ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 12, December 2017

Human Membrane Protein Classification Using KNN Multi Label Classifier (KNNMLC)

Nijil Raj N[1], Dr.T.Mahalekshmi[2]

Associate Professor, Department of Computer Science and Engineering, Younus College of Engineering and Technology, Vadakkevila P.O, Kollam, India[1]

Principal, Sree Narayan Institute of Technology, Vadakkevila, Kollam, India[2]

ABSTRACT: The Membrane proteins were found to be involved in various cellular processes performing various important functions, which are mainly associated to their types and location. One membrane protein can have several types at the same time ,that is a multi label class approach. This paper proposes, the K- Nearest Neighbor multi-label classification(KNNMLC) algorithm to classify the Membrane proteins. The KNN classifier classifies a membrane protein into the 6 classes of membrane protein types. Three set of data are used in this paper,which are D-I ,D- II,D-III respectively. An essential set of features were extracted from the membrane protein sequences.which are used for the proposed method(KNNMLC).KNNMLC method revealed an accuracy of 72.87%, 71.12%, 72.89% respectively,whereas the existing methods revealed an accuracy of network based 66.68%,62.46%,58.75%and shortest path 54.97%,48.75%,44.99%.The accuracy got in the existing methods are not for the full set of protein datasets,but it is achieved after removal of few unannotated protein.Both accuracy wise and complexity wise our proposed method seems to be better than existing method.

KEYWORDS: Membrane type classification,Multi-label classification,KNNMLC

I. INTRODUCTION

Human Membrane proteins are important parts of proteins playing various roles involved in cellular process Almen et.al[1]in the immune response serves as enzymes. About 30% of human genomes have been encoded from membrane protein. Knowledge of a given membrane protein type is helpful in determining its function. Membrane proteins also called membrane-bound proteins or membrane associated proteins, are classified according to two different schemes: one based on their interaction modes with membranes, and the other on their cellular locations. According to Gao et.al[2] estimated the number of membrane proteins was about 8000 in human. Almost 20-30% of all genes in most genomes encode membrane proteins[3]. In addition membrane protein constitute 60% of drug targets[4] which are crucial to new drug discovery as well as to understand the mechanism of cellular activities [4][5][6]. Wang et.al[7] observed that the functions of a membrane protein are closely associated with its type and location. However, it is time consuming and costly to determine types of uncharacterized membrane proteins by using traditional biophysical methods[8]. Thus there is a growing need for effective computational methods to predict the membrane protein types. Traditionally, depending upon the interactions between membrane proteins and membrane, some studies [9] broadly classified membrane proteins into two categories, namely integral (intrinsic) membrane proteins and peripheral (extrinsic) membrane proteins.
Integral membrane proteins are permanently bound to the biological membrane. Peripheral membrane proteins are temporarily attached to a membrane or integral membrane proteins. Integral membrane proteins are classified as Transmembrane proteins and Anchored membrane proteins. Transmembrane proteins are Single pass type I, Single- pass type II, Multi-pass, and Anchored membrane proteins are Lipid and GPI. According to their intramolecular arrangements and positions in a cell, membrane proteins are generally classified into the following six types[10], (1) Single -pass type I, (2)Single-pass type II, (3) Multi-pass, (4) Lipid-anchor, (5) GPI

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 12, December 2017

(Glycosylphosphatidylinisotol)- anchor, (6) Peripheral membrane proteins, as shown in Fig.1. 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[11].one membrane proteins can have several types at the same time ie,multi label classification.In multi label classification,each sample can be associated with a set of class labels. This paper proposes a KNNMLC approach to human membrane proteins . For that three dataset are constructed from UniProt database. It is reported from the performance of this method that it could be quite effective to classify membrane protein types.

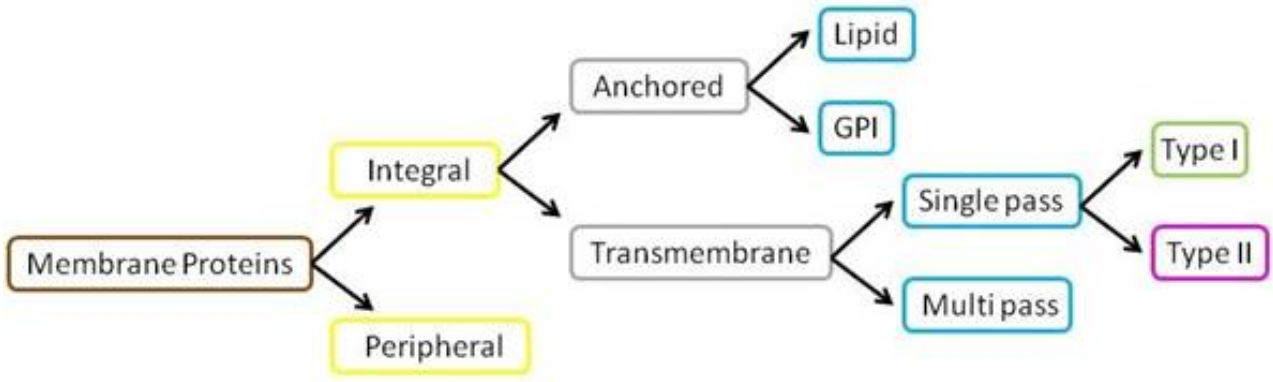


Fig. 1: Classification of Membrane Proteins

II. RELATED WORK

Several machine learning methods are used in the field of computational biology for classification and analysis of biological data. One of the major area is classifications of membrane proteins. Membrane proteins are classified based on their types. Some related works are explained for supporting our proposed method.

Membrane proteins are classified according to two different schemes by Kuo et al [12]which are based on protein types and location. Their dataset was constructed from the SWISS PROT (release 35) database. The overall rates of correct prediction thus obtained by both self-consistency and jackknife tests, as well as by an independent dataset test, are around 76-81% for the classification of five types, and 66-70% for the classification of nine cellular locations.

The above method is improved by using N-Terminal Amino Acid Sequence [13]. This method also uses the dataset from SWISS-PROT database, from which all sequences are extracted and some of the inappropriate sequences are removed before redundancy reduction, which is undertaken to avoid problems related to redundant data during Neural Networks training and testing. They implemented a Neural Network based tool Target for large scale subcellular localization prediction of newly identified proteins, with a success rate of 85% (plant) or 90% (nonplant) on redundancy reduced test were observed.

Another method incorporates a new strategy for the prediction of the types of membrane proteins using support vector machine, based on the concept of functional domain. Membrane proteins are generally classified into five types by Yu-Dong et al [14]. The dataset constructed by Chou and Elord[12] is used to demonstrate this method. They performed the prediction on only 2059 proteins and achieved a prediction accuracy of 86.3%.

Meng Wang et al. introduced Weighted-Support Vector Machines For Predicting Membrane Protein Types [15], which is devoted to combining the concept of pseudo amino acid composition and Weighted SVM to develop a new predictor for predicting the five types of membrane proteins. They used the same dataset constructed by Chou and Elrod (1999) [12], The dataset contains 2059 membrane protein sequences. Chou and Elrod classified the 2059 sequences into five groups and their prediction accuracy is about 82.3%.

Garg at el. [16] introduced a systematic approach for predicting subcellular localizations(cytoplasm, mitochondrial, nuclear, and plasma membrane) of human proteins. The dataset of human proteins with experimentally annotated subcellular localization has been derived from release 44.1 of the SWISS-PROT database[17]. The 7910 sequences

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 12, December 2017

from this database are screened strictly in order to develop a high quality dataset for predicting subcellular localization of human proteins. The final dataset consists of 3780 protein sequences that belong to 11 subcellular locations. The support vector machine (SVM)-based modules for predicting subcellular localization using traditional amino acid and dipeptide (i+1) composition achieved overall accuracy of 76.6% and 77.8%, respectively. PSI-BLAST, when carried out using a similarity-based search against a nonredundant database of experimentally annotated proteins, yielded 73.3% accuracy.

Yu-Dong at el. [8] proposed a new strategy for the prediction of the types of membrane proteins using the Nearest Neighbor Algorithm. They used a manually constructed dataset from Swiss-Prot (http://cn.expasy.org/, release 51.2) [18] mainly according to the annotation line stated as subcellular location, to classify the six types of membrane proteins. The predictor with 56 most contributive features achieved an acceptable prediction accuracy of 87.02%.

Lipeng at el.[19] proposed a new method in which, protein can be represented by a high dimensional feature vector by using Dipeptide composition method. They used only 2059 membrane protein sequences from the dataset constructed by Chou and Elord, based on the reduced low dimensional features. KNN classifier was introduced to identify the types of membrane proteins, with prediction accuracy of 82.0% [12].

Jei Lein at el. [20] classified protein based on chou's pseudo amino acid compostion with an Ensemble classifier. The proteins locations are generally classified into 5 types. The training and testing dataset that they used originally is constructed by Cedano et al. (1997)[21]. The composite KNN classifier predicted the proteins with location types (1) integral membrane proteins, (2) anchored membrane proteins, (3) extracellular proteins, (4) intracellular proteins (non-nuclear), and (5) nuclear proteins (M, A, E, I, N) with accuracy of 90.0%, 70.8%, 74.2%, 81.5%, 82.5% respectively.

According to Huang at el [22] for classifying six types of membrane proteins by using Network-Based Method and Shortest-Distance Method . They proposed an integrated approach to predict multiple types of membrane proteins by employing sequence homology and protein-protein interaction network. 3789 protein sequences of experimentally verified membrane proteins of human are downloaded from UniProt database[23]. According to their intramolecular arrangements and positions in a cell, membrane proteins are generally classified into the following six types 1.The network-based method achieved the highest Accuracy, i.e. 66.68%, 62.46%, 58.75% from the three datasets, respectively. Since no interactive proteins can be found in the corresponding datasets, there were 86, 38, 41 proteins unannotated. In the shortest-distance method, the lowest Accuracy was achieved (54.97%, 48.75%, 44.99% on the three datasets, respectively). However, all proteins can be annotated. The shortest distance method was capable of annotating all proteins, although it was least effective. Therefore the proposed method with KNNMLC adopted the same three dataset from this integrated method and perform multi label classification of membrane proteins with 967 features. The proposed method is capable of annotating all proteins from the three dataset.

III. PROPOSED METHODOLOGY

A. Dataset

In KNNMLC method using three sets of data which are downloaded from the UNIPORT database [23].Totally 3789 human membrane protein sequence were download and evaluate the performance of the prediction method.The sequence clustering program CD-HIT(Cluster Database at Height Identity Tolerance)[24] use to prepare the benchmark datasets D-I,D-II,D-III from 3789 protein sequence .Data set D-I comprises 2874 protein sequence in which protein had less than 70% sequence identity,D-II contain 2072 protein sequence in which protein had sequence identity lower than 40%,D-III have 1462 protein sequences with sequence identity less than 25%.

B. Methodology

The proposed methodology is illustrated in the Fig.2 and the step by step methods are following :

1) Retrieving the protein sequence from th uniport database by using protein id from the dataset D-I,D-II,D-III.
2) Preprocess the data sets and create the actual position specific scoring matrix(PSSM)
3) Extract the feature set vector(totally 967 features)from the dataset.
4) Perform a 10_Fold Cross Validation on these datasets and the features.
5) Apply K_Nearest Neighbor Classification on the above 10 Folds.
6) Combine the Results from the above step, which is the classified output.
7) Predict the result based on the classifier result
8) Evaluate the performance metrics.
9) Compare the actual matrix with the predicted matrix.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 12, December 2017

1) Preprocessing of Data: The three datasets of proteins are preprocessed according to their types and Protein id from the training dataset. For that create a Position Specific Scoring Matrix (PSSM) of each of the three datasets. The PSSM is the numerical representation of proteins in the dataset, which are presented in the 6 types of membrane proteins. The PSSM matrix consists of zeros and ones. If a Protein is presented in one or more membrane protein type, its entry in PSSM matrix is represented with ones, otherwise it is represented as zeros. This PSSM matrix is used for the evaluation of performance metrics, after creating the PSSM of classified output.
2) Feature Extraction: Features are extracted from the protein sequence which are collected from database. A protein sequence consists of 20 unique amino acids. The 20 amino acids are A, C, D, E, F, G, H, I, J, K, L, M, N, P, Q, R, S, T, V, W, Y. All amino acids have a common basic chemical structure, but possess different chemical properties due to differences in their side chains. A protein can be represented by a chain of amino acids. Different proteins have different amino acid string, in terms of the ordering and total number(length of the sequence). The proposed sequence based KNNML classifier used 967 distinct features. Extracted features are as follows:

a) Sequence length: Total number of amino acids in the given amino acid sequence.For example :the sequence of 'dcafgyhrdmscvsa' is 15.

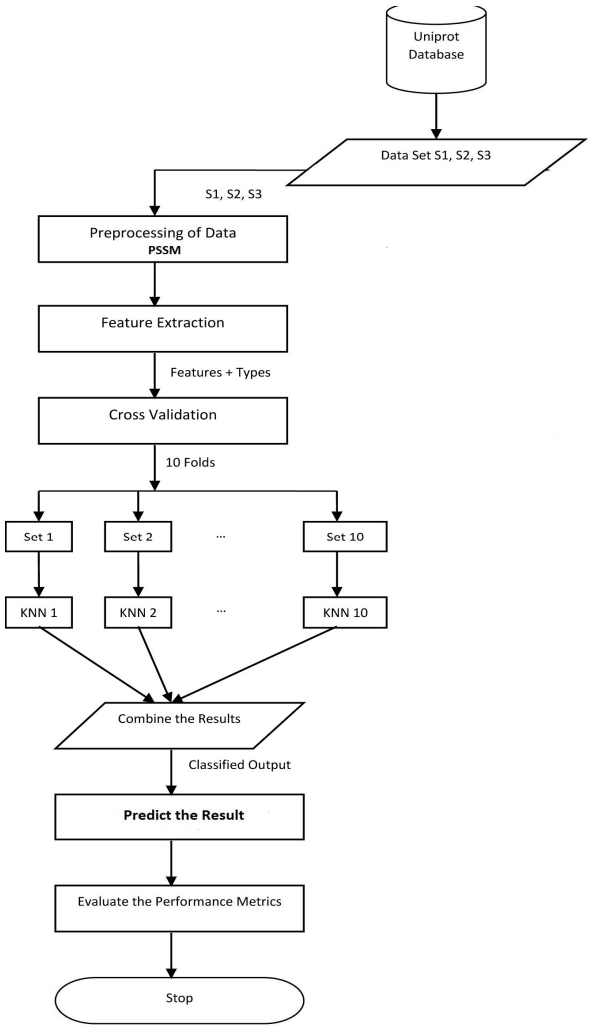


Fig. 2: Proposed KNN Multi-label Classification(KNNMLC)

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 12, December 2017

b) 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). (http://www.sciencegateway.org/tools/)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.

c) Count Of Each Amino Acid Residues: Amino Acid residues are building block of proteins. Count of each amino acid residue is one of the feature used. For Example, let 'AANDCC' be a amino acid sequence, count of amino acid residue A is 2, D is 1, C is 2 and N is 1.,as a total of 20 features are collected as count of each amino acid.

d) Di-amino acid: Di-Amino acids 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 (20 numbers) are counted, as a total of 400 features are generated as frequency for a particular Protein sequence.

e) AAindex: It is a database of numerical indices representing various physicochemical and biochemical properties of amino acids and pairs of amino acids.AAindex [25] for the amino acid index of numerical values gives a total of 544 features. The AAIndex is released approximately annualy. The latest version of the AAIndex is 9.2 realease Therefore a total of 967 features are extracted from each of the protein sequences from the three dataset as shown in Table.I.

3) KNN-Classification: Classification of membrane proteins using machine learning methods, such as K-Nearest Neighbor(KNN) classification. It is carried out by using MatlabR2012a. The 10-Fold Cross Validation is performed with KNN classification. K-Nearest Neighbor classification can be used for both classification and regression predictive

TABLE I: List of Features

FEATURES	COUNT
SEQUENCE LENGTH	1
MOLECULAR WEIGH	1
COUNT OF EACH AMI	20
DI-AMINO ACID	400
AAINDEX	544
TOTAL	967

problems. It is more widely used in classification problems. KNN makes decision based on the entire training data set according to their extrated features. The Cross validation is almost an inherent part of machine learning. Cross validation is used to compare the performance of different predictive modelling techniques. K-fold cross validation [26] is one way to improve over the holdout method. In K-fold cross validation, sometimes called rotation estimation, the dataset D is randomly split into K mutually exclusive subsets (the folds) D_1 , D_2 , ..., D_K of approximately equal size. Then the dataset is trained and tested K times. This method used a 10-fold cross validation, after that KNN classification is performed to each of the 10 sets from the 10-fold cross validation. Then combine the results from classification on 10 folds and create position specific scoring matrix of both the input dataset and the classified output types, and evaluate the performance with self consistency test. The three dataset D-I,D-II and D-III are revealed the better accuracies.

4) Performance Metrics: The overall classification accuracy of a classification model is evaluated using Self Consistency test. It involves training and testing the model with same dataset. This test is also termed as Resubstitution test, which is used to test the three dataset. For multi-label classification, the concepts such as Precision, Recall, Accuracy[28] is used to measure the performance of methods. The performance of the classifiers[29] is accessed

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 12, December 2017

through the following standard parameters. In order to find the values of Precision, Recall, Accuracy, calculate the True Positive, True Negative, False Positive, False Negative. For that calculate the count of 1 values and 0 values in actual score matrix. Then generate the total count of 0 and 1 as N. Next calculate True positive (tp) is the count of 1 values in the intersection of actual score matrix and predicted score matrix. Similarly True Negative (tn) is the count of 0 values in the intersection of actual score matrix and predicted score matrix. False Positive (fp) and False Negative (fn) are calculated using the formula,

$$fp = n - tn \quad (1)$$

$$fn = p - tp \quad (2)$$

Using these values, calculate Accuracy, Precision, Recall from the following equations.

a) Accuracy: It is the percentage prediction of true examples namely, True prediction divided by the total number of examples. Then the accuracy is defined by the equation(3), but more generalised form is shown in the equation (4)

$$\text{Accuracy} = (tp + tn)/N \quad (3)$$

Let D is a dataset with N instances. Let Y_i and Z_i are the set of original and predicted labels, respectively, where i D, then the accuracy becomes,

$$Acc_i = 1/N \left(\sum_{i=1}^{n} (|Y_i \cap Z_i|)/(|Y_i \cup Z_i|) \right) \quad (4)$$

b) Precision: It is the number of correct predictions divided by the number of all returned prediction. It is calculated using the following equation (5), but more generalised form is shown in the equation (6)

$$\text{Precision} = tp/(tp + fp) \quad (5)$$

$$Pre_i = \sum_{i/ie D \wedge k \epsilon Z_j} (|Y_i \cap Z_i|)/(|Z_i|) \quad (6)$$

TABLE II: Performances of KNNMLC approach on dataset D-I,D-II,D-III

DATASET	ACCUR	PRECISIO	RECALL
D-I	72.87%	0.2169	0.2197
D-II	71.12%	0.1441	0.1367
D-III	72.89%	0.2033	0.1944

c) Recall: It is the number of correct predictions divided by the number of predictions. It is calculated using the following equation (7), but more generalised form is shown in the equation (8)

$$\text{Recall} = tp/p \quad (7)$$

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

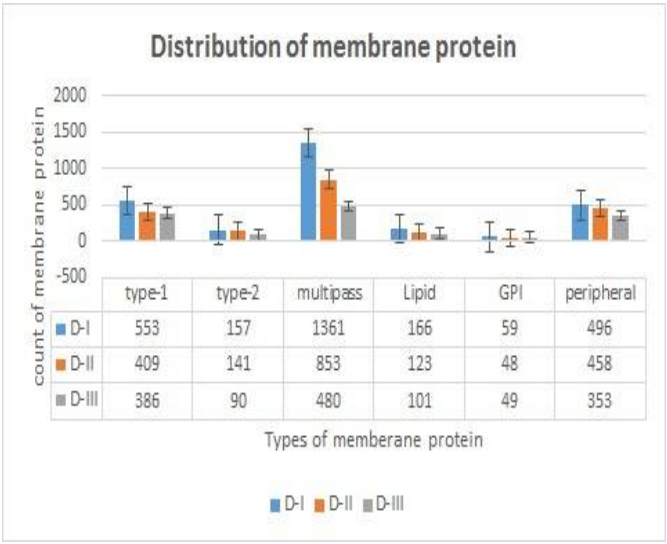
Website: www.ijircce.com

Vol. 5, Issue 12, December 2017

$$Rec_i = \sum_{i/i\epsilon D \wedge k\epsilon Y_i} (|Y_i \cap Z_i|)/(|Y_i|) \qquad (8)$$

IV. RESULTS AND DISCUSSION

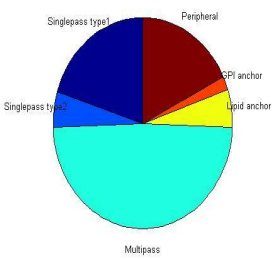
This section depicts the results of both existing Network Based Method, Shortest Distance Method and proposed KNNMLC. The results of proposed methods are compared with results of existing methods. From the analysis, the KNNMLC is an efficient multi-label classifier for classifying the human membrane proteins into the following six classes, (1) Single -pass type I, (2)Single-pass type II, (3) Multi-pass, (4) Lipid-anchor, (5) GPI (Glycosylphosphatidylinisotol)-anchor, (6) Peripheral membrane proteins.



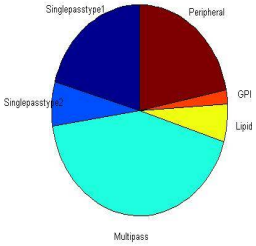
	type-1	type-2	multipass	Lipid	GPI	peripheral
D-I	553	157	1361	166	59	496
D-II	409	141	853	123	48	458
D-III	386	90	480	101	49	353

Fig. 3: The Distribution of Types of Membrane Proteins on Dataset D-I,D-II,D-III

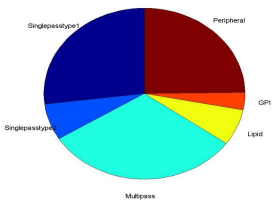
Fig.4 demonstrate the distribustion of different types of membrane proteins in different datset D-I,D-11 andD-III by using pie chart digram. Figure 3(a),(b),(c) shows the output of KNN classification on dataset D-I, D-II, D-III respectively.



(a) D-I



(b) D-II



(c) D-III

Fig. 4: KNN classification on D-I, D-II & D-III

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 12, December 2017

TABLE III: Comparison Of Accuracy: Existing Vs Proposed
(NU* represents Number of Unannotated Proteins)

DATASET	CLASSIFICATIONMETHODS					
	NETWORK		SHORTEST		KNNMLC	
	ACC	NU*	ACC	NU*	ACC	NU*
D-I	66.68%	86	54.97%	0	72.87%	0
D-II	62.46%	38	48.75%	0	71.12%	0
D-III	58.75%	41	44.99%	0	72.89%	0

The multipass, lipid, GPI, peripheral, type1, type2 membrane proteins are represented by the colours, green, yellow, orange, brown, dark blue, light blue respectively.
The distribution of membrane proteins to their types in three dataset by KNN classification are shown in Fig.3. Therefore in KNN classification the more number of proteins are classified as Multipass and the less number of proteins as GPI in all the three dataset with all annotated proteins. The Accuracy, Precision, Recall, are calculated and the results are shown in Table.II The Multilabel KNN classification gives the better results when compared to the existing methods. The classification accuracies are reached 72.87%, 71.12%, 72.89% on D-1, D-II, D-III respectively. Therefore from the results KNN classification achieves better classification accuracy.

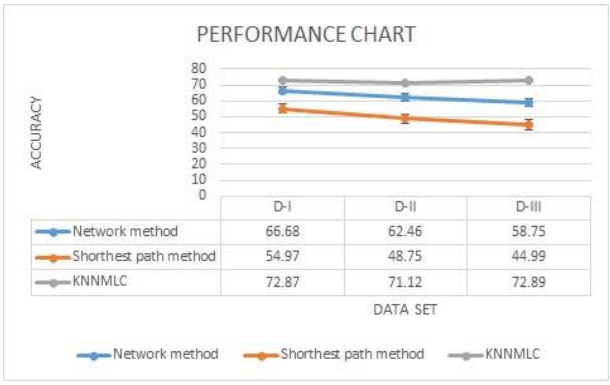


Fig. 5: Accuracies of Existing method verses proposed KNNMLC

In figure.5 represents the performance chart for the existing method and the proposed KNNMLC method.Its shown that the accuracy wise our proposed method is better than the existing method.

V. CONCLUSION

The multi-label classification methods are increasingly required by modern applications, such as protein function classification, text categorization, music categorization etc. In multi-label classification, each sample can be associated with a set of class labels. This paper proposed an efficient K-Nearest Neighbor multi-label classification algorithm. It is used to classify the Membrane proteins according to their types, based on the 967 protein sequence based features. The 2874, 2072, 1462 membrane proteins of three datasets D-I, D-II, D-III are classified using KNNMLC based on the features extracted from these proteins. As a result, the KNNMLC , revealed an acceptable accuracies of three dataset 72.87%, 71.12%, 72.89% respectively, by using 10_fold cross validation . Therefore the KNNMLC method is anticipated to be a better method for classifying multi label membrane protein types with acceptable accuracies on three datasets.Complexity and accuracy wise our proposed KNNMLC method is better than the existing method.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 12, December 2017

REFERENCES

[1] M. S. Almén, K. J. Nordström, R. Fredriksson, and H. B. Schiöth, "Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin," BMC biology, vol. 7, no. 1, p. 1, 2009.
[2] Q.-B. Gao, X.-F. Ye, Z.-C. Jin, and J. He, "Improving discrimination of outer membrane proteins by fusing different forms of pseudo amino acid composition," Analytical biochemistry, vol. 398, no. 1, pp. 52–59, 2010.
[3] A. Krogh, B. Larsson, G. Von Heijne, and E. L. Sonnhammer, "Predicting transmembrane protein topology with a hidden markov model: application to complete genomes," Journal of molecular biology, vol. 305, no. 3, pp. 567–580, 2001.
[4] Y. Arinaminpathy, E. Khurana, D. M. Engelman, and M. B. Gerstein, "Computational analysis of membrane proteins: the largest class of drug targets," Drug discovery today, vol. 14, no. 23, pp. 1130–1135, 2009.
[5] J. Davey, "G-protein-coupled receptors: new approaches to maximise the impact of gpcrs in drug discovery," Expert opinion on therapeutic targets, vol. 8, no. 2, pp. 165–170, 2004.
[6] G. C. Terstappen and A. Reggiani, "In silico research in drug discovery," Trends in pharmacological sciences, vol. 22, no. 1, pp. 23–26, 2001.
[7] J. Wang, Y. Li, Q. Wang, X. You, J. Man, C. Wang, and X. Gao, "Proclusensem: predicting membrane protein types by fusing different modes of pseudo amino acid composition," Computers in biology and medicine, vol. 42, no. 5, pp. 564–574, 2012.
[8] P. Jia, Z. Qian, K. Feng, W. Lu, Y. Li, and Y. Cai, "Prediction of membrane protein types in a hybrid space," Journal of proteome research, vol. 7, no. 3, pp. 1131–1137, 2008.
[9] H. Lodish, D. Baltimore, A. Berk, S. L. Zipursky, P. Matsudaira, and J. Darnell, Molecular cell biology. Scientific American Books New York, 1995, vol. 3.
[10] K.-C. Chou and Y.-D. Cai, "Prediction of membrane protein types by incorporating amphipathic effects," Journal of chemical information and modeling, vol. 45, no. 2, pp. 407–413, 2005.
[11] J. P. Overington, B. Al-Lazikani, and A. L. Hopkins, "How many drug targets are there?" Nature reviews Drug discovery, vol. 5, no. 12, pp. 993–996, 2006.
[12] K.-C. Chou and D. W. Elrod, "Prediction of membrane protein types and subcellular locations," Proteins: Structure, Function, and Bioinformatics, vol. 34, no. 1, pp. 137–153, 1999.
[13] O. Emanuelsson, H. Nielsen, S. Brunak, and G. Von Heijne, "Predicting subcellular localization of proteins based on their n-terminal amino acid sequence," Journal of molecular biology, vol. 300, no. 4, pp. 1005–1016, 2000.
[14] Y.-D. Cai, G.-P. Zhou, and K.-C. Chou, "Support vector machines for predicting membrane protein types by using functional domain composition," Biophysical journal, vol. 84, no. 5, pp. 3257–3263, 2003.
[15] 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 Engineering Design and Selection, vol. 17, no. 6, pp. 509–516, 2004.
[16] 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 similarity search," Journal of biological Chemistry, vol. 280, no. 15, pp. 14 427–14 432, 2005.
[17] A. Bairoch and R. Apweiler, "The swiss-prot protein sequence database and its supplement trembl in 2000," Nucleic acids research, vol. 28, no. 1, pp. 45–48, 2000.
[18] 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 knowledgebase and its supplement trembl in 2003," Nucleic acids research, vol. 31, no. 1, pp. 365–370, 2003.
[19] L. Wang, Z. Yuan, X. Chen, and Z. Zhou, "The prediction of membrane protein types with npe," IEICE Electronics Express, vol. 7, no. 6, pp. 397–402, 2010.
[20] J. Lin, Y. Wang, and X. Xu, "A novel ensemble and composite approach for classifying proteins based on chous pseudo amino acid composition," African Journal of Biotechnology, vol. 10, no. 74, pp. 16 948–16 952, 2011.
[21] J. Cedano, P. Aloy, J. A. Perez-Pons, and E. Querol, "Relation between amino acid composition and cellular location of proteins," Journal of molecular biology, vol. 266, no. 3, pp. 594–600, 1997.
[22] G. Huang, Y. Zhang, L. Chen, N. Zhang, T. Huang, and Y.-D. Cai, "Prediction of multi-type membrane proteins in human by an integrated approach," PloS one, vol. 9, no. 3, p. e93553, 2014.
[23] U. Consortium et al., "The universal protein resource (uniprot) in 2010," Nucleic acids research, vol. 38, no. suppl 1, pp. D142–D148, 2010.
[24] W. Li, "Fast program for clustering and comparing large sets of protein or nucleotide sequences," in Encyclopedia of Metagenomics. Springer, 2015, pp. 173–177.
[25] S. Kawashima and M. Kaneshisa., "Aaindex:amino acid index database," Nucleic acids research, vol. 28(1), p. 374, 2012.
[26] R. Kohavi et al., "A study of cross-validation and bootstrap for accuracy estimation and model selection," in Ijcai, vol. 14, no. 2, 1995, pp. 1137–1145.
[27] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped blast and psi-blast: a new generation of protein database search programs," Nucleic acids research, vol. 25, no. 17, pp. 3389–3402, 1997.
[28] M. Deng, K. Zhang, S. Mehta, T. Chen, and F. Sun, "Prediction of protein function using protein-protein interaction data," Journal of Computational Biology, vol. 10, no. 6, pp. 947–960, 2003.
[29] M. Hayat and A. Khan, "Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition," Journal of Theoretical Biology, vol. 271, no. 1, pp. 10–17, 2011.