Classification of Membrane Proteins using Four Phase Split Amino Acid Composition Descriptor over Random Forest

Abdul Jamil
Department of Computer Science, Islamia College Peshawar, Pakistan
tariqjamil845@gmail.com

Waseem Ullah
Department of Computer Science, Islamia College Peshawar, Pakistan
khattakullah31@gmail.com

Samee Ullah
Department of Computer Science, Islamia College Peshawar, Pakistan
samimarwat2017@gmail.com

Muhammad Sajjad
Department of Computer Science, Islamia College Peshawar, Pakistan
muhammad.sajjad@icp.edu.pk

Abstract—**Recently information regarding the classification of uncharacterized membrane-proteins is more appreciated in the last decade. The degree at which the varieties of proteins are uploaded in the Pole-Genomic age and the firm process determines the function of plasma proteins in different biological experiments. However, it is necessary to have an automatic method for classification of membrane proteins into their corresponding classes. In this paper, a novel multi-classification framework is presented that classifies the given membrane-proteins into Single-pass Type1, Single-pass Type2, GPI, lipid anchored, Multipass and Peripheral. The proposed framework combines four Phase SAAC Descriptor model with Random Forest (RF) classifier to predict the Membrane Proteins. These different descriptors in combination with various machine learning techniques are experimentally used for the classification of plasma proteins types. Among the existing techniques, the proposed framework shows impressive performance over Datasets-1, obtaining 94.39%, and 94.41%, accuracy with 70-30 and 10 fold cross validation ratio respectively.**

Keywords— Membrane proteins, Random Forest, SVM, KNN, Four Phase SAAC,

I. Introduction

Membranes are generally continuous, unbroken sheets and enclose compartments. It is the first layer in animal cell, while it is situated next to cell wall in plant. Cell membrane is permeable that transmit fatty acid, water, amino acid, glucose, glycerol, ion and other gases[1]. Cell membrane is composed of about 60 to 80% of proteins which are called membrane proteins or plasma protein. The pharmaceutical companies and drug designers are of the opinion that these proteins play a vital role in the drug interaction[2]. Generally there are six types of membrane proteins namely Peripheral, Multi-pass, Single-pass Type I, Single-pass Type II, GPI and Lipid Anchored proteins as classified[3].

In Bioinformatics, many efforts have been made to develop the accurate discrimination classifiers system, but finding the right classifier with suitable feature extraction method and higher accuracy is still challenging. M.Hayat et al., [4] method consist the Mem-EnsSAAC as a features descriptor in combination of ensemble classifiers such are probabilistic neural network, support vector machine, random forest, nearest neighbor and Adaboost. Some recent papers like Y-Kuang Chen et al., [5] have used some sequence attribute with physicochemical properties of amino acids, proteins topology domain, the cationic patch size, the presence of signal anchors by using the one-versus –one based SVM classifier. Chao Huang et al., [6] presented the method for classifying the Membrane Proteins Types by using the multilabel KNN algorithm and Pseudo Amino Acid Composition. In the same way J-K Leman et al., [7] used the de novo modeling technique in combination of molecular dynamics simulation to modulate the Membrane Proteins. In 2013 M.Hayat et al., [8] predict the Membrane Proteins Types by using the Pseudo Amino Acid Composition method as a numerical features descriptor and ensemble classification method which consist neural network and K-Nearest Neighbor. Similar in 2015 F-Ali et al., [1] predict the Membrane Proteins Types using the machine learning techniques called the Pseudo Amino Acid as feature descriptor and Voting feature interval as a classifier. Another contribution was attempted by E.Siva Shankari et al., [9] used the Pseudo Amino Acid Composition as a feature extractor method and used the different decision methods like random under sampling boost tree, classification and regression tree, rotation forest, reduced error pruning tree and random forest tree. Similarly in 2018 M.Hayat et al., [10] used the SVM as a classifier and extended notion of Split Amino Acid into Chou's Pseudo Amino Acid Composition method as a features extractor.

A lot of work has been done in the field of protein classification but there is still remains some challenges that need further improvements. Thus a promising classification model is necessary to correctly classify the membrane proteins approximately in combination of features extraction method.

In this paper, we used the Four Phase Split Amino Acid Composition Descriptor as a features extraction method in combination with the Random Forest as a classifier to classify the membrane proteins types into six classes called the GPI, Single-pass Type I, Single-pass Type II, Peripheral, Multi-pass and Lipids. We used the Four Phase descriptor model instead of existing model because of obtaining the large feature vector that consist of 8o D- Length for better accuracy. To achieve higher accuracy, we used the standard 70-30% split up ratio of data and 10 fold cross validation test method for producing result separately.

II. Proposed Framework

The proposed methodology is tested on benchmark datasets provided by http://www.csbio.sjtu.edu.cn/bioinfo

SWISS PROT databank research community over the years. The used [11] dataset has 3249 membrane proteins sequences that are divided into eight membrane proteins types. The mention dataset consist of 610 single-pass type I, 312 single-pass Type II, 24 single- pass Type III, 44 single-pass Type IV, 151 lipid anchored chain, 316 Multipass, 182 GPI and 610 peripheral membrane proteins sequence. We have used the one locally independent dataset that consists of 720 sequences of peripheral, single-pass Type I, single-pass Type II, Multipass, Lipid anchored and GPI. For formulation of membrane proteins sequence, the features extraction is the basic and main part for the learning process. There are various features extraction methods like AAC, DPC and Pseudo Amino Acids Composition etc. The salient features are very important in order to improve the consistent hypothesis and get generalization capability. In our proposed study we used the Four Phase Split Amino Acid Composition descriptor to get the most important salient features for prediction of membrane proteins types into their specific types. In this method, the protein sequence is divided into four regions that are called A, B, C and the region between these sides. The Amino Acid Composition of A, B, C and region among these sides are calculated. Hence the 80 Dimension length vector is obtained to get the most important salient features for prediction of membrane proteins types into their specific types. Four phase SAAC is very important for the extraction of complementary informative peptides. This numerical descriptor is most used by different investigator due to its supervisory upon other existing features methods [12-15]. The Four Phase SAAC achieved the information from four parts of proteins independently that is why it analyzes the proteins sequence with great depth rather of other mention features extraction methods. It is not easy to directly identify informative peptides. For this purpose the Four Phase Split Amino Acid method is proposed [16]. The simple formula for the representation of numerical values of the Four Phase Split Amino Acid Composition can be shown as,

$$P = S_1^A \text{ K } S_{20}^A, S_1^B \text{ K } S_{20}^B, , S_1^{mid} \text{ K } S_{20}^{mid} , , S_1^C \text{ K } S_{20}^C$$

The Four Phase SAAC gives the best result by using the Randomforest classifier to accurately classify the membrane proteins types into their respective six classes. The description of the above proposed method have been illustrated by the following desired framework as,

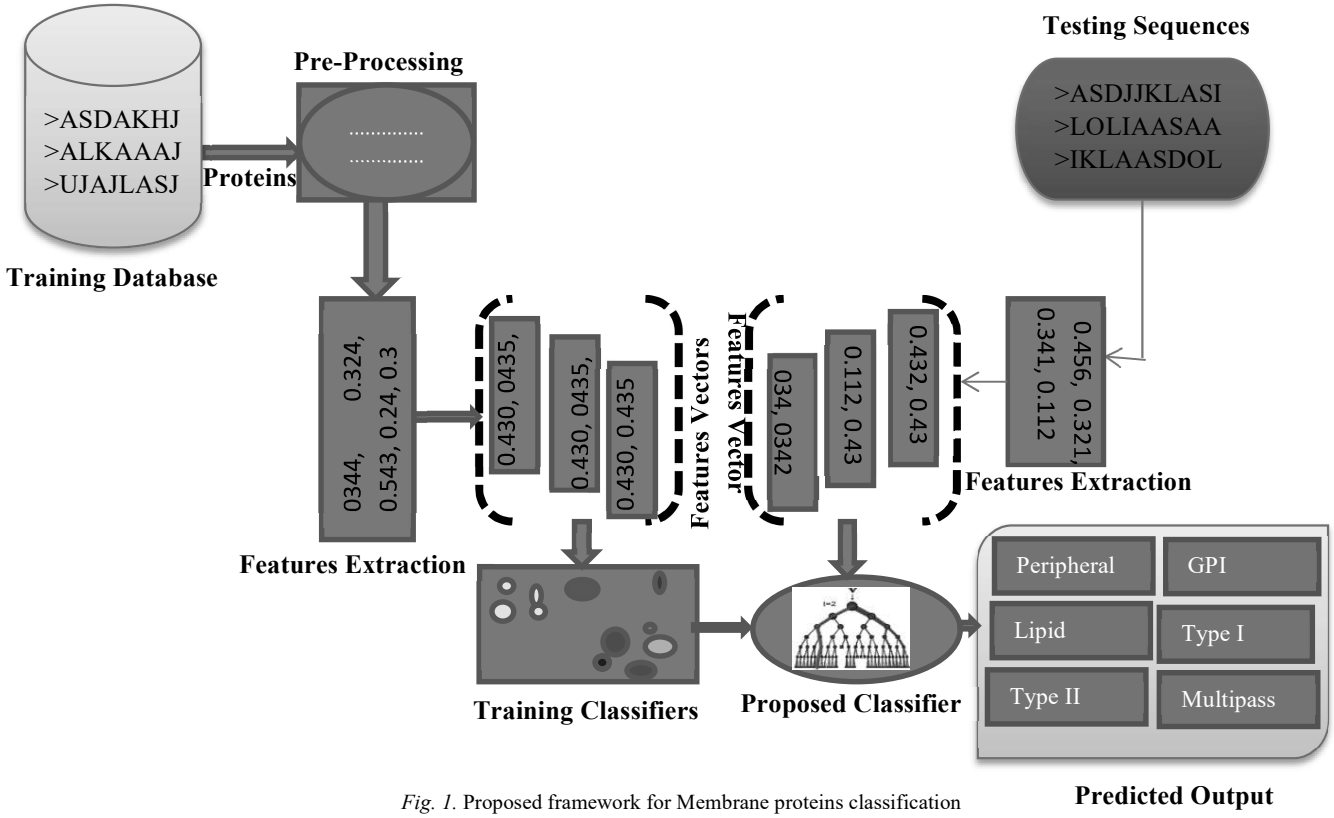

Fig. 1. Proposed framework for Membrane proteins classification

III. Experimental Results

We evaluate the performance result of different classifiers in order to achieve the best classification methods for the classification of membrane proteins types. The prediction results of the each discussed classifiers are listed in the form of Tables. But using the Four Phase SAAC Descriptor as a features extraction method, 70-30% split up ratio and using the RF as the classifier, the achieved Accuracy is **94.39%**, Sensitivity is 0.944, Specificity is 0.012 and Precession is 0.948.While using the Jackknife 10 fold cross validation test, Accuracy is **94.41%**, Sensitivity is 0.946, Specificity is 0.013 and Precession is 0.949.

TABLE I. Accuracy result of our proposed method

70 and 30% split up Test					Jackknife 10 fold cross validation Test			
	Accuracy	Sensitivity	Specificity	Precession	Accuracy	Sensitivity	Specificity	Precession
Four Phase SAAC								
KNN N=1	90.61	0.926	0.015	0.911	90.88	0.937	0.013	0.937
SVM	92.36	0.921	0.019	0.918	92.64	0.935	0.012	0.935
RF	**94.39**	0.944	0.012	0.948	**94.41**	0.946	0.013	0.949

IV. Conclusion and Future Work

In our proposed work, we have established a good promising predicting model for the classification of membrane proteins types into six classes accordingly. We have used the Four Phase SAAC Descriptor as numerical extractor for the protein sequences representation in combinations of different classifiers like KNN, SVM and Random Forest. The best result of our proposed methodology has been achieved by using the Four Phase SAAC descriptor as feature extraction methods in combination with Random Forest as a classifier. We have used the one locally independently datasets by applying the 70-30% splitting ratio and Jackknife 10 fold cross validation test on the dataset. The obtained accuracy by using the 70-30% split up ratio on Dataset 1 is **94.39%**, similarly by using the Jackknife 10 cross fold validation test, the achieved accuracy on Dataset 1 is **94.41%**. As further and future work, we plan to more collaborate with our proposed framework to bring it on broad model which will capable to handle the varieties of protein classification for the best accurate predicting result.

References

[1] Ali, F. and M. Hayat, Classification of membrane protein types using Voting Feature Interval in combination with Chou' s Pseudo Amino Acid Composition. Journal of theoretical biology, 2015. 384: p. 78-83.

[2] 2. Tusnady, G.E. and I. Simon, The HMMTOP transmembrane topology prediction server. Bioinformatics, 2001. 17(9): p. 849-850.

[3] Gao, Q.-B., et al., Improving discrimination of outer membrane proteins by fusing different forms of pseudo amino acid composition. Analytical biochemistry, 2010. 398(1): p. 52-59.

[4] Hayat, M., A. Khan, and M. Yeasin, Prediction of membrane proteins using split amino acid and ensemble classification. Amino acids, 2012. 42(6): p. 2447-2460.

[5] Chen, Y.-K. and K.-B. Li, Predicting membrane protein types by incorporating protein topology, domains, signal peptides, and physicochemical properties into the general form of Chou's pseudo amino acid composition. Journal of Theoretical Biology, 2013. 318: p. 1-12.

[6] Huang, C. and J.-Q. Yuan, A multilabel model based on Chou's pseudo–amino acid composition for identifying membrane proteins with both single and multiple functional types. The Journal of membrane biology, 2013. 246(4): p. 327-334.

[7] Koehler Leman, J., M.B. Ulmschneider, and J.J. Gray, Computational modeling of membrane proteins. Proteins: Structure, Function, and Bioinformatics, 2015. 83(1): p. 1-24.

[8] Hayat, M. and A. Khan, Prediction of membrane protein types using pseudo-amino acid composition and ensemble classification. International Journal of Computer and Electrical Engineering, 2013. 5(5): p. 456.

[9] Sankari, E.S. and D. Manimegalai, Predicting membrane protein types using various decision tree classifiers based on various modes of general PseAAC for imbalanced datasets. Journal of theoretical biology, 2017. 435: p. 208-217.

[10] Arif, M., M. Hayat, and Z. Jan, iMem-2LSAAC: A two-level model for discrimination of membrane proteins and their types by extending the notion of SAAC into chou's pseudo amino acid composition. Journal of theoretical biology, 2018. 442: p. 11-21.

[11] Chou, K.-C. and H.-B. Shen, MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochemical and biophysical research communications, 2007. 360(2): p. 339-345.

[12] Hayat, M. and A. Khan, Discriminating outer membrane proteins with fuzzy K-nearest neighbor algorithms based on the general form of Chou's PseAAC. Protein and Peptide Letters, 2012. 19(4): p. 411-421.

[13] Shen, H.-B. and K.-C. Chou, Ensemble classifier for protein fold pattern recognition. Bioinformatics, 2006. 22(14): p. 1717-1722.

[14] Mishra, A.K., R.K. Prajapati, and S. Verma, Probing structural consequences of N9-alkylation in silver-adenine frameworks. Dalton Transactions, 2010. 39(42): p. 10034-10037.

[15] Afridi, T.H., A. Khan, and Y.S. Lee, Mito-GSAAC: mitochondria prediction using genetic ensemble classifier and split amino acid composition. Amino Acids, 2012. 42(4): p. 1443-1454.

[16] 1Hayat, M. and A. Khan, MemHyb: predicting membrane protein types by hybridizing SAAC and PSSM. Journal of theoretical biology, 2012. 292: p. 93-102.