Membrane Protein Types Prediction by using a Two-Stage SVM Method

Anjali Devi G
PG Scholar
Department Of Computer Science and Engineering
Younus College of Engineering and Technology
Kollam, India-691010

Prof. Nijil Raj N
Head Of The Department
Department of Computer Science and Engineering
Younus College of Engineering and Technology
Kollam, India-691010

Abstract—Membrane proteins perform a variety of important functions in many biochemical processes that are vital to the survival of organisms and therefore they are attractive targets of drug discovery for many illness. The details of membrane protein types helps us to understand the structure and function of proteins [1]. Here is a method for predicting membrane protein types by combining amino acid properties and physicochemical properties used by a two stage support vector machine (SVM). Also for this a two step feature selection process is used. In addition to the thirty physicochemical properties, the 566 AAindex properties were also used in feature extraction. The hydrophobicity value of twenty numbers that is for each amino acids were also calculated for each sequences. In the two step feature extraction, firstly each feature extraction were considered separately and in the next step the more optimal feature sets were combined to obtain the final feature set. The method is evaluated based on six types of membrane proteins. The classifiers classifies a membrane protein into the following 6 classes, such as 1.) Type I membrane proteins, 2.) Type II membrane proteins, 3.) Multipass transmembrane proteins, 4.) Lipid chain anchored-membrane proteins, 5.) GPI-anchored-membrane proteins and 6.) Peripheral membrane proteins.

Index Terms—Physicochemical, SVM, AAindex, Hydrophobic-ity, .

I. INTRODUCTION

Cells are ”building blocks” of life: all living things, whether plants, animals, people, or tiny microscopic organisms, are made up of cells. Proteins consist of three main classes which are classified as globular, fibrous and membrane proteins. A cell is enveloped by a membrane which makes the boundary of a cell and enables it to maintain the distinction between cytosolic and extracellular environment. Cells consist of various organelles such as golgi body, endoplasmic reticulum, mitochondria and several other membrane bound organelles. The difference between cytosol and these organelles are maintained by individual membranes. These biological membranes are made up of mainly lipid bilayers whereas functions are carried out by membrane proteins [2]. Membrane associated proteins can be classified in the following two ways : Mode of interaction with the membranes & Cellular locations.

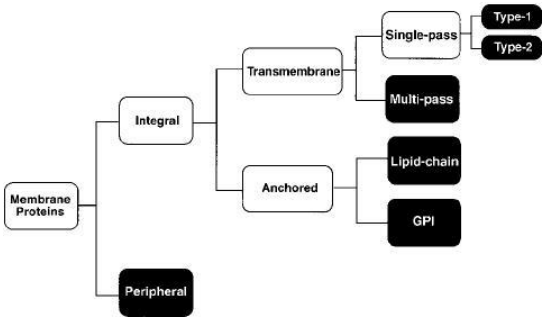

Fig. 1. Membrane Protein Types

A. Membrane Protein Types

The membrane proteins can be categorized as follows:

1.) Type I membrane proteins. 2.) Type II membrane proteins. 3.) Multipass transmembrane [3] proteins. 4.) Lipid chain anchored-membrane proteins. 5.) GPI-anchored-membrane proteins. 6.) Peripheral membrane proteins.

Integral or intrinsic membrane proteins : These proteins are associated with membranes and interact strongly with the hydrophobic part of the phospholipid bilayer. Presence of one or more apolar regions accounts for the span of lipid bilayer (-helix and -sheet as well). They interact mainly through van der Waals interaction with the hydrophobic core of the lipid bilayer. Thus they can be extracted from the membrane only through membrane disruption by detergents.

Transmembrane Protein : A transmembrane [3] protein (TP) is a type of integral membrane protein that spans the entirety of the biological membrane to which it is permanently attached. Many transmembrane proteins function as gateways to permit the transport of specific substances across the biological mem-brane. They frequently undergo significant conformational changes to move a substance through the membrane. Type-I and Type-II transmembrane proteins are called single-pass transmembrane proteins.

Type I membrane proteins : The N-terminus of this pro-tein is extracellular (luminal) and C-terminus remains in the cytoplasmic region for a cell (or organelle) membrane.

Type II membrane proteins : The C-terminus of this pro-

tein is extracellular (luminal) and N-terminus remains in the cytoplasmic region for a cell (or organelle) membrane.

Multipass transmembrane [3] proteins :These proteins are able to cross the lipid bilayer multiple times compared to Type I and Type II single pass membrane proteins which can cross the lipid bilayer only once. Membrane straddling region of polypeptide chains possess mostly -helical conformation as in the lipid environment hydrogen bonding between polypeptide chains would be maximum if it form helical conformation.

Anchored proteins are proteins that are anchored in membranes. Lipid chain anchored-membrane proteins : Lipid chain anchored-membrane proteins are related with lipid bilayer via one or greater than one covalently attached fatty acid chains or prenyl groups (other type of lipid chains). GPI-anchored-membrane proteins : GPI-anchored-membrane proteins are associated with lipid bilayer via glycosylphosphatidylinositol (GPI) anchor.

Peripheral or extrinsic membrane proteins : These proteins are known to interact either non covalently with the membrane surface through electrostatic or hydrogen bonds or with covalent bonds through lipids or GPI (glycosylphosphatidylinositol) anchors. They interact with the hydrophilic surfaces of the bilayer through electrostatic interaction. They can be isolated from the membrane using strong salt or change in pH.

B. Problem Definition

One of the earliest medical applications of bioinformatics has been in aiding rational drug design [4]. Membrane proteins are a common type of proteins along with soluble globular proteins, fibrous proteins, and disordered proteins. They are targets of all modern medicinal drugs. It is estimated that 20-30% of all genes in most genomes encode membrane proteins. Knowledge of a given type of cell membrane protein is crucial for determining its function. Based on their function, membrane proteins are classified into different types. For classifying membrane types different classification methods are used. Since classification is crucial, best classification methods are needed for classifying. Here, a two-stage SVM [5] is used for classification. Three benchmark datasets are used for classification. Datasets were downloaded from UniProt database (release 2012 09 - Oct 3, 2012).

II. RELATED WORKS

Many researches are going in the field of bioinformatics. One among them is prediction and classification of membrane proteins.Algorithms that have been used for protein sequence classification can be classified roughly into several types, depending on whether they are based on the K-Nearest Neighbor (K-NN) approach, the Hidden Markov Model (HMM) approach,Elastic Net(EN) and so on.

A. Geometrical Approach to Distinguish Between Transmembrane & Globular Proteins

Despite their great importance, transmembrane proteins (TMPs) are highly underrepresented in the protein structure database, due to difficulties in crystallizing them in an aqueous environment. The TMPs are usually larger than globular proteins, making their structure determination quite difficult by the NMR technique as well. This gives an explanation for their relatively low occurrence among the more than 20 000 structures deposited into the Protein Data Bank (PDB) so far. Currently, more than 300 membrane protein structure

files can be found in the PDB, representing around 3040 different folds. The size of this subset is approaching the level, where an automatic procedure is required to construct and maintain a database specific for TMPs.

There are several reasons that the separation between the two groups is not unequivocal. The surface of globular proteins is usually not entirely hydrophilic as apolar atoms of polar residues may be exposed and larger hydrophobic patches involved in ligand binding can also be found on the surface. Analogously, the membrane embedded parts of a TMP may contain polar and charged residues playing role in enzymatic activity or ion transport. While the surface of transmembrane [3] and globular proteins are adapted to their different environment, the inside of the two groups is commensurable in their hydrophobicity [6]. This can make distinguishing short fragments located inside an intact globular or transmembrane protein quite difficult. Similar problems can occur in the case of large multichain complexes. The interface of globular oligomers is often hydrophobic, while in the case of transmembrane chains it is more likely to be polar. Thus, when considering individual chains without the valid quaternary structure, the difference between the surface compositions can also diminish.

Another factor influencing the discrimination of transmembrane and globular proteins is related to the quality of the structure. The crystal structure of several membrane proteins is of low resolution, often reflected in distorted secondary struc-tures with incomplete hydrogen bond network or a structure with C atoms only. Determination of the structure by the NMR in a detergent solvent instead of the lipid bilayer can result in a highly flexible structure with the structural boundaries of the membrane regions melted. As a result of all these factors, the objective function aiming to distinguish between transmem-brane and globular proteins should not only account for the physical difference in their environments, but it should also incorporate practical limitations associated with the structure determination. These difficulties necessitate the development of a new automated algorithm to distinguish transmembrane and globular proteins by their atomic coordinates as well as to identify the transmembrane segments of the TMPs using only their atomic coordinates.

B. Predicting The Type or location of a Given Membrane Protein Based on Its Amino Acid Composition

Membrane proteins are classified according to two different schemes. In scheme 1, they are discriminated among the following five types: (1) type I single-pass transmembrane,

(2) type II single-pass transmembrane, (3) multipass transmembrane, (4) lipid chain-anchored membrane, and (5) GPI-

anchored membrane proteins. In scheme 2, they are discriminated among the following nine locations: (1) chloroplast, (2) endoplasmic reticulum, (3) Golgi apparatus, (4) lysosome, (5) mitochondria, (6) nucleus, (7) peroxisome, (8) plasma, and (9) vacuole. An algorithm is formulated for predicting the type or location of a given membrane protein based on its amino acid composition. The overall rates of correct prediction thus obtained by both self-consistency and jackknife tests, as well as by an independent dataset test, were around 76-81 the classification of five types, and 66-70 classification of nine cel-lular locations. Furthermore, classification and prediction were also conducted between inner and outer membrane proteins; the corresponding rates thus obtained were 88-91proteins, as well as their cellular locations and other attributes, are closely correlated with their amino acid composition. It is anticipated that the classification schemes and prediction algorithm can expedite the functionality determination of new proteins. The concept and method can be also useful in the prioritization of genes and proteins identified by genomics efforts as potential molecular targets for drug design.

C. Pseudo Amino Acid Composition (PSEAAC)

With the accomplishment of human genome sequencing, the number of sequence-known proteins has increased explosively. Incontrast, the pace is much slower in determining the irbiological attributes. As a consequence, the gap between sequence-known proteins and attribute-known proteins has become increasingly large.The unbalanced situation, which has critically limited the ability to timely utilize the newly discovered proteins for basic research and drug development, has called for developing computational methods or high-throughput automated tools for fast and reliably identifying various attributes of uncharacterized proteins based on their sequence information alone. Actually, during the last two decades or so, many methods in this regard have been established in hope to bridge such a gap. In the course of developing these methods, the following things were often needed to consider- (1)bench- mark dataset construction, (2)protein sam-ple formulation,(3)operating algorithm, (4)anticipated accu-racy, and (5)web-server establishment. Reviewing particularly in how to use the general formulation of PseAAC [7], [8], [9], [10], [11] to reflect the core and essential features that are deeply hidden in complicated protein sequences.

D. Prediction Of Protein Cellular Attributes Using Pseudo Amino Acid Composition

The cellular attributes of a protein, such as which compartment of a cell it belongs to and how it is associated with the lipid bilayer of an organelle, are closely correlated with its biological functions. The success of human genome project and the rapid increase in the number of protein sequences entering into data bank have stimulated a challenging frontier: How to develop a fast and accurate method to predict the cellular attributes of a protein based on its amino acid sequence? The existing algorithms for predicting these attributes were all based on the amino acid composition in which no sequence order effect was taken into account. To improve the prediction quality, it is necessary to incorporate such an effect. However, the number of possible patterns for protein sequences is extremely large, which has posed a formidable difficulty for realizing this goal. To deal with such a difficulty, the pseudo-amino acid composition is introduced. It is a combination of a set of discrete sequence correlation factors and the 20 components of the conventional amino acid composition. A remarkable improvement in prediction quality has been observed by using the pseudo amino acid composition. The success rates of prediction thus obtained are so far the highest for the same classification schemes and same data sets. It has not escaped from our notice that the concept of pseudo-amino acid composition as well as its mathematical framework and biochemical implication may also have a notable impact on improving the prediction quality of other protein features.

III. MATERIALS AND METHODS

A. Dataset

6,417 sequences of experimentally verified membrane proteins of homo sapiens were downloaded from the Uniprot database [16]. To evaluate the performance of classifiers three benchmark dataset S1,S2 and S3 are constructed from 6,415 membrane proteins. Dataset S1 contains 2880 membrane proteins , S2 contains 2075 membrane proteins , S3 contains 1460 membrane proteins. Accession numbers are used to represent membrane proteins in datasets. Sequence based features of membrane proteins are used for membrane protein classifica-tion.

B. Feature-based Sequence Representation

There are 20 unique amino acids that are used as a pro-teins building blocks. All amino acids have a common basic chemical structure, but different chemical properties due to differences in their side chains. A protein can be represented by a string of amino acids. Different proteins have different sequences, in terms of the ordering of their amino acids and length of the sequence. The first step in classifying proteins is to find a common way to represent the sequences. In this work, a feature vector is adopted to represent protein chains. Any protein, regardless of the length or composition of its sequence, can be mapped to a feature vector representation. 9 feature sets are used within the feature vector.

1) Local Amino Acid Composition (LAAC): Amino acid Composition is the normalized frequency of occurrence of each of the twenty amino acids in the given proteins amino acid sequence. Therefore, this feature set includes 20 features.

2) Local Dipeptide Composition (LDC): Dipeptide Composition [12] describes the proportion of each common amino acid pair within a sequence. It gives 400 features.

3) Hydrobhobicity: Each amino acid has an associated hydrophobic affinity, which is often measured using a hydrophobic index.The hydrophobicity index [6] is a measure of the relative hydrophobicity, or how soluble an amino acid is in water. 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. here the feature set includes 20 features.

4) AAIndex: It is a database of numerical indices representing various physicochemical and biochemical properties of amino acids and pairs of amino acids. AAindex [13] for the amino acid index of 20 numerical values. iT gives a total of 566 features.

5) Autocorrelation Descriptors (Moreau-Broto, Moran & Geary): Autocorrelation descriptors (AD)[14] are defined based on the distribution of amino acid properties along the sequence, and amino acid properties are amino acids indices taken from AAIndex database. For each type of index, three types of autocorrelation descriptors (Moreau-Broto, Moran and Geary) are defined. It gives a total of 240 features for each of the descriptors.

6) Sequence-order Descriptors (Sequence-order-Coupling & Quasi-Sequence-order Descriptors): The sequence-order descriptors (SD) [8] are proposed by K.C.Chou, et.al. They are derived from the distance matrix between the 20 amino acids. The feature set produced is 60 and 100 respectively.

After the phase of initial feature extraction, a final optimal feature set is obtained by the union of some of the initial feature set. Supposing that for the initial stage the number of feature extraction methods used is M, there are M optimal feature subsets constructed. In the second step, for each classification, we extract the final optimal feature subset on the union of M optimal feature subsets obtained in the first step.

IV. PROPOSED SYSTEM

In this section, we discuss our work in which a two-stage SVM is used for the membrane protein prediction. Support vector machines, SVM's [5] are supervised learning models that are used for classification. It is formally defined by a separating hyperplane. All optimal feature subsets are obtained by the two-step optimal feature selection procedure. When a two-stage SVM [15] is used for membrane protein classification, the outputs of the first stage of SVM are used as inputs for the next stage. In the second stage we use multi-class SVMs to predict membrane protein types. The two-step optimal feature selection method along with the two-stage support vector machine are considered to be effective than other previous classification methods.

A. Computational Framework

Figure.2 displays the architectural framework for the membrane protein types prediction. Firstly the protein sequences are collected from the Uniport database [16]. Three benchmark dataset S1,S2 and S3 are constructed from 6,415 membrane proteins. Dataset S1 contains 2880 membrane proteins ,S2 contains 2075 membrane proteins ,S3 contains 1460 mem-brane proteins.

After that the protein sequences are converted to feature vectors of particular dimensions corresponding to each type of feature extraction processes.

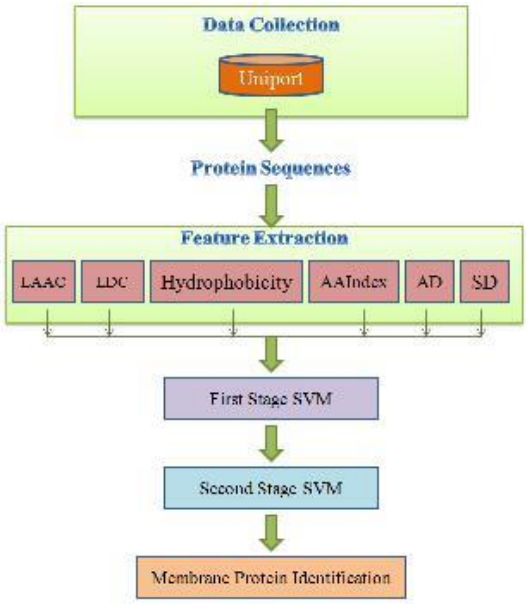

Fig. 2. The Framework for Membrane Protein Types Prediction

For a two-step optimal feature extraction process we are combining the feature sets obtained in the initial feature extrac-tion process. Here a combination of LAAC, Hydrophobicity and SD is taken with a feature set dimension of 200 (see table I).

TABLE I

FEATURES & DIMENSIONS

No.	Feature	Feature Dimension
Step 1		
1	LAAC	20
2	LDC	400
3	Hydrophobicity	20
4	AAIndex	566
5	AD	240 + 240 + 240
6	SD	60 + 100
Step 2		
7	LAAC + Hydrophobicity + AD + SD	200

Finally, a classification technique, a two-stage SVM based classifier, is used to classify the membrane protein types. In the second stage, we use conventional multi-class SVM to predict membrane protein types. We use LIBSVM to implement SVMs.

Here Naive Bayes classifier was also used to classify the membrane proteins in addition to two-stage SVM method. Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Three benchmark dataset is fed as input to naive bayes classifier for training. 5-fold cross validation is done on each dataset. After training naive bayes classifier is tested with same datasets itself. Table II shows the accuracy for both the classifiers.

TABLE II
COMPARISON OF ACCURACY OF TWO-STAGE SVM WITH NAIVE BAYES CLASSIFIER

Sl. No	Data set	Two-Stage SVM	Naive Bayes Classifier
Step 1			
1	S1	82.44%	69.51%
2	S2	84.19%	71.23%
3	S3	87.68%	74.49%
Step 2			
1	S1	91.14%	72.49%
2	S2	93.78%	79.51%
3	S3	95.48%	83.24%

B. Evaluation

The overall prediction accuracy A_{cc}, sensitivity S_{sn}, specificity S_{sp} and Matthew's correlation coefficient MCC are used to evaluate the prediction of performance of our work. Here, TP is the number of positive events that are correctly predicted; true negatives TN is the number of negative events that are correctly predicted; false positives FP is the number of negative events that are incorrectly predicted; false negatives FN is the number of subjects that are predicted to be negative despite they are positive;. In addition, MCC ranges from -1 to 1. A value of MCC = 1 indicates the best possible prediction; while MCC = -1 indicates the worst possible prediction. The equations are shown below:

$$A_{cc} = \frac{(TP + TN)}{(TP + FP + TN + FN)} \quad (1)$$

$$S_{sn} = \frac{TP}{(TP + FN)} \quad (2)$$

$$S_{sp} = \frac{TN}{(TN + FP)} \quad (3)$$

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TN + FN)(TP + FN)(TN + FP)}} \quad (4)$$

The two-stage SVM classifier obtained an accuracy of 86.11%, 88.95% 94.24% for each of the three datasets S1, S2, S3 respectively (Table III).

V. DISCUSSION

This section portay the results of both existing methods and proposed methods. Sequence based feature extraction is adopted for proposed system.Extracted features are local amino acid composition, local dipeptide composition [12], hydrophobicity [6], Amino Acid Index(AAIndex), Autocorrelation Descriptors and Sequence-order Descriptors. weight

TABLE III
DATA SET AND ACCURACY

Sl. No	Data set	Method	Accuracy	MCC
Step 1				
1	S1	5-fold cross validation	89.44%	76.12%
2	S2	5-fold cross validation	91.19%	79.78%
3	S3	5-fold cross validation	93.69%	85.72%
Step 2				
1	S1	5-fold cross validation	90.11%	79.83%
2	S2	5-fold cross validation	92.95%	81.53%
3	S3	5-fold cross validation	94.24%	89.46%

and atomic sum of binding domain.Proposed system contains 9 feature sets within the feature vectors. Input to the classifiers is three dataset S1,S2 and S3. Performance metrics for the classifier will be compared with existing methods. method. Proposed method classifies homo sapiens membrane proteins into the following six classes,(1) Single -pass type I , (2)Single-pass type II, (3) Multi-pass, (4) Lipid-anchor, (5) GPI-anchor, (6) Peripheral membrane proteins. After the 5-fold cross validation only the data sets are fed to the classifier. The ROC curve for the work is shown in figure 3. The comparison of the proposed system with existing system is shown in table IV.

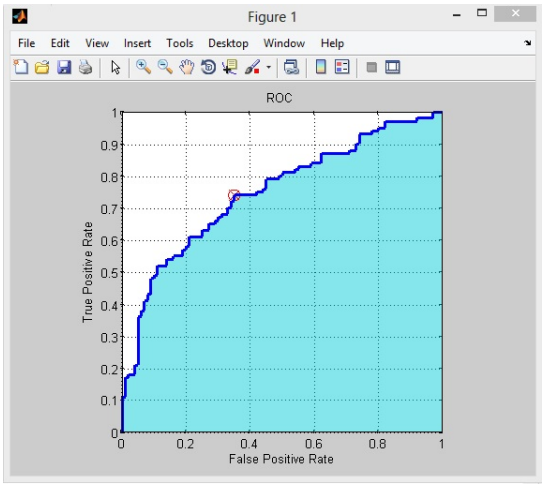

Fig. 3. The ROC curve for Membrane Protein Types Prediction Framework

TABLE IV
COMPARISON OF OVERALL PREDICTION ACCURACY OF TWO-STAGE SVM WITH EXISTING SYSTEM

Method	Accuracy
Two-Stage SVM	89.11%
Existing System	86.62%

VI. CONCLUSION

Membrane proteins perform a variety of important functions in many biochemical processes that are vital to the survival of organisms. They are attractive targets of drug discovery for many illness. Membrane proteins are a common type of

proteins along with soluble globular proteins, fibrous proteins, and disordered proteins. Based on their function, membrane proteins are classified into different types. Here introduced a two-stage SVM method for the classification of membrane proteins. Training is done on three benchmark datasets S1,S2 and S3. In feature extraction amino acid classifications, physicochemical and biochemical properties of amino acids are incorporated. We are hoping better prediction accuracy for the method when compared to all other existing method.

In the future, we will try to develop a novel method for the prediction of non-membrane proteins along with membrane proteins with addition of more feature set.

REFERENCES

[1] 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.

[2] A. Bruce, B. Dennis, L. Julian, R. Martin, R. Keith, and D. James, "Molecular biology of the cell. garland publishing," *New York*, vol. 19832, pp. 255–317, 1994.

[3] G. E. Tusnády, Z. Dosztányi, and I. Simon, "Transmembrane proteins in the protein data bank: identification and classification," *Bioinformatics*, vol. 20, no. 17, pp. 2964–2972, 2004.

[4] P. R. Sanders, L. M. Kats, D. R. Drew, R. A. O'Donnell, M. O'Neill, A. G. Maier, R. L. Coppel, and B. S. Crabb, "A set of glycosylphosphatidyl inositol-anchored membrane proteins of plasmodium falciparum is refractory to genetic deletion," *Infection and immunity*, vol. 74, no. 7, pp. 4330–4338, 2006.

[5] 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.

[6] G. D. Rose, A. R. Geselowitz, G. J. Lesser, R. H. Lee, and M. H. Zehfus, "Hydrophobicity of amino acid residues in globular proteins," *Science*, vol. 229, pp. 834–839, 1985.

[7] Y.-D. Cai, P.-W. Ricardo, C.-H. Jen, and K.-C. Chou, "Application of svm to predict membrane protein types," *Journal of Theoretical Biology*, vol. 226, no. 4, pp. 373–376, 2004.

[8] K.-C. Chou, "Prediction of protein subcellular locations by incorporating quasi-sequence-order effect," *Biochemical and biophysical research communications*, vol. 278, no. 2, pp. 477–483, 2000.

[9] ——, "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," *Bioinformatics*, vol. 21, no. 1, pp. 10–19, 2005.

[10] 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.

[11] K.-C. Chou, H.-B. Shen *et al.*, "Review: recent advances in developing web-servers for predicting protein attributes," *Natural Science*, vol. 1, no. 02, p. 63, 2009.

[12] I. Dubchak, I. Muchnik, S. R. Holbrook, and S.-H. Kim, "Prediction of protein folding class using global description of amino acid sequence," *Proceedings of the National Academy of Sciences*, vol. 92, no. 19, pp. 8700–8704, 1995.

[13] S. Kawashima and M. Kanehisa, "Aaindex: amino acid index database," *Nucleic acids research*, vol. 28, no. 1, pp. 374–374, 2000.

[14] Z.-R. Li, H. H. Lin, L. Han, L. Jiang, X. Chen, and Y. Z. Chen, "Profeat: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence," *Nucleic Acids Research*, vol. 34, no. suppl 2, pp. W32–W37, 2006.

[15] G.-S. Han, Z.-G. Yu, and V. Anh, "A two-stage svm method to predict membrane protein types by incorporating amino acid classifications and physicochemical properties into a general form of chou's pseaac," *Journal of Theoretical Biology*, vol. 344, pp. 31–39, 2014.

[16] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane *et al.*, "Uniprot: the universal protein knowledgebase," *Nucleic acids research*, vol. 32, no. suppl 1, pp. D115–D119, 2004.