Journal of Theoretical Biology 270 (2011) 56–62

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties

Krishna Kumar Kandaswamy [a,b], Kuo-Chen Chou [c], Thomas Martinetz [a], Steffen Möller [a], P.N. Suganthan [d], S. Sridharan [e], Ganesan Pugalenthi [d,f,*]

[a] Institute for Neuro- and Bioinformatics, University of Lübeck, 23538 Lübeck, Germany
[b] Graduate School for Computing in Medicine and Life Sciences, University of Lübeck, 23538 Lübeck, Germany
[c] Gordon Life Science Institute, San Diego, CA 92130, USA
[d] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore
[e] Bharathidasan University, Tiruchirappalli, Tamilnadu 620 024, India
[f] Laboratory of Structural Biochemistry, Genome Institute of Singapore, 60 Biopolis Street, Singapore 138672, Singapore

ARTICLE INFO

Article history:
Received 18 August 2010
Received in revised form
29 October 2010
Accepted 29 October 2010
Available online 4 November 2010

Keywords:
Thermal hysteresis proteins
Ice binding proteins
Freeze tolerance
Physicochemical properties
Machine learning method

ABSTRACT

Some creatures living in extremely low temperatures can produce some special materials called "antifreeze proteins" (AFPs), which can prevent the cell and body fluids from freezing. AFPs are present in vertebrates, invertebrates, plants, bacteria, fungi, etc. Although AFPs have a common function, they show a high degree of diversity in sequences and structures. Therefore, sequence similarity based search methods often fails to predict AFPs from sequence databases. In this work, we report a random forest approach "AFP-Pred" for the prediction of antifreeze proteins from protein sequence. AFP-Pred was trained on the dataset containing 300 AFPs and 300 non-AFPs and tested on the dataset containing 181 AFPs and 9193 non-AFPs. AFP-Pred achieved 81.33% accuracy from training and 83.38% from testing. The performance of AFP-Pred was compared with BLAST and HMM. High prediction accuracy and successful of prediction of hypothetical proteins suggests that AFP-Pred can be a useful approach to identify antifreeze proteins from sequence information, irrespective of their sequence similarity.

1. Introduction

The surrounding environment plays a key role in the survival of living organisms. Extremely cold temperature causes intracellular ice formation which is considered to be lethal to the cell. Initially, it was thought that the coldest regions like Antarctica are uninhabitable due to extremely cold temperature which is lower than the freezing point of body fluids. In 1957, Scholander et al. observed that certain fish species were able to survive in the conditions where the temperature is lower than the freezing point of their body fluids (Scholander et al., 1957). Later it was reported that some overwintering plants such as Silene acaulis and Carex firma can survive at temperatures of less than −50 °C (Sakai and Larcher,1987; Yoshida et al., 1997; Moriyama et al., 1995). These findings suggest 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" (AFPs) (Logsdon and Doolittle, 1997; Ewart et al., 1999; Cheng, 1998; Davies and Sykes, 1997).

AFPs have the ability to adsorb onto the surface of ice crystals. The interaction between AFPs and ice crystals has significant effects on the overall growth of ice (Davies et al., 2002). Firstly, AFPs inhibit ice crystal growth and lower the freezing temperature of the water without altering the melting point. This process creates a difference between the freezing temperature and melting point which is known as thermal hysteresis (Urrutia et al., 1992). Each antifreeze protein has its own characteristic values for thermal hysteresis. Secondly, AFPs inhibit the recrystallization of ice, which involves the growth of larger ice crystals at the expense of smaller ice crystals (Yu and Griffith, 2001). Larger ice crystals increase the possibility of physical damage within frozen plant tissues (Griffith et al., 1997). Finally, AFPs also have the ability to interact with ice nucleators, which may result in either the inhibition or the enhancement of ice nucleation activity. Overwintering plants and animals adopt two strategies namely freeze tolerance and freeze avoidance to survive at low and subzero temperatures (Sformo et al., 2009; Lewitt, 1980). Freeze tolerance involves the activation or synthesis of ice-nucleating agents (INAs) in winter in freeze-tolerant species whereas freeze avoidance involves the inactivation or removal of ice-nucleating agents in freeze-avoiding species.

* Corresponding author at: Laboratory of Structural Biochemistry, Genome Institute of Singapore, 60 Biopolis Street, Singapore 138672, Singapore.
E-mail address: pugalenthig@gis.a-star.edu.sg (G. Pugalenthi).

0022-5193/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2010.10.037

AFPs have been discovered in various fish, insects, bacteria, fungi and overwintering plants including ferns, gymnosperms, monocotyledonous, dicotyledonous, angiosperms, etc. (Scholander et al., 1957; Moriyama et al., 1995; Logsdon and Doolittle, 1997; Ewart et al., 1999; Cheng, 1998; Davies and Sykes, 1997; Davies et al., 2002; Urrutia et al., 1992; Yu and Griffith, 2001). Analyses of AFPs from fish, insects and plants have shown that there is no consensus sequence or structure for an ice-binding domain. Some AFPs undergo structural changes at low temperatures (Davies et al., 2002). One explanation for AFP diversity is that ice can present many different surfaces with different geometric arrangements of oxygen atoms (Davies et al., 2002). The ice binding domains and their interaction with ice varies from species to species. For example, the ice binding domains of fish and insect AFPs are relatively hydrophobic and their adsorption onto ice is a hydrophobic interaction whereas plant antifreeze proteins have multiple, hydrophilic ice binding domains (Davies et al., 2002).

In fish, AFPs are classified into five known types namely AFGPs, AFP I, AFP II, AFP III and AFP IV (Fig. 1) (Davies et al., 2002; Davies and Hew, 1990; Chou, 1992). AFGPs are made up of 4 to more than 50 tandem repeats of Ala-Ala-Thr with a disaccharide attached to each Thr OH. It has an amphipathic polyproline type II helix fold. Type I AFPs are made up of alanine-rich, amphipathic helices. Type II AFPs are globular proteins with mixed secondary structure. Type III AFPs are made up of short beta-strands and one helix turn that gives it a unique flat-faced globular fold. Type IV AFPs are helix-bundle protein. Insect AFPs shows a beta helical structure (Graether et al., 2000). So far, crystal structure is not available for plant AFPs.

AFPs have potential industrial, medical, biotechnological and agricultural application in different fields, such as food technology, 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. With the rapid increase of sequenced genomic data, the need for an automated and accurate tool to recognize AFP becomes increasingly important. Encouraged by the overwhelming success of machine learning methods in an engineering, medical and financial applications, many research groups have been using neural networks (NN), support vector machines (SVM), KNN, random forest and other machine learning algorithms in the biological field especially in the classification and prediction of protein structure and functional profile (Anand et al., 2008; Cai et al., 2004; Chou, 2001, 2005; Chou and Cai, 2005; Chou and Shen, 2009; Huang et al., 2009; Qiu et al., 2009). So far, bioinformatics and statistical learning methods like support vector machine and random forest have not been explored for the prediction of antifreeze proteins. In this paper, we report a random forest approach to identify antifreeze proteins from sequence information, irrespective of the sequence similarity.

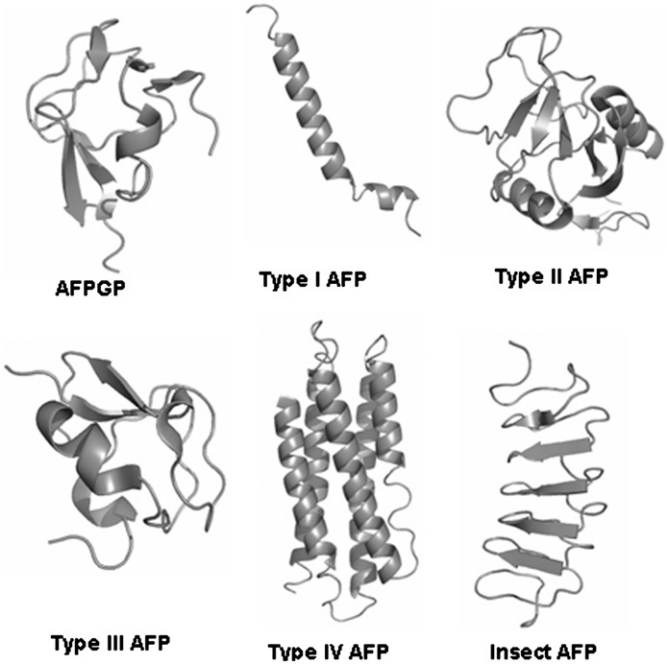

Fig. 1. Types of antifreeze proteins.

2. Materials and methods

2.1. Dataset

We obtained 221 antifreeze protein sequences from seed proteins of the Pfam database (Sonnhammer et al., 1997). To enrich the dataset, we performed PSI-BLAST search for each sequence against non-redundant sequence database with stringent threshold (*E*-value 0.001) (Altschul et al., 1997). Each sequence was subjected to manual inspection to retain only antifreeze proteins. Proteins with incomplete sequences were excluded. The sequences with $>=$40% sequence similarity were removed from the dataset using CD-HIT (Li et al., 2001). The final positive dataset contained 481 non-redundant antifreeze proteins. The negative dataset was constructed from 9493 seed proteins (representative members) of Pfam protein families, which are unrelated to antifreeze proteins (Sonnhammer et al., 1997).

Training set: 300 antifreeze domains were randomly selected from 481 antifreeze proteins for the positive dataset. Similarly, 300 non-antifreeze proteins were randomly taken from 9493 non-antifreeze proteins for the negative dataset.

Test set: The remaining 181 antifreeze proteins domains were served as a positive dataset for testing. Remaining 9193 non-antifreeze proteins (after excluding 300 non-antifreeze proteins that are used for training) were used as a negative dataset.

2.2. Features

In this work, each sequence is encoded by 119 features (Table 1) (please see http://www3.ntu.edu.sg/home/EPNSugan/index_files/AFP-Pred.htm for the complete list of 119 features).

Frequency of functional groups: We categorized 20 amino acids into 10 functional groups based on the presence of side chain chemical groups such as phenyl (F/W/Y), carboxyl (D/E), imidazole (H), primary amine (K), guanidino (R), thiol (C), sulfur (M), amido (Q/N), hydroxyl (S/T) and non-polar (A/G/I/L/V/P) (Pugalenthi et al., 2008).

Table 1
List of 119 sequence derived features.

Features	No. of features
Frequencies of 10 functional group	10
Frequency of hydrophobic, hydrophilic, neutral , positive, negative, polar and non-polar amino acids	7
Overall composition of helix (H), stand (E) and coil (C)	3
Frequencies of 10 functional groups in helix, strand and coil regions	30
Frequencies of hydrophobic, hydrophilic, neutral, positive, negative, polar and non-polar amino acids in helix, strand and coil regions	21
Frequencies of short peptides	17
Physicochemical properties	31
Total	119

The frequency of 10 functional groups (number of occurrences of functional group "X" divided by length of the protein) was computed for each sequence.

Frequency of physico-chemical groups: On the basis of physico-chemical properties, we classified 20 amino acids into 7 groups such as hydrophobic, hydrophilic, neutral, positive, negative, polar and non-polar amino acid groups. The frequencies of physico-chemical groups were computed for each sequence.

Frequency of short peptides: We incorporated the frequency of short peptides (10 residue length, in this case) which are rich in 10 functional groups and 7 physico-chemical groups. For example, if a short peptide has more than six hydrophobic residues, then we consider this peptide as hydrophobic rich short peptide. Similarly, we calculated short peptides which are rich in functional groups and physico-chemical groups. The frequency of short peptides for each group was calculated.

Content of secondary structural element (SSE): SSE (helix: H, strand: E and coil: C) information was assigned to all the sequences in the alignment using a secondary structure prediction program, PSIPRED (McGuffin et al., 2000). The overall composition of helix (H), beta sheet (E), coil (C) and the frequencies of 10 amino acid group and 7 physico-chemical groups at helix, sheet and coil regions were calculated.

Physicochemical properties: Physicochemical properties derived from AAINDEX database is used to compute the following properties: hydrophobicity, hydrophilicity, isoelectric point, flexibility, mol wt, polarity, refractivity, accessibility, normalized frequency of β-strand, normalized frequency of α-helix, melting point, heat capacity, side chain volume, side chain hydrophobicity, signal sequence helical potential, membrane buriability, conformational parameter of inner helix, retention coefficient, free energy change, steric hindrance, parameter of charge transfer capability, parameter of charge transfer donor capability, relative mutability, average membrane preference, optical rotation, number of hydrogen bond donors, positive charge, negative charge, net charge, buriability and amphiphilicity index (Kawashima et al., 1999). For each sequence, physico-chemical property value was calculated as the sum of physicochemical property value for all residues of the sequence, divided by the length of the sequence.

2.3. Classification protocol

Random forest (RF), introduced by Breiman, has been applied successfully in various biological problems (Breiman, 2001; Wu et al., 2003; Lee et al., 2005; Uriarte and Andres, 2006; Kandaswamy et al., 2010; Kumar et al., 2009; Masso and Vaisman, 2010). RF is an ensemble method which uses recursive partitioning to generate many trees and then aggregate the results. Each tree is independently constructed using a bootstrap sample of the training data. For each tree, two-third of the training samples are used for tree construction and the remaining one-third of the samples are used to test the tree. This left out data, named "Out of Bag", is used to calibrate the performance of each tree. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. Compared with the decision tree classifier, random forests have better classification accuracy, are more tolerant to noise and are less dependent on the training datasets (Han et al., 2009; Ho, 1998, 2002).

The most machine learning methods that need to resort to cross-validation for the estimation of classification error, the random forest can natively estimate an out-of-bag (OOB) error in the process of constructing the forest, and this estimate is claimed to be unbiased in many tests (Breiman, 2001; Hua et al., 2010). The training set of antifreeze proteins leaves out a significant portion of the samples, thus called 'out-of-bag' (OOB) data. These excluded samples, serve to provide an unbiased estimate of classification error. All OOB data is input through their respective trees and a classification for each OOB case is voted on. The collective OOB data serve as a formative antifreeze test set, and each OOB case or sample assigned a series of test set classifications, the number equivalent to how many times that particular case was left out of the training set for a tree. OOB error is calculated by taking the proportion of classifications for a OOB case that do not agree with the true, 'gold-standard' classification over the total number of cases. The RF algorithm was implemented by the randomForest R package (Liaw and Wiener, 2002).

2.4. Relief-F

ReliefF is an efficient procedure to estimate the quality of attributes in problems with strong dependencies between attributes (Zhang et al., 2008). In practice, ReliefF is usually applied in data pre-processing as a feature subset selection method. The basic idea of Relief-F is to draw instances at random, compute their nearest neighbors, and adjust a feature weighting vector to give more weight to features that discriminate the instance from neighbors of different classes. This method was implemented using Weka 3.5 (Frank et al., 2004).

2.5. Evaluation parameter

The performance of various models developed in this study was computed by using threshold-dependent as well as threshold-independent parameters. In threshold-dependent parameters, we used sensitivity, specificity, overall accuracy and Matthew's correlation coefficient (MCC) using following equations. These measurements are expressed in terms of true positive (TP), false negative (FN), true negative (TN) and false positive (FP).

Sensitivity: This parameter allows computation of the percentage of correctly predicted antifreeze proteins

$$Sensitivity = \frac{TP}{TP+FN} \tag{1}$$

Specificity: This parameter allows computation of the percentage of correctly predicted non-antifreeze proteins

$$Specificity = \frac{TN}{TN+FP} \tag{2}$$

Accuracy: Percentage of correctly predicted antifreeze and non-antifreeze proteins

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{3}$$

Matthews's Correlation Coefficient (MCC): It is the statistical parameter to assess the quality of prediction and to take care of the unbalancing in data. The Matthew's correlation coefficient ranges from $-1 \le MCC \le 1$. A value of $MCC=1$ indicates the best possible prediction while $MCC=-1$ indicates the worst possible prediction (or anti-correlation). Finally, $MCC=0$ would be expected for a random prediction scheme

$$MCC = \frac{TPTN-FPFN}{\sqrt{(TP+FP)(TN+FN)(TP+FN)(TN+FP)}} \tag{4}$$

Area under the Curve (AUC): Most of above measures have a common drawback that their performance depends on threshold selected. A known threshold independent parameter is Receiver

Operating Curve (ROC). It is a plot between true positive rate (TP/TP+FN) and false positive rate (FP/FP+TN).

3. Results and discussion

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, subsampling test and jackknife test (Chou and Zhang, 1995). In the independent dataset test, although none of the proteins to be tested occurs in the training dataset used to train the predictor, the selection of proteins for the testing dataset could be quite arbitrary unless it is sufficiently large. This kind of arbitrariness may directly affect the conclusion. For instance, a predictor yielding higher success rate than the others for a testing dataset might fail to remain so when applied to another testing dataset (Chou and Shen, 2008). For the subsampling test, the practical procedure often used in literatures is the 5-fold, 7-fold or 10-fold cross-validation. The problem with the subsampling examination as such is that the number of possible selections in dividing a benchmark dataset is an astronomical figure even for a very simple dataset (see Eq. (50) of Chou and Shen, 2007 or Eq. (1) of Chou and Shen, 2010a). Therefore, any practical result by the subsampling test only represents one of many possible results, and hence cannot avoid the arbitrariness either. In the jackknife cross-validation, each of the protein samples in the benchmark dataset is in turn singled out as a tested protein and the predictor is trained by the remaining proteins. During the jackknifing process, both the training dataset and testing dataset are actually open, and a protein will in turn move from one to the other. The jackknife cross-validation can exclude the memory effects during entire testing process and also the result thus obtained is always unique for a given benchmark dataset. Therefore, of the above three examination methods, the jackknife test is deemed the most objective (Chou and Shen, 2008), and has been widely recognized and increasingly used by investigators to examine the accuracy of various predictors (see, e.g., Chen et al., 2008, 2009; Chou and Shen, 2010b, 2010c; Ding et al., 2009; Du and Li, 2008; Ji et al., 2010; Jiang et al., 2008; Joshi and Sekharan, 2010; Li and Li, 2008; Li et al., 2009; Lin, 2008; Lin et al., 2008, 2009; Liu et al., 2010; Lu et al., 2009; Mohabatkar, 2010; Nanni and Lumini, 2009; Shao et al., 2009; Shi et al., 2008; Tian et al., 2008; Vilar et al., 2009; Wang et al., 2010; Yang et al., 2009, 2010; Zeng et al., 2009; Zhou, 1998; Zhou and Assa-Munt, 2001; Zhou and Doctor (2003); Zhou et al., 2007). Accordingly, the jackknife test will also be used in this study to evaluate our method.

3.1. Prediction using PSI-BLAST

PSI-BLAST is an extensively used local similarity search tool for identifying homologous sequences (Altschul et al., 1997). The performance of PSI-BLAST was evaluated using jackknife cross validation method, where each sequence in the positive dataset (481 antifreeze proteins) was used as a BLAST query sequence and remaining sequences (480 antifreeze proteins) were used as a BLAST database. Three iterations of PSI-BLAST were carried out at *E* value 0.001.It was observed that only 280 antifreeze proteins showed similarity (BLAST hit) with other antifreeze proteins (*E* value—0.0001) and no hits were obtained for the remaining 201 AFPs. This implies that similarity-based search methods alone may not be the best choice for the annotation of antifreeze proteins. Therefore, we decided to explore machine learning method to predict AFPs from sequence derived features such as frequency of amino acid groups, secondary structural element and physiochemical properties, etc.

3.2. Prediction of antifreeze proteins by AFP-Pred

In this work, we reported a random forest method for the prediction of antifreeze proteins from protein sequence using 119 sequence-derived properties. We trained our random forest model on the dataset containing 300 antifreeze proteins and 300 non-antifreeze proteins. AFP-Pred achieved 81.33% training accuracy using all the features. In order to examine the performance of the newly developed model, we tested our training models on dataset containing 181 antifreeze proteins and 9193 non-antifreeze proteins. As shown in Table 2, AFP-Pred achieved 83.38% accuracy with 84.67% sensitivity, 82.32% specificity and MCC of 0.6674.

To identify the prominent features, we carried out feature selection using ReliefF based feature subset selection method. The number of features was reduced from 119 to 10 features. Table 2 shows the performance of our method on the test dataset using different feature subsets. As seen in Table 2, feature selection generally does not deteriorate the classification performance much until the number of features decreases to 10. AFP-Pred achieved 83.38% accuracy with 84.67% sensitivity, 82.32% specificity and MCC of 0.6674. The prediction accuracy was slightly improved when the features were reduced from 119 to 25. Using 25 features, AFP-Pred obtained 84.29% accuracy with 84.67% sensitivity, 83.98% specificity and MCC of 0.6846. This result suggests that our feature reduction approach selected useful features by eliminating the uncorrelated and noisy features.

We also investigated the influence of the feature reduction by plotting Receiver Operating Characteristic (ROC) curves (Fig. 2) derived from the sensitivity (true positive rate) and specificity (false positive rate) values for the classifiers using all the features and the 25 best performing features, respectively. The area under

Table 2
Performance of random forest model on test dataset containing 181 AFPs and 9193 non-AFPs using different feature subsets.

Feature subset	Sensitivity (%)	Specificity (%)	MCC	Accuracy (%)
10 features	80.00	80.66	0.6052	80.36
25 features	84.67	83.98	0.6846	84.29
50 features	86.00	81.77	0.6749	83.69
75 features	84.00	82.87	0.6667	83.38
100 features	84.00	81.77	0.6553	82.78
All features	84.67	82.32	0.6674	83.38

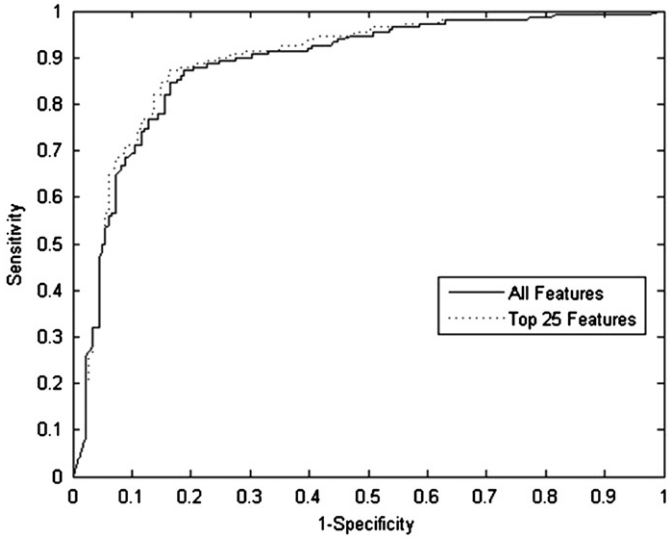

Fig. 2. ROC Plot for random forest models developed using all features and top 25 features.

Table 3
Prediction result for 16 potential antifreeze proteins.

GI Code	AFP-Pred	BLAST	HMM	Source of annotation	NCBI definition
26325086	–	AFP	AFP	INTERPRO	Unnamed protein product
26344193	AFP	AFP	–	INTERPRO	Unnamed protein product
74221639	AFP	AFP	–	INTERPRO	Unnamed protein product
12843602	AFP	–	–	INTERPRO	Unnamed protein product
257049854	AFP	–	–	KEGG	Hypothetical protein
30249105	AFP	AFP	AFP	INTERPRO	Type I antifreeze protein
226941159	AFP	AFP	AFP	INTERPRO	Type I antifreeze protein
126464034	AFP	–	–	KEGG	Type I antifreeze protein
45435722	AFP	AFP	–	INTERPRO	Hypothetical protein
281341260	AFP	AFP	AFP	INTERPRO	Hypothetical protein
2315605	AFP	–	–	INTERPRO	Hypothetical protein
260817607	AFP	AFP	AFP	INTERPRO	Hypothetical protein
26388908	AFP	–	–	INTERPRO	Unnamed protein product
26348120	AFP	–	–	INTERPRO	Unnamed protein product
26333557	AFP	–	–	INTERPRO	Unnamed protein product
26332695	AFP	AFP	–	INTERPRO	Unnamed protein product

curve for all features was 0.87 and for the top 25 features was 0.89, respectively.

3.3. Performance of AFP-Pred, BLAST and HMM

To test the capability, our algorithm was evaluated by an independent dataset obtained from INTERPRO and KEGG databases (Hunter et al., 2009; Kanehisa and Goto, 2000). The sequences that are present in the positive training dataset were removed from the list. Finally, we got 16 proteins which are annotated as "antifreeze proteins" (Table 3). Our approach correctly predicted 15 proteins as antifreeze proteins. The performance of our algorithm was compared with PSI-BLAST and HMM (Altschul et al., 1997; Eddy, 1998). PSI-BLAST search for each sequence was carried out against the swissprot database with an E value of 0.1 HMM search for each query sequence was performed against the HMM profile obtained from Pfam database (Pfam release 23) (Sonnhammer et al., 1997). Out of 16 proteins, BLAST search retrieved antifreeze protein hits from swissprot database for only 9 proteins. No hits were found for the remaining 7 proteins. Similarly, HMM search against Pfam database returned no hits for 11 proteins. As seen in Table 3, AFP-Pred, BLAST and HMM predicted 15, 9 and 5 proteins, respectively. This result indicates that AFP-Pred is a useful approach to predict AFPs from sequence information in the absence of sequence similarity. Out of 16 proteins, 3 proteins are annotated as "antifreeze proteins" and the remaining 14 proteins are annotated as "unnamed protein product" or "hypothetical proteins" in NCBI database. AFP-Pred correctly predicted all the hypothetical proteins as antifreeze proteins. This shows that AFP-Pred can be efficiently used to annotate hypothetical proteins.

3.4. Comparison with other methods

The proposed random forest method was compared with several state-of-the-art classifiers such as SVM, Naïve Bayes, MLP and K-nearest neighbor classifiers (George and Langley, 1995; Aha and Kibler, 1991; Vapnik, 1998). The performance comparison of these different classifiers was obtained by 5-fold cross-validation (Table 4). We compared the performance of AFP-Pred with the other models using the same feature subsets. All models were tested on the test dataset containing 181 positive and 9193 negative sequences. The prediction accuracy of random forest is about 7% and 6% higher than Naïve Bayes and K-nearest neighbor classifiers, respectively. Although the performance of random forest, SVM and MLP is comparable, there is a slight improvement in the sensitivity and specificity values of random forest classifier.

Table 4
Comparison of AFP-Pred with other machine learning methods.

Method	Sensitivity (%)	Specificity (%)	MCC	Accuracy (%)
Naïve Bayes	66.6	84.53	0.5233	76.44
MLP	80.00	80.66	0.6052	80.36
IBK	78.67	75.69	0.5413	77.04
SVM	82.67	80.11	0.6254	81.27
AFP-Pred	84.67	82.32	0.6674	83.38

This result shows that AFP-Pred can be used to predict antifreeze protein with higher accuracy.

4. Conclusion

Identification of antifreeze proteins from sequence databases is difficult due to poor sequence similarity. We reported a random forest based approach, AFP-Pred, for the prediction of antifreeze proteins from sequence using sequence derived properties. Very high prediction accuracies on the training and testing datasets show that AFP-Pred is potentially useful tool for the prediction of antifreeze from protein primary sequence. Because of its simplicity, 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. The AFP-Pred program and dataset is available at http://www3.ntu.edu.sg/home/EPNSugan/index_files/AFP-Pred.htm. Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful prediction methods (Chou and Shen, 2009), we shall make efforts in our future work to provide a web-server for the method presented in this paper.

Acknowledgments

KKK acknowledges the support by the Graduate School for Computing in Medicine and Life Sciences funded by Germany's Excellence Initiative [DFG GSC 235/1]. KKK acknowledges Dr. Kai-Uwe Kalies and Prof. Enno Hartmann, University of Luebeck, Germany for their support.

References

Aha, D., Kibler, D., 1991. Instance-based learning algorithms. Mach. Learn. 6, 37–66.
Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25 (17), 3389–3402.

Anand, A., Pugalenthi, G., Suganthan, P.N., 2008. Predicting protein structural class by SVM with class-wise optimized features and decision probabilities. J. Theor. Biol. 253 (2), 375–380.
Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32.
Breton, G., Danyluk, J., 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 (4), 373–376.
Chen, C., Chen, L.X., Zou, X.Y., Cai, P.X., 2008. Predicting protein structural class based on multi-features fusion. J. Theor. Biol. 253, 388–392.
Chen, C., Chen, L., Zou, X., Cai, P., 2009. Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine. Protein Peptide Lett. 16, 27–31.
Cheng, C.H., 1998. Evolution of the diverse antifreeze proteins. Curr. Opin. Genet. Dev. 8 (6), 715.
Chou, K.C., 1992. Energy-optimized structure of antifreeze protein and its binding mechanism. J. Mol. Biol. 223 (2), 509–517.
Chou, K.C., 2001. Prediction of protein cellular attributes using pseudo amino acid composition. Proteins: Struct. Funct. Genet. 43, 246–255.
Chou, K.C., 2005. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21, 10–19.
Chou, K.C., Cai, Y.D., 2005. Prediction of membrane protein types by incorporating amphipathic effects. J. Chem. Inf. Modeling 45, 407–413.
Chou, K.C., Shen, H.B., 2007. Review: recent progresses in protein subcellular location prediction. Anal. Biochem. 370, 1–16.
Chou, K.C., Shen, H.B., 2008. Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. Nat. Protocols 3, 153–162.
Chou, K.C., Shen, H.B., 2009. Recent advances in developing web-servers for predicting protein attributes. Nat. Sci. 1, 63–92.
Chou, K.C., Shen, H.B., 2010a. Cell-PLoc 2.0: an improved package of web-servers for predicting subcellular localization of proteins in various organisms. Nat. Sci. 2, 1090–1103.
Chou, K.C., Shen, H.B., 2010b. A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. PLoS ONE 5, e9931.
Chou, K.C., Shen, H.B., 2010c. Plant-mPLoc: a top-down strategy to augment the power for predicting plant protein subcellular localization. PLoS ONE 5, e11335.
Chou, K.C., Zhang, C.T., 1995. Review: prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. 30, 275–349.
Davies, P.L., Sykes, B.D., 1997. Antifreeze proteins. Curr. Opin. Struct. Biol. 7 (6), 828–834.
Davies, P.L., Baardsnes, J., Kuiper, M.J., Walker, V.K., 2002. Structure and function of antifreeze proteins. Philos. Trans. R. Soc. London B Biol. Sci. 357 (1423), 927–935.
Davies, P.L., Hew, C.L., 1990. Biochemistry of fish antifreeze proteins. FASEB J. 4, 2460–2468.
Ding, H., Luo, L., Lin, H., 2009. Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition. Protein Peptide Lett. 16, 351–355.
Du, P., Li, Y., 2008. Prediction of C-to-U RNA editing sites in plant mitochondria using both biochemical and evolutionary information. J. Theor. Biol. 253, 579–589.
Eddy, S.R., 1998. Profile hidden Markov models. Bioinformatics 14 (9), 755–763.
Ewart, K.V., Lin, Q., Hew, C.L., 1999. Structure, function and evolution of antifreeze proteins. Cell Mol. Life Sci. 55 (2), 271–283.
Frank, E., Hall, M., Trigg, L., Holmes, G., Witten, I.H., 2004. Data mining in bioinformatics using Weka. Bioinformatics 20, 2479–2481.
George, H.J., Langley, P., 1995. Estimating Continuous Distributions in Bayesian Classifiers. Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, pp. 338–345.
Graether, S.P., Kuiper, M.J., Gagne, S.M., Walker, V.K., Jia, Z., Sykes, B.D., Davies, P.L., 2000. Beta-helix structure and ice-binding properties of a hyperactive antifreeze protein from an insect. Nature 406, 325–328.
Griffith, M., Ewart, K.V., 1995. Antifreeze proteins and their potential use in frozen foods. Biotechnol. Adv. 13 (3), 375–402.
Griffith, M., Antikainen, M., Hon, W.C., Pihakaski-Maunsbach, K., Yu, X.-M., Chun, J.U., Yang, 1997. Antifreeze proteins in winter rye. Physiol. Plant. 100, 327–332.
Han, P., Zhang, X., Norton, R.S., Feng, Z.P., 2009. Large-scale prediction of long disordered regions in proteins using random forests. BMC Bioinf. 10, 8.
Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20, 832–844.
Ho, T.K., 2002. A data complexity analysis of comparative advantages of decision forest constructors. Pattern Anal. Appl. 5, 102–112.
Hua, L., Li, D.G., Lin, H., Li, L., Li, X., Liu., Z.C., 2010. The correlation of gene expression and co-regulated gene patterns in characteristic KEGG pathways. J. Theor. Biol. 266 (2), 242–249.
Huang, R.B., Du, Q.S., Wei, Y.T., Pang, Z.W., Wei, H., Chou, K.C., 2009. Physics and chemistry-driven artificial neural network for predicting bioactivity of peptides and proteins and their design. J. Theor. Biol. 256, 428–435.
Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R.D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A.F., Selengut, J.D., Sigrist, C.J., Thimma, M., Thomas, P.D., Valentin, F., Wilson, D., Wu, C.H., Yeats, C., 2009. InterPro: the integrative protein signature database. Nucleic Acids Res. 37 (Database Issue), D224–D228.
Ji, G., Wu, X., Shen, Y., Huang, J., Quinn, Li, Q., 2010. A classification-based prediction model of messenger RNA polyadenylation sites. J. Theor. Biol. 265, 287–296.
Jiang, X., Wei, R., Zhang, T.L., Gu, Q., 2008. Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy. Protein Peptide Lett. 15, 392–396.
Joshi, R.R., Sekharan, S., 2010. Characteristic peptides of protein secondary structural motifs. Protein Peptide Lett. 17, 1198–1206.
Kandaswamy, K.K., Pugalenthi, G., Hartmann, E., Kalies, K.U., Möller, S., Suganthan, P.N., Martinetz, T., 2010. SPRED: a machine learning approach for the identification of classical and non-classical secretory proteins in mammalian genomes. Biochem. Biophys. Res. Commun. 391, 1306–1311.
Kanehisa, M., Goto, S., 2000. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 27–30.
Kawashima, S., Ogata, H., Kanehisa, M., 1999. AAindex: amino acid index database. Nucleic Acids Res. 27, 368–369.
Kumar, K.K., Pugalenthi, G., Suganthan, P.N., 2009. DNA–Prot: identification of DNA binding proteins from protein sequence information using random forest. J. Biomol. Struct. Dyn. 26 (6), 679–686.
Lee, J.W., Lee, J.B., Park, M., Song., S.H., 2005. An extensive comparison of recent classification tools applied to microarray data. Comput. Stat. Data Anal. 48, 869–885.
Lewitt, J., 1980. Responses of Plants to Environmental Stresses, vol. 1. Academic Press, New York.
Li, F.M., Li, Q.Z., 2008. Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach. Protein Peptide Lett. 15, 612–616.
Li, S., Li, H., Li, M., Shyr, Y., Xie, L., Li, Y., 2009. Improved prediction of lysine acetylation by support vector machines. Protein Peptide Lett. 16, 977–983.
Liaw, A., Wiener, M., 2002. Classification and regression by random forest. R News 2, 18–22.
Li, W., Jaroszewski, L., Odzik, G.A., 2001. Clustering of highly homologous sequences to reduce the size of large protein database. Bioinformatics 17, 282–283.
Lin, H., 2008. The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. J. Theor. Biol. 252, 350–356.
Lin, H., Ding, H., Feng-Biao Guo, F.B., Zhang, A.Y., Huang, J., 2008. Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. Protein Peptide Lett. 15, 739–744.
Lin, Z.H., Wang, H.L., Zhu, B., Wang, Y.Q., Lin, Y., Wu, Y.Z., 2009. Estimation of affinity of HLA-A*0201 restricted CTL epitope based on the SCORE function. Protein Peptide Lett. 16, 561–569.
Liu, T., Zheng, X., Wang, C., Wang, J., 2010. Prediction of Subcellular location of apoptosis proteins using pseudo amino acid composition: an approach from auto covariance transformation. Protein Peptide Lett. 17, 1263–1269.
Logsdon, J.M., Doolittle, W.F., 1997. Origin of antifreeze protein genes: a cool tale in molecular evolution. Proc. Natl. Acad. Sci. USA 94 (8), 3485–3487.
Lu, J., Niu, B., Liu, L., Lu, W.C., Cai, Y.D., 2009. Prediction of small molecules' metabolic pathways based on functional group composition. Protein Peptide Lett. 16, 969–976.
Masso, M., Vaisman, I.I., 2010. Knowledge-based computational mutagenesis for predicting the disease potential of human non-synonymous single nucleotide polymorphisms. J. Theor. Biol. 266 (4), 560–568.
McGuffin, L.J., Bryson, K., Jones, D.T., 2000. The PSIPRED protein structure prediction server. Bioinformatics 16 (4), 404–405.
Mohabatkar, H., 2010. Prediction of cyclin proteins using Chou's pseudo amino acid composition. Protein Peptide Lett. 17, 1207–1214.
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. Grassland Sci. 41, 21–25.
Nanni, L., Lumini, A., 2009. A further step toward an optimal ensemble of classifiers for peptide classification, a case study: HIV protease. Protein Peptide Lett. 16, 163–167.
Pugalenthi, G., Kumar, K.K., Suganthan, P.N., Gangal, R., 2008. Identification of catalytic residues from protein structure using support vector machine with sequence and structural features. Biochem. Biophys. Res. Commun. 367, 630–634.
Qiu, J.D., Luo, S.H., Huang, J.H., Liang, R.P., 2009. Using support vector machines to distinguish enzymes: approached by incorporating wavelet transform. J. Theor. Biol. 256 (4), 625–631.
Sakai, A., Larcher, W., 1987. Frost Survival of Plants. Springer-Verlag, Heidelberg, Germany.
Scholander, P.F., VanDam, L., Kanwisher, J.W., Hammel, H.T., Gordon, M.S., 1957. Supercooling and osmoregulation in Arctic fish. J. Cell. Comp. Physiol. 49, 5–24.
Sformo, T., Kohl, F., McIntyre, J., Kerr, P., Duman, J.G., Barnes, B.M., 2009. Simultaneous freeze tolerance and avoidance in individual fungus gnats, *Exechia nugatoria*. J. Comp. Physiol. B. 179 (7), 897–902.
Shao, X., Tian, Y., Wu, L., Wang, Y.L.J., Deng, N., 2009. Predicting DNA- and RNA-binding proteins from sequences with kernel methods. J. Theor. Biol. 258, 289–293.
Shi, M.G., Huang, D.S., Li, X.L., 2008. A protein interaction network analysis for yeast integral membrane protein. Protein Peptide Lett. 15, 692–699.
Sonnhammer, E.L., Eddy, S.R., Durbin, R., 1997. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 28 (3), 405–420.
Tian, F., Lv, F., Zhou, P., Yang, Q., Jalbout, A.F., 2008. Toward prediction of binding affinities between the MHC protein and its peptide ligands using quantitative structure-activity relationship approach. Protein Peptide Lett. 15, 1033–1043.
Urrutia, M.E., Duman, J.G., Knight, C.A., 1992. Plant thermal hysteresis proteins. Biochim. Biophys. Acta 1121 (1–2), 199–206.
Uriarte, R.D., Andres, S.A., 2006. gene selection and classification of microarray data using random forest. BMC Bioinformatics 3.
Vapnik, V., 1998. Statistical Learning Theory. Wiley-Interscience, New York, USA.

Vilar, S., Gonzalez-Diaz, H., Santana, L., Uriarte, E., 2009. A network-QSAR model for prediction of genetic-component biomarkers in human colorectal cancer. J. Theor. Biol. 261, 449–458.
Wang, T., Xia, T., Hu, X.M., 2010. Geometry preserving projections algorithm for predicting membrane protein types. J. Theor. Biol. 262, 208–213.
Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., Ward, D., Williams, K., Zhao, H., 2003. Comparison of statistical methods for classification of ovarian cancer using a proteomics dataset. Bioinformatics 19, 1636–1643.
Yoshida, M., Abe, J., Moriyama, M., Shimosakawa, 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.
Yang, J.Y., Peng, Z.L., Yu, Z.G., Zhang, R.J., Anh, V., Wang, D., 2009. Prediction of protein structural classes by recurrence quantification analysis based on chaos game representation. J. Theor. Biol. 257, 618–626.
Yang, X.Y., Shi, X.H., Meng, X., Li, X.L., Lin, K., Qian, Z.L., Feng, K.Y., Kong, X.Y., Cai, Y.D., 2010. Classification of transcription factors using protein primary structure. Protein Peptide Lett. 17, 899–908.
Yu, X.M., Griffith, M., 2001. Winter rye antifreeze activity increases in response to cold and drought, but not abscisic acid. Physiol. Plant. 112, 78–86.
Zhang, Y., Ding, C., Li, T., 2008. Gene selection algorithm by combining reliefF and mRMR. BMC Genomics 9 (2), S27.
Zeng, Y.H., Guo, Y.Z., Xiao, R.Q., Yang, L., Yu, L.Z., Li, M.L., 2009. Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. J. Theor. Biol. 259, 366–372.
Zhou, G.P., 1998. An intriguing controversy over protein structural class prediction. J. Protein Chem. 17, 729–738.
Zhou, G.P., Assa-Munt, N., 2001. Some insights into protein structural class prediction. Proteins: Struct. Funct. Genet. 44, 57–59.
Zhou, G.P., Doctor, K., 2003. Subcellular location prediction of apoptosis proteins. Proteins: Struct. Funct. Genet. 50, 44–48.
Zhou, X.B., Chen, C., Li, Z.C., Zou, X.Y., 2007. Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. J. Theor. Biol. 248, 546–551.