Motif analysis and identification of antifreeze protein sequences

Huan Wen
Laboratory of Theoretical Biophysics,
School of Physical Science and Technology, Inner Mongolia University,
Hohhot, China
e-mail:hwenpy@163.com

Jun-Jie Liu*, Qian-Zhong Li
Laboratory of Theoretical Biophysics,
School of Physical Science and Technology, Inner Mongolia University,
Hohhot, China
*e-mail: pyljj@imu.edu.cn

***Abstract*—Antifreeze proteins (AFPs), which are also known as thermal hysteresis proteins, are ice-binding proteins. AFPs can adsorb to ice crystal surface and inhibit the growth of ice crystals in solution. But the interaction between AFPs and ice crystal is not known completely. Analyzing physicochemical characteristics of AFPs sequences is very significant to understand the ice-protein interaction. Through the analysis of the sequence motif by MEME, hydrophobic amino acids shown blue are most. According to the hydropathy, acid-base property, the chemical structure of the R group of amino acid and the polarity of the amino acids, the amino acids are respectively divided into 6 groups, 3 groups, 6 groups, 4 groups. In this study, based on the n-Peptide compositions and these physicochemical characteristics, an algorithm of Support Vector Machine (SVM) is proposed for predicting antifreeze proteins. The best results of the jackknife test show that the sensitivity, the specificity, the overall identification accuracy and the Mcc value are 93.14%, 96.08%, 94.62% and 0.8927, respectively. The hydropathy and the chemical structure of the R group of amino acid are important physicochemical characteristics for identifying AFPs.**

Keywords-motif; AFPs; physicochemical characteristics; SVM; amino acid composition

I. INTRODUCTION

Some organisms 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 were first identified in Antarctic teleost fishes, which were found to lower the freezing point by more than 1°C to match that of seawater. Later studies demonstrate AFPs are ice-binding proteins. Because of the interaction with the ice crystal within the interfacial region, AFPs inhibit the growth of ice crystal and depress the freezing point of solution below the melting point[1]. The difference between the freezing and melting temperatures is referred to as thermal hysteresis. Therefore, AFPs is also called thermal hysteresis protein. The binding sites of AFPS and ice are relatively flat. The portion of the protein exposed to the solvent are also somewhat hydrophobic [2].With the development of scientific technology and test conditions, different types of antifreeze proteins are isolated from the organisms of ocean fishes, terrestrial insects, plants, bacteria, fungi and algae[3].

AFPs found in fish were categorized into five classes according to their structural diversity, namely AFGPs, AFP I, AFP II, AFP III and AFPIV. AFGPs are made up of many tandem repeats of 3 peptides [-Ala-Ala-Thr (disaccharide -based)-][4].AFPI is the 37 amino acid long high performance ance liquid chromatography (HPLC)-6, which is typical of the isoforms that contain three 11 amino acid repeats of Thr–X2–Asx–X7 where X is generally alanine[5,6]. AFP II comprises two twisted anti-parallel β-sheets with a helix on either side. All cysteines of AFP II are paired to form disulfide bonds [7]. AFP III comprises a compact, globular, single domain with two internal tandem motifs, which includes four short β-strands and a one-turn helix. The motifs are connected by a 15-residue loop that contains a two-turn helix[8]. AFP IV are helix-bundle protein.

Insect AFPs from Choristoneura fumiferana retains the extremely regular left-handed β-helical structure. The β-helix is formed by a series of 15-amino acid turns, which result in an elongated protein with a triangular cross-section[9].

The first crystal structure of an Antarctic bacterial AFP, Which is the largest AFP structure, folds as a Ca^{2+}-bound parallel beta-helix with an extensive array of ice-like surface waters that are anchored via hydrogen bonds directly to the polypeptide backbone and adjacent side chains[10].

So far, there is only a crystal structure about plant AFPs. It folds as a novel left-handed beta-roll with eight 14- or 15-residue coils and is stabilized by a small hydrophobic core and two internal Asn ladders. The ice-binding site is formed by a flat beta-sheet on one surface of the beta-roll[11].

The Motif analysis and identification of antifreeze proteins is helpful for us to study the antifreeze mechanism of the AFPs. Sequence motifs are short conservative patterns, they maybe connective with biological function. These sequence motifs will be important for searching ice-binding site. With the rapid increase of sequenced genomic data, the need for an effective tool to recognize AFPs becomes also increasingly important. In this paper,

Support Vector Machine is used to recognize AFPs by analyzing the n-peptide compositions of antifreeze protein sequences and their physical and chemical characteristics.

II. DATASET

A. The positive dataset

The positive dataset was constructed from the UniProtKB database [12]. We respectively restricted the term “antifreeze” and “thermal hysteresis protein” to the protein

name.612 antifreeze proteins and 22 thermal hysteresis proteins were obtained. Removing 2 duplicate proteins, 632 proteins were got. To construct a good quality benchmark dataset, a winnowing procedure was taken to remove those proteins which were annotated with "predicted", "putative" and ''antifreeze-like proteins'' in the protein name field [13]. The final positive dataset contained 597 non-redundant antifreeze proteins. The unclassified proteins are also removed .According to the organism species, these 594 antifreeze proteins are divided into fish, insect, plant, bacteria, fungi and algae. Distribution of the 594 AFP sequences by the types of organism is shown in TABLE I.

TABLE I. DISTRIBUTION OF THE 594 AFP SEQUENCES BY THE TYPES OF ORGANISM

Organism	Number of Proteins
fish	158
insect	117
plant	29
bacteria	254
fungi	17
algae	19

B. *Motif analysis of antifreeze protein sequences from the positive dataset*

MEME is a tool for discovering motifs in a group of related protein or DNA sequences. MEME represents motifs as position-dependent letter-probability of each possible letter at each position in the pattern. Individual MEME motifs do not contain gaps. The amino acids of different physicochemical characteristic are shown by different color in the MEME motifs. Patterns with variable-length gaps are split by MEME into two or more separate motifs. MEME uses statistical modeling techniques to automatically choose the best width, number of occurrences, and description for each motif[14]. We use MEME (Multiple Em for Motif Elicitation) search for sequence motif in sets of antifreeze protein sequences from fish, insect, plant, bacteria, fungi and algae. In the insect AFPs, there are two typical motifs, which are shown in Figure 1. One contains three 12 amino acid repeats of GTXSXXCXXAXT where X is uncertain amino acid, the other motif from 12 to 23 shown blue are most hydrophobic.In the AFGP from fish, there is a motif that contains two 12 amino acid repeats of TPALIFAATAAT in Figure 2.

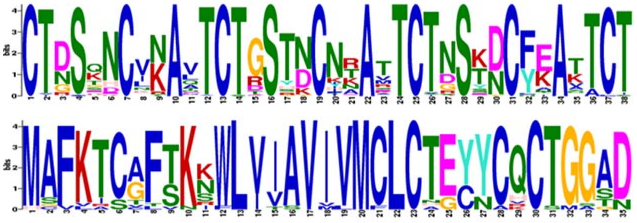

Figure 1. Sequence motifs of insect antifreeze protein

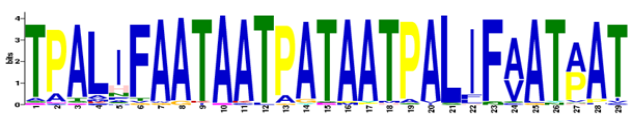

Figure 2. Sequence motif of AFGP from fish

C. *The negative dataset*

The negative dataset was constructed from the PISCES server [15]. To include as many representative structures as possible, the non-AFPs had <25% sequence identity, a crystallographic resolution of at least 2 Å, R-factors of 0.25 and the length of sequence of 40 to 1000. Removing the AFPs, we contained 5124 non-AFPs. In order to maintain the balance of the positive set and negative set, 600 non-AFPs were randomly extracted to be used to calculate.

III. METHODS

A. *The n-peptide compositions of amino acid sequence*

The description of a protein sequence can be based on the n-peptide composition coding (denoted by An). In case of n =1, the A_1 coding is amino acid composition, which can be considered as the first-order approximation to the complete protein sequence. For n= 2, the A_2 coding gives the dipeptide composition. As n increases, the coding provides progressively more detailed sequential information. But in the case of n>3, the number of information becomes too much, and computation becomes not only impractical but also susceptible to the danger of overfitting. So we chose the case o f n≤2.

B. *The physicochemical characteristics of amino acid sequence*

Through the analysis of the sequence motif, hydrophobic amino acids shown blue are most. We can overcome the size problem by regrouping the amino acids into smaller number of classes based on their physicochemical characteristics [16]. The classifications of Amino acids are shown in TABLE II .According to the hydropathy(denoted by H), 20 amino acids were reduced to 6 groups, namely strongly hydrophilic (RDENQKH), strongly hydrophobic (LIVAMF), weakly Hydrophilic or weakly hydrophobic (STYW), proline (P), cysteine (C) and clutamic acid (G) [17]. According to the acid-base property (denoted by B), 20 amino acids were reduced to 3 groups, namely acidic (DE), alkaline (RKH) and neutral (NQLIVAMFSTYWPGC). According to the chemical structure of the R group of amino acid (denoted by S), 20 amino acids were reduced to 6 groups, namely aromatic (FYW), heterocyclic (HP), aliphatic and neutral (GAVLI), aliphatic containing hydroxyl or sulfur (STCM), aliphatic and alkaline (RK), aliphatic containing carboxyl or amide (DENQ). According to the polarity (denoted by P), 20 amino acids were reduced to 4 groups, namely Non-polarity (AVLIFWMP), polarity and positive charge (RKH), polarity and negative charge(DE), polarity and non-charge (STYGNQC).

TABLE II. THE CLASSIFICATIONS OF AMINO ACIDS BASED ON PHYSICOCHEMICAL PROPERTIES.

	Chassification property	Amino acid
H	strongly hydrophilic	RDENQKH

	strongly hydrophobic	LIVAMF
	weakly hydrophilic or weakly hydrophobic	STYW
	proline	P
	cysteine	C
	clutamic acid	G
B	acidic	DE
	alkaline	RKH
	neutral	NQLIVAMFSTYWPGC
S	aromatic	FYW
	heterocyclic	HP
	aliphatic and neutral	GAVLI
	aliphatic containing hydroxyl or sulfur	STCM
	aliphatic and alkaline	RK
	aliphatic containing carboxyl or amide	DENQ
P	Non-polarity	AVLIFWMP
	polarity and positive charge	RKH
	polarity and negative charge	DE
	polarity and non-charge	STYGNQC

C. *Support vector machine*

In recent years, support vector machine (SVM) has been widely used in various recognition problems. In this work, the SVM is used to identify antifreeze protein. All SVM calculations were performed using LibSVM [18]. The publicly available LibSVM software is developed by Lin's lab, and it can be freely downloaded from: http://www.csie.ntu.edu.tw/~cjlin/libsvm.

IV. RESULTS AND DISCUSSION

A. *performance assessment*

In order to evaluate the correct prediction rate and reliability of a predictive method, the sensitivity (Sn), the specificity (Sp)，the overall identification accuracy(Acc) and the Matthews Correlation Coefficient (MCC) are defined by

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

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

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

$$Mcc = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TN + FN) \times (TP + FN) \times (TN + FP)}}$$

where TP represents the number of the correctly recognized AFPs, FN represents the number of AFPs recognized as non-AFPs, TN represents the number of the correctly recognized non-AFPs, and FP represents the number of non-AFPs recognized as AFPs.

B. *Result and discussion*

In statistical prediction, the following three cross-validation tests are often used to examine the power of a predictor: independent dataset test, sub-sampling test and jackknife test. Of these three, the jackknife test is accepted as the most rigorous and objective one, and hence has been used by more and more investigators in examining the power of various prediction methods. Only a characteristic is considered, all the predictive results are shown in TABLE III. In this table, A1 indicates amino acid composition; A2 indicates dipeptide composition; H1 indicates amino acid hydropathy composition; H2 indicates hydropathy dipeptide composition;B1 indicates amino acid acid-base property composition; B2 indicates acid-base property dipeptide composition; S1 indicates amino acid composition based on the chemical structure of the R group of amino acid; S2 indicates dipeptide composition based on the chemical structure of the R group of amino acid. P1 indicates amino acid polarity composition; P2 indicates polarity dipeptide composition.

TABLE III. IDENTIFICATION OF AFP BY SVM BASED ON A CHARACTERISTIC IN THE JACKKNIFE TEST

characteristics	Sn(%)	Sp(%)	Acc(%)	Mcc
A1	**91.75**	**93.80**	**93.66**	**0.8740**
A2	90.29	96.83	93.58	0.8734
H1	88.44	88.28	88.36	0.7673
H2	92.48	91.61	92.04	0.8409
B1	78.11	84.78	81.46	0.6305
B2	80.47	88.17	84.34	0.6886
S1	83.22	85.94	84.59	0.6921
S2	87.88	95.67	91.79	0.8383
P1	80.70	86.06	83.39	0.6689
P2	85.86	90.50	88.19	0.7645

Only a characteristic is considered. The best predictive results are obtained by selecting the amino acid composition as imputing parameters. The results of jackknife test show that the sensitivity, specificity, the overall identification accuracy and the Mcc value are 91.75%, 93.80% , 93.66% and 0.8740, respectively. The Acc and Mcc obtained with the dipeptide composition as input feature was similar to that obtained with the amino acid composition. The performance of SVM obtained with the physicochemical characteristics as input feature is lower than that obtained with the amino acid composition and the dipeptide composition, respectively. However, the result obtained with hydropathy and the chemical structure of the R group of amino acid are relatively high. And for the same physicochemical characteristics, the results obtained with dipeptide composition of the physicochemical characteristics is higher than that of amino acid composition

In order to enhance the identification of antifreeze proteins, we also consider many characteristics as imputing parameters. The results are shown in the TABLE IV.The best predictive results are obtained by selecting the characteristic (A1+H1+H2+S1+S2)as imputing parameters. The results of jackknife test show that the sensitivity, specificity and Mcc value are 93.14%, 96.08% 94.62% and 0.8927, respectively. It indicated that the amino acid composition and dipeptide composition of the hydropathy and the chemical structure of the R group of amino acid can improve the identification of antifreeze proteins.

TABLE IV. IDENTIFICATION OF AFP BY SVM BASED ON MANY CHARACTERISTICS IN THE JACKKNIFE TEST

characteristics	Sn(%)	Sp(%)	Acc(%)	Mcc
A1+H1	90.74	96.83	93.80	0.8776
A1+S1	91.41	95.33	93.38	0.8683
A1+H1+ S1	93.10	93.83	93.47	0.8694
A1+H1+S1+B1+P1	92.76	94.83	93.80	0.8762
A1+H1+H2+S1+S2	**93.14**	**96.08**	**94.62**	**0.8927**
A1+H1+S1+B1+P1 +H2+ S2+B2+P2	92.42	96.00	94.41	0.8889

C. Conclusion

Choosing a set of reasonable information parameters from protein sequence is very helpful for predicting the antifreeze protein. The hydropathy and the chemical structure of the R group of amino acid are important physicochemical characteristics for identifying AFPs. AFPs are also predicted with high level of accuracy, from their primary amino acid sequence. It also evidences that the primary sequence contains important information which determines protein advance structure. Through the analysis of AFPs amino acid sequence motifs by MEME, we can further identify the key functional residues of the ice-binding surfaces. For the same physicochemical characteristics, the predictive results of dipeptide compositions are relatively higher than that obtained by amino acid composition.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China（31106188）and Technology Foundation of Ministry of Education of China (211030).

REFERENCES

[1] Y.Yeh and R.E.Feeney,“Antifreeze Proteins: Structures and Mechanism of Function,”Chemical Reviews,Vol. 96(2), pp. 601-618, March 1996.

[2] A.C.Doxey,M.W.Yaish,M.Griffith and B.J.McConkey, “Ordered surface carbons distinguish antifreeze proteins and their ice-binding regionsm,” Nature Biotechnology, Vol. 24, pp. 852-855, July 2006.

[3] K.K.Kandaswamy,K.C.Chou, and T.Martinetz etal. “AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties,” Journal of Theoretical Biology vol. 270, pp. 56-62, February 2011.

[4] L.Chen,A.L.Devries and C.H.C.Cheng, “Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish,” Proc. Natl. Acad. Sci. USA, Vol. 94, pp. 3811-3816, April 1997.

[5] S. N. Patel and S. P. Graether, “Structures and ice-binding faces of the alanine-rich type I antifreeze proteins,”Biochem. Cell Biol.， Vol. 88,pp. 223-229, March 2010.

[6] J.J.Liu and Q.Z.Li, “A teoreticalodel on thermal hysteresis activity of the winter flounder protein ‘HPLC-6’ ,” Chemical Physics Letters， vol. 378 , pp. 238-243, July 2003.

[7] Y.Liu,Z.Li,Q.Lin,J.Kosinski and J.Seetharaman, “Structure and Evolutionary Origin of Ca^{2+}-Dependent Herring Type II Antifreeze Protein,”PLoS ONE,vol. 2(6): e548, June 2007.

[8] T.P.Ko,H.Robinson,Y.G.Gao,C.H.C.Cheng and A.L.DeVries etal., “The Refined Crystal Structure of an Eel Pout Type III Antifreeze Protein RD1 at 0.62-Å Resolution Reveals Structural Microheterogeneity of Protein and Solvation,”Biophysical Journal,Vol.84, pp.1228-1237, February 2003.

[9] E.K.Leinala,P.L.Davies,D.Doucet,M.G.Tyshenko and V.K.Walker etal.,“A β-Helical Antifreeze Protein Isoform with Increased Activity,”The American Society for Biochemistry and Molecular Biology, Vol. 277,pp. 33349-33352, September 2002.

[10] C.P.Garnham,R.L.Campbell and P.L.Davies, “Anchored clathrate waters bind antifreeze proteins to ice,” PNAS, vol. 108, pp. 7363-7367, May 2011.

[11] A.J.Middleton,C.B.Marshall,and Frédérick Faucher,etal., “Antifreeze Protein from Freeze-Tolerant Grass Has a Beta-Roll Fold with an Irregularly Structured Ice-Binding Site,”J.Mol.Biol., vol. 416, pp.713-724, January 2012.

[12] A.Bairoch,R.Apweiler,C.H.Wu and W.C.Barker etal., “The Universal Protein Resource (UniProt), ”Nucleic Acids Research, Vol. 33, pp. D154-D159, October 2005.

[13] C. S. Yu and C. H. Lu, “Identification of Antifreeze Proteins and Their Functional Residues by Support Vector Machine and Genetic Algorithms based on n-Peptide Compositions,”PLoS ONE,vol. 6(5):e20445, May 2011.

[14] Timothy L. Bailey, Nadya Williams1, Chris Misleh1 and Wilfred W. Li, “MEME: discovering and analyzing DNA and protein sequence motifs,” Nucleic Acids Research, Vol. 34, Web Server issue W369–W373,2006.

[15] G.Wang and R.L.Dunbrack Jr., “PISCES: a protein sequence culling server,”Bioinformatics，Vol. 19(12), pp. 1589-1591, March 2003.

[16] C.S.Yu,Y.C.Chen,Y.C.Chen,C.H.Lu and J.K.Hwang, “Prediction of Protein Subcellular Localization,” PROTEINS: Structure, Function, and Bioinformatics, vol.64, pp.643-651, June 2006.

[17] Y.L.Chen and Q.Z.Li, “Prediction of the subcellular location of apoptosis proteins,” Journal of Theoretical Biology, vol. 245, pp. 775-783, 2006.

[18] C.C. Chang and C.J. Lin, LIBSVM: a library for support vector machines.Software available from http://www.csie.ntu.edu.tw/~cjlin/ Libsvm.