+^-}{N^-}\right)}} & -1 \le \mathrm{MCC} \le 1 \end{cases}, \tag{17}$$

where N^+ represents the total number of AFPs investigated, whereas N_-^+ the number of true AFPs incorrectly predicted to be of non-AFPs; N^- the total number of the non-AFPs investigated, whereas N_+^- the number of non-AFPs incorrectly predicted to be of AFPs.

According to Eq. (17), it is crystal clear to see the following. When $N_-^+ = 0$ meaning that none of the true AFPs are incorrectly predicted to be of non-AFPs, we have the sensitivity Sn = 1. When $N_-^+ = N^+$ meaning that all the AFPs are incorrectly predicted to be of non-AFPs, we have the sensitivity Sn = 0. Likewise, when $N_+^- = 0$ meaning that none of the non-AFPs are incorrectly predicted to be of AFPs, we have the specificity Sp = 1, whereas when $N_+^- = N^-$ meaning that all the non-AFPs are incorrectly predicted to be of AFPs, we have the specificity Sp = 0. When $N_-^+ = N_+^- = 0$ meaning that none of AFPs in the positive dataset and none of the non-AFPs in the negative dataset are incorrectly predicted, we have the overall accuracy Acc = 1 and MCC = 1; when $N_-^+ = N^+$ and $N_+^- = N^-$ meaning that all the AFPs s in the positive dataset and all the non-AFPs in the negative dataset are incorrectly predicted, we have the overall accuracy Acc = 0 and MCC = − 1; and when $N_-^+ = N^+/2$ and $N_+^- = N^-/2$, we have Acc = 0.5 and MCC = 0 meaning no better than random guess. As we can see from the above discussion, it would make the meanings of sensitivity, specificity, overall accuracy, and Mathew's correlation coefficient much more intuitive and easier to understand by using Eq. (17) to examine a predictor for its four metrics, particularly for the meaning of MCC. It should be pointed out that the metrics as defined in Eq. (17) are valid for single-label systems; for multi-label systems, a set of more complicated metrics should be used as given in Chou (2013).

Due to the evaluation metrics available, the next step is which validation method should be used to generate the metrics values. In statistical prediction, the following three cross-validation methods are often used to derive the metrics values for predictor: independent dataset test (Chou and Shen 2008), subsampling or *K*-fold cross-validation (such as five-fold, seven-fold, or tenfold) test (Chou 2013), and jackknife test (Chou and Zhang 1995). Of the three methods, however, the jackknife test is deemed the least arbitrary that can always yield a unique outcome for a given benchmark dataset as elucidated in (Cai et al. 2005; Y-d et al. 2004; Shi et al. 2008). Therefore, the jackknife test has been widely recognized and increasingly utilized by investigators to examine the quality of various predictors (Min et al. 2013; Xiao et al. 2013; Fan and Li 2013). However, to reduce the computational time, in this study, we adopted the tenfold cross-validation, as done by most investigators with SVM and random forests algorithms as the prediction engine (Gu et al. 2016; Gu et al. 2015; Wen et al. 2015). In *K*-fold cross-validation, the original sample is randomly partitioned into *K* subsamples. Of the *K* subsamples, a single subsample is retained as the validation data for testing the model, and the remaining subsamples are used as training data. The cross-validation process is then repeated *K* times (the folds), with each of the *K* subsamples used exactly once as the validation data. The advantage of this method is that all observations are used for both training and validation, and has high computational efficiency. Thus, the cross-validation has been used in many protein attributes prediction model checking. However, as can be seen, the partition is random, and the result is variable.

Web-Server and User Guide

To enhance the value of its practical applications, a web-server for AFPs predictor was established at the web-site http://www.jci-bioinfo.cn/iAFP-Ense. Moreover, for the convenience of the vast majority of experimental scientists,

here a step-to-step guide is provided for how to use the web-server predictors as follows.

Step 1: Go to the internet at http://www.jci-bioinfo.cn/iAFP-Ense and you will see the top page of the predictor on your computer screen, as shown in Fig. 3. Click on the **Read Me** button to see a brief introduction about AFPs predictor.

Step 2: When the predicted i**AFP-Ense** only in several protein sequences, you can either type or copy/paste the query protein sequence into the input box at the top half of Fig. 3. It is important to note that the input sequence should be in the FASTA format. A sequence in FASTA format consists of a single initial line beginning with a greater than symbol (''>'') in the first column, followed by lines of sequence data. The words right after the ''>'' symbol in the single initial line are optional and only used for the purpose of identification and description.

Step 3: To get the predicted result, you only need to click on the **Submit** button. For example, if you use the query amino acids sequences in the Example window as the input, you will see on your screen that the status of your job. When the job was done, the result will be displayed in the page.

As regards the computational time, the work will be accomplished within 15 s for most cases. However, the length of sequence is the key crucible of time-consuming, the longer the query protein sequence is, the more time it is usually needed.

Step 4: As shown on the lower panel of Fig. 3, you may also choose the batch prediction by entering your e-mail address and your desired batch input file (in FASTA format) via the ''**Browse**'' button. To see the sample of batch input file, click on the button **Batch-example**.

Step 5: By clicking the Citation button, you will find the relevant papers that document the detailed development of the predictor.

Step 6: Click on the Supporting Information button to download the benchmark dataset used to train and test the i**AFP-Ense** predictor.

Results and Discussion

The number of AFPs 481 positive examples were much lesser than 9193 negative examples, and we investigated for bias in identification due to selection process, by keeping the number of positive examples constant to 300. To deal with situation, an asymmetric boostrap approach was adopted as elaborated in (Jia et al. 2011) and illustrated in Fig. 1. Briefly, we selected 300 positive examples and 300 negative examples, eleven times randomly for model generation, and evaluated the prediction using the above-mentioned mathematical formulae after tenfold cross-validation. Each of the times can be used to train the RF predictor. Listed in Table 1 are the values of the four metrics Eq. (17) obtained by the current iATP-Ense predictor using the tenfold cross-validation on the new random subdataset. For facilitating comparison, the corresponding results

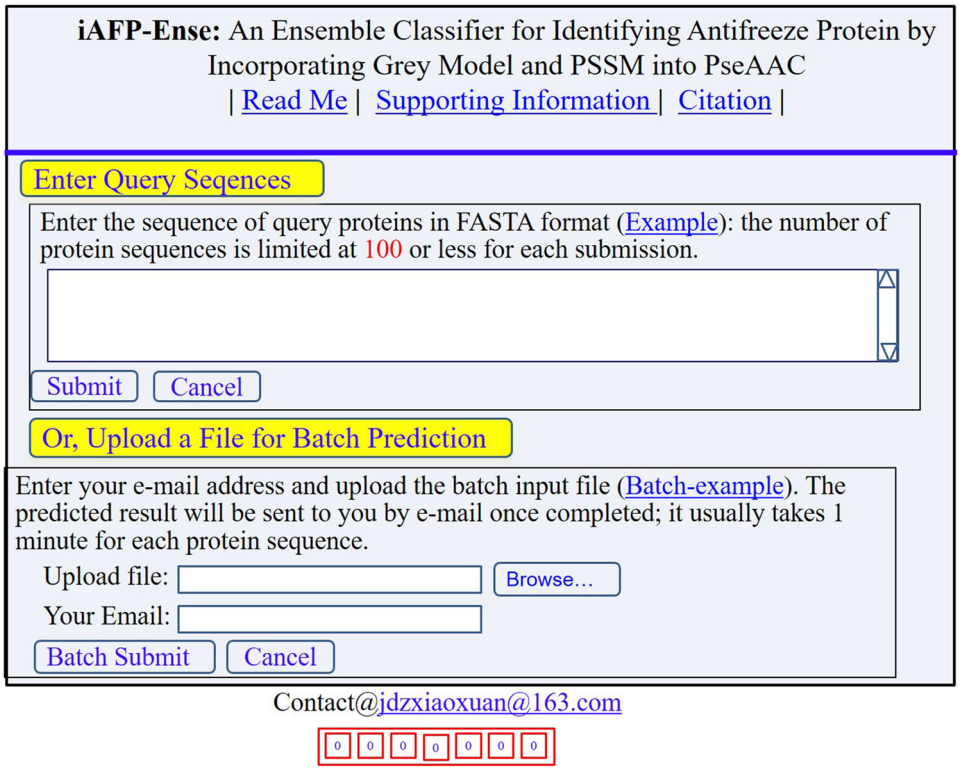

Fig. 3 A semi-screenshot to show the top page of the iAFP-Ense web-server at http://www.jci-bioinfo.cn/iAFP-Ense

Table 1 The comparison of AFP-Ense[b] with AFP-PseAAC[a] via the tenfold cross-validation on the same benchmark dataset whose ratio of positive sample and negative sample is 1:1

Method	Acc (%)	Sn (%)	Sp (%)	MCC
AFP-PseAAC[a]	89.69	88.89	91.00	0.800
AFP-Ense[b]	98.33	97.72	96.67	0.9672

[a] Results reported by Mondal and Pai (2014)

[b] Results obtained by the current predictor using the same cross-validation method on the same benchmark dataset

Table 2 The comparison of AFP-Ense[b] with AFP-PseAAC[a] via the tenfold cross-validation on the benchmark dataset whose ratio of positive sample and negative sample is 1:3

Method	Acc (%)	Sn (%)	Sp (%)	MCC
AFP-PseAAC[a]	91.25	77.67	96.78	0.762
AFP-Ense[b]	99.17	98.92	98.94	0.9761

[a] Results reported by Mondal and Pai (2014)

[b] Results obtained by the current predictor using the same cross-validation method on the same benchmark dataset

obtained by the existing methods (Mondal and Pai 2014) are also given there Because Mondal's method is the newest, and it has compared the previous method.

From the result shown in Table 1, we can see that our predict model achieves a good performance on the new dataset. The average results of the model are 98.33% for accuracy, 0.9672 for MCC, 97.72% for sensitivity, and 96.67% for specificity. The result output by AFP-PseAAC is as follows: 89.69% for accuracy, 0.800 for MCC, 88.89% for sensitivity, and 91.00% for specificity.

Then, we also have researched the influence of the number of negative examples on the AFP-Ense predictor. We did this by selecting 900 negative examples (three times higher than the number of positive samples 300) instead of the original ratio 1:1 for development of models followed by performance assessment using same evaluation parameters as mentioned above after tenfold cross-validation. We gained insights into the selection bias and influence of negative examples on prediction; the average performing results of the predictor are listed in Table 2; meanwhile, for the convenience to the comparison with the existing methods (Mondal and Pai 2014), the result of Mondal is also given in Table 2.

Conclusion

The operation engine to run AFPs prediction is an ensemble classifier formed by a voting system which can remarkably improve the prediction quality of a predictor. Very high prediction metricses on the training datasets show that iAFP-Ense is potentially useful tool for the prediction of antifreeze from protein primary sequence. Because of its simplicity and in this model, we can get the individual conservative genetic information for each of samples based on evolutionary information features. This approach can be easily extended to recognizing other specific functional properties and should be a useful tool for the high-throughput and large-scale analysis of proteomic and genomic data. It was confirmed via cross-validations that the new predictor established with the above procedures is very powerful and promising. We anticipate that iAFP-Ense predictor will become a very useful high-throughput tool for identifying AFPs or at the very least, a complementary tool to the existing prediction methods in this area. The iAFP-Ense program and dataset are available at http://www.jci-bioinfo.cn/iAFP-Ense. Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful prediction methods, we shall make efforts in our future work to provide a web-server for the method presented in this paper.

It is anticipated that the current strategy and novel technique can also be used to improve all those existing statistical predictors that were trained by highly unbalanced training datasets.

Acknowledgements This work was partially supported by the National Nature Science Foundation of China (No. 31260273, 61261027), the Jiangxi Provincial Foreign Scientific and Technological Cooperation Project (No. 20120BDH80023), Natural Science Foundation of Jiangxi Province, China (No. 20114BAB211013, 20122BAB211033, 20122BAB201044, 20122BAB201020), the Department of Education of JiangXi Province (GJJ12490, GJJ4642, GJJ14651, GJJ14640), the LuoDi plan of the Department of Education of JiangXi Province (KJLD12083), and the JiangXi Provincial Foundation for Leaders of Disciplines in Science (20113BCB22008).

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res 25:3389–3402

Anand A, Pugalenthi G, Suganthan P (2008) Predicting protein structural class by SVM with class-wise optimized features and decision probabilities. J Theor Biol 253:375–380

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

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

Cai Y-D, Ricardo P-W, Jen C-H, Chou K-C (2004) Application of SVM to predict membrane protein types. J Theor Biol 226:373–376

Cai Y-D, Zhou G-P, Chou K-C (2005) Predicting enzyme family classes by hybridizing gene product composition and pseudo-amino acid composition. J Theor Biol 234:145–149

Cao D-S, Xu Q-S, Liang Y-Z (2013) propy: a tool to generate various modes of Chou's PseAAC. Bioinformatics 29:960–962

Chen W, Lin H, Feng P-M, Ding C, Zuo Y-C et al (2012) iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties. PLoS ONE 7:e47843
Chen W, Feng P-M, Lin H, Chou K-C (2013) iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucl Acids Res 41:e68
Cheng C-HC (1998) Evolution of the diverse antifreeze proteins. Curr Opin Genet Dev 8:715–720
Chou K-C (1992) Energy-optimized structure of antifreeze protein and its binding mechanism. J Mol Biol 223:509–517
Chou KC (2001a) Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 43:246–255
Chou KC (2001b) Prediction of protein signal sequences and their cleavage sites. Proteins 42:136–139
Chou K-C (2001c) Using subsite coupling to predict signal peptides. Protein Eng 14:75–79
Chou K-C (2001d) Prediction of signal peptides using scaled window. Peptides 22:1973–1979
Chou K-C (2005a) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21:10–19
Chou K-C (2005b) Prediction of G-protein-coupled receptor classes. J Proteome Res 4:1413–1418
Chou K-C (2009) Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Curr Proteom 6:262–274
Chou K-C (2011) Some remarks on protein attribute prediction and pseudo amino acid composition. J Theor Biol 273:236–247
Chou K-C (2013) Some remarks on predicting multi-label attributes in molecular biosystems. Mol BioSyst 9:1092–1100
Chou K-C, Cai Y-D (2005) Prediction of membrane protein types by incorporating amphipathic effects. J Chem Inf Model 45:407–413
Chou K-C, Cai Y-D (2006) Prediction of protease types in a hybridization space. Biochem Biophys Res Commun 339:1015–1020
Chou K-C, Shen H-B (2006) Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. J Proteome Res 5:1888–1897
Chou K-C, Shen H-B (2007a) Recent progress in protein subcellular location prediction. Anal Biochem 370:1–16
Chou K-C, Shen H-B (2007b) Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. J Proteome Res 6:1728–1734
Chou K-C, Shen H-B (2008) Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. Nat Protoc 3:153–162
Chou K-C, Shen H-B (2009) Review: recent advances in developing web-servers for predicting protein attributes. Nat Sci 1:63
Chou K-C, Zhang C-T (1995) Prediction of protein structural classes. Crit Rev Biochem Mol Biol 30:275–349
Davies PL, Baardsnes J, Kuiper MJ, Walker VK (2002) Structure and function of antifreeze proteins. Philos Trans Royal Soc B 357:927–935
Du P, Wang X, Xu C, Gao Y (2012) PseAAC-Builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. Anal Biochem 425:117–119
Ewart K, Lin Q, Hew C (1999) Structure, function and evolution of antifreeze proteins. Cell Mol Life Sci CMLS 55:271–283
Fan G-L, Li Q-Z (2013) Discriminating bioluminescent proteins by incorporating average chemical shift and evolutionary information into the general form of Chou's pseudo amino acid composition. J Theor Biol 334:45–51
Feng K-Y, Cai Y-D, Chou K-C (2005) Boosting classifier for predicting protein domain structural class. Biochem Biophys Res Commun 334:213–217
Griffith M, Ewart KV (1995) Antifreeze proteins and their potential use in frozen foods. Biotechnol Adv 13:375–402
Gu B, Sun X, Sheng V-S (2016) Structural Minimax Probability Machine. IEEE Transactions on Neural Networks and Learning Systems, doi: 10.1109/TNNLS.2016.2544779
Gu B, Sheng V-S, Wang Z, Ho D, Osman S, Li S (2015) Incremental learning for ν-support vector regression. Neural Networks, 67:140–150
Huang R-B, Du Q-S, Wei Y-T, Pang Z-W, Wei H et al (2009) Physics and chemistry-driven artificial neural network for predicting bioactivity of peptides and proteins and their design. J Theor Biol 256:428–435
Huang T, Wang J, Cai Y-D, Yu H, Chou K-C (2012) Hepatitis C virus network based classification of hepatocellular cirrhosis and carcinoma. PLoS ONE 7:e34460
Jia Z, Davies PL (2002) Antifreeze proteins: an unusual receptor–ligand interaction. Trends Biochem Sci 27:101–106
Jia J, Xiao X, Liu B, Jiao L (2011) Bagging-based spectral clustering ensemble selection. Pattern Recogn Lett 32:1456–1467
Jia J, Xiao X, Liu B (2015) Prediction of protein–protein interactions with physicochemical descriptors and wavelet transform via random forests. J Lab Autom 22:368–377
Jiang Y, Huang T, Chen L, Gao Y-F, Cai Y et al (2013) Signal propagation in protein interaction network during colorectal cancer progression. BioMed Res Int 2013:9
Kandaswamy KK, Chou K-C, Martinetz T, Möller S, Suganthan P et al (2011) AFP-Pred: a random forest approach for predicting antifreeze proteins from sequence-derived properties. J Theor Biol 270:56–62
Levitt J (1980) Responses of plants to environmental stresses. Volume II. Water, radiation, salt, and other stresses, Academic Press, New York
Li B-Q, Huang T, Liu L, Cai Y-D, Chou K-C (2012) Identification of colorectal cancer related genes with mRMR and shortest path in protein-protein interaction network. PLoS ONE 7:e33393
Lin W-Z, Fang J-A, Xiao X, Chou K-C (2011) iDNA-Prot: identification of DNA binding proteins using random forest with grey model. PLoS ONE 6:e24756
Liu B, Zhang D, Xu R, Xu J, Wang X et al (2014) Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. Bioinformatics 30:472–479
Min J-L, Xiao X, Chou K-C (2013) iEzy-Drug: a web server for identifying the interaction between enzymes and drugs in cellular networking. BioMed Res Int 2013:13
Mondal S, Pai PP (2014) Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction. J Theor Biol 356:30–35
Moriyama M, Abe J, Yoshida M, Tsurumi Y, Nakayama S (1995) Seasonal changes in freezing tolerance, moisture content and dry weight of three temperate grasses [Dactylis glomerata, Lolium perenne, Phleum pratense]. J Jpn Soc Grassl Sci
Qiu W-R, Xiao X, Chou K-C (2014) iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components. Int J Mol Sci 15:1746–1766
Sakai A, Larcher W (1987) Frost survival of plants. Responses and adaptation to freezing stress. Springer, Berlin
Schäffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL et al (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucl Acids Res 29:2994–3005
Scholander P, Van Dam L, Kanwisher J, Hammel H, Gordon M (1957) Supercooling and osmoregulation in Arctic fish. J Cell Comp Physiol 49:5–24
Sformo T, Kohl F, McIntyre J, Kerr P, Duman J et al (2009) Simultaneous freeze tolerance and avoidance in individual fungus gnats, Exechia nugatoria. J Comp Physiol B 179:897–902

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

Shen H-B, Chou K-C (2009) A top-down approach to enhance the power of predicting human protein subcellular localization: hum-mPLoc 2.0. Anal Biochem 394:269–274

Shen H-B, Yang J, Chou K-C (2006) Fuzzy KNN for predicting membrane protein types from pseudo-amino acid composition. J Theor Biol 240:9–13

Shi J-Y, Zhang S-W, Pan Q, Zhou G-P (2008) Using pseudo amino acid composition to predict protein subcellular location: approached with amino acid composition distribution. Amino Acids 35:321–327

Sonnhammer EL, Eddy SR, Durbin R (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins-Struct Funct Genet 28:405–420

Wang M, Yang J, Xu Z-J, Chou K-C (2005) SLLE for predicting membrane protein types. J Theor Biol 232:7–15

Wen X, Shao L, Xue Y, Fang W (2015) A rapid learning algorithm for vehicle classification. Inform Sciences.295:395–406

Xiao X, Min J-L, Wang P, Chou K-C (2013a) iGPCR-Drug: a web server for predicting interaction between GPCRs and drugs in cellular networking. PLoS ONE 8:e72234

Xiao X, Wang P, Lin W-Z, Jia J-H, Chou K-C (2013b) iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. Anal Biochem 436:168–177

Xu Y, Ding J, Wu L-Y, Chou K-C (2013a) iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. PLoS ONE 8:e55844

Xu Y, Shao X-J, Wu L-Y, Deng N-Y, Chou K-C (2013b) iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. PeerJ 1:e171

Xu Y, Bäumer A, Meister K, Bischak CG, DeVries AL et al (2016) Protein–water dynamics in antifreeze protein III activity. Chem Phys Lett 647:1–6

Y-d Cai, Zhou G-P, Jen C-H, Lin S-L, Chou K-C (2004) Identify catalytic triads of serine hydrolases by support vector machines. J Theor Biol 228:551–557

Yoshida M, Abe J, Moriyama M, Shimokawa S, Nakamura Y (1997) Seasonal changes in the physical state of crown water associated with freezing tolerance in winter wheat. Physiol Plant 99:363–370

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

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