ISSN 0974 - 9373

Vol. 15(2011) Special Issue 2 **Journal of International Academy of Physical Sciences** pp. 507-513

Accurate Prediction of Fish Antifreeze Proteins Using Artificial Neural Networks*

Abhigyan Nath and Radha Chaube
Bioinformatics section
MMV, Banaras Hindu University, Varanasi -22100
Email: radhachaube72@gmail.com

S. Karthikeyan
Department of Computer Science
Banaras Hindu University, Varanasi-221005
Email: karthinikita@gmail.com

(Received June 14, 2011)

Abstract: Antifreeze proteins (AFP) prevent the growth of ice-crystal in order to enable certain organism to survive under sub-zero temperature surroundings. These AFPs have evolved from different types of proteins and having very different sequences and structure. But they all perform the same function and become the classical example of convergent evolution. Inspired by the success of machine learning algorithms we used ANN for their prediction. A feature vector was prepared using different physicochemical property groups of amino acids, amino acid composition and dipeptide composition. Though our AFP dataset was small, the ANN is able to correctly classify the AFPs and non-AFPs. A larger dataset, incorporation of structural information and better selection of amino acid physicochemical properties for making the feature vectors will further validate better accuracy in prediction of AFPs using ANN.
Keywords: Antifreeze Proteins, Artificial Neural Networks, physicochemical property groups, amino acid composition, and dipeptide composition

1. Introduction

Antifreeze proteins or the ice structuring proteins or thermal hysteresis proteins are a diverse group of proteins which inhibit the growth of ice crystals[1]. These proteins have evolved as an adaptation to cold temperatures and are found in different organisms like fish, insect, bacteria, fungi and plants[2]. Antifreeze proteins results in non-colligative (i.e. lowering of

*Paper presented in CONIAPS XIII at UPES, Dehradun during June 14-16, 2011.

freezing point is not in proportion to its concentration), non-equilibrium lowering of the freezing point of the extracellular fluids to safe level[3]. Based on their origins and properties, fish AFPs have been classified into five distinct types –AFGP, Type I, Type II, Type III and Type IV, each of them are unrelated and possesses distinct characteristics both in structure and sequence composition , although all of them perform the same function of antifreeze activity. These proteins were first discovered by De Varies in the blood plasma of marine teleosts. AFPs play an important role in protecting the fishes from freezing in ice-laden sea water[4]. During geologically recent cooling and glaciations events, there might be intense selective pressure to avoid freezing under progressively cooler conditions. These environmental changes must have favored any means of lowering the freezing point to avoid any physiological changes and that a number of different proteins have adapted to the task of antifreeze .The surprising diversity and distribution of the AFPs has led to the hypothesis that they have each evolved recently and independently as an adaptation to cooling during freezing of Antarctic oceans about 10-30 million years ago and Arctic oceans about 1-2 million years ago[5].

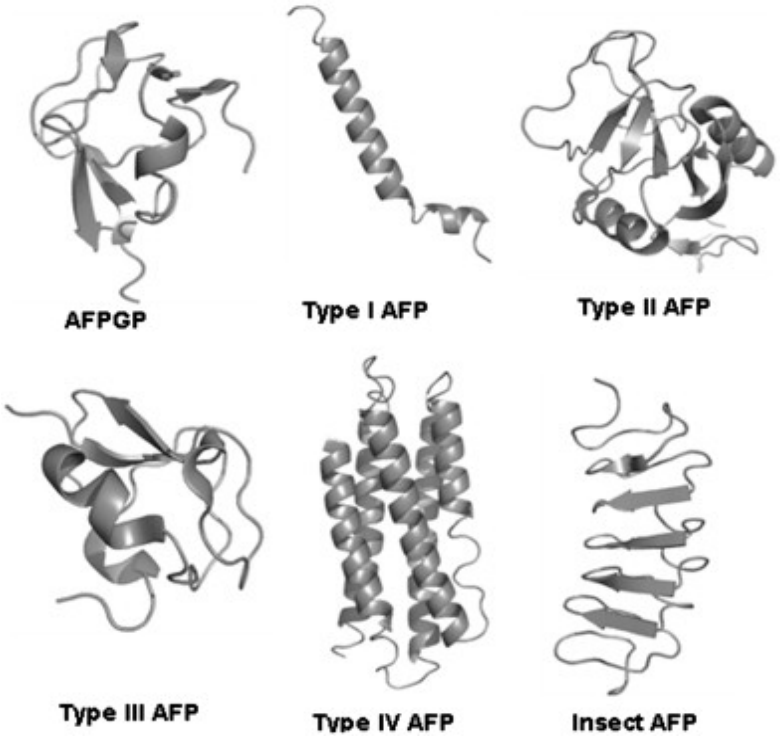

Figure 1: Types of antifreeze proteins[6]

2. Background

AFPs are a remarkable example of parallel and convergent evolution. A series of different proteins have independently evolved a common function (ice binding) despite having no amino acid or sequence similarity among them. There is an absence of a consensus ice binding motif which has made a PCR based study of these proteins impossible [7]. 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. A popular similarity search program such as BLAST [8] fails to detect putative antifreeze proteins. Inspired by the success of machine learning algorithms we used ANN for their prediction.

3. Method

A. Dataset

The dataset of Kandaswamy et al[6] is used for our experiment. The data set contains 481 Antifreeze Proteins and 481 non-Antifreeze Proteins. All these AFPs were having less than 40% sequence identity with each other.

The training set was created with 674 proteins .The validation and the testing set consists of 144 proteins each. All the three dataset consist equal number of both antifreeze and non-antifreeze proteins.

B. Selection of Features Vectors

The success of any prediction method depends on the quality of the input data. The quality is related to extraction of the relevant features from the input data. Here we had taken three features from the protein sequences to create input feature vector. The three features are amino acid composition, residues property groups and dipeptide counts. So each sequence is encoded by 431 input features as listed below:

Table 1: Selected Features

Name of the Feature	Size
Different Residues	20
Residues Property Groups	11
Dipeptide counts	400
Total	431

We have considered 11 property groups in generating the input feature vector. The following table provides the details regarding the property

groups in which an amino acid belongs. Some of the amino acids fall in more than one property group[6].

Table 2. Residue Property Groups

Residue Group	Residues in the Specific Group
Tiny amino Residues	Ala, Cys, Gly, Ser, Thr
Small Residues	Ala, Cys, Asp, Gly, Asn, Pro, Ser, Thr and Val
Aliphatic Residues	Ile, Leu and Val.
Non-polar Residues	Ala, Cys, Phe, Gly, Ile, Leu, Met, Pro, Val, Trp and Tyr
Aromatic Residues	Phe, His, Trp and Tyr
Polar Residues	Asp, Glu, His, Lys, Asn, Gln. Arg, Ser, and Thr.
Charged Residues	Asp, Glu, His, Arg, Lys
Basic Residues	His, Lys and Arg
Acidic Residues	Asp and Glu
Hydrophobic Residues	Ala, Cys, Phe, Ile, Leu, Met, Val, Trp, Tyr
Hydrophilic Residues	Asp, Glu, Lys, Asn, Gln

C. Classification protocol

Artificial neural networks have been applied successfully for classification and prediction in biological data. We used scaled conjugate gradient Back-error-propagation algorithm [10] for our prediction and used 70% of our data for training and 15% of the data for validation and the remaining 15% for testing.

D. Evaluation parameter

The performance of artificial neural network prediction method used in our study was computed by using sensitivity, specificity, overall accuracy, using the 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

$$Sensitivity = \frac{TP}{(TP + FN)}.$$

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

$$Specificity = \frac{TN}{(TN + FP)}.$$

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

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

4. Results and Discussion

Experiment was conducted by training the neural network using scaled conjugate gradient backpropogation algorithm with 674 samples and validating with 144 samples. The testing with 144 samples, we achieved an overall accuracy of 85.4%, sensitivity of 86.9% and specificity of 85%.

We have also investigated the performance of our prediction method by plotting Receiver Operating Characteristic (ROC) curves derived from the sensitivity and specificity values.

Results

	Samples	MSE	%E
Training:	674	9.28629e-2	13.05637e-0
Validation:	144	1.00177e-1	15.27777e-0
Testing:	144	1.40916e-1	20.83333e-0

Fig.2.Training Parameters

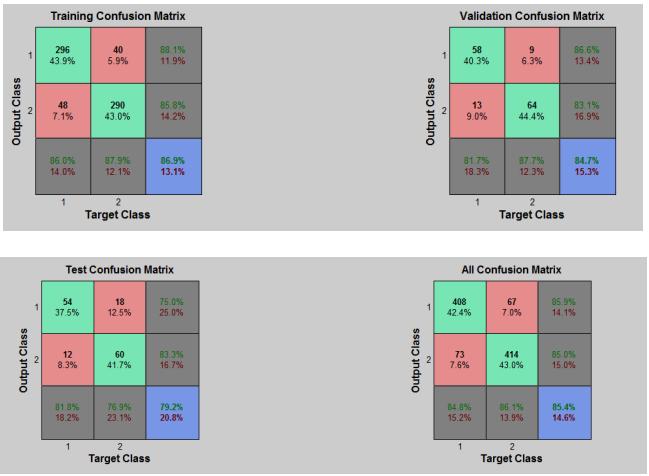

Fig.3.Confusion Matrix

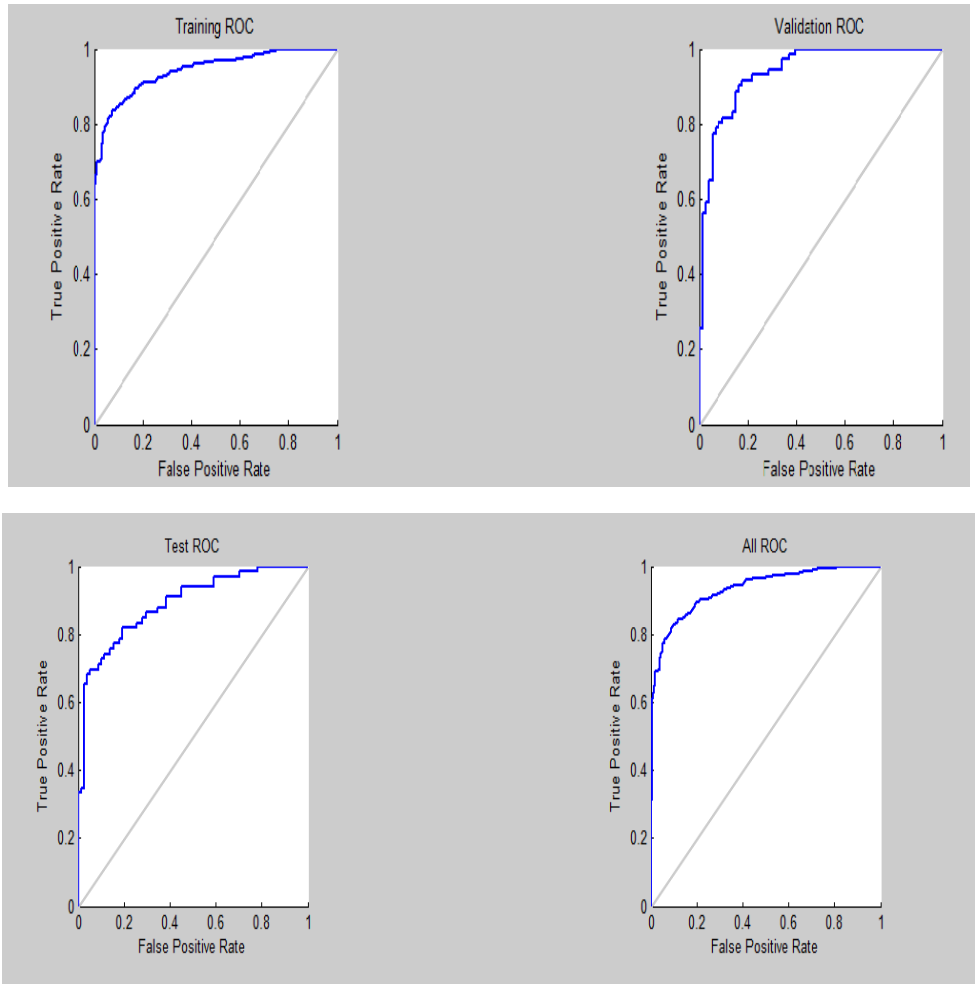

Fig.4. ROC curves

The confusion matrices and the ROC curve obtained during the training, validating, and testing are shown in the figure 3 and figure 4

5. Conclusion

We have got enhanced accuracy using three input features namely amino acid composition, 11 residues property groups and dipeptide counts. This accuracy can further be enhanced by using a larger dataset, incorporation of structural information (which is presently limited due to paucity of AFP structures) and better selection of physicochemical properties of amino acids for making the feature vector in predicting AFPs. We will further explore the possibilities of application of various feature selection techniques to reduce

the dimensionality of the feature vectors to avoid over-fitting data, which we believe will improve the classification accuracy.

References

1. Y. Yeh and R. E. Feeney, Antifreeze proteins, Structures and mechanisms of function, *Chemical Reviews*., **96**(1996) 601-617.
2. P. L. Davies and B. D. Sykes, Antifreeze proteins, *Current Opinion in Structural Biology*, **7**(1997) 828-834.
3. P. L. Davies, J. Baardsnes, M. J. Kuiper and V. K. Walker, Structure and function of antifreeze proteins, *Phil. Trans. R. Soc. Lond*., **B 357**(2002) 927-935.
4. G. L. Fletcher, C. L. Hew and P. L. Davis, Antifreeze Proteins of Teleost Fishes, *Annu Rev. Physiol*, **63**(2001) 359-90.
5. J. M. Logsdon, Jr. and W. Ford Doolittle, Origin of antifreeze protein genes: A cool tale in molecular evolution, *Proc. Natl. Acad. Sci*., **94**(1997)3485-3487.
6. AFP-Pred, A random forest approach for predicting antifreeze proteins from sequence-derived properties Krishna Kumar Kandaswamy, Kuo-Chen Chou, Thomas Martinetz, Steffen Möller, P.N. Suganthan, S. Sridharan and Ganesan Pugalenth.
7. J. Barret, Thermal hysteresis proteins, *The International Journal of Biochemistry & Cell Biology*., **33**(2001)105-117.
8. F. S. Altschul et al.,.BLAST., *J. Mol.Biol*., **215** (1990) 403-410.
9. S. Jahandideh, P. Abdolmaleki, M. Jahandideh and A. E. Barzegari, Sequence and structural parameters enhancing adaptation of proteins to low temperatures, *J. Theor. Biol*., **12(1)** (2007)159-166.
10. Stuart Russell and Peter Norvig, *Artificial Intelligence A Modern Approach*. p. 578.