Human Membrane Protein Classification using Multi class Support Vector Machine (MCSVM)

Nijil Raj N, Dr. T. Mahalekshmi

Abstract — Membrane proteins are essential for the sustenance of organisms. They play crucial part in many biochemical processes, and it has attractive tar- gets of drug discovery for many illnesses. Technically the types of membrane protein help us to find out the function and structure of proteins. This paper proposes a new approach to classify human membrane protein types by combining amino acid properties and physicochemical properties by using Multi class Support Vector Machine (MCSVM).Features are extracted in two step method. In the beginning each feature extraction were considered separately and after that the more optimal feature sets were combined to retrieve the final feature set. The method is evaluated based on three different set of membrane proteins namely S1,S2 and S3. The two-step method classifies the membrane protein into six classes, in second step combination-3 MCSVM classifier revealed an accuracy of 83.33%, 86.11% and 88.89% for the datasets S1, S2 and S3 respectively.

Index Terms — MCSVM, Membrane protein

I. Introduction

Membrane proteins play crucial role in many biochemical processes and it is essential for the sustenance of organisms. They are considered as attractive targets of drug discovery for many illnesses. This membrane serves to separate and protect a cell from its surrounding environment and is made mostly from a double layer of phospholipids, which are amphiphilic. These biological membranes are made up of mainly lipid bilayers whereas functions are carried out by membrane proteins [3]. Proteins consist of three main classes which are classified as globular, fibrous and membrane proteins. Membrane associated proteins can be categories in the following two ways : Mode of interaction with the membranes & Cellular locations. Mem- brane proteins perform a variety of functions vital to the survival of organisms a) Membrane receptor proteins convey signals between the cell's internal and external environments. b) Membrane transport proteins shift molecules and ions across the membrane. c) Membrane enzymes have numerous actions. d) Cell linkage molecules allow cells to identify each other and interact. Membrane proteins are those that are found in biological membranes. As their name suggests, membrane proteins are anchored in membranes. They are the protein component of plasma membrane and may be broadly classified into: (1) integral membrane proteins and (2) peripheral membrane proteins Fig.1.

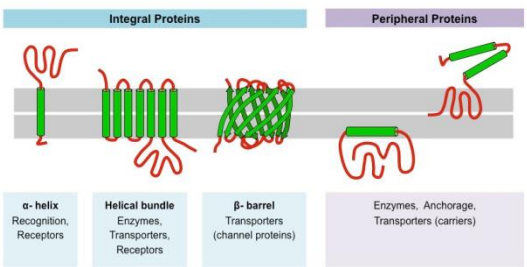

Figure. 1: Membrane Proteins

The membrane proteins can be categorized in to six types[18] as shown in Fig.2: 1.) Type I membrane proteins. 2.) Type II membrane proteins. 3.) Multipass transmembrane proteins.4.) Lipid chain anchored-membrane proteins. 5.) GPI-anchored-membrane proteins. 6.) Peripheral membrane proteins.

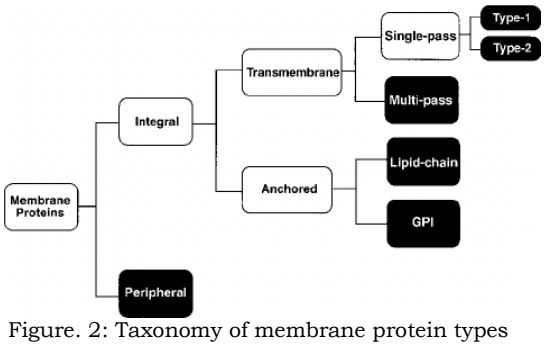

Figure. 2: Taxonomy of membrane protein types

Support vector machine (SVM) are generally used for binary classification. However, real- world problems often require the discrimination for more than two categories. Thus, the multi- class pattern recognition has a wide range of applications including optical character recognition [15], intrusion detection [11], speech recognition [7], and bioinformatics [2].In general the multi-class classification problems are commonly decomposed into a series of binary problems such that the standard SVM can be directly applied. Two representative ensemble schemes are one-versus-rest (1VR) [19] and one-versus one (1V1) [12] approaches. Both
1VR and 1V1 are special cases of the Error Correcting Output Codes (ECOC)[5]which decomposes the multi-class problem into a predefined set of binary problems.
In this paper we propose to classify the membrane protein into different classes. A two-step method is introduced and MCSVM can be used. It reveals better result compared to the existing method.

II. Related Works

Literature survey revealed existence of various methods to classify membrane protein into five types. The methods that use the amino acid composition apply various classifiers such as Hamming distance, Euclidian distance, ProtLock, and covariant discriminant analysis[13] result shown in Table-I. The methods based on pseudo amino acid composition are generally more accurate. They apply Hamming distance, Euclidian distance, ProtLock, covariant discriminant analysis, fuzzy K-nearest neighbor,

TABLE I: Rate of correct prediction of the membrane protein type by different test methods and algorithms

Algorithms				
Testing method	Least Hamming distance	Least Euclidean distance	Prot-Lock	Covariant discriminant
self-consistency	62.8%	63.5%	66.6%	81.1%
Jackknife	62.1%	62.8%	65.5%	76.4%
Independent data set	66.7%	69.2%	63.8%	79.4%

optimized evidence-theoretic K-nearest neighbor, supervised locally linear embedding, and various ensembles of classifiers to classify the protein membrane sequences.

Ensemble method for predicting subnuclear localizations from primary protein structures [8] consist of a novel two-stage multiclass support vector machine. It only considers those feature extraction methods based on amino acid classifications and physicochemical properties. The novel method which incorporates amino acid classifications and physicochemical properties into a general form of Chou's PseAAC by using a two-stage SVM method [9] is used to classify membrane proteins into five types. Two popular benchmark datasets were used to analyze the performance of the method, the training dataset and the independent dataset. The training dataset consist of 2059 protein sequences and 435 ,152,1311,51,110 sequences are type- I, type-II, multi-pass transmembrane, lipid-chain- anchored, and GPI anchored membrane proteins respectively. The independent dataset contains 2625 membrane protein sequences, which consists of 487 type I, 180 type II, 1867 multi- pass, 14 lipid-chains anchored, and 86 GPI anchored membrane proteins.

Here the accuracy is based on the sample size of data. The large sample sizes of two classes such as type I and multipass transmembrane proteins, in the dataset achieve highest prediction results.

The Table I represent the rate of correct pre- diction of membrane protein in existing methods and algorithms [4]

III. Materials And Methods

A. Dataset

3789 sequences of experimentally verified membrane proteins of homosapiens were down- loaded from the Uniprot database [1].

To analyze the performance of classifiers three set of data S1, S2 and S3 are constructed from 3789 membrane proteins sequences. Dataset S1 comprises 2883 membrane proteins, S2 contains 2081 membrane proteins and S3 have 1469 membrane proteins, shown in Fig.3.

Accession numbers are used to represent membrane proteins in datasets. Sequence based features of membrane proteins are used for membrane protein classification.

B. Feature-based Sequence Representation

A protein can be represented by a string of amino acids. Different proteins have different sequences, in terms of the ordering of their amino acids and length of the sequence. The first step in classifying proteins is to find a common way to represent the sequences. In this paper, a feature vector is adopted to represent protein chains. Any protein, regardless of the length or composition of its sequence, can be mapped to a feature vector representation. In this work, eight feature sets are used within the feature vector.

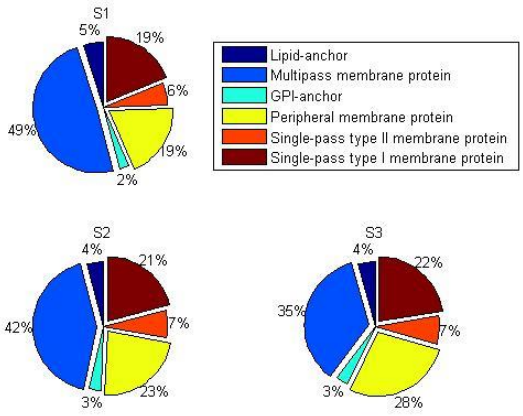

Figure: 3 Detailed view of S1, S2 & S3

1) Amino Acid Composition (AAC):

AAC is the normalized frequency of occurrence of each of the twenty amino acids in the given protein sequences. Thus, this feature set includes 20 features and it can be calculated from the Equation(1)

$$P = (\frac{N1}{N}, \frac{N2}{N}, \frac{N3}{N} \dots\dots \frac{N20}{N},)$$

Where N_i represents the number of type i amino acid, and N is the length of the sequence

2) Dipeptide Composition (DPC):

Dipeptide Composition [6] describes the proportion of each common amino acid pair within a sequence. Thus, this feature set includes 400 features and it can be calculated from the Equation(2)

$$P_{r,s} = \left(\frac{N_{rs}}{N-1}\right)$$

Where N_{rs} is the sum of dipeptides containing amino type r and types, and N is the length of the sequence.

3) Hydrophobicity:

Each amino acid has an associated hydrophobic index

TABLE II: Features & Dimensions

No.	Feature	Feature Dimension
Step 1		
1	AAC	20
2	DPC	400
3	Hydrophobicity	394
4	AAIndex	566
5	Daubechies wavelets	401
Step 2		
6	AAC + DPC + Hydro (COM1)	814
7	AAC + DPC + Hydro + AAIndex (COM2)	1380
8	AAC + DPC + Hydro + AAIndex + DB (COM3)	1781

affinity, which is often measured using a hydrophobic [16]. In a protein, hydrophobic amino acids are likely to be found in the interior, whereas hydrophilic amino acids are likely to be in contact with the aqueous environment. Several hydrophobicity scales have been published for various uses. Of this commonly used hydrophobicity scale is Kyte-Doolittle scale and this scale is used in this work. Here the feature set includes 394 features and is obtained by replacing the normalized protein sequences with the Kyte-Doolittle scale.

4) AAIndex :

It is a database of numerical indices representing various physicochemical and biochemical properties of amino acids and pairs of amino acids. AAindex [10] for the amino acid index of 20 numerical values. It gives a total of 566 features. The AAindex is released approximately annually. The latest version is the 9.0 release.

5) Daubechies Wavelet Transform:

The continuous wavelet transform based on Daubechies wavelets function is used to extract wavelet coefficients from the protein sequences. The Daubechies wavelets are chosen due to their successful applications in biological sequences analysis [14] [17]. The Matlab wavelet toolbox provides the tool for wavelet analysis. Here the Daubechies wavelets function gives total of 401 features .For this protein sequences are converted to numerical numbers based on the mapping scheme with respect to each amino acids in hydrophobic index [16].

In the initial phase, the feature extraction is done using the individual feature extraction method. In the next phase an optimal feature set is created by the union of feature sets obtained during the initial phase. For example, for the initial phase, if the number of feature extraction methods used was M, then M optimal feature subsets will be constructed. In the second step, for each classification, we can extract the optimal feature subset on the union of M optimal feature subsets obtained in the first step.

IV. Methodology

In this paper, a Multi class Support Vector Machine (MCSVM) is used for the membrane protein prediction. MCSVM's [4] are supervised learning models that are used for classification. It is formally defined by a separating hyper- plane. All optimal feature subsets are obtained by the two-step optimal feature selection procedure. A multi-class SVMs is used to predict the membrane protein types. The two-step optimal feature selection methods along with the multi- class support vector machine are considered to be effective method.

A. Computational Framework

The architectural framework for the membrane protein type's prediction is illustrated in Fig.4.

1) Data Collection :

The protein sequences are collected from the Uniport database [1]. Three benchmark dataset S1,S2 and S3 are constructed from 3,789 membrane proteins. Dataset

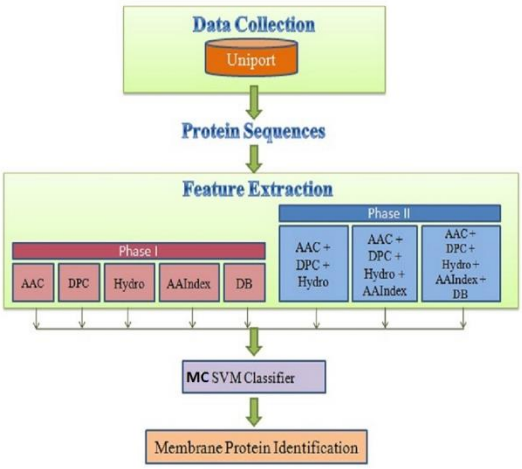

Figure. 4: The Framework for Membrane Protein Types classification

S1 contain 2883 membrane proteins, S2 posses 2081 membrane proteins, and S3 has 1469 membrane proteins.

2) Sequence - Based Feature Extraction:

In the two-step optimal feature extraction process combination of feature sets obtained in the initial feature extraction process was done.

Step One : After that the protein sequences are converted to feature vectors of particular dimensions corresponding to each type of individual feature extraction processes such as AAC, DPC, Hydrophobicity, AAIndex and Daubechies wavelets. Here the feature set dimensions are 20, 400, 394, 566,

TABLE III: Feature Set And Accuracy Of MCSVM

No	Feature Extraction Method	S1	S2	S3	Overall Accuracy
First stage					
1	ACC	53.33%	57.78%	46.67%	52.59%
2	DPC	70.56%	70%	70.67%	70.41%
3	AAINDEX	40%	42.22%	47.22%	43.15%
4	HYDRO	75.56%	77.78%	72.5 %	75.28%
5	DB	46.67%	42%	47.22%	45.3%
Second stage					
1	COM1	59.44%	63.33%	64.44%	62.4%
2	COM2	66.67%	72.22%	77.78%	72.22%
3	COM3	83.33%	86.11%	88.89%	86.11%

and 401 respectively.

Step Two: The three combinations of feature sets were used in the second step, which is given below.

- Combination I : The first combination consist of AAC, DPC and Hydrophobicity with a feature set dimension of 814.
- Combination II : The second combination consist of AAC, DPC, Hydrophobicity and AAIndex with a feature set dimension of 1380.
- Combination III : The third combination contains

AAC, DPC, Hydrophobicity, AAIndex and Daubechies wavelets with a feature set dimension of 1781 (refer table II)

3) Support Vector Machine :

Support vector machine (SVM) was initially designed for bi- nary classification. To extend SVM to the multi-class scenario, a number of classification models were proposed such as the one by Crammer and Singer (J Mach Learn Res 2:265292, 2001). However, the number of variables in Crammer and Singers dual problem is the product of the number of samples (l) by the number of classes Finally, a classification technique, the multi-class SVM(MCSVM) based classifier[20], is used to classify the membrane protein types. Training and testing is done on three benchmark datasets S1, S2 and S3.

B. Evaluation

The overall prediction accuracy A_{cc} , sensitivity S_{sn} , and specificity S_{sp} are used to evaluate the prediction of the performance of the work. The equations (3),(4)and(5) are shown below:

$$A_{cc} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

$$S_{sn} = \frac{(TP)}{(TP + FN)}$$

$$S_{sp} = \frac{(TN)}{(TN + FP)}$$

TP denotes the number of positive events that are correctly predicted; TN denotes the no of negatives events that are correctly predicted; FP denotes the number of negative events that are incorrectly predicted and FN is the number of subjects that are predicted to be negative despite they are positive.

V. Result & Discussion

Proposed method classifies homo sapiens membrane proteins into the following six classes, (1) Single -pass type I , (2) Single- pass type II, (3) Multi-pass, (4) Lipid-anchor, (5) GPI-anchor and (6) Peripheral membrane proteins. Sequence based feature extraction is adopted for proposed system which is achieved in two steps. In the first step, features are extracted using individual feature extraction method like Amino acid composition (AAC), Dipeptide composition(DPC) [6], Hydrophobicity [16], Amino Acid Index(AAIndex) and Daubechies wavelets. In the second step a three set of combination of feature sets are used. The first combination consists of AAC, DPC and Hydrophobicity with a feature set dimension of 814. The second combination consist of AAC, DPC, Hydrophobicity and AAIndex with a feature set dimension of 1380. And the third combination contains AAC, DPC, Hydrophobicity, AAIndex and Daubechies wavelets with a feature set dimension of 1781.The overall accuracy vs feature set is illustrated in Fig.5. Fig.6 illustrates the overall accuracy of eight feature set with different dataset

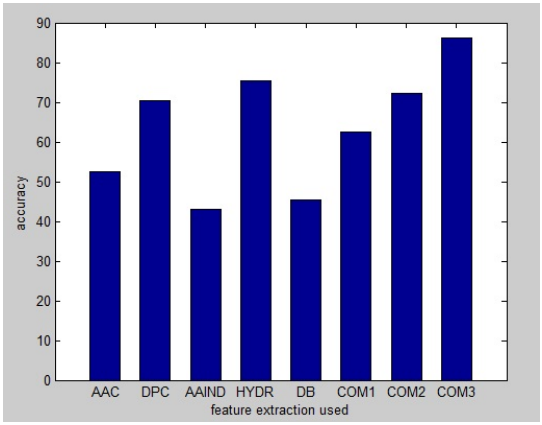

Figure 5: Overall Accuracy Vs Feature Extraction Method

The Given Table. III illustrates the accuracy of each dataset S1,S2,S3 obtained from the proposed method.

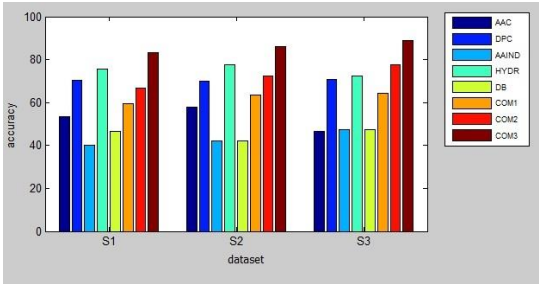

Figure. 6: Accuracy Vs Dataset

The two stage MCSVM classifier method revealed that a better result in combination 2 of second stage are 83.33%, 86.11% and 88.89% of datasets S1, S2 and S3 respectively.

Conclusion

In this paper we have described the multi class SVM(MCSVM) method for classifying the membrane proteins types based on the two step feature extraction method. Training and testing is done on three benchmark datasets S1, S2 and S3. All the 566 AAindex properties were used in feature extraction along with Amino acid composition, Dipeptide composition, and Hydrophobicity. In addition to this, the continuous wavelet transform based on Daubechies wavelets function is also used to extract wavelet coefficients from the protein sequences. The method is evaluated based on six types of membrane proteins. The multi-class SVM (MCSVM) classifier revealed an accuracy of 83.33%, 86.11% and
88.89% each of the three datasets S1, S2 and S3 respectively from the final feature set combination 3 in step 2.Our proposed method seems to be better than the existing method in accuracy wise and complexity wise.

Acknowledgment

The authors would like to acknowledge to all the faculties in dept of CSE,YCET.
Conflict of Interest: The authors declare that they have no competing interests for publishing this paper.

References

[1] Rolf Apweiler, Amos Bairoch, Cathy H Wu, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, et al. Uniprot: the universal protein knowledgebase. Nucleic acids research, 32 (suppl 1):D115–D119, 2004.

[2] Pierre Baldi and Gianluca Pollastri. A machine learning strategy for protein analysis. IEEE Intelligent Systems,17:28–35, 2002.

[3] Alberts Bruce, B Dennis, L Julian, R Martin, R Keith, and DW James. Molecular biology of the cell. garland publishing. New York, 19832:255–317, 1994.

[4] Yu-Dong Cai, Guo-Ping Zhou, and Kuo-Chen Chou. Support vector machines for predicting membrane protein types by using functional domain composition. Biophysical journal-Elsevier, 84:3257–3263, 2003.

[5] Bakiri G.: Dietterich, T. Solving multiclass learning problems via error-correcting output codes. J. Artif. Intell. Res, 2:263286, 1995.

[6] Inna Dubchak, Ilya Muchnik, Stephen R Holbrook, and Sung-Hou Kim. Prediction of protein folding class using global description of amino acid sequence. Proceedings of the National Academy of Sciences, 92(19):8700–8704,1995.

[7] Hamaker J.-Picone J. Ganapathiraju, A. Applications of support vector machines to speech recognition. IEEE Trans. Signal Process, 52(8):2348–2355, 19-july-2004.

[8] Guo Sheng Han, Vo Anh, Anaththa PD Krishnajith, Yu-Chu Tian, et al. An ensemble method for predicting subnuclear localizations from primary protein structures. PLoS One, 8(2):e57225, 2013.

[9] Guo-Sheng Han, Zu-Guo Yu, and Vo Anh. A two-stage svm method to predict membrane protein types by incorporating amino acid classifications and physicochemical properties into a general form of chou's pseaac. Journal of Theoretical Biology, 344:31–39, 2014.

[10] Shuichi Kawashima and Minoru Kanehisa. Aaindex: amino acid index database. Nucleic acids research,28(1):374–374, 2000.

[11] Awad M.-Thuraisingham B Khan, L. A new intrusion detection system using support vector machines and hier- archical clustering. The VLDBJ, 16(4):507–521, 2007.

[12] C. Smola-A. Kreel, U.Burges. Pairwise classification and support vector machines. MIT Press, Cambridge (1999),

[13] Yu-Dong Cai Kuo-Chen Chou. Prediction of mem- brane protein types by incorporating amphipathic effects. j.chem.inf.model, 45:407–413, 2005.

[14] JK Meher, MK Raval, PK Meher, and GN Dash. Wavelet transform for detection of conserved motifs inprotein sequences with ten bit physico-chemicalproperties. International Journal of Information and Electronics Engineering,2(2):200, 2012.

[15] Suen C.-Yamamoto K. Mori, S. Historical review of ocr research and development,. IEEE Computer Society Press, Los Alamitos, pages 244–273, 1995.

[16] George D Rose, Ari R Geselowitz, Glenn J Lesser, Richard H Lee, and Micheal H Zehfus. Hydrophobicity of amino acid residues in globular proteins. Science,229:834–839, 1985.

[17] M Sifuzzaman, MR Islam, and MZ Ali. Application of wavelet transform and its advantages compared to fourier transform. 2009.

[18] Ga´bor E Tusna´dy, Zsuzsanna Doszta´nyi, and Istva´n Simon.

[19] Transmembrane proteins in the protein data bank: identification and classification. Bioinformatics, 20(17):2964–2972, 2004.

[20] V Vapnik. Statistical Learning Theory. Wiley,New York,1998.

[21] Zhe Wang and Xiangyang Xue. chapter-2 multi- class support vector machine. In support vector ma- chine applications, volume VII, page 302. SPRINGER, http://www.springer.com/978-3-319-02299-4, 2014,

Authors' Detail

Nijil Raj N
Associate Professor,
Department of Computer Science and Engineering, Younus College of Engineering and Technology, Vadakkevila P.O, Kerala, Kollam–691010,india
Email: nijilrajn@ymail.com

Dr.T.Mahalekshmi
Principal, Sree Narayan Institute of Technology, Vadakkevila P.O, Kerala, Kollam-691010, India,
Email: mlakshmi.t@gmail.com

Cite this Article As :

Nijil Raj N, Dr. T. Mahalekshmi, "Human Membrane Protein Classification using Multi class Support Vector Machine," International Journal of Technology and Science, vol. 5, Issue. 1, pp. 1-5, 2018