AFP-CKSAAP: Prediction of Antifreeze Proteins Using Composition of k-Spaced Amino Acid Pairs with Deep Neural Network

Muhammad Usman and Jeong A Lee
Chosun University, Korea
Email: usman@chosun.kr, jalee@chosun.ac.kr

***Abstract*—Antifreeze proteins (AFPs) are the sub-set of ice binding proteins indispensable for the species living in extreme cold weather. These proteins bind to the ice crystals, hindering their growth into large ice lattice that could cause physical damage. There are variety of AFPs found in numerous organisms and due to the heterogeneous sequence characteristics, AFPs are found to demonstrate a high degree of diversity, which makes their prediction a challenging task. Herein, we propose a machine learning framework to deal with this vigorous and diverse prediction problem using the manifolding learning through composition of k-spaced amino acid pairs. We propose to use the deep neural network with skipped connection and ReLU non-linearity to learn the non-linear mapping of protein sequence descriptor and class label. The proposed antifreeze protein prediction method called AFP-CKSAAP has shown to outperform the contemporary methods, achieving excellent prediction scores on standard dataset. The main evaluator for the performance of the proposed method in this study is Youden's index whose high value is dependent on both sensitivity and specificity. In particular, AFP-CKSAAP yields a Youden's index value of 0.82 on the independent dataset, which is better than previous methods.**

***Index Terms*—Antifreeze protein, deep neural network, amino acid composition**

I. Introduction

Organisms living in the subzero temperature cope with the extreme cold environment using antifreeze proteins (AFPs). The temperature falling below the freezing point of water can cause the phenomenon of ice creation in the cells of the organisms which can be fatal [1]. Also because of the property of re-crystallization of ice, the smaller crystals grows to become large ice lattice and cause physical damage to the cell membrane. AFPs hinders the ice crystal growth by binding onto the ice surface [2]. Antifreeze proteins are the heterogenous group of glycoproteins or polypeptides found in fish and other plants and animals [3]. They are also responsible for the process of thermal hysteresis in which the freezing point is lowered without any alteration in the melting point of water, thus the gap between freezing and melting points is enlarged [4].

Two strategies that are observed for the cold hardiness adaptation or prevention by the overwintering plants and animals are reported in [5]; the freeze tolerance in which the synthesis of ice nucleating agent is involved and the freeze avoidance in which ice-nucleating agents are removed in the species. To understand the process of interaction of AFP with ice, the utmost requirement is to be able to identify them reliably [6]. There is lack of similarity in the structure and sequences of the ice-binding proteins of even closely related species though they contribute to similar engagements, which in turn results in a complex prediction problem [7]. Antifreeze proteins have variety of applications in the food industry, biotechnological applications, medicine, preservation of organs and cell lines etc. Therefore, their reliable prediction is of utter importance [8].

Several researchers have successfully used machine learning approaches to address the classification of AFPs and non-AFPs. In [6], Kandaswany *et.al.* proposed an anti-freeze prediction model also known as (AFP-Pred), which is considered as the first computational framework in this direction. They explored the usage of sequence information along with the physicochemical properties, short peptides and functional groups. Random forest approach was utilized for the identification of proteins. In AFP_PSSM [9], the evolutionary information of the AFPs were explored for classification with support vector machine (SVM). In AFP-PseAAC [10], using the features based on the Chou's pseudo amino acid composition, the information based on sequence order was explored as an additional feature. SVM classifier was used for the prediction of AFPs to achieve reasonable accuracy. In iAFP [11], n-peptide composition of AFP from variety of organisms were analyzed using SVM and genetic algorithm. In particular, iAFP utilizes n=3 tripedtide composition features, however the experimental results are shown on very limited dataset. Another recently proposed method is afpCOOL [12], in which four types of features were used to identify protein sequence through support vector machine. The most recent work in this direction is RAFP-Pred [2], in which protein sequence was equally divided into two segments and each segment is analysed for amino acid and di-peptide composition features. The subset of significant features is selected using information gain and ranker method to design a random forest classifier.

In this study we propose to use composition of k spaced amino acid pairs (CKSAAP) to construct the feature vector for the prediction of AFPs. Application of CKSAAP can be found in variety of bioinformatics [13], [14], [15]. The encoding in this scheme is based on the frequency of amino acid pairs that

2471-7819/19/$31.00 ©2019 IEEE
DOI 10.1109/BIBE.2019.00016

are separated by k number of residues. In general, large values of k contributes to better prediction results [16]. However, increasing the number of attributes eventually increases the computation complexity. The conventional machine learning algorithms such as support vector machine or random forest can only generate competent prediction results when number of attributes are low [17]. On the other hand a well designed neural network can efficiently deal with large number of attributes for highly nonlinear classification tasks such as protein secondary structure prediction etc [18]. Due to the exponentially large expressiveness, deep neural network model can easily learn highly non-linear mapping using ReLU and skipped connections [19]. To pre-process data for the deep neural network, a localized analysis is performed on the amino acid pairs using the notion of CKSAAP to achieve a manifold learning approach for reliable prediction of AFPs. Remainder of the paper is structured as follows: In section II, the proposed approach is discussed in detail. In section III, the description of dataset, performance measures and evaluations are presented, which is followed by discussion in section IV. The paper is finally concluded in section V.

II. Proposed Approach

Machine learning approaches have been efficiently utilized in the area of protein prediction. The two vital steps to successfully address the prediction problem includes the formation of a significant feature set and selection of an appropriate machine learning algorithm. The later sections discuss the aforementioned issues.

A. Dataset

For the evaluation of the suggested method in this study we use the standard dataset from [6]. The dataset is composed of protein sequences that are obtained from the Pfam database [20]. A stringent threshold was chosen (E=0.001) during the PSI-BLAST to remove any redundancy from the data. Furthermore, to retain only antifreeze proteins in the dataset, a manual search is performed. The final positive dataset contains the proteins which have less than 40% similarity in the sequence. The non-antifreeze protein sequences in the dataset are obtained from seed proteins of Pfam family. The final dataset is composed of 9974 protein sequences including 481 antifreeze proteins and 9493 non-antifreeze proteins. Two subsets are formed by dividing dataset into Tr for training and Ts for testing. Tr contains a total of 600 protein sequences including 300 proteins from the AFP class and 300 proteins from non-AFP class. For generalization the sequences are selected randomly. Testing was done on Ts that is composed of leftover protein sequences from both positive and negative datasets i.e 181 antifreeze proteins and 9193 sequences of non-antifreeze proteins. We also consider the effect of imbalanced training dataset on validation performance, details of which is given in Section III.

B. Features

For feature construction, a sparse encoding scheme named CKSAAP is employed which has been used by researchers in various protein prediction problems [13] [15], [21]. CKSAAP stands for **C**ompostion of **k**- **s**paced **A**mino **A**cid **P**airs in which the frequency of the amino acid pair is calculated that are separated by k-residues (k = 0,1,2,3,4,5...). This scheme basically represents the short and long range interactions amongst the residues along the sequence resulting in competitive performance in protein prediction problems [13], [22]. Selecting the value of k = 0, CKSAAP encoding becomes identical to the di-peptide composition. The feature vector (FV) for k = 0 can be defined as:

$$FV = \left(\frac{F_{AA}}{N}, \frac{F_{AC}}{N}, \frac{F_{AD}}{N}, \ldots, \frac{F_{YY}}{N} \right)_{400} \qquad (1)$$

where N is the sequence length.

The composition of residue pair in the protein sequence is represented by the value of each descriptor. For example if the residue pair AC appears j times, the composition of the corresponding residue pair is equal to j/N. For the protein length L, when k = 0, 1, 2, 3, 4, 5 the values of N is $L-1$, $L-2$, $L-3$, $L-4$, $L-5$ and $L-6$ respectively. Figure 1 shows the encoding scheme for $k = 2$.

C. Classification

Several machine learning based approaches have been proposed for the classification of proteins, for instance, random forest [2], sparse representation classifier [23], SVM [12] and multi-layer perceptron [24]. The classical machine learning algorithm show reasonable performance gain in most of the situations. However, conventional machine learning algorithms e.g., multi-layer perceptron neural network (MLP-NN) have inherited issues which hinder their scalability to higher layers. Due to the recent advancements in the deep learning theories [25], we are now able to solve most of the issues with conventional MLP-NN. Therefore in this work we develop a deep neural network.

Deep learning has gained a lot of recognition in the area of machine learning in the recent era. We develop a four layered deep neural network architecture that is composed of an input layer, two hidden layers, and an output layer. The number of neurons in the input layer are equal to the feature's dimension while the first and second hidden layer has 70 and 128 neurons respectively. To avoid vanishing gradient issue in deep layers, we used rectified linear unit (ReLU) as the activation function and cross entropy as loss function. For better generalization, after every layer a dropout layer with 30% is used. The idea of utilizing dropout is to arbitrarily drop some of the units alongwith the connections which in turns significantly reduces the overfitting issues [26]. Additionally we use mini-batch gradient descent algorithm which selects a small portion of training sample at each iteration. Furthermore we utilize short and long skip connections to achieve faster convergence rate [27]. The model is trained using Adam optimizer for 200 epochs.

Figure 2 shows the workflow of the proposed classification model. Implementation of the model is done using Python on

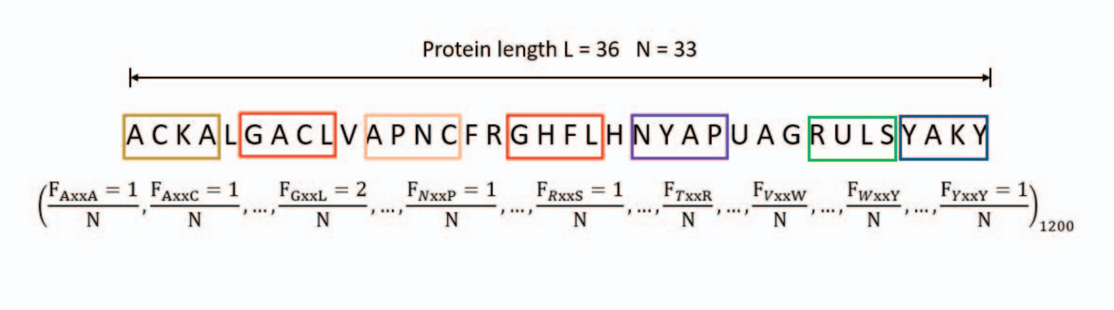
Protein length L = 36   N = 33
ACKALGACLVAPNCFRGHFLHNYAPUAGRULSYAKY
$\left(\frac{F_{AxxA}=1}{N}, \frac{F_{AxxC}=1}{N}, \ldots, \frac{F_{GxxL}=2}{N}, \ldots, \frac{F_{NxxP}=1}{N}, \ldots, \frac{F_{RxxS}=1}{N}, \ldots, \frac{F_{TxxR}}{N}, \ldots, \frac{F_{VxxW}}{N}, \ldots, \frac{F_{WxxY}}{N}, \ldots, \frac{F_{YxxY}=1}{N}\right)_{1200}$

Fig. 1. Illustration of CKSAAP descriptor calculation for $k = 2$

Keras (Tensor flow) and is freely available online for validation (https://github.com/musman88/AFPCKSAAP.git).

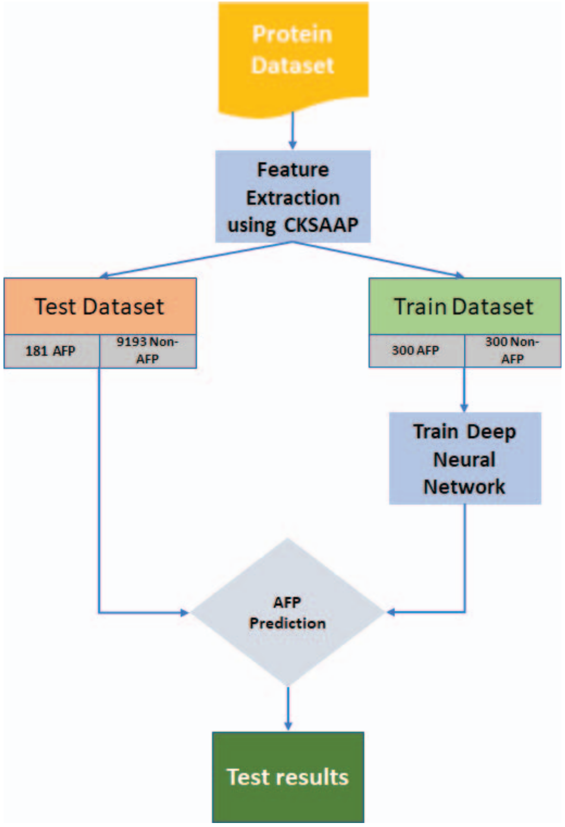

Fig. 2. Work-flow of the proposed approach.

III. Experimental Setup

A. Performance measures

For the performance evaluation of the proposed framework we consider the widely used threshold-dependent biostatistics nomenclatures including sensitivity, specificity and accuracy. Furthermore Youden's index and Matthew's correlation coefficient (MCC) are employed to measure the performance of the proposed method. The measurements can be defined using following equations:

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$MCC = \frac{TPTN - FPFN}{\sqrt{(TP+FP)(TN+FN)(TP+FN)(TN+FP)}} \tag{5}$$

$$Youden'sIndex = Sensitivity + Specificity - 1 \tag{6}$$

Here TP, FP, TN and FN constitute to True Positive (correctly classified AFP), False Positive (incorrect classification of non-AFP as AFP), True Negative (correctly classified non-AFP) and False Negative (incorrect classification of AFP as non-AFP) respectively. Therefore, specificity computes the fraction of non-AFPs that are correctly classified as non-AFP and sensitivity computes the fraction of AFPs being correctly classified as AFPs. The accuracy constitutes for the ratio of total number of correctly classified samples and the total number of samples. This indicates that in our case where the test dataset is highly imbalanced, a prediction model would yield a high accuracy value even if it classifies all the samples to the majority class which in our case is non-AFP. Therefore the prediction problems with imbalanced dataset should not be ranked according to the accuracy value they produce, instead they must be evaluated for the class specific performance measures such as Youden's index and MCC which are measured for joint performance gain. Another comprehensive threshold-independent parameter is receiver operating characteristic (ROC). The ROC curve is plotted for the true positive rate (sensitivity) versus the false positive rate (1-specificity). Higher values of area under the curve (AUC) indicates better classification abilities of the predictor model. The ROC curve for the proposed prediction model is shown in Figure 3.

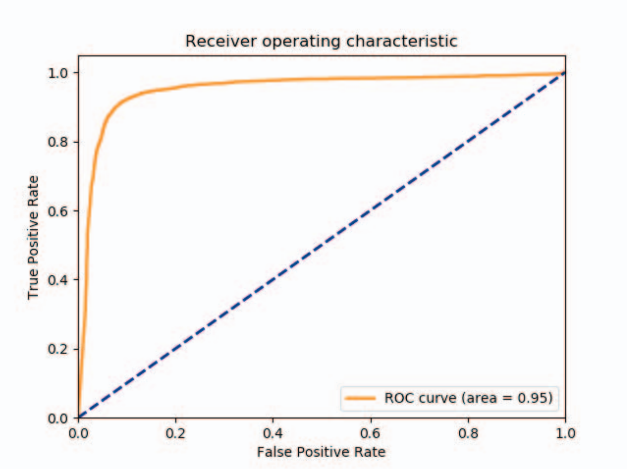

Fig. 3. ROC curve for the proposed algorithm

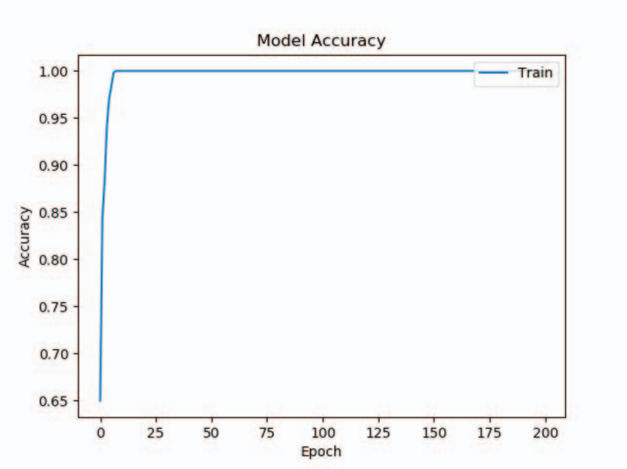

Fig. 4. Learning curve of the model

B. Performance evaluation of the proposed AFP-CKSAAP

Figure 4 shows the training curve of the algorithm over 200 epochs. The graph depicts that the model achieves 100% accuracy in less than 10 iterations. To study the intervention of the value of k on the performance, we evaluated the proposed classifier for different values of k ranging from 0 to 13. The results reveal a significant increase in accuracy for the values between 0 to 8. However, values higher than 8 does not contribute to better performance for DNN. Therefore in proposed AFP-CKSAAP method we choose an optimum value of k = 8 for the prediction purpose.

For comprehensive performance evaluation of the proposed method, we tested different values of k using the balanced training dataset on random forest and SVM classifiers and compared with proposed classifier. In Figure 5, an improvement in the prediction's Youden's index with subsequent values of k for all the classifiers can be observed. Deep learning method however, outperforms random forest and SVM approaches by a mean value of 0.10 and 0.07 respectively. Furthermore to observe the effect of the distribution of dataset we evaluated the performance of the algorithm for imbalanced training datasets by selecting 3 times more negative samples i.e 300 AFPs and 900 non-AFPs as shown in Figure 6. The increase in the negative samples contributes to increasing trend in the values of accuracy as the classifier associates most of the samples to the majority class but the sensitivity measure is negatively affected. This proves that the accuracy should not be considered as an evaluation parameter for imbalanced datasets. Table I shows the performance comparison of the proposed classifier with the conventional machine learning classifiers. All the classifiers have been trained on the balanced dataset containing 300 sequences of both AFPs and non-AFPs using CKSAAP encoding scheme. Finally in Table II, the efficacy of the proposed method is compared with other machine learning methods in the literature. The results show that the proposed algorithm provides superior performance using the balanced training dataset.

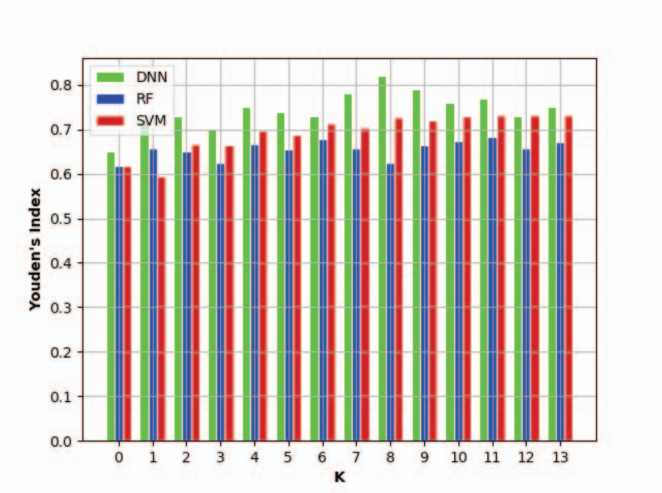

Fig. 5. Effect of varying values of (k) using DNN, Random Forest and SVM

C. Comparison of AFP-CKSAAP with contemporary methods

In this section we compare the results of our proposed method with other methods reported previously. RAFP-Pred has been reported to produce good results in terms of Youden's index which is better than any previous approach. Table II shows that the proposed prediction method can predict the AFPs with the accuracy of 88.0% which is better than AFP-Pred, AFP-PseAAC and comparable to CryoProtect. The algorithm demonstrates highest sensitivity value of 94.4% which is almost 10% greater than the next in order contributing in yielding the best value of 0.82 on the scale of Youden's index.

IV. Discussion

In this study a predictor for antifreeze protein has been proposed that produced better results. The classifier is composed of a deep learning neural network and the features are

TABLE I
COMPARISON OF SEVERAL MACHINE LEARNING APPROACHES ON BALANCED TRAINING DATASET USING CKSAAP FEATURES WITH $k = 8$

Classifier	**Sensitivity**	**Specificity**	**Accuracy**	**Youden's index**	**MCC**
SVM (Linear)	80.1%	80.3%	80.3%	0.60	0.20
Random Forest	76.8%	79.2%	79.1%	0.56	0.18
J48	76.8%	77.0%	77.0%	0.53	0.17
IBk	75.1%	74.0%	74.0%	0.49	0.15
Attribute Selected Classifier (J48)	76.2%	69.6%	69.8%	0.45	0.13
SVM (RBF)	69.6%	74.7%	74.6%	0.44	0.14
Naive Bayes	62.3%	80.0%	71.2%	0.42	0.19
BayesNet	68.5%	73.6%	73.5%	0.42	0.13
Proposed (Deep Neural Network)	94.4%	87.8%	88.0%	0.82	0.33

TABLE II
COMPARISON OF AFP-CKSAAP WITH CONTEMPORARY APPROACHES

Methods	**Classifier**	**Accuracy**	**Specificity**	**Sensitivity**	**Youden's Index**	**AUC**
iAFP [11]	SVM	95.6%	97.2%	9.9%	0.07	NA
AFP-Pred [6]	RF	79.0%	79.0%	82.3%	0.61	0.89
AFP_PSSM [9]	SVM	93.0%	93.2%	75.8%	0.69	0.93
AFP-PseAAC [10]	SVM	87.5%	87.6%	82.8%	0.70	NA
afpCOOL [12]	SVM	96.0%	98.0%	72.0%	0.70	NA
RAFP-Pred [2]	RF	91.0%	91.0%	84.0%	0.75	0.95
CryoProtect [8]	RF	88.2%	88.3%	87.2%	0.76	NA
AFP-CKSAAP	DNN	88.0%	87.8%	94.4%	0.82	0.95

iAFP features: n-peptide compositions
AFP-Pred features: Frequency of several functional groups + Physicochemical properties
AFP˙PSSM features: AAC + DPC + PseAAC + Evolutionary information
AFP-PseAAC features: PseAAC
afpCOOL features: Hydropathy + Physicochemical + AAC + Evolutionary information
RAFP-Pred features: AAC + DPC
CryoProtect features: AAC + DPC
AFP-CKSAAP features: CKSAAP

extracted using CKSAAP encoding scheme. Many previous approaches have used binary encoding schemes to predict the AFPs. The binary encoding only represents the amino acids at distinct positions and loses the information of coupling of amino acids at different locations with in a sequence, whereas the CKSAAP encoder retrieves the information of amino acid correlation at different locations and also characterizes long and short interactions of amino acids within a sequence window. This information turns out to be extremely useful for achieving best prediction.

The conventional machine learning approaches used in the past does not show satisfactory performance when data with large number of attributes is to be processed. Previously, researchers have addressed the prediction problem using random forest and support vector machines. The deep neural network have gained immense popularity in the recent years and has been adopted in the fields of bioinformatics, computational biology, image and speech recognition. The deep layers can effectively extract complex functions and features from large data that can improve the classification of the prediction algorithms.

V. CONCLUSION

Antifreeze proteins are responsible for the antifreeze effect of organisms living in subzero temperatures. The AFPs are useful in many applications of food industry and bioinformatics. Due to a great degree of diversity in the structure and sequences of protein the classification of AFPs and non-AFPs is considered a challenging task. Herein, we propose the use of deep neural network to serve the purpose of AFP prediction. The features are extracted from a sparse encoding scheme named CKSAAP which by varying window size describes the short and long range interactions of amino acid pairs in a sequence. Extensive experiments on the standard dataset have been performed. The proposed AFP-CKSAAP approach exhibit outstanding performance on the scale of Youden's index compared to the prior predictors such as RAFP-Pred, iAFP, AFP_PSSM, afpCool, AFP-Pred, CryoProtect and AFP-PseAAC. In our future work we aspire to come up with a systematic strategy for the calibration of the neural network and also to explore the structural information of the proteins. The Python implementation of AFP-CKSAAP can be downloaded from (https://github.com/musman88/AFPCKSAAP.git).

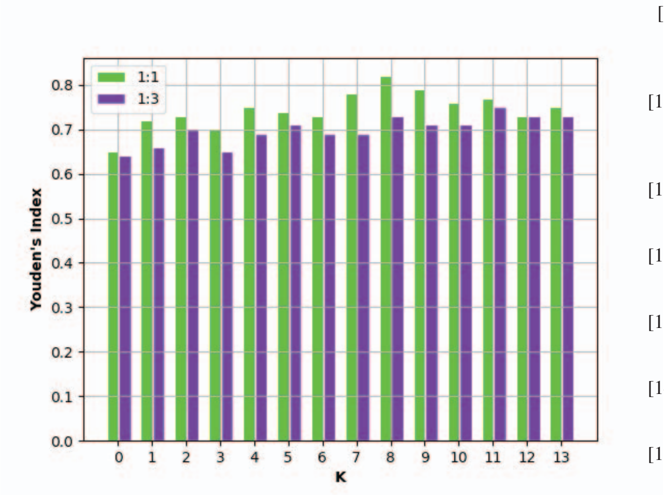

Fig. 6. Effect of varying amino acid pair gap (k) for distributed dataset using DNN

VI. Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (No. 2019R1I1A3A01058887) and in part by Korea Institute of Energy Technology Evaluation and Planning, and the Ministry of Trade, Industry and Energy of the Republic of Korea (No. 20184010201650).

References

[1] M. Kuramochi, C. Takanashi, A. Yamauchi, M. Doi, K. Mio, S. Tsuda, and Y. C. Sasaki, “Expression of ice-binding proteins in caenorhabditis elegans improves the survival rate upon cold shock and during freezing,” *Scientific reports*, vol. 9, no. 1, p. 6246, 2019.

[2] S. Khan, I. Naseem, R. Togneri, and M. Bennamoun, “Rafp-pred: Robust prediction of antifreeze proteins using localized analysis of n-peptide compositions,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 15, no. 1, pp. 244–250, Jan 2018.

[3] J. C. Iovine, P. N. Gallo, K. A. Callaway, P. L. Davies, and N. V. Nucci, “Studying the dynamic motions of water surrounding the ice-binding face of m1. 1 antifreeze protein,” *Biophysical Journal*, vol. 116, no. 3, p. 186a, 2019.

[4] M. Masud, M. U. Joardder, and M. Karim, “Effect of hysteresis phenomena of cellular plant-based food materials on convection drying kinetics,” *Drying Technology*, vol. 37, no. 10, pp. 1313–1320, 2019.

[5] T. Sformo, F. Kohl, J. McIntyre, P. Kerr, J. Duman, and B. Barnes, “Simultaneous freeze tolerance and avoidance in individual fungus gnats, exechia nugatoria,” *Journal of Comparative Physiology B*, vol. 179, no. 7, pp. 897–902, 2009.

[6] K. K. Kandaswamy, K.-C. Chou, T. Martinetz, S. Möller, P. Suganthan, S. Sridharan, and G. Pugalenthi, “Afp-pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties,” *J.Theor. Biol.*, vol. 270, no. 1, pp. 56–62, 2011.

[7] M. Bhattacharya, A. Hota, A. Kar, D. S. Chini, R. C. Malick, B. C. Patra, and B. K. Das, “In silico structural and functional modelling of antifreeze protein (afp) sequences of ocean pout (zoarces americanus, bloch & schneider 1801),” *Journal of Genetic Engineering and Biotechnology*, vol. 16, no. 2, pp. 721–730, 2018.

[8] R. Pratiwi, A. A. Malik, N. Schaduangrat, V. Prachayasittikul, J. E. Wikberg, C. Nantasenamat, and W. Shoombuatong, “Cryoprotect: a web server for classifying antifreeze proteins from nonantifreeze proteins,” *Journal of Chemistry*, vol. 2017, 2017.

[9] Z. Xiaowei, M. Zhiqiang, and Y. Minghao, “Using support vector machine and evolutionary profiles to predict antifreeze protein sequences,” *International Journal of Molecular Science*, vol. 13, pp. 2196–2207, 2012.

[10] P. Du, X. Wang, C. Xu, and Y. Gao, “PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou’s pseudo-amino acid compositions,” *Analytical biochemistry*, vol. 425, no. 2, pp. 117–119, 2012.

[11] 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*, 2011.

[12] M. Eslami, Z. Takalloo, G. Mahdevar, A. Emamjomeh, R. H. Sajedi, J. Zahiri *et al.*, “afpcool: An accurate tool for antifreeze protein detection,” *bioRxiv*, p. 231761, 2017.

[13] D. Chen, X. Tian, B. Zhou, and J. Gao, “Profold: Protein fold classification with additional structural features and a novel ensemble classifier,” *BioMed research international*, vol. 2016, 2016.

[14] Y. Xu, Y. Yang, J. Ding, and C. Li, “iglu-lys: A predictor for lysine glutarylation through amino acid pair order features,” *IEEE transactions on nanobioscience*, vol. 17, no. 4, pp. 394–401, 2018.

[15] L. Wei, M. Liao, X. Gao, and Q. Zou, “An improved protein structural classes prediction method by incorporating both sequence and structure information,” *IEEE transactions on nanobioscience*, vol. 14, no. 4, pp. 339–349, 2014.

[16] X. Zhao, W. Zhang, X. Xu, Z. Ma, and M. Yin, “Prediction of protein phosphorylation sites by using the composition of k-spaced amino acid pairs,” *PloS one*, vol. 7, no. 10, p. e46302, 2012.

[17] A. Singh, R. Lakshmiganthan *et al.*, “Impact of different data types on classifier performance of random forest, naive bayes, and k-nearest neighbors algorithms,” *International Journal of Advanced Computer Science and Applications, (IJACSA)*, vol. 8, no. 12, 2017.

[18] J. J. A. Armenteros, K. D. Tsirigos, C. K. Sønderby, T. N. Petersen, O. Winther, S. Brunak, G. von Heijne, and H. Nielsen, “Signalp 5.0 improves signal peptide predictions using deep neural networks,” *Nature biotechnology*, vol. 37, no. 4, p. 420, 2019.

[19] Y. H. Yoon, S. Khan, J. Huh, and J. C. Ye, “Efficient b-mode ultrasound image reconstruction from sub-sampled rf data using deep learning,” *IEEE transactions on medical imaging*, vol. 38, no. 2, pp. 325–336, 2018.

[20] E. L. Sonnhammer, S. R. Eddy, R. Durbin *et al.*, “Pfam: a comprehensive database of protein domain families based on seed alignments,” *Proteins-Structure Function and Genetics*, vol. 28, no. 3, pp. 405–420, 1997.

[21] Z. Chen, Y. Zhou, Z. Zhang, and J. Song, “Towards more accurate prediction of ubiquitination sites: a comprehensive review of current methods, tools and features,” *Briefings in bioinformatics*, vol. 16, no. 4, pp. 640–657, 2014.

[22] Z. Chen, Y.-Z. Chen, X.-F. Wang, C. Wang, R.-X. Yan, and Z. Zhang, “Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs,” *PloS one*, vol. 6, no. 7, p. e22930, 2011.

[23] I. Naseem, S. Khan, R. Togneri, and M. Bennamoun, “Ecmsrc: A sparse learning approach for the prediction of extracellular matrix proteins,” *Current Bioinformatics*, vol. 12, no. 4, pp. 361–368, 2017.

[24] H. He, J. Zhao, and G. Sun, “Prediction of morfs in protein sequences with mlps based on sequence properties and evolution information,” *Entropy*, vol. 21, no. 7, p. 635, 2019.

[25] J. C. Ye, Y. Han, and E. Cha, “Deep convolutional framelets: A general deep learning framework for inverse problems,” *SIAM Journal on Imaging Sciences*, vol. 11, no. 2, pp. 991–1048, 2018.

[26] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[27] A. E. Orhan and X. Pitkow, “Skip connections eliminate singularities,” *arXiv preprint arXiv:1701.09175*, 2017.