Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 14, no. 7, July 2018, pp. 1628-1634
DOI: 10.23940/ijpe.18.07.p27.16281634

A Novel Multi-Label Predictor for Identifying Multi-Functional Classes of Human Membrane Proteins

Xiao Wang*, Guoqing Li, Weiwei Zhang, Hongwei Tao, and Yinghui Meng

School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou, 450002, China

Abstract

Knowing which types of functionality that human membrane proteins belong to is very helpful for understanding their functions. However, most existing online prediction methods have some disadvantages, including: 1) they obtain very low prediction accuracy, and 2) they can only predict single-functional classes of cytomembrane proteins in humans. To overcome the drawbacks, a new multi-label predictor, namely mMem-Hum, is proposed. In addition to predicting types of single-function membrane proteins, it can also predict multi-functional types. Specifically, discriminative features of membrane proteins are generated by using amino acid sequence information and evolutionary information, and then they are classified by a new multi-label classifier that utilizes label correlations. Experimental results reveal that the performance of mMem-Hum is significantly better than other existing forecasting methods. This indicates that mMem-Hum may become a promising prediction tool for classifying functional classes of cytomembrane proteins in humans.

Keywords: cytomembrane proteins; type prediction; multi-label classification; label correlation

(Submitted on April 4, 2018; Revised on May 17, 2018; Accepted on June 23, 2018)

1. Introduction

The proteins on the cell cytomembrane or organelle cytomembrane are called cytomembrane proteins, which are involved in various biological processes [1]. The accurate and rapid identification of functional classes of cytomembrane proteins serves an important role in disease treatment and drug design. Based on the interaction between cytomembrane proteins and the lipid layer, cytomembrane proteins can be divided into eight types [3]. Detailed information about these eight types is included in [9].

Understanding the functional classes of cytomembrane proteins provides an important clue for determining the function of a membrane protein sequence [6, 7]. Therefore, it is indispensable to develop computational tools for efficiently and accurately identifying the functional classes of cytomembrane proteins. In the last decade, researchers at home and abroad have made many efforts in classifying functional classes of cytomembrane proteins [4, 5]. While a lot of membrane protein predictors have been developed, they still have some inadequacies. These existing predictors can only deal with cytomembrane proteins with single-label functional classes. However, many cytomembrane proteins simultaneously belong to multiple functional classes. At present, in the existing research, only a small number of predictors are able to predict multi-label cytomembrane proteins [6, 9]. Moreover, these predictors perform poorly and can still be improved.

In order to deal with the aforementioned problems, this article puts forward a novel multi-label predictor called mMem-Hum. The predictor can identify single label and multi-label human membrane proteins with functional types. mMem-Hum extracts features for membrane proteins by using PseAAC and PSSM-AC approaches and makes predictions by harnessing a new multi-

* Corresponding author.
E-mail address: pandaxiaoxi@163.com

label classifier that utilizes label correlations. By performing a leave-one-out cross-validation (LOOCV) test on the benchmark dataset, experimental results show that mMem-Hum is superior compared to existing up-to-date predictors in terms of various multi-label performance metrics. This indicates that mMem-Hum is an efficient and useful tool for predicting functional classes of cytomembrane proteins in humans.

2. Materials and Methods

2.1. Benchmark Dataset

The prediction performance of the mMem-Hum predictor will be tested on a common benchmark dataset [10]. Table 1 displays the breakdown of the benchmark dataset. The dataset is constructed to be specialized for human cytomembrane proteins, where cytomembrane proteins were collected from UniProtKB/Swiss-Prot (released in April 2014). The sequence similarity of the dataset was cut off at 60%. The dataset contains 3166 proteins distributed in 8 different functional types, of which 3069 belong to one functional type, there are 93 of 2 types, and no membrane proteins belong to 4 or more than 4 functional types.

2.2. Feature Extraction

For machine learning based membrane protein type predictions, it is an indispensable but difficult step to extract discriminative features for training an effective and efficiency classifier. In this work, we used sequence composition and evolutionary information such as input features to discriminate functional types of membrane proteins. Specifically, we used the PseAAC method to extract composition information of membrane protein sequences and the transformation PSSM-AC method to obtain sequence evolutionary information. Thus, this section includes the following two subsections: (1) PseAAC; (2) PSSM-AC.

2.2.1. PseAAC

The conventional AAC is a classical simple method for structuring statistical features of proteins; however, it cannot obtain sequence order information in proteins. Chou [2] proposed the improved PseAAC by adding the sequence order effects in proteins. PseAAC transfers a protein sequence into a $(20 + \xi \cdot \lambda)$ dimensional vector, among the first 20 amino acid residues represent the original amino acid composition, and the latter $\xi \cdot \lambda$ components approximately express the sequence order information of a protein sequence. The latter $\xi \cdot \lambda$ components, also named sequence-order correlated factors, can be formulated by the following equations:

$$\begin{cases} \eta_1 = \frac{1}{Lin-1}\sum_{i=1}^{Lin-1}\Psi_1(W_i, W_{i+1}) \\ \eta_2 = \frac{1}{Lin-1}\sum_{i=1}^{Lin-1}\Psi_2(W_i, W_{i+1}) \\ \vdots \\ \eta_\xi = \frac{1}{Lin-1}\sum_{i=1}^{Lin-1}\Psi_\xi(W_i, W_{i+1}) \\ \eta_{\xi+1} = \frac{1}{Lin-2}\sum_{i=1}^{Lin-2}\Psi_1(W_i, W_{i+2}) \\ \eta_{\xi+2} = \frac{1}{Lin-2}\sum_{i=1}^{Lin-2}\Psi_2(W_i, W_{i+2}) \\ \vdots \\ \eta_{2\xi} = \frac{1}{Lin-2}\sum_{i=1}^{Lin-2}\Psi_\xi(W_i, W_{i+2}) \\ \cdots \\ \eta_{\xi(\lambda-1)+1} = \frac{1}{Lin-\lambda}\sum_{i=1}^{Lin-\lambda}\Psi_1(W_i, W_{i+\lambda}) \\ \eta_{\xi(\lambda-1)+2} = \frac{1}{Lin-\lambda}\sum_{i=1}^{Lin-\lambda}\Psi_2(W_i, W_{i+\lambda}) \\ \vdots \\ \eta_{\xi\lambda} = \frac{1}{Lin-\lambda}\sum_{i=1}^{Lin-\lambda}\Psi_\xi(W_i, W_{i+\lambda}) \end{cases} \quad (\lambda < Lin), \tag{1}$$

where $\eta_1, \eta_2, \ldots, \eta_\xi$ are the first-tier sequence-order correlated factor, $\eta_{\xi+1}, \eta_{\xi+2}, \ldots, \eta_{2\xi}$ the 2nd-tier sequence-order correlated factor**,** ξ the number of physicochemical properties selected, λ the maximum number of correlation layers between membrane protein sequences, Lin the length of the amino acid sequence of a protein, W_i the amino acid residue at position i of a protein chain, $\Psi_t(t=1,\ldots,\xi)$ the correlation function based on the t-th physicochemical property selected, as defined below by:

$$\Psi_t(W_i, W_j) = S_t(W_i) \cdot S_t(W_j), \tag{2}$$

where $S_t(W_i)$ is the t-th physicochemical property value of the amino acid W_i, $S_t(W_j)$ the value for the amino acid W_j, and the dot (·) represents the multiplication sign.

Table 1. The detail of 8 different functional classes of cytomembrane protein

Order	Functional Class	Number of Proteins
1	single-pass class I	605
2	single-pass class II	195
3	single-pass class III	25
4	single-pass class IV	27
5	multi-pass	1444
6	lipid-anchor	251
7	GPI-anchor	83
8	peripheral	637
The number of proteins		3166

In this work, the following six physicochemical properties were selected: (1) hydrophobicity; (2) hydrophilicity; (3) pK1 (C^{α}-COOH); (4) pK2 (NH3); (5) PI (25℃); (6) side-chain mass. Thus, the value of ξ is equal to 6. Six physical and chemical properties of each original amino acid can be obtained. After a standard conversion procedure was used in the physicochemical values according to Eq. (4) of [2], they were submitted to Eq. (2).

In view of this, P represents a protein sequence, which can be formulated as a $(20 + \xi \cdot \lambda)$ dimensional vector by the following equation:

$$\mathrm{P}_{\mathrm{pseaac}} = [e_1, e_2, \cdots, e_{20}, e_{20+1}, \cdots, e_{20+\xi\cdot\lambda}]^T, \tag{3}$$

where

$$e_j = \begin{cases} \dfrac{f_j}{\sum_{i=1}^{20} f_i + w\sum_{k=1}^{\xi\cdot\lambda}\eta_k} & , \quad (1 \le j \le 20) \\ \dfrac{w\eta_{j-20}}{\sum_{i=1}^{20} f_i + w\sum_{k=1}^{\xi\cdot\lambda}\eta_k} & , \quad (20+1 \le j \le 20+\xi\cdot\lambda; \lambda < Lin) \end{cases}, \tag{4}$$

where f_i is the normalized occurrence frequency of the 20 native amino acids in the protein P, η_k is the sequence-order correlated factors computed by Eq. (1), and w is the weight factor.

2.2.2. PSSM-AC

The position-specific scoring matrix is usually used as a feature extraction model for obtaining protein sequence information. PSSM-related feature extraction methods have been used in many subfields of bioinformatics. For a protein sequence P that has L amino acid residues, its PSSM can be formulated by a $L \times 20$ matrix, as follows:

$$\mathrm{P}_{\mathrm{pssm}}^{(0)} = \begin{bmatrix} E_{1\to1}^{(0)} & E_{1\to2}^{(0)} & \cdots & E_{1\to20}^{(0)} \\ E_{2\to1}^{(0)} & E_{2\to2}^{(0)} & \cdots & E_{2\to20}^{(0)} \\ \vdots & \vdots & \vdots & \vdots \\ E_{L\to1}^{(0)} & E_{L\to2}^{(0)} & \cdots & E_{L\to20}^{(0)} \end{bmatrix}, \tag{5}$$

where $E_{i\to j}^{(0)}$ indicates the initial probability value of the i-th amino acid residue in the protein is mutated to the amino acid type j during the process of change. The numbers $1,2,\cdots,20$ represent the 20 primitive amino acid types. The $L \times 20$ matrix was structured by running the PSI-BLAST program against the Swiss-Prot database with three iterations and an E-value cutoff value of 0.001 for multiple sequence alignment against the protein sequence P. To obtain more accurate results, a standardization procedure was applied to the original PSSM. As a result, Eq. (5) will be converted to a new matrix as follows:

$$\mathrm{P}_{\mathrm{pssm}}^{(1)} = \begin{bmatrix} E_{1\to 1}^{(1)} & E_{1\to 2}^{(1)} & \cdots & E_{1\to 20}^{(1)} \\ E_{2\to 1}^{(1)} & E_{2\to 2}^{(1)} & \cdots & E_{2\to 20}^{(1)} \\ \vdots & \vdots & \vdots & \vdots \\ E_{L\to 1}^{(1)} & E_{L\to 2}^{(1)} & \cdots & E_{L\to 20}^{(1)} \end{bmatrix}, \tag{6}$$

As can be seen from Eq. (6), the number of rows corresponding to different sequence length proteins is also different. Therefore, in order to obtain the same size vectors in the PSSM matrix with different rows, a protein sequence P can be represented by a 20-D feature vector by the following equation:

$$\mathrm{P}_{\overline{\mathrm{pssm}}} = [\bar{E}_1, \bar{E}_2, \cdots, \bar{E}_{20}]^T, \tag{7}$$

where

$$\overline{E}_j = \frac{1}{L}\sum_{i=1}^{L} E_{i\to j}^{(1)} \ (j = 1,2,\cdots,20), \tag{8}$$

where $\overline{\mathrm{E}}_{\mathrm{j}}$ on behalf of the average probability value of the amino acids in the protein P being mutated to amino acid type j during the process of change. $\mathrm{P}_{\overline{\mathrm{pssm}}}$ lost the sequence-order information in PSSM during the process of change. Therefore, to avoid this problem, each column of PSSM was transferred to contain sequence-order information by the auto-covariance transformation, denoted by Φ_j^{λ}:

$$\Phi_j^{\lambda} = \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda}\left(E_{i\to j}^{(1)} - \bar{E}_j\right) \times \left(E_{(i+\lambda)\to j}^{(1)} - \bar{E}_j\right) (j = 1,2,\cdots,20, \lambda < L), \tag{9}$$

where $\Phi_{\mathrm{j}}^{\lambda}$ represents the average correlated factor among the λ-th most contiguous PSSM probability values with respect to the amino acid type j.

Then, a protein sequence P can be formulated as a $(20 + 20 \cdot \lambda)$ dimensional vector by the following equation:

$$\mathrm{P}_{\mathrm{pssm-ac}} = \left[\bar{E}_1, \bar{E}_2, \cdots, \bar{E}_{20}, \Phi_1^1, \cdots, \Phi_{20}^1, \cdots, \Phi_1^{\lambda}, \cdots, \Phi_{20}^{\lambda}\right]^T, \tag{10}$$

where $\overline{\mathrm{E}}_{\mathrm{j}}$ reflects the evolutionary information of the protein sequence P and Φ_j^{λ} approximately represents its sequence-order effects.

In this work, a large number of preliminary experiments have been done. The parameter n = 2 is found to be the best choice. Therefore, in this paper, we will use this parameter.

2.3. Multi-Label Classifier

In this work, to achieve a much better prediction accuracy for recognizing multi-functional classes of human cytomembrane proteins, a novel multi-label classifier that considers correlations among these functional types is proposed.

Suppose a training dataset T contains M human cytomembrane proteins which are distributed in the N subsets. Specifically, $\mathrm{T} = \mathrm{T}_1 \cup \mathrm{T}_2 \cup \ldots \cup \mathrm{T}_{\mathrm{i}} \ldots \cup \mathrm{T}_{\mathrm{N}}$, where T_{i} (i=1, 2 …, N) is a subset containing M_{i} human cytomembrane proteins. It is worth noting that a human cytomembrane protein may attribute to two or more subsets and $\mathrm{M} \le \mathrm{M}_1 + \mathrm{M}_2 + \ldots + \mathrm{M}_{\mathrm{i}} + \ldots + \mathrm{M}_{\mathrm{N}}$. The proposed

multi-label prediction algorithm has two levels structure for prediction, where each level consists of N SVM classifiers each corresponding to one functional type. L_1 represents the first level, and the second level is denoted as L_2, i.e.

$$\begin{cases} L_1 = \{SVM_1^1, SVM_1^2, \ldots, SVM_1^N\} \\ L_2 = \{SVM_2^1, SVM_2^2, \ldots, SVM_2^N\} \end{cases} \tag{11}$$

where SVM_1^1 and SVM_2^1 are the prediction classifier for the first functional type, SVM_1^2 and SVM_2^2 are for the second functional type, and so on. The first level classifier plays an assistant role in the whole algorithm, and then the second level classifiers make the last prediction by using middle prediction probabilities obtained by the first level classifiers.

The first level can be constructed by the binary relevance method using the above-mentioned features on the opposite training set. Thus, for the second level, we first obtain prediction probabilities of each functional type for all human membrane proteins by using the first level classifiers. Then, we append prediction probabilities of all functional types on top of the additional features that are currently predicted for the original feature space. Finally, classifiers in the second level are constructed by the binary relevance method using the above-augmented features on the corresponding training set.

Setting *p* as a query protein, query its function types according to the following steps:

1) Using the PseACC+PSSM-AC model to extract the feature vector from the query protein *p*.
2) Feeding the feature vector of the protein *p* into the classifiers in the first level, and then obtaining *N* prediction probabilities for *p*.
3) Augmenting *p*'s feature space using prediction probabilities for *p* in step 2.
4) Putting the augmented feature from step 3 into classifiers in the second level, and then obtaining final prediction outputs for *p*.

In order to express the prediction process more clearly, Figure 1 provides a detailed illustration of how to combine the two levels of classifiers to produce the final result.

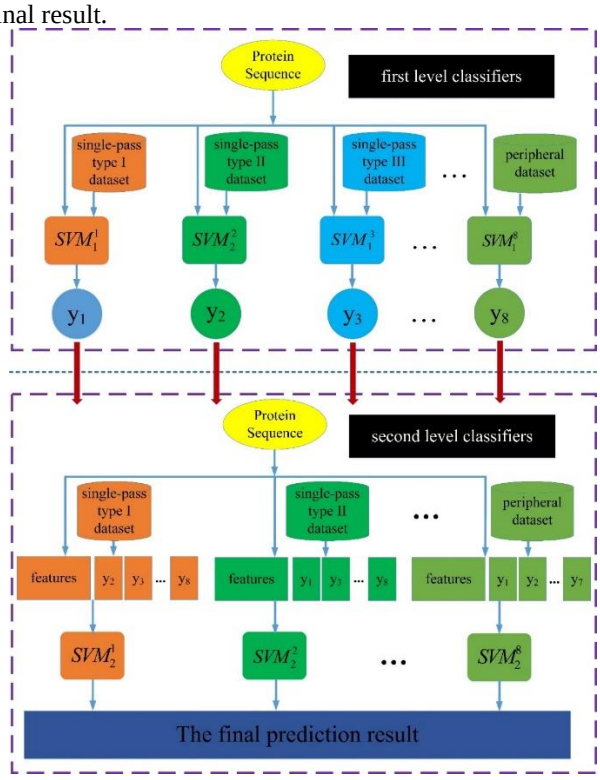

Figure 1. A framework describes how the algorithm works. The binary relevance method is used to train classifiers for the eight different functional classes of human cytomembrane proteins datasets. The prediction probabilities of the first level classifiers, i.e., y1 to y8, play an auxiliary role in the algorithm framework. The second level classifiers utilize y1 to y8 as the additional features to retrain the eight classifiers and improve the final prediction result.

3. Results and Discussion

The jackknife test was used in this work to test the prediction performance of mMem-Hum. Predictions of multi-functional classes of human cytomembrane proteins belong to the multi-label classification system. It is more difficult to test the prediction performance of a multi-label classifier, and the multi-label evaluation metrics are very different from the single-label ones. Finally, we use the following evaluation indicators to test the predictive performance of our predictor, defined as:

$$\begin{cases} \text{MLhloss} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\|L_i \cup L_i^*\| - \|L_i \cap L_i^*\|}{M}\right) \\ \text{MLacc} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\|L_i \cap L_i^*\|}{\|L_i \cup L_i^*\|}\right) \\ \text{MLpre} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\|L_i \cap L_i^*\|}{\|L_i^*\|}\right) \\ \text{MLrec} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\|L_i \cap L_i^*\|}{\|L_i\|}\right) \\ \text{ABStrue} = \frac{1}{N}\sum_{i=1}^{N} 1(L_i \equiv L_i^*) \end{cases}, \tag{12}$$

where L_i represents the subset of true functional classes of the i-th human cytomembrane protein and L_i^* the subset of predicted functional classes of the i-th one. $\|.\|$ represents the number of its elements. If true functional types are entirely identical to the predicted types, $1(L_i \equiv L_i^*)$ equals 1, 0 otherwise. Among the five evaluation metrics, the score of MLhloss is opposite to those of the other four. The lower the MLhloss is, the better the prediction performance will be. However, the higher the scores of the other metrics are, the better the prediction performance will be.

In fact, there are only three predictors, namely MemHum-PseAAC [10], MemHum-AAC-IE [10] and MemHum-D-PSSM [10], which can predict multi-label functional classes of human cytomembrane proteins. According to extracting feature vectors, MemHum-PseAAC uses pseudo-amino acid compositions to construct feature vectors, MemHum-AAC-IE utilizes amino acid compositions and information entropy, and MemHum-D-PSSM uses distribution and position-specific score matrices. According to training multi-label classifiers, the aforementioned three predictors train their prediction models by using the mlknn multi-label learning algorithm. Particularly, MemHum-D-PSSM achieves the best prediction results because position-specific score matrices are applied to extract discriminative features. To demonstrate the prediction performance of our proposed predictor mMem-Hum, the mMem-Hum predictor will be compared with the aforementioned three predictors in the work.

Table 2 displays the prediction scores obtained by mMem-Hum and the other three predictors on the dataset through the jackknife cross-validation test. From Table 2, mMem-Hum significantly outperforms MemHum-PseAAC, MemHum-AAC-IE and MemHum-D-PSSM according to all the five evaluation metrics. More specifically, compared to the other three methods, the proposed method mMem-Hum reveals better results. The value of the ABStrue is 80.36%, while the MLhloss is 5.32%. The ABStrue rate is quite high, and the MLhloss rate is very low. This indicates that the method is helpful for classifying multi-functional classes of human membrane proteins, and it will be a useful and efficient tool in the process of classifying multi-functional classes of human cytomembrane proteins.

Table 2. The prediction scores by mMem-Hum and other up-to-date predictors using the jackknife test in the aforementioned dataset

Predictors	Evaluation metrics				
	MLhloss	MLacc	MLpre	MLrec	ABStrue
MemHum-PseAAC	0.08	0.64	0.64	0.64	0.64
MemHum-AAC-IE	0.09	0.61	0.61	0.61	0.61
MemHum-D-PSSM	0.06	0.75	0.75	0.75	0.74
mMem-Hum	0.05	0.79	0.81	0.82	0.80

How can the proposed predictor significantly improve the prediction performance? First, the proposed multi-label prediction algorithm is a two-level prediction algorithm. The first level is responsible for the auxiliary prediction and the second level obtains the final prediction results of each functional type for all human membrane proteins. Second, it extracts new features which are

the prediction probabilities of all functional types except the current one to be predicted as additional features; that is, it takes into account the relationships among these functional types in the second level, which is the key to improving the final prediction performance. Third, it uses PseACC+PSSM-AC methods to construct the discriminative features from the protein sequences, which enhances the ability of feature representation.

4. Conclusions

A novel multi-functional predictor called mMem-Hum is proposed to classify single-functional and multi-functional classes of human cytomembrane proteins. For the uncharacterized proteins, the discriminative features are first generated by fusing protein sequence information and evolution information, and then they are fed into our proposed novel multi-functional classifier with class label correlations for predicting potential functional type(s) of the uncharacterized protein. The Jackknife test reveals that the prediction performance of our proposed mMem-Hum predictor is much better than other up-to-date multi-functional predictors in terms of various multi-label metrics. It is estimated that mMem-Hum may be a useful tool for the prediction of functional classes of cytomembrane proteins in humans.

Acknowledgments

This work was partially supported by the Natural Science Foundation of China (No. 61402422, 61501405), Doctoral Research Fund of Zhengzhou University of Light Industry (2013BSJJ082), and Postgraduate Technology Innovation Project in Zhengzhou University of Light Industry (No. 2017032).

References

1. M. S. Almén, K. J. V. 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, article 50, 2009
2. K.-C. Chou, "Prediction of Protein Cellular Attributes Using Pseudo-Amino Acid Composition," *Proteins: Structure, Function, and Bioinformatics*, vol. 43, no. 3, pp. 246-255, 2001
3. K. C. Chou and H. B. Shen, "MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM," *Biochemical and Biophysical Research Communications*, vol. 360, no. 2, pp. 339-345, 2007
4. 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
5. C. Ding, L. F. Yuan, S. H. Guo, H. Lin, and W. Chen, "Identification of mycobacterial membrane proteins and their types using over-represented tripeptide compositions," *Journal of Proteomics*, vol. 77, no. 24, pp. 321-328, 2012
6. C. Huang and J.-Q. Yuan, "A Multilabel Model Based on Chou's Pseudo–Amino Acid Composition for Identifying Membrane Proteins with Both Single and Multiple Functional Types," *Journal of Membrane Biology*, vol. 246, no. 4, pp. 327-334, 2013
7. L. Nanni and A. Lumini, "An ensemble of support vector machines for predicting the membrane protein type directly from the amino acid sequence," *Amino Acids*, vol. 35, no. 3, pp. 573-580, 2008.
8. C.-T. Su, C.-Y. Chen, and Y.-Y. Ou, "Protein Disorder Prediction by Condensed PSSM Considering Propensity for Order or Disorder," *BMC Bioinformatics*, vol. 7, article 319, 2006.
9. S. Wan, M.-W. Mak, and S.-Y. Kung, "Mem-mEN: predicting multi-functional types of membrane proteins by interpretable elastic nets," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 4, pp. 706-718, 2016.
10. H. L. Zou and X. Xiao, "A New Multi-label Classifier in Identifying the Functional Types of Human Membrane Proteins," *Journal of Membrane Biology*, vol. 248, no. 2, pp. 179-186, 2015.