Prediction of Membrane Protein Types Using Pseudo-Amino Acid Composition and Ensemble Classification

Maqsood Hayat and Asifullah Khan

Abstract—**Predicting membrane protein types is an important and challenging research in current molecular and cellular biology. The knowledge of membrane proteins types often provides crucial hints for determining the function of uncharacterized membrane proteins. It is thus highly desirable to develop an automated method that can serve as a high throughput tool in identifying the types of newly found membrane proteins by their primary sequence information only. In this paper, features are extracted from membrane protein sequences using pseudo-amino acid *(PseAA)* composition. An ensemble classification approach is developed using K-nearest neighbor and Probabilistic Neural Network as the basic learning mechanisms. Each basic classifier is trained using *PseAA* composition with different tiers. The success rate has been obtained by the ensemble classifier on all the tests such as self-consistency, jackknife, and independent dataset test is quite promising and indicating that the ensemble classifier may become a useful and high performance tool in identifying membrane proteins and their types.**

Index Terms—**Ensemble classification, K-nearest neighbor, pseudo-amino acid (PseAA) composition, probabilistic neural network.**

I. Introduction

Cell membrane is vital to living organisms. Biologically, membrane proteins are the most imperative proteins because they control the cell processes inside or outside the cell. It also permits the cell to communicate with their environment. In addition, they determine whether the immune system identifies the foreign cell or not. Research on membrane proteins is interesting due to their key roles in organizing the processes of life. Membrane proteins can generally be classified into five types [1]: 1) Type-I transmembrane proteins, 2) Type-II transmembrane proteins, 3) Multipass transmembrane proteins, 4) lipid chain-anchored membrane proteins, and 5) *GPI* anchored membrane proteins (Fig. 1). Type- I and Type-II are single pass transmembrane proteins because they pass the lipid bilayer only once. Type-I transmembrane proteins have extracellular on N-terminus and Cytoplasmic on C-terminus, while type-II extracellular is on C-terminus and Cytoplasmic on N-terminus. Multipass transmembrane proteins are multipass transmembrane proteins because the polypeptides pass the lipid bilayer multiple times. Lipid chain-anchored and *GPI*-anchored membrane proteins are also called anchored membrane proteins. However, the lipid chain anchored is attached with the bilayer only by means of one or more covalently attached fatty acid while the *GPI* is associated to the membrane by a Glycolsylphosphatidylinositol (*GPI*) anchor.

A great deal of research has been carried out on prediction of membrane protein types in order to establish a model, which can efficiently predict the type of membrane proteins from their sequences. Chou and Elrod [1] have introduced the covariant discriminant algorithm (*CDA*) to predict the types of membrane proteins based on the amino acid composition. Then Chou [2] has proposed the *CDA* in conjunction with pseudo-amino acid (*PseAA*) composition. Liu et al. have used Low-frequency Fourier spectrum [3], Wavelet based features and cascaded neural network [4] are also used for prediction of membrane protein types. Other research works related to this area are reported in [5]-[13].

The present study, focus on the development of a prediction system for membrane protein types. The propose approach based on *PseAA* composition and ensemble classification approach that uses nearest neighbor *(KNN)* and Probabilistic Neural Network *(PNN)* as basic learners. In the remaining parts of this paper, Section II discusses Materials and Methods, while Section III presents the Results and Discussions. Finally, Conclusions are drawn in Section IV.

II. Materials and Methods

A. Dataset

The dataset used for training and testing is the same as originally constructed by Chou and Elrod [1]. It contains 2,059 membrane protein sequences, of which 435 are type I transmembrane proteins, 152 type II transmembrane proteins, 1311 Multipass transmembrane proteins, 51 lipid-chain-anchored membrane proteins, and 110 GPI anchored membrane proteins. The sequences used in the independent dataset test are 2,625 in which 478 are type I transmembrane proteins, 180 type II transmembrane proteins, 1867 Multipass transmembrane proteins, 14 lipid-chain-anchored membrane proteins, and 86 are *GPI* anchored membrane proteins

B. Pseudo-Amino Acid (PseAA) Composition

The concept of pseudo-amino acid composition has been proposed by Chou [2]. According to the classical definition, amino acid composition comprised of 20 discrete numbers with each representing the occurrence frequency of one of the 20 native amino acids in the protein sequences. Thus, in terms of amino acid composition, a protein sequence can be expressed by a vector of 20D (dimensional) space [14]-[18].

Manuscript received March 3, 2013; revised April 28, 2013.

The authors are with Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences, Nilore, Islamabad, Pakistan (e-mail: Maqsood.hayat@gmail.com, asif@pieas.edu.pk).

DOI: 10.7763/IJCEE.2013.V5.752

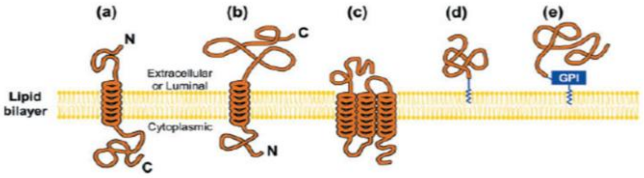

Fig. 1. Five types of membrane proteins (a)type I transmembrane proteins with C-terminal region in cytoplasmic side,(b)type II transmembrane proteins with N-terminal region in cytoplasmic side, (c)Multipass transmembrane, (d) lipid-chain anchored membrane proteins and (e) Glycosylphosphatidylinositol(GPI) anchored membrane proteins Reproduced from Chou et al. [2] with permission.

$$P = [p_1, p_2, ..., p_{20}]^T \tag{1}$$

where p_1, p_2, p_3... p_{20} are the composition components of 20 amino acids for the protein P and T denotes transpose. However, the main problem with amino acid composition to represent a protein sequence is that, all its sequence-order and sequence-length effects would be lost. To compensate this problem, Chou has proposed to represent a protein sequence by pseudo-amino acid composition [2], which is represented in a (20 + λ) D space as formulated:

$$P = [p_1, p_2, ...p_{20}, p_{20+1}, p_{20+2}, ...p_{20+\lambda}]^T \tag{2}$$

The first 20 components are the same as those in the conventional amino acid composition, where p_{20+1}... $p_{20+\lambda}$ are the factors related to λ different ranks of sequence-order correlations that can be easily computed by Eqs. (2)– (6) developed by Chou [2]. In this study the λ = 21 means taking the first 21 ranks of sequence-order correlations into consideration. Thus according to Eq. (2), a protein sample is represented by a (20 + λ) D = 62D vector.

C. Ensemble Classifier

In this paper, first *KNN* and *PNN* based ensemble classifiers are developed. *KNN* is a learning algorithm that is based on the concept of proximity in the feature space [19], while *PNN* is based on Bayes theory to estimates the likelihood of a sample being part of a learned category [20]. Then these two ensembles are combined to form a composite ensemble. The framework of ensemble classifier system works by combining numerous basic classifiers together in order to reduce the variance, caused by the peculiarities of a single training set [21]. This also enables the ensemble classifier to learn a more expressive concept in classification than a single classifier for example; the composite ensemble can be represented as,

CEnsb=$KNN\ (\lambda_1)\ \forall KNN\ (\lambda_2)\ \ldots\ \forall KNN\ (\lambda_m)\ \forall PNN\ (\lambda_1)\ \forall PNN\ (\lambda_2)\ \ldots\ \forall PNN\ (\lambda_m)$.

CEnsb is a composite ensemble classifier, *KNN*(λ_1)and *PNN*(λ_1) are the individual *KNN* and *PNN* classifier trained by proteins based on (20 + λ_1) components, *KNN*(λ_2) and , *PNN*(λ_2) are the classifiers based on (20 + λ_2) components, and so forth; the symbol ∀ denotes the combination operator. The prediction result is determined according to the voting scores of all the constituent classifiers: *KNN* (λ_1), *KNN* (λ_2)… *KNN* (λ_m), *PNN* (λ_1), *PNN* (λ_2)… *PNN* (λ_m). The ensemble classifier thus formed can better reflect the sequence-order effects and reduce the variance caused by the peculiarities of some individual subsets. The ensemble classifier works as shown in Fig. 2. The final output of the ensemble classifier is actually a ''fusion'' of the outputs produced by a set of basic classifiers: *KNN* (λ_1), *KNN* (λ_2)… *KNN* (λ_m), *PNN* (λ_1), *PNN* (λ_2)… *PNN* (λ_m). The outcome of the fusion is a voting result among the constituent individual classifiers operated independently with different λ (1…*m*) respectively. The operator max means taking the maximum one among those in the brackets. If there is a tie for the voting results, the query protein will be randomly assigned to one of the locations associated with the tie case.

III. Results and Discussions

In this section, three typical statistical tests are explained. These include self-consistency, Jackknife, and Independent dataset test.

A. Self-consistency Test

Self-consistency test is designed to examine the self-consistency of an identification method [22]-[25]. When the self-consistency test is performed for the current classifier, the same dataset is used for training and testing. The success rate for the 2,059 membrane proteins is listed in Table I. However, the self-consistency test is definitely necessary because any algorithm whose self-consistency performance is poor cannot be deemed a good one. On the other hand, it is not sufficient for evaluating the performance of a classifier.

B. Jackknife Test

The independent dataset, Self-consistency, and jackknife test are the three most common methods of cross-validation in statistical prediction. Among these three, the jackknife test is regarded as the most objective and effective one. During jackknifing, each membrane protein in the dataset is in turn taken out and the entire rule parameters are calculated based on the remaining proteins. During the process of jackknifing, both the training data set and testing data set are actually open and a protein will move from one to the other in turn. The results of the jackknife test thus obtained for the 2,059 membrane proteins are given in Table I.

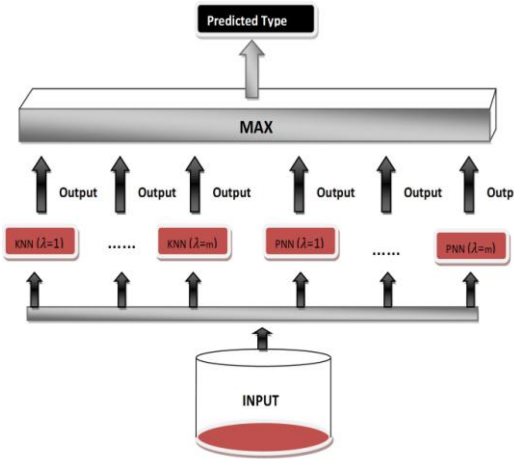

Fig. 2. Composite ensemble classifier

TABLE I: COMPARISON OF THE RESULTS OF THE PROPOSED WORK WITH PREVIOUS STUDIES

Methods	Self-consistency test	Jackknife test	Independent dataset test
Proposed Composite Ensemble using *PseAA.*	99.95	84.12	94.93
Proposed Ensemble *PNN* using *PseAA*	99.95	84.70	94.59
Proposed Ensemble *KNN* using *PseAA*	99.95	82.41	94.40
Wavelet & Cascade Neural Network[4]	96.8	81.3	91.4
Low Frequency Fourier Spectrum[3]	99.0	78.0	87.0
Covariant-discriminant Algorithm & *PseAA* [2].	90.9	80.9	87.5
Covariant-discriminant Algorithm [1].	76.4	79.4	81.1

C. Independent Dataset Test

Furthermore, prediction is also conducted for the 2,625 independent membrane proteins based on the rule parameter derived from the 2,059 proteins in the training dataset. The 2,625 independent proteins were those used by Chou and Elrod [1].

As shown in Table I, the obtained results in self-consistency, jackknife, and independent dataset test are 99.95%, 84.12% and 94.93% using *PseAA* composition and Composite Ensemble *KNN* & *PNN* base approach while in simple ensemble *KNN* and ensemble *PNN* results are 99.95%, 82.41%, 94.40% and 99.95, 84.70%, and 94.59%, respectively. Comparison of these results with similar work of covariant discriminant algorithm [1], Low-frequency Fourier spectrum [3], cascade neural network and wavelet analysis [4] is also shown in Table I. Fig. 3 shows the graphs of all proposed techniques.

In jackknife test, ensemble *PNN* is better than composite ensemble classification but composite ensemble classifier is strong than ensemble *PNN* because it is the combination of two classifiers.

IV. CONCLUSIONS

In this work, it is shown that the types of membrane protein are predictable with considerable accuracy, using *PseAA* composition and the ensemble classification approach of *KNN & PNN*. It is observed that *PseAA* composition based feature extraction strategy can significantly discriminate the different types of membrane protein. The composite ensemble made up of *KNN* and *PNN* based ensembles provides promising prediction performance.

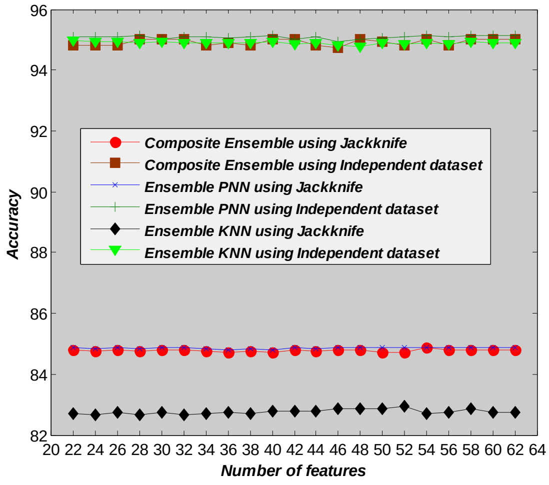

Fig. 3. Ensemble classifier performance

The overall success rates obtained by the current method are 99.95%, 84.12%, and 94.93% for the self-consistency, jackknife, and independent dataset tests, respectively.

ACKNOWLEDGMENT

This work is supported by the Higher Education Commission of Pakistan under the indigenous Ph.D. scholarship program 17-5-3(Eg3-045)/HEC/Sch/2006). The authors would like to thank professor Kuo-Chen Chou for providing the datasets of membrane protein sequences and granting permission for reprinting Fig. 1.

REFERENCES

[1] K. C. Chou and D. W. Elrod, *PROTEINS:Struct,funct.,Genet*., 4, pp. 137-153. 1999

[2] K. C. Chou, “Prediction of protein subcellular attributes using pseudo-amino acid composition,” *Proteins:Struct, Funct .Genet,* vol. 43, pp. 246-255, 2001.

[3] H. Liu, M. Wang, and K. C. Chou, “Low-frequency Fourier spectrum for predicting membrane protein types,” *Biochem, Res .Commun*., vol. 336, pp. 737-739, 2005.

[4] M. A. Rezaei, P. A. Maleki, Z. Karami, E. B. Asadabadi, M. A. Sherafat, K. A. Moghaddam, M. Fadaie, and M. Forouzanfar, “Prediction of membrane protein types by means of wavelet analysis and cascaded neural network,” *Journal of theoraticaly biology*, vol. 255, pp. 817-820, 2008.

[5] H. S. Shen HB and K. C. Chou, “Using ensemble classifier identify membrane protein types,” *Amino Acid,* vol. 32, pp. 483-488, 2007.

[6] Y. D. Cai and K. C. Chou, “Predicting membrane protein type by functional domain composition and pseudo amino acid composition,” *J. Theor. Biol.*, vol. 238, pp. 395-400, 2006.

[7] Y. D. Cai, G. P. Zhou, and K. C. Chou, “Support vector machines for prediction membrane proteins types by suing functional domain composition,” *J. Biophys*, vol. 84, pp. 3257-3263, 2003.

[8] K. C. Chou and Y. D. Cai, “Prediction of membrane protein types by incorporating amphipathic effects,” *J. Chem, imf. Model*, vol. 45, pp. 407-413, 2005.

[9] K. C. Chou and Y. D. Cai, “Using GO-PseAA predictor to indentify membrane proteins and their types,” *Biochem, Biophys, Res. Commun.*, vol. 327, pp. 845-847, 2005.

[10] K. C. Chou and H. B. Shen, “Memtype-2L a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM,” *Biochem, Biophys, Res. Commun*, vol. 360, pp. 339-345, 2007.

[11] M. Wang, J. Yang, G. P. Liu, Z. J. Xu, and K. C. Chou, “Weighted-support vector machines for predicting membrane protein types based on pseudo-amino acid composition,” *Protein Eng. Des. Sel.* vol. 17, pp. 509-516, 2004.

[12] M. Wang, J. Yang, Z. J. Xu, and K. C. Chou, “SLLE for predicting membrane protein types,” *J. Theor. Biol*, vol. 232, pp. 7-15, 2005.

[13] S. Q. Wang, J. Yang, and K. C. Chou, “Using stacked generalization to predict membrane protein types based on pseudo- amino acid,” *J. Theor, Biol*, vol. 242, pp. 941-946, 2006.

[14] K. C. Chou and C. T. Zhang, “Predicting protein folding types by distance functions that make allowances for amino acid interactions,” *J. Biol. Chem*., vol. 269, pp. 22014–22020, 1994.

[15] K. C. Chou, “A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space,” *Proteins: Structure. Function & Genetics*, vol. 21, pp. 319–344, 1995

[16] J. J. Chou and C. T. Zhang, “A joint prediction of the folding types of 1490 human Proteins from their Genetic Codons,” *J. Theor. Biol*., vol. 161, pp. 51–262, 1993.

[17] P. Y. Chou, *Prediction of Protein Structure and the Principles of Protein Conformation*, Plenum Press, G. D. Fasman, Ed. New York, 1989, pp. 549–586.

[18] H. Nakashima, K. Nishikawa, and T. Ooi, *J. Biochem*, vol. 99, pp. 152–162, 1986.

[19] A. Khan, M. Fayyaz, and T. S. Choi, "Proximity based GPCRs prediction in transform domain," *Biochem, Biophys, Res. Commun*, vol. 371, pp. 411-415, 2008.

[20] A. Khan, A. Majid, and T. S. Choi, "Predicting protein Sub-cellular Location: Exploiting amino acid based sequence of feature spaces and fusion of diverse classifiers," *Amino Acids*, vol. 38, pp. 347-350, 2010.

[21] A. Khan, A. Majid, and A. M. Mirza, "Combination and Optimization of Classifiers in Gender Classification Using Genetic Programming," Int. *J. of Know. Int. Engg. Sys*, vol. 9, pp. 1-11, 2005.

[22] G. P. Zhou, "An intriguing controversy over protein structural class prediction," *J Protein Chem.*, vol. 17, pp. 729–738, 1998.

[23] G. P. Zhou and N. Assa-Munt, "Some insights into protein structural class prediction," *Proteins*, vol. 44, pp. 57–59, 2001.

[24] G. P. Zhou and K. Doctor, "Subcellular location prediction of apoptosis proteins," *Proteins Struct Funct Genet.*, vol. 50, pp. 44-48, 2003.

[25] Y. D. Cai, "Is it a paradox or misinterpretation," *Proteins Struct. Funct. Genet*, vol. 43, pp. 336–338, 2001.

Maqsood Hayat received his MCS degree from Gomal University, D I Khan, in 2004 and his MS degree in Software & System Engineering from Mohammad Ali Jinnah University (MAJU), Islamabad, in 2009. Currently, he received his Ph.D. degree in the Department of Computer & Information Sciences, Pakistan Institute of Engineering & Applied Sciences, Islamabad, Pakistan. His main research includes Machine learning and its application in Bioinformatics.

Asifullah Khan received his MSc degree in Physics from University of Peshawar, Pakistan in 1996 and his MS degree in Nuclear Engineering from Pakistan Institute of Engineering and Applied Sciences (PIEAS), Islamabad, Pakistan, in 1998. He received his MS and Ph.D. degrees in Computer Systems Engineering from Ghulam Ishaq Khan Institute of Engineering Sciences and Technology (GIK Institute), Topi, Pakistan, in 2003 and 2006 respectively. From October 1998 to June 2006, he had been working as a Senior Scientist at Pakistan Institute of Nuclear Sciences and Technology (PINSTECH). Currently, he is working as an associate professor in Department of Computer and Information Sciences at PIEAS. His research is as include Digital Watermarking, Pattern Recognition, Genetic Programming, Data Hiding, Machine Learning, and Computational Materials Science.