National Conference on Innovations in Computing and Networks (NCICN) - 2019
© 2019 IJSRSET | Volume 5 | Issue 9 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099

Classification of Membrane Protein Types by Using Machine Learning Approach

Dr. NijilRaj*[1] ,Y asir A[2], Siyad S[3], Arun kumar A[3], Keerthi Krishna R[3]

*[1]Professor & Head, Department of Computer Science and Engineering, Younus College of Engineering and Technology, Pallimukku, Vadakkevila, Kollam, Kerala , India

[2]Assistant Professor, Department of Computer Science and Engineering, Younus College of Engineering and Technology, Pallimukku, Vadakkevila, Kollam, Kerala, India

[3] B. Tech, Department of Computer Science and Engineering, Younus College of Engineering and Technology, Pallimukku, Vadakkevila, Kollam, Kerala, India

ABSTRACT

The Membrane proteins are performing different cellular processes and important functions, which are based on the protein types. Each membrane protein have different roles at the same time this is called multi class classification. A general form of multi class classification is Multi-label classification. Each membrane proteins are lies in different classes at the same time that is known as multi label classification. The main feature of multilabel problem is that the instance can be assigned to any number of classes. Our proposed method is a multi label classification of membrane proteins by implementing machine learning algorithm like Logistic Regression Classification, Random Forest Classification and Neural Network Classification. An essential set of features are extracted from the homo-sapiens dataset S1 which are used for the proposed method, and it was revealed an accuracy of 89. 176%, whereas existing methods are revealed an accuracy is 58. 923%, 40. 769% for the Decision tree and Support vector machine respectively. Both accuracy wise and complexity wise, the proposed method seems to be better than the existing method.

Keywords : Multilabel Classification, Membrane Protein Type, Machine Learning

I. INTRODUCTION

Proteins are essential nutrients for the human body. They are one of the building blocks of body tissue or they are polymer chains made of Amino acids linked together by peptide bonds. Protein type classification methods are progressively used in various research fields. In protein type classification, one of the major types of protein is membrane protein. Membrane proteins[1] are proteins that are part of or interact with biological membranes. In our proposed method is a multi label classification[2]of different types of membrane proteins by implementing various machine learning approach. Membrane proteins play different roles in cellular biology. About 30% of human genomes have been encoded from membrane proteins. Information of a given membrane protein type helps to determine its function. Membrane proteins are referred as membrane associated proteins or membrane-bound proteins. Membrane proteins participate in important reactions of the cell, including transporting the substance into and out of the cell as a carrier, acting as a specific receptor for the hormone, carrying the recognition function of the cell and being responsible for signal transduction and cell-cell interactions[3]. In

addition, membrane proteins are of particular importance in drug therapy as the targets for many drugs [4].

On the basis of the interactions between membrane proteins and membrane. H. Lodish et. al[5]membrane protein are divided in to two types intrinsic and extrinsic membrane proteins The closely relation between the type and function of membrane proteins, knowing the type can provide clues for the structure and function of the protein . With the incredibly growing number of protein sequences discovered in the post genomic era, there is an urgent need for an effective method to predict membrane proteins and the introduction of machine learning methods greatly solve the problems. Based on their functions, membrane proteins can be classified into three classes: integral, peripheral and lipid-anchored. Membrane proteins are a common type of proteins along with soluble globular proteins, fibrous proteins, and disordered proteins. They are targets of over 50% of all modern medicinal drugs[6] classification of membrane proteins into eight types is a resource intensive and time consuming task. Therefore, developing an effective computational method is an urgent need for the protein functional type prediction. For that datasets S1 is constructed from Swiss prot database. It is reported from the performance of this method that it could be quite effective to classify membrane protein types

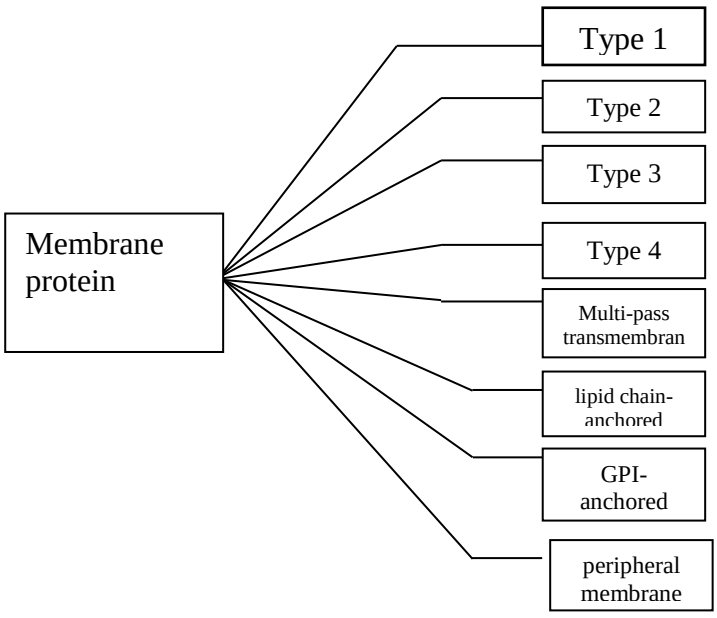

Figure1 : Membrane protein types

Based on the direct interaction relation between membrane proteins and lipid bilayers, the three classes can be further extended into eight basic types: (1) Type I membrane proteins, (2) Type II membrane proteins, (3) Type III membrane proteins, (4) Type IV membrane proteins, (5) Multi-pass trans membrane proteins, (6) Lipid chain-anchored membrane proteins, (7) GPI-anchored membrane proteins, (8) Peripheral membrane proteins. Among them, Types I, II, III, and IV are of single-pass trans membrane proteins.

II. Related work

M. Hayat, A. Khan, [7]Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition. In this paper, neural networks based membrane protein type prediction system is proposed. Composite protein sequence representation(CPSR) is used to extract the features of a protein sequence. The SVM success rates obtained using self consistency, jackknife, and independent dataset test are99. 9%, 86. 01%, and 95. 23% accuracy respectively, while that of PNN are 99. 9%, 82. 51%, and 95. 73%, respectively. These are the best prediction results reported so far and thus show the effectiveness of neural networks based classification strategies using CPSR based feature extraction for membrane protein type prediction

A. Garg, M. Bhasin, G. P. Raghava[8]Support vector machine based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. Garg *et al*[8]introduced a systematic approach for predicting subcellular localizations(SL) of human proteins. A set of human proteins with experimentally annotated SL has been retrieved from the SWISS-PROT database[9]. The SVM-based modules for predicting SL using traditional amino acid and di-peptide (i+1) composition achieved accuracy of 76. 6% and 77. 8%. PSI-BLAST, when carried out using a similarity-based search against a non

redundant database of experimentally annotated proteins, yielded 73. 3% accuracy.

Yu-Dong at el[10]proposed a new method for predicting the membrane protein types using the Nearest Neighbor Algorithm. They used manually constructed dataset from Swiss Prot [11]mainly according to the annotation line stated as SL, to classify the six types of membrane proteins. The predictor achieved the accuracy of 87. 02 %by using the 56 most contributive features .

Lipeng at el[12]proposed a new method in which, protein can be represented by a high dimensional feature vector by using Dipeptide composition method. They used membrane protein sequences from the dataset prepared by Chou and Elord, [13]with prediction accuracy of 82. 0%.

NijilRajNandT. Mahalekshmi[2]Multilabel Classification of Membrane Protein in Human by Decision Tree(DT)Approach. " In multi-label classification, each sample can be associated with a set of class labels. In protein type classification, one of the major types of protein is membrane protein. In this study proposes membrane protein type classification using Decision Tree (DT) classification algorithm. The DT classifies a membrane protein into six types . An essential set of features are extracted from the membrane protein dataset S1 which are used for the proposed method, and it was revealed an accuracy of 69. 81%

Author name	Method	Accuracy
A. Garg, M. Bhasin, G. P. Raghava, 2005	Support vector machine	73. 3%
Yu-Dong 2008	Nearest Neighbor Algorithm	87. 02 %
Lipeng 2010	Dipeptide composition method	82. 0%.
M. Hayat, A. Khan 2012	Neural networks	86. 01%
NijilRajN and T. Mahalekshmi 2018	Decision Tree classification algorithm	69. 81%

Table 1: Comparison of exisiting method with corresponding accuracies

Table 1 shows the comparison of existing method with corresponding accuracies

III. MATERIALS AND METHODOLOGY

Dataset

A total of 3249 human membrane protein sequence were downloaded and verified from Swiss prot Protein database. . In the proposed method use the dataset S1 for classification.

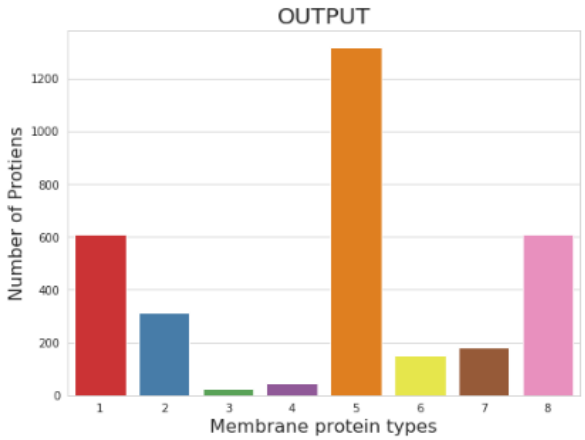

Figure 2 : Different Types of Membrane Proteins on Dataset S1

The figure 2 shows that the bar chart of different types of membrane proteins on dataset S1. From the bar chart Type 5 is the highest number of protein in dataset S1. Second highest is both Type 1 and Type 8. Type 3 is the least number of proteins in the dataset S1.

1. Feature Extraction

To establish an effective membrane protein prediction system, the key point is how to convert an original membrane protein sequence into a feature vector. To capture as much information of protein samples as possible and apply such feature extraction methodProteins are represented by a chain of amino acids. Features are usually extracted from the protein sequence. A sequence comprises of 20 unique amino acids namely A, C, D, E, F, G, H, I, J, K, L, M, N, P, Q, R, S, T, V, W, and Y.

Features	Dimension
Di-Amino Acid	400
Isoelectric point	1
Molecular weight	1
Aromaticity	1
Count	20
Total	423

Table 2 : Feature Vector and its Dimension

From the table2 represent the list of feature vector with dimension. There are five set of feature are extracted from the dataset S1. Totaly 423 feature are extracted for the each protein sequence of the dataset S1.

a. Count

Count of each amino acid residue is one of the feature of protein. For example, let 'AANDCC' be a amino acid sequence, count of amino acid residue A is 2, N is 1, D is 1 and C is 2. A total 20features are collected as count for each amino acid.

b. Di-Amino Acid Count

Amino acids frequency is the number of combinations of amino acid residue. The count of the combination of sequence pattern AA, AC, . . , AY, CA, CC, . . . CY, and. . , YA, YC, . . , YY in the protein sequence is called the amino acid frequency. From this, only count the combination of sequence patterns of Amino acid A, C, D, E. For example the sequence AA, AC, AD, AE, . . AY (20 numbers) and CA, CC, CD, CE. . . CY(20 numbers), and DA, DC, DD, . . . , DY (20 numbers) and EA, EC, ED, . . . EY (20numbers) are counted. As a total of 400 features are generated as frequency for a particular Protein sequence.

c. Molecular Weight

Molecular weight is the mass of a molecule. The size of a protein can be represented with the number of amino acids contained in that protein or by using molecular weight. It is represented by unit of Daltons or in KiloDaltons(KDa). tools used for finding the molecular weight of a protein from its protein sequence. For example, molecular weight of the sequence 'ACDEFGHIKLMNPQRSTVWY' is 2. 4 kilodaltons, and protein with protein id Q9P299 has the molecular weight of 23679. 0820 KDa.

d. Aromaticity

Aromaticity is a property of conjugated cycloalkenes in which the stabilization of the molecule is enhanced due to the ability of the electrons in the orbitals to delocalize.

This act as a framework to create a planar molecule. In organic chemistry, the term aromaticity is used to describe a cyclic (ring-shaped), planar (flat) molecule with a ring of resonance bonds that exhibits more stability than other geometric or connective arrangements with the same set of atoms. Aromatic molecules are very stable, and do not break apart easily to react with other substances. Organic compounds that are not aromatic are classified as aliphatic compounds they might be cyclic, but only aromatic rings have special stability (low reactivity).

e. Isoelectric point

The isoelectric point (pI, pH(I), IEP), is the pH at which a particular molecule carries no net electrical charge or is electrically neutral in the statistical mean. The standard nomenclature to represent the isoelectric point is pH(I), although pI is also commonly seen, [2] and is used in this article for brevity. The net charge on the molecule is affected by pH of its surrounding environment and can become more positively or negatively charged due to the gain or loss, respectively, of protons(H+). The pH at which the electrolyte concentration of an amphoteric substance such as protein is electrically zero because the concentration of its cation form equals the concentration of its anion form.

2. Methods

a. Logistic Regression Classification

Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to

describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

b. Random forest classification

Random forest or Random decision forests are ensemble learning method for classification and other task that are operated by constructing a multiple of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of individual trees.

c. Neural Network Classification

The neural network alogorithm are inspired by the human brain. It is interconnected and communicate with each other. Each connection is weighted by previous learning events and with each new input of data more learning takesplace

3. METHODOLOGY

The step by step explanation for membrane protein prediction by using various Classifier algorithms is shows in figure 3. In the first step 3249 membrane proteins from dataset S1 are used as input. In the next step the features are extracted from the dataset S1. Then applying various classifying algorithms such as logistic regression, neural network and random forest. Then evaluate the performance matrices and count corresponding accuracies.

Algorithm for predicting membrane protein types by applying various machine learning approach is follows;
Step1: Start.
Step2: Input Dataset S1 (3249 membrane protein sequence)
Step3: Pre processing the data from the data set S1
Step4: Extract the feature set from the dataset S1.
Step5: Apply various machine learning approaches for classifying membrane protein types.
step6: Evaluate the performance matrices.
Step7:prediction of membrane protein types
Step8:stop.

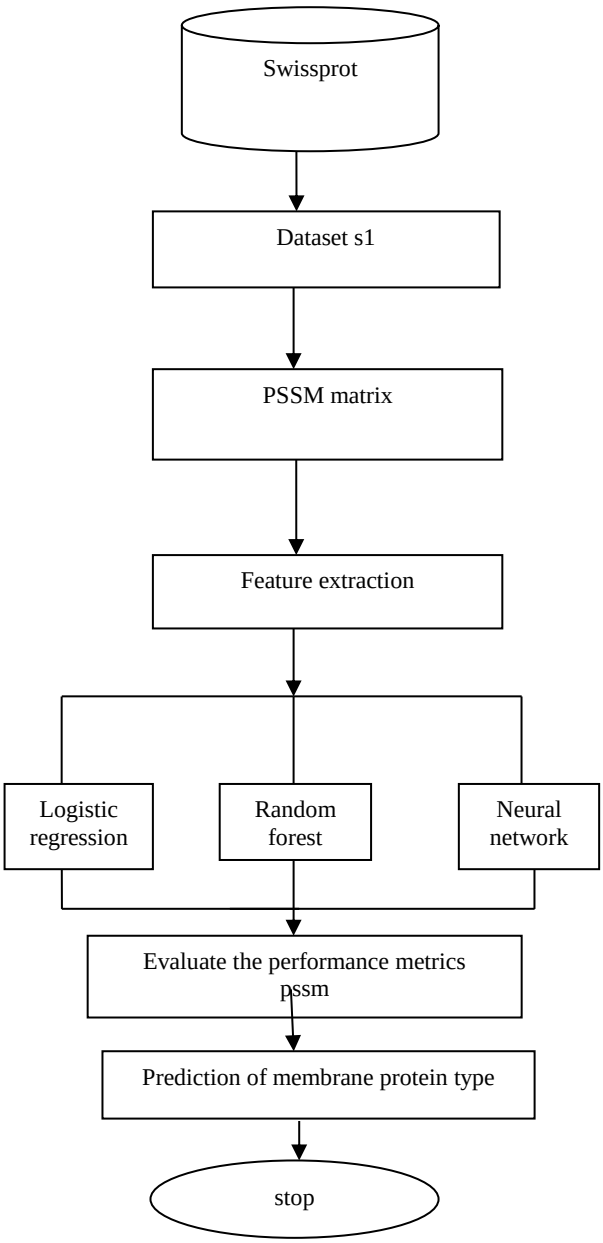

Figure 3 : Work flow diagram

IV. RESULTS AND DISCUSSION

This section depicts the results of both existing Method, and proposed classification algorithms using methods. The proposed classification performs classification on the dataset S1. This method uses the whole number of proteins from the dataset for the classification purpose. It is obvious that the Neural network method contributed the most, annotating 3249 proteins and achieved accuracy of 89. 176%, on datasets S1and the random forest classification with accuracy of 70. 154%, on the dataset S1 and logistic regression classification with accuracy 66. 769% on the dataset S1

	Author name	Method	Accuracy
Existing method	A. Garg, M. Bhasin, G. P. Raghava, 2005	Supportvector machine	73. 3%
	Yu-Dong 2008	Nearest Neighbor Algorithm	87. 02 %
	Lipeng2010	Dipeptide composition method	82. 0%.
	M. Hayat, A. Khan 2012	neuralnetworks	86. 01%
	NijilRajN and T. Mahalekshmi, 2018	Decision Tree classification algorithm	69. 81%
Proposed method		Logistic regression	66. 769
		Random forest	70. 154
		Neural network	**89. 176**

Table 3 : comparison between existing and proposed method with corresponding accuracies

The proposed classification accuracy Results are shown in the Table 3. From Table 3 it is obvious that the Neural network method contributed the most, annotating 3249 proteins and achieved accuracy of 89. 176%, on datasets S1, random forest classification with Accuracy of 70. 154%, on the dataset S1 and logistic regression classification with accuracy 66. 769% on the dataset S1.

A. Figures and Table

Algorithm	**Precision**	**Recall**	**Accuracy**
Logistic regression	62. 5	62. 5	66. 769
Random forest	76. 8	66. 5	70. 154
Neural N/W MLP	**84. 86**	**80. 30**	**89. 176**

V. CONCLUSION

In our proposed method, the system can predict the membrane protein type based on the effective accuracies followed by different classification algorithms. In the future, more features can be added than the existing system. This helps to improve the prediction accuracy. The use of classification algorithms helps to get the prediction type more accurate. As per the literature reveals that different classification algorithms are accurately predict that 5 types of membrane proteines. But in our proposed method predicting the 8 different types of membrane protein by machine learning approach. Our proposed method reveals that the Neural N/W MLP as the better classification algorithm for membrane protein type.

ACKNOWLEDGMENT

The Authors would like to acknowledge Dr. P. SreeRaj, Principal of younus college of engineering&technology, Dr. Nijilraj. N, Head of CSE Department in ycet, Prof. Yasir. A, Guide and project co-ordinator of cse department in ycet and also thanks to staffs from CSE department and younus college management, supporting for this work.

VI. REFERENCES

[1]. Lei GUO1, Shunfang WANG 1, Zhenfeng LEI 2, AND XUEREN WANG3" Prediction for Membrane Protein Types Based on Effective Fusion Representation and MIC-GA Feature Selection"

[2]. NijilRajNandT. Mahalekshmi, "MultilabelClassificationOf "MembraneProtein in Human byDecision Tree(DT)Approach. "Biomedical&PharmacologyJournalVol. 11(1), 113-121(2018)

[3]. M. S. Alme´n, K. J. Nordstro¨ m, R. Fredriksson, and H. B. Schio¨ th, "Mapping the human membrane proteome: a majority of the human

membrane proteins can be classified according to function and evolutionary origin, ” BMC biology, 7(1): p. 1: (2009).

[4]. J. P. Overington, B. Al-Lazikani, and A. L. Hopkins, ``How many drug targetsare there?'' Nature Rev. Drug Discovery, vol. 5, no. 12, pp. 993_996, Dec. 2006

[5]. H. Lodish, D. Baltimore, A. Berk, S. L. Zipursky, P. Matsudaira, and J. Darnell, Molecular cell biology. Scientific American Books New York, 3; (1995).

[6]. J. P. Overington, B. Al-Lazikani, and A. L. Hopkins, “How many drug targets are there?” Nature reviews Drug discovery, 5(12): 993–996 (2006).

[7]. M. Hayat and A. Khan, ``MemHyb: Predicting membrane protein types byhybridizing SAAC and PSSM, '' J. Theor. Biol. , vol. 292, no. 1, pp. 93_102, Jan. 2012

[8]. A. Garg, M. Bhasin, and G. P. Raghava, “Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and simi- larity search, ” Journal of biological Chemistry, 280(15): 14 427–14 432: (2005).

[9]. A. Bairoch and R. Apweiler, “The swiss-prot protein sequence database and its supplement trembl in 2000, ” Nucleic acids research, 28(1): 45–48 (2000).

[10]. P. Jia, Z. Qian, K. Feng, Yu-Dong W. Lu, Y. Li, and Y. Cai, “Prediction of membrane protein types in a hybrid space, ” Journal of proteome research, 7(3), : 1131–1137 (2008).

[11]. B. Boeckmann, A. Bairoch, R. Apweiler, M. -C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O’Donovan, I. Phan et al. , “The swiss-prot protein knowl- edgebase and its supplement trembl in 2003, ” Nucleic acids research, 31(1): 365–370 : (2003).

[12]. L. Wang, Z. Yuan, X. Chen, and Z. Zhou “The prediction of membrane protein types with npe, ” IEICE Electronics Express, 7(6): 397–402: (2010).

[13]. K. -C. Chou and Y. -D. Cai, ``Using GO-PseAA predictor to identifymembrane proteins and their types, '' Biochem. Biophys. Res. Commun. , vol. 327, no. 3, pp. 845_847, Feb. 2005.