Multilabel Classification Of Membrane Protein in Human by Decision Tree(DT) Approach

N. NIJIL RAJ1 and T.MAHALEKSHMI2

1Department of Computer Science and Engineering, Younus College of Engineering and Technology, Vadakkevila,Kollam-691010,India.
2Principal, Sree Narayan Institute of Technology, Vadakkevila, Kollam-691010, India.
*Corresponding author Email: nijilrajn@ymail.com

http://dx.doi.org/10.13005/bpj/1353

(Received: December 28, 2017; accepted: February 15, 2018)

ABSTRACT

Multi-label classification methods are important in various fields,such as protein type,protein function, semantic scene classification and music categorization. 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. The Membrane proteins are performing different cellular processes and important functions, which are based on the protein types. Each membrane protein have different rolls at the same time. In this study we 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%, whereas existing methods network based and shortest path revealed an accuracy of 66.78%,54.97%.The accuracy got in the existing methods are not for the full set of protein in dataset S1, but it is achieved after removal of few unannotated protein. Both accuracy wise and complexity wise, the proposed method seems to be better than the existing method.

Keywords: Multi-label classification,DT,Membrane type classification.

INTRODUCTION

Multilabel classification methods are progressively used in recent research works, protein function,protein type,semantic scene classification and music categorization.A general form of multi class classification is Multi-label classification.It is single-label problem of grouping instances into one of more than two classes. The main feature of multi-label problem is that the instance can be assigned to any number of classes. We proposes a multi label classification of different types of membrane proteins by implementing DT classifier algorithm. 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 refereed as membrane associated proteins or membrane-bound proteins . They are classified on the basis of their interaction modes with membranes, and cellular locations. Membrane proteins play important role in-involved in various cellular processes. The number of membrane proteins in humans is to 8000 as per the estimation of Gao et al. According to Krogh et.al 20-30% of genes are involved in encoding membrane proteins. The role of membrane proteins the discovery of new drugs as well as in the analyses.
of the mechanism of cellular activities is worth mentioning\(^4,5,6\). All membrane protein functions are usually related with its type\(^7\). Application of traditional biophysical methods\(^8\) are time consuming and costly while determining the types of uncharacterized membrane proteins. On the basis of the interactions between membrane proteins and membrane, H. Lodish et al.\(^9\) membrane protein are divided into two types intrinsic and extrinsic membrane proteins. (Fig. 1)

Transmembrane protein (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 type I, type II, and Multi-pass, whereas Anchored membrane proteins are Lipid and GPI. Based on the positions and intramolecular arrangements in a cell, membrane proteins are classified into six types\(^10\), shown in Fig. 2.

In MPT, the polypeptide crosses the lipid bilayer multiple times, i.e., spanning the membrane more than once. LCM are covalently linked to a lipid molecule and serve to anchor them to either the cytoplasmic or extracellular surface of a biological membrane. GPI-anchored membrane protein is also called membrane-anchored proteins. It is bound to the membrane by a glycosylphosphatidylinositol (GPI) anchor.

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\(^12\). It is estimated that 20-30% of all genes in most genomes encode membrane proteins\(^3\). Thus classification of membrane proteins into six types is a resource intensive and time consuming task. Therefore, developing a reliable and effective computational method is an urgent need for the protein functional type prediction. This paper proposed a multi-label classification of membrane proteins in humans, using DT classifier algorithm. For that datasets S1 is constructed from UniProt database. It is reported from the performance of this method that it could be quite effective to classify membrane protein types.

**Related work**

The computational methods used for the classification of membrane proteins include analytical methods, mathematical modeling and simulation. The bioinformatics application generally use strategical analysis methods, like machine learning methods for the classification and prediction of membrane proteins.

For the successful implementation of machine learning techniques are equally important. both feature extraction and learning algorithms are equally required. Feature predictions commonly used are: amino acid composition (AAC), position-specific scoring matrices (PSSMs), pseudo amino acid composition (PseAAC), physicochemical properties of amino acids and functional domains. AAC was simplest and most efficient representation of protein sequence. Membrane proteins are classified according to two different schemes by Kuo et al.\(^13\) which are based on protein types and its location. Their dataset was constructed from the SWISS PROT (release 35) database. The rate of correct
prediction of membrane proteins type and cellular location revealed that 76-81% and 66-70% by using the self consistency, jackknife tests, as well as by an independent dataset test. This method was improved by using N-Terminal Amino Acid Sequence. It also used the dataset from SWISS-PROT database, from which all sequences were extracted and some of the inappropriate sequences were removed before redundancy reduction. It was undertaken to avoid problems related to redundant data during Neural Networks training and testing. A success rate of 85% (plant) or 90% (non plant) on redundancy reduced test sets were observed. Garg et al. 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. The final dataset consists of 3780 protein sequences that belong to 11 SL. The SVM-based modules for predicting SL using traditional amino acid and dipeptide (i+1) composition achieved accuracy of 76.6% and 77.8%. 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 proposed a new method for predicting the membrane protein types using the Nearest Neighbor Algorithm. They used manually constructed dataset from Swiss-Prot (http://cn.expasy.org/, release 51.2) 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 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 prepared by Chou and Elord. Based on the reduced low dimensional features KNN classifier was introduced to identify the membrane protein types, with prediction accuracy of 82.0%. Jei Lein et el classified protein based on Chou’s pseudo amino acid composition with an Ensemble classifier. The protein locations are classified into 5 types. The testing and training dataset that they used originally was prepared by Cedano et al. (1997). The composite KNN classifier predicted the proteins with location types (1) nuclear proteins, (2) intracellular proteins (non-nuclear), (3) extracellular proteins, (4) anchored membrane proteins, and (5) integral membrane proteins (M, A, E, I, N) with accuracy of 90.0%, 70.8%, 74.2%, 81.5%, 82.5% respectively.

For classifying 6 types of membrane proteins 3 methods such as, BLAST/PSI-BLAST Method, Network-Based Method, Shortest-Distance Method were introduced by Huang at el. They proposed an integrated approach to predict multiple types of membrane proteins by employing sequence homology and protein-protein interaction network. According to their positions and intramolecular arrangements in a cell, membrane proteins are classified into six types: (1) GPI (Glycosylphosphatidylinositol) - anchor; (2) Lipid-anchor (LCM); (3) Multi-pass (MPT); (4) Peripheral (PM); (5) Single-pass type I; (6) Single-pass type II membrane proteins shown in Fig.3. To evaluate the performance of classification method, the sequence clustering program CD-HIT was employed (Cluster Database at High Identity with Tolerance) to construct three datasets: S1, S2, S3 from 3789 proteins. S1 contained 2935 protein sequences in which protein had less than 70% sequence similarity. S2 contained 2120 protein sequences in which protein had sequence similarity lower than 40%. S3 contained 1475 protein sequences with sequence identity less than 25%. The BLAST/PSI-BLAST method achieved the best performance with the highest accuracy 94.71%, 91.15% and 85.02% on datasets S1, S2, and S3, respectively. However, 481, 529 and 620 proteins cannot be annotated from data set. The network-based method achieved the second highest accuracy, i.e. 66.68%, 62.46%, 58.75% on the three datasets, S1, S2, S3 respectively. Since no interactive proteins can be found in the corresponding datasets, there were 86, 38, 41 proteins unannotated. The shortest distance method was capable of annotating all proteins, although it was least effective with lowest Accuracy achieved (54.97%, 48.75%, 44.99% on the three datasets, respectively). The proposed method is capable of annotating all proteins from the dataset S1. It uses 967 features from each of the membrane protein sequences.

MATERIAL AND METHODOLOGY

Dataset
A total of 3789 human membrane protein sequences were downloaded and verified from
Uniprot Protein database (release 2012). To evaluate the performance of the prediction method, W.Li et al\textsuperscript{23} use the sequence clustering program CD-HIT (Cluster Database at Height Identity Tolerance)\textsuperscript{24} to prepare the benchmark set of data S1 from 3789 containing 2935 proteins sequences with sequence similarity less than 70%. In our proposed method we use the dataset S1(2935 proteins) used for classification.

Methodology

The flow diagram for the proposed methodology is in Fig: 4 and the step by step procedures are as follows:

Step1: Start.
Step2: Input Dataset S1 (2935 membrane protein sequence)
Step3: Preprocessing the data from the data set S1, and create position specific scoring matrix (PSSM)
Step4: Extract the feature set from the dataset S1.
Step5: Apply the DT classifier algorithm for classifying membrane protein types.
Step6: Evaluate the performance matrices.
Step7: Stop.

Preprocessing of Data

The S1 datasets of proteins are preprocessed according to their types and Protein id from the training dataset. For this create a Position Specific Scoring Matrix (PSSM) of the 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.

Feature Extraction

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. Even though all amino acids have a common basic chemical structure, they exhibit different chemical properties because of the differences in their side chains. Proteins are represented by a chain of amino acids. The difference in the amino acid string among proteins is due to their order and total number (length of the sequence). The proposed DT classification used 968 distinct features. Extracted features are as follows:

Sequence length

The total number of amino acids in the given protein sequence. For example: the sequence length of ‘acdfyrsmeacvss’ is 15
Table 1: List of Features from Dataset S1

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Di-Amino Acid</td>
<td>400</td>
</tr>
<tr>
<td>Hydrophobicity</td>
<td>1</td>
</tr>
<tr>
<td>Aaindex</td>
<td>544</td>
</tr>
<tr>
<td>Count of Each Amino Acids</td>
<td>20</td>
</tr>
<tr>
<td>Sequence Molecular Weight</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>967</td>
</tr>
</tbody>
</table>

Table 2: Performances of Decision Tree classification tested on dataset s1

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>69.81%</td>
<td>0.1165</td>
<td>0.1157</td>
</tr>
</tbody>
</table>

Table 3: Comparison Of Classification Method

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Network</th>
<th>Shortest</th>
<th>DT</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>AC C</td>
<td>AC C</td>
<td>AC C</td>
</tr>
<tr>
<td></td>
<td>NU*</td>
<td>NU*</td>
<td>NU*</td>
</tr>
<tr>
<td></td>
<td>66.68%</td>
<td>54.97%</td>
<td>69.81%</td>
</tr>
<tr>
<td></td>
<td>86</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Hydrophobicity

The hydrophobicity of an amino acid is related to its transfer free energy from a polar medium (such as the cytoplasm) to another polar medium (like a membrane). The transfer free energy depends on the chemical nature of the two solvents, as well as on the structural context of the amino acid residue. The hydrophobicity index is a measure of the relative hydrophobicity. This index is used to measure the hydrophobic affinity of a protein sequence or an amino acid sequence. 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.

AA index

It is a database representing numerical values of various physicochemical and biochemical properties of amino acids and pairs of amino acids. AAINDEX is for the amino acid index of 20 numerical values. It gives a total 544 features. Every year the updated version (9.0) of AAINDEX is released.

Di-Amino Acid

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 (20 numbers) are counted. As a total of 400 features are generated as frequency for a particular Protein sequence.

Count Of Each Amino Acid Residues

Amino Acid residues are the 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. A total 20 features are collected as count for each amino acid.

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, molecule weight of the sequence ‘ACDEFGHIKLMN-PQRSTVWY’ is 2.4 kilodaltons, and protein with protein id Q9P299 has the molecular weight of 23679.0820 KDa.

Decision Tree Classification (DT)

A DT is a decision support tool that uses a tree structure graph or model of decisions and their
possible consequences, including chance event outcomes, resource costs, and utility. Decision trees are commonly used in operations research, specifically in decision analysis, to identify a strategy most likely to reach a goal. A DT is a flowchart-like structure: internal node represents a test on an attribute, branch represents the outcome of the test and leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represents classification rules. The extracted features are provided as input to DT classifier. The DT classifies the protein types into six types according to the rule, with classification accuracies 69.81%, on the dataset S1.

Performance Metrics: The overall classification accuracy of a classification model is evaluated using Self Consistency test. It use training and testing the model with same dataset. This test is also termed as Resubstitution test, which is used to test the dataset. For multi-label classification, the concepts such as Precision, Recall, Accuracy are used to measure the performance of methods. The following standard parameters are used to evaluate the performance of classifier. In order to find the values of Precision, Recall, Accuracy, calculate the True Positive(tp), True Negative(tn), False Positive(fp), False Negative(fn). 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 \( tp \) which is the count of 1 values in the intersection of actual score matrix and predicted score matrix. Similarly \( tn \) is the count of 0 values in the intersection of actual score matrix and predicted score matrix. \( fp \) and \( fn \) are calculated using the equation (1) and (2).

\[
fp = n - tn \quad \text{ ...(1)} \\
fn = p - tp \quad \text{ ...(2)}
\]

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

**Accuracy**

It is the percentage prediction of true examples i.e., True prediction divided by the total number of examples. The accuracy is defined by the equation (3), but more generalised form is shown in the equation (4).

\[
\text{Accuracy} = \frac{tp + tn}{N} \quad \text{ ...(3)}
\]

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

\[
\text{Acc}_i = \frac{1}{N} \left( \sum_{i=1}^{n} \frac{(|Y_i \cap Z_i|)}{|Y_i \cup Z_i|} \right) \quad \text{ ...(4)}
\]

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 generalized form is shown in the equation (6).

\[
P = tp + fp \\
P \text{ precision} = \frac{tp}{p} \quad \text{ ...(5)}
\]

\[
P \text{ rec} = \sum_{i \in D \land \forall j \neq i} \frac{|Y_i \cap Z_i|}{|Z_i|} \quad \text{ ...(6)}
\]

**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).

\[
p = tp + fn \\
\text{Recall} = \frac{tp}{p} \quad \text{ ...(7)}
\]

\[
Rec_i = \sum_{i \in D \land \forall j} \frac{|Y_i \cap Z_i|}{|Z_i|} \quad \text{ ...(8)}
\]

**RESULTS AND DISCUSSION**

This section depicts the results of both existing Network Based Method, Shortest Distance Method and proposed DT classification. The results of proposed method is compared with results of existing methods. From the analysis, the Decision Tree classifier 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 (Glycosylphosphatidylinositol)-anchor, (6) Peripheral membrane proteins.

The proposed DT classification Results are shown in the Table II. The Fig.5 illustrate the pie chart representation of decision tree classification on dataset S1. The multipass, lipid, GPI, peripheral, type1, type2 membrane proteins are represented by the colours, green, yellow, orange, brown, dark blue, light blue respectively. From the 2935 proteins from S1, more number of proteins are classified as multipass membrane proteins and less number of proteins as GPI anchored membrane proteins.

Each membrane protein can have labeled in one or more types, Fig.6 shows the number of proteins having 1-6 types of the dataset S1. In dataset S1, almost 510 membrane proteins are classified as Type1, 119 proteins as Type2, 1543 proteins as Multipass, 157 proteins as Lipid, 57 as GPI, and 549 proteins as Peripheral. Therefore in DT classification, the more number of proteins are classified as Multipass and the less number of proteins as GPI in all the dataset s1. The Accuracy, Precision, Recall, are calculated and the results are shown in Table II. The Multi label Protein classification using DT gives better results with all the annotated proteins, when compared to the existing methods with few number of unannotated proteins. The classification accuracy is reached 69.81% on dataset S1.

The proposed DT classification performs classification on the dataset S1. This method uses the whole number of proteins from the dataset for the classification purpose. Its classification accuracies
are presented in the Table. It is obvious that the DT method contributed the most, annotating 2935 proteins and achieved Accuracy of 69.81%, on datasets S1, but the network-based method number of annotated protein are 467 from the dataset S1, and obtained Accuracy of 66.68%, and shortest-distance method with Accuracy of 54.97%, on the dataset S1.

The Fig.7 shows the correct prediction of different multi-type membrane proteins in dataset S1. X axis represent the types of membrane proteins like ONE type, TWO types, THREE types. Y axis represented the total count of correct predicted membrane proteins in each type. 2482 one type membrane proteins and 34 two type membrane proteins. Very few of them are partially predicted and some of them are not correctly predicted.

The Fig 8 shows the performance of existing and proposed method accuracies. This bar graph shows the classification methods like Network method (NWM), Shortest distance method (SDM), and the proposed Decision tree (DT) classification in Y axis and the corresponding accuracies in the X axis.

CONCLUSION

In multi-label classification, each sample can be associated with a set of class labels. This paper proposed a DT classification algorithm. The 2935 membrane proteins of the datasets S1 are classified using Decision Tree based on the 967 features extracted from these proteins. As a result, the Decision Tree classifier with most contributive features achieved an acceptable accuracy of 69.81% of the dataset compared to the existing network based and shortest path method.

ACKNOWLEDGMENT

The authors would like to acknowledge to Guohua Huang, Institute of system biology, China, Yuchao Zhang, Department of Maths, Shoyang University, China, Lei Chan, CIE, SM University, China, Ning Zhang, Dept. of BME, Tianjin University, China, supporting for this work.

REFERENCES

9. H. Lodish, D. Baltimore, A. Berk, S. L.


