Predicting Membrane Protein Types Using Naïve Bayesand K-nearest Neighbor Classifier

A. Haleema Fazal[1], E. Siva Sankari[2], D. Manimegalai[3]

[1,2] Department of CSE, Government College of Engineering, Tirunelveli, Tamil Nadu, India
[3]Department of IT, National Engineering College, Kovilpatti, Tamil Nadu, India

ABSTRACT

Knowledge of membrane protein's structure and function has significance in biological and pharmacological studies since more than 50% of drugs are targeted against membrane proteins. The experimental determination of membrane protein types, despite being more accurate and reliable, is not always feasible due to the high prices of laboratory procedures, there by creating a need for the development of automated bioinformatics methods. Consequently, an automated approach is particularly effective, that could help in identifying the new membrane protein types. In an existing work, three benchmark datasets (i.e.) data set1, data set2 and data set3 of primary sequence of membrane proteins were used. From that primary sequence of the membrane proteins 89 features were extracted and it is available online. SVM classifier was implemented and had obtained an accuracy of 88.2% for the data set3. In the proposed work, only data set3 with the readily available 89 features is used and two classifiers Naïve Bayes and K Nearest Neighbor are implemented. The overall accuracy achieved in this work is 93.51% for Naïve Bayes classifier and 95.72% for KNN classifier.

Keywords: K-nearest Neighbor (KNN), Naïve Bayes, Membrane protein types, Prediction, Extracted Features.

1. INTRODUCTION

According to cellular anatomy, a cell consists of different functional units or organelles, most of which are enveloped by membranes and they are necessary to form any biological functions. Although the lipid bilayer is the basic structure of membranes, most of the specific functions of the cell membrane are performed by the membrane proteins(Alberts et al.1994;Lodish et al.1995)[1,2].Many researchers believe that membrane proteins constitute approximately 50% of possible targets for novel drugs[3]. Membrane proteins can generally be classified into eight types[1]: 1) Type-I transmembrane proteins, 2) Type-II transmembrane proteins, 3)Type-III transmembrane proteins 4)Type-IV transmembrane proteins 5) Multipass transmembrane proteins, 6) lipid chain-anchored membrane 7) GPI anchored membrane proteins 8) peripheral membrane proteins (Fig. 1) [4] .

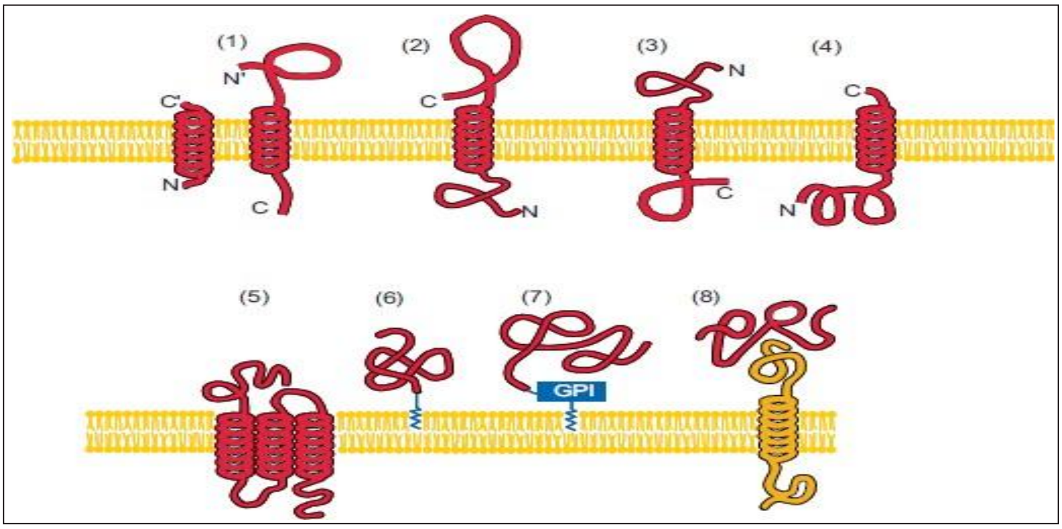

Figure 1: Graphical illustration shows the eight types of membrane proteins

In the last few years, numerous approaches have been evolved for predicting membrane protein types. Chou and Elrod (1999) first achieved the prediction of membrane protein types using Covariant Discriminant Algorithm(CDA) and Amino Acid Composition(AAC). Further, Caietal.(2004) used AAC together with Support Vector Machine (SVM) to predict membrane protein types. But sequence- order and sequence-length effects are lost. To avoid this short coming, the pseudo-amino acid composition (PseAAC) was proposed[17,18,19] to improve the prediction accuracy of membrane protein types. Then, numerous forms of the PseAAC method were proposed[5-15] for predicting membrane protein types and related tasks. Chou and Shen (2007)[4] proposed the Pse-PSSM method and developed a web server for predicting membrane protein types.

2. MATERIALS AND METHODS

Protein sequences were collected from the Swiss-Prot data base http://www.ebi.ac.uk/swissprot/ (Version 72 released on 2011- October-19) by using the annotation of „„„protein subcellular localization‟‟‟. The dataset contains 6677 sequence with both training and testing dataset. Number of proteins in each testing and training dataset has been shown in Table 1. The following exclusion criteria were implemented to ensure the quality of the dataset: (1) Proteins that are either annotated as fragments, or are shorter than 50 amino acid residues in length were excluded; (2) proteins that are either annotated with non-experimental qualifiers in topology, or with more than one topology were removed; (3) homologous sequences were removed by CD-hit if they share a high sequence identity (greater than 40%) with any sequence in the dataset. The resulting sequences were classified into their respective membrane protein types based on the annotation of the topology before they were randomly assigned into the training and the testing set by using the percentage distribution method[17] .

Table 1: Training and Testing dataset of Membrane protein types

Membrane Protein types	Training Dataset	Testing Dataset
Single pass type I	561	245
Single pass type II	316	79
Single pass type III	32	9
Single pass type IV	65	17
Multipass transmembrane	1119	2478
Lipid-chain-anchored membrane	142	36
Gpi-anchored membrane	174	41
Peripheral membrane	674	699
Total	**3073**	**3604**

Sequence features used in this study includes:(1) the length and the number of transmembrane segments; (2) the presence or absence of lipid-binding domains, signal peptides, signal anchors, GPI-anchoring signals, or intracellular N-terminal sequences; and (3) the composition of surface amino acids, the size of cationic patches, the protein flexibility, the degree of hydrophobicity of a protein segment, the propensity of protein lipidation, and the charge difference between the two flanking segments across the cell membrane. The sequence partition and the five-level grouping composition strategy to represent the aforementioned physical and biochemical properties for protein sequences was also used. The dataset along with features extracted are available in http://bsaltools.ym.edu.tw/predmpt/down load.html as Dataset3[17] .

A. Machine Learning Algorithm

The basic premises of all the machine learning algorithms are the same by using a set of known examples to obtain information about unknown example. The known examples are usually called training set and unknown examples are called testing set. Machine learning algorithms can be classified into supervised and unsupervised learning. Supervised learning usually consists of concerning a series of attributes of the data to a specific class or numerical value known as a label of that specific example. In contrast, in unsupervised learning, there are no predefined classes or labels.

3. CLASSIFICATION

Classification is the sub-discipline of data mining where data is assigned to the predefined groups. It is usually known as supervised learning because the labels of these groups or classes are known in advance. In a classification process, classes are determined on the basis of data attribute values and characteristics of already known data for which these classes are defined.

A. Naïve Bayes

The Bayesian classification is a probabilistic learning approach .This classification technique is based totally on Bayes‟ Theorem. It assumes that the presence of a specific feature in a class is unrelated to the presence of any other feature. Bayes theorem provides a way of calculating posterior probability P(c|x) from prior probability P(c), evidence P(x) and likelihood P(x|c.

Steps

1. D : Set of tuples,
2. Each Tuple is an ‘n’ dimensional attribute vector
3. X : ($x1,x2,x3,\ldots.xn$)
4. Let there be ‘m’ Classes :C1,C2,C3…Cm
5. Naïve Bayes classifier predicts X belongs to Class C iff

P(Ci/X) >P(Cj/X) for 1<= j <=m

6. Maximum Posteriori Hypothesis

P(Ci/X) = P(X/Ci) P(Ci) /P(X)

7. Maximize P(X/Ci) P(Ci) as P(X) is constant
8. With many attributes, it is computationally expensive to evaluate P(X/Ci).
9. Naïve Assumption of “class conditional independence”
10. P(X/Ci) = P(x1/Ci) * P(x2/Ci) *…* P(xn/Ci)

B. K-Nearest Neighbor

KNN is the most well known classifier in the area of pattern recognition, regression, and classification owing to its simplicity, adaptability, high performance, and simple to realize. Irrespective of its simplicity, it is able to provide competitive and incredible performance as compared to many different learning algorithms. KNN is a non-parametric classification algorithm and has no ahead information about the distribution of the data. It has no explicit training phase, while keeping all the training data in testing phase. The uncategorized instances are classified through nearest neighbors in the feature space. It is also called as instance base learner or lazy learner. The KNN learner primarily based at the perception of distance, which calculates the distance between the protein query and the training instances. Finally, the specified number of K instances from the feature space is selected, which has closest distance from protein query. Atlast, the most frequently occurring class is assigned to the protein query. When the number of K=1 it is called nearest neighbor classifier, otherwise, it refers to KNN classifier, which makes the decision on majority voting scheme. In case of a tie, the decision is made by assigning randomly one of the associated classes with the tie to protein query. However, this situation happens very rarely, because the number of K is mostly odd. The overall prediction performance of KNN classifier improves and reduces the impact of noise in the classification when the variety of nearest associates will increases. On the other hand, the computational cost rises and also makes the boundaries less distinct between classes.

Pseudo code

```
Classify (X,Y,x)//X: training data, Y: class labels of X, x: unknown sample,
For i = 1 to m do
Compute euclideandistance ,d(Xi,x)

$(Xi,x) = \frac{Xi,x}{1-(|Xi||x|)}$

End For
Compute set I containing indices for the k-smallest distances d(Xi,x) Return majority label for {Yi} Where, ||Xi||||x|| is the dot product of the two vectors
||Xi|| and ||x|| are the moduli
```

4. PERFORMANCE EVALUATION

In this present study metrics such as accuracy, sensitivity, specificity and MCC are used.

A. Accuracy

Accuracy determines the degree of true prediction of a model either true positive or true negative. It is the proportion of true predictions. It can be calculated as,

$Acc = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-}$…(1)

where, $N_-^+$is false negative, $N_+^-$is false positive, N^+ is total number of positive samples, N^- is the total number of negative samples[15,16]. The accuracy of classifiers is shown in Table 2.

Table 2: Comparison of prediction accuracy among four methods of membrane protein identification

Membrane Protein Types	SVM(%) [17]	MEMTYPE-2L(%) [17]	NB(%)	KNN(%)
Single pass type I	85.369.0		94.8	94.5
Single pass type II	64.658.2		87.4	95.7
Single pass type III	22.255.6		97.3	99.1
Single pass type IV	58.852.9		98	99.1
Multipass transmembrane	92.590.7		97.5	98.9
Lipid-chain-anchored membrane	50.033.3		96.8	95.8
Gpi-anchored membrane	85.465.9		82.2	91.4
Peripheral membrane	80.343.9		94.1	91.3
Overall	**88.278.3**		**93.52**	**95.72**

B. Sensitivity

Itshowstheratiobetweenthepredictedtruepositiveinstancesandtotalnumberoftruepositiveinstances,

$S_n = 1 - \frac{N_-^+}{N^+}$□ □ □□ □

where, $N_-^+$is false negative, N^+ is total number of positive samples[16]. The sensitivity of classifiers is shown in Table 3.

Table 3: Comparison of sensitivity of membrane protein identification

Membrane Protein Types	SVM(%) [17]	NB(%)	KNN(%)
Single pass type I	85.3	75.9	79.6
Single pass type II	64.6	48.1	43.0
Single pass type III	22.2	33.3	11.1
Single pass type IV	58.8	70.6	23.5
Multipass transmembrane	92.5	56.1	80.5
Lipid-chain-anchored membrane	50.0	50	22.2
Gpi-anchored membrane	85.4	74.3	89.5
Peripheral membrane	80.3	78.4	70.8
Overall	**88.2**	**60.8**	**52.5**

C. Specificity

It shows the ratio between the predicted true positive instances and total number of true positive instances,

$s_p = 1 - \frac{N_+^-}{N^-}$…(3) □

where, N_+^-is false negative, N– is the total number of negative samples [16].The specificity of classifiers is shown in Table 4.

Table 4: Comparison of specificity of membrane protein identification

Membrane Protein Types	SVM(%) [17]	NB(%)	KNN(%)
Single pass type I	75.7	96.2	95.6
Single pass type II	31.7	88.3	96.9
Single pass type III	33.3	97.5	99.3
Single pass type IV	41.7	98.1	99.4
Multipass transmembrane	98.2	97.9	99.1
Lipid-chain-anchored membrane	20.7	97.3	96.5
Gpi-anchored membrane	71.4	99.5	95.7
Peripheral membrane	84.4	97.9	96.2
Overall	**57.1**	**96.6**	**97.3**

D. Mathew's Correlation Coefficient (MCC)

It is considered as one of the most effective and rigorous performance parameters for any prediction method. MCC takes values in the interval [-11], where by 1 indicates that the classifier predicts the entire examples as correct and -1 indicates that the classifier predicts all the examples as incorrect.

MCC is able to recover the disadvantage of accuracy concerning an unbalance data. For example, if positive examples are more than negative examples in the dataset, due to bias the classifier can predict most of the positive examples. Thus, the performance of classifier will be affected because it predicts majority of the negative examples incorrectly. Therefore, the MCC is deemed as the best performance parameter for the classification of unbalanced data.

$$MCC = \frac{1-\left(\frac{N^{+}_{-}}{N^{+}}-\frac{N^{-}_{+}}{N^{-}}\right)}{\sqrt{\left(1+\frac{N^{-}_{+}N^{+}_{-}}{N^{+}}\right)\left(1+\frac{N^{+}_{-}N^{-}_{+}}{N_{-}}\right)}} \ldots(4)$$

where, $N^{+}_{-}$is false negative, $N^{-}_{+}$is false positive, N^{+} is total number of positive samples, N^{-} is the total

number of negative samples[15]The MCC of the classifiers is shown in Table 5.

Table 5: Overall Mathew's Correlation Coefficient

Method	MCC
SVM[17]	0.57
NB	0.80
KNN	0.71

5. RESULTS AND DISCUSSION

From Table 2 it can be seen that KNN and Naïve Bayes works well for all protein types compared to previous works. KNN is the first best classifier as it handles multiclass problems. For each protein type accuracy have increased ranging from 82% to 98% for NaïveBayes and 91% to 99% for KNN. TypeIV and III accuracy is highest as Naïve bayes can predict well with small amount of data. Multipass transmembrane has larger data samples yet Naïve Bayes can predict it with greater accuracy because the feature considered for training transmembrane i.e. the number of transmembrane segment and hydrophobicity of protein sequence, can provide better knowledge for the classifier likewise the feature used for idenfifying type III is signal peptide which also provides greater accuracy for its type.

The overall sensitivity of SVM is greater than Naïve Bayes and KNN as they have less overfitting. Only type III and IV is increased by 11.1 and 11.8 for Naïve bayes as it works well with smaller samples of data. KNN is sensitive to all types except GPI as they are sensitive to neighbourhood structure.

Overall specificity of KNN is greater as they work well in larger dataset and more or less same for Naïve Bayes as it can do well in practice with enough representative data. Type III, IV, multipass have 99% specificity because of less misclassification rate in KNN. Specificity of Gpi-anchor is increased because it is identified using False Positive rate of

GPI anchoring signals in NaïveBayes.

The MCC of Naïve Bayes and KNN is greater when compared to SVM as they are robust.

CONCLUSION

Classification of membrane protein types is an important task drug discovery and enzymes. It even helps researchers to discover a new drug. In this paper, we applied different data mining classification approach like KNN, Naïve Bayes. The used KNN and Naïve Bayes shows the better accuracy than the SVM approach and it has shown 7.52% and 5.32 improvements.

REFERENCES

[1] Alberts,B.,Bray,D.,Lewis,J.,Raff,M.,Roberts,K.,Watson,J.D.,1994.*MolecularBiologyoftheCell. Garland Publishing,* New York &London.

[2] Lodish, H., Baltimore, D., Berk, A., Zipursky, S.L., Matsudaira, P., Darnell, J., 1995. *Molecular Cell Biology*. Scientific American Books, NewYork.

[3] Gao,Q.B.,Ye,X.F.,Jin,Z.C.,He,J.,2010.Improving discrimination of outer membrane proteins by fusing different forms of pseudo-aminoacid composition. *Anal.Biochem.*,**398**: 52–59.

[4] K.C. Chou and H.B. Shen, 2007. "MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM," *Biochemical and Biophysical Research Communications,* **360**(2):339–345.

[5] Wang,M.,Yang,J.,Liu,G.P.,Xu,Z.J.,Chou,K.C.2004.Weighted-supportvec to machines for predicting membrane protein types based on pseudo amino acid composition. *Protein Eng. Des. Sel.,* **17**: 509–516.

[6] Wang, M., Yang, J., Xu, Z.J., Chou, K.C. 2005. SLLE for predicting membrane protein types. *J..Biol.*,**232**: 7–15.

[7] Wang,S.Q.,Yang, J.,Chou, K.C. 2006. Using tacking generalization to predict membrane proteintypes based on pseudo-aminoacid composition. *J. Theor. Biol.,* **242**: 941–946.

[8] Shen, H.B., Chou, K.C. 2005. Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo-aminoacid composition to predict membrane protein types. *Biochem. Biophys. Res. Commun.,* **334**: 288–292.

[9] Shen,H.B.,Yang,J.,Chou,K.C.2006.FuzzyKNNforpredictingmembraneproteintypesfrompseudo-aminoacid composition. *J. Theor. Biol.,* **240**: 9–13.

[10] Lin, H. 2008. ThemodifiedMahalanobis discriminant for predicting outer membrane proteins byusing Chou's pseudo-aminoacidcomposition.*J.Theor. Biol.,* **252**: 350–356.

[11] Mahdavi,A.,Jahandideh,S.2011.Applicationofdensitysimilaritiestopredictmembraneproteintypes based on pseudo-aminoacid composition. *J. Theor. Biol.,* **276**:132–137.

[12] Wang,J.Y.,Li,Y.P.,Wang,Q.Q.,You,X.G.,Man,J.J.,Wang,C.,Gao,X.2012.ProClusEnsem:predicting membrane protein types by fusing different modes of pseudo-aminoacid composition. *Comput. Biol. Med.,* **42**: 564–574.

[13] Hayat, M., Khan,A. 2012a. MemHyb: predicting membrane protein types by hybridizing SAAC and PSSM.*J. Theor. Biol.,* **292**: 93–102.

[14] Hayat, M., Khan, A. 2012b. Discriminating outer membrane proteins with fuzzy K-nearestneighbour algorithms based on the general form of Chou's Pse AAC. *Protein Pept.Lett.*,**19**: 411–421.

[15] Cheng, X., Zhao, S.G., Xiao, X. 2017. iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals. *Bioinformatics,* **33**: 341-346.

[16] E.S. Sankari and D. Manimegalai, 2017. "Predicting membrane protein types using various decision tree classifiers based on various modes of general PseAAC for imbalanced datasets," *J. Theor. Biol.,* **435**: 208–217.

[17] Yen-KuangChen and Kuo-Bin Li 2013. "Predicting membrane protein types by incorporating protein topology,domains,signalpeptidesandphysiochemicalpropertiesintothegeneralformofChou'spseudo amino acid composition" *J. Theor. Biol.,* **318:**1-12.

[18] Hayat,Khan,2011."Predictingmembraneproteintypesbyfusingcompositeproteinsequencefeatures into pseudo amino acid composition" *J. Theor. Biol.,* **271**:10-17.

[19] Chou "Prediction of protein cellular attributes using pseudo-amino acid composition" **43**:246-55.