Identification of Membrane Protein Types Based Using Hypergraph Neural Network

Zhiqiang Hui, Meiling Qian

Suzhou University of Science and Technology

Abstract. The problem in membrane protein classification and prediction is an important topic of membrane proteomics research because the function of proteins can be quickly determined if membrane protein types can be discriminated. most current methods to classify membrane proteins are labor-intensive and require a lot of resources. In this study, the hypergraph neural network model (HGNN) was used to predict membrane protein types.

1 Methods

To address the above issues, we have proposed an innovative hypergraph neural network model (HGNN). This model constructs a multi feature fusion hypergraph correlation matrix by combining various feature extraction methods, including Average Block Method (AvBlock), Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT), Directional Gradient Histogram (HOG), and Pseudo Position Specific Matrix (PsePSSM). Finally, by inputting these feature matrices and hypergraph correlation matrices into the HGNN model, the classification and prediction of membrane protein types were achieved. The proposed method in this paper is shown in Figure 1.

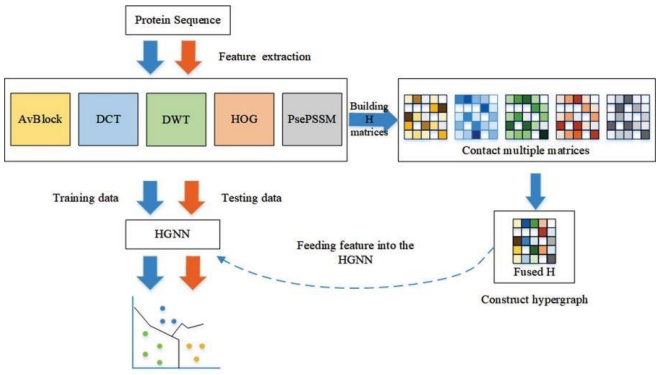

Figure 1. Schematic diagram of our proposed method.

2 Datasets

In this study, we used four datasets to test the performance of the proposed hypergraph neural network model (Table 1). Dataset 1: From Chou and Shen's research, it contains 3249 training sequences and 4333 testing sequences, totaling 8 types of membrane proteins. Dataset 2: Obtained based on Dataset 1 after removing redundant data, containing 2288 training sequences and 2306 testing sequences .Dataset 3: Expanded the dataset size to include 3073 training sequences and 3604 testing sequences .Dataset 4: Research from Chou and Elrod, containing 2059 training sequences and 2625 testing sequences.

Table 1. Statistics of different types of membrane proteins on 4 datasets.

Specific Types	Dataset 1		Dataset 2		Dataset 3		Dataset 4	
	Train	Test	Train	Test	Train	Test	Train	Test
Single-span type 1	610	444	388	223	561	245	435	478
Single-span type 2	312	78	218	39	316	7	152	180
Single-span type 3	24	6	19	6	32	9	-	-
Single-span type 4	44	12	35	10	65	17	-	-
Multi-span type 5	1,316	3,265	936	1,673	1,119	2,478	1,311	1,867
Lipid-anchor type 6	151	38	98	26	142	36	51	14
GPI-anchor type 7	182	46	122	24	164	41	110	86
Peripheral type 8	610	444	472	305	674	699	-	-
Overall	3,249	4,333	2,288	2,306	3,073	3,604	2,059	2,625

3 Results

The model proposed in this paper (HGNN) and the Memtype-2L model were compared on datasets 1, 2, and 3, respectively. The results of the test set comparison are shown in Table 7. It was found that the overall accuracy of HGNN was better than the other methods on the three datasets (92.8%, 88.6%, and 88.2%). Compared with MemType-2L (91.6%, 85.3%, 78.3%), the overall accuracy of HGNN was improved by 1.2%, 3.3%, and 9.9% (Table 2).

Table 2. Prediction accuracy of different classifiers on the dataset.

Specific Types	LR(%)	RF(%)	DNNE(%)	Our method(%)
Single-span type 1	67.6(300/444)	85.6(380/444)	92.6(411/444)	92.6(411/444)
Single-span type 2	62.8(49/78)	61.5(48/78)	76.9(60/78)	79.5(62/78)

Single-span type 3	0(0/6)	0(0/6)	0(0/6)	16.7(1/6)
Single-span type 4	66.7(8/12)	41.7(5/12)	41.7(5/12)	75.0(9/12)
Multi-span type 5	97.0(3166/3265)	92.1(3006/3265)	92.6(3024/3265)	94.8(3094/3265)
Lipid-anchor type 6	39.5(15/38)	31.6(12/38)	34.2(13/38)	44.7(17/38)
GPI-anchor type 7	8.3(36/46)	43.5(20/46)	67.4(31/46)	82.6(38/46)
Peripheral type 8	52.9(235/444)	75.2(334/444)	80.9(359/444)	87.4(388/444)
Overall	87.9(3809/4333)	87.8(3805/4333)	90.1(3903/4333)	92.8(4020/4333)

Finally, on dataset 4, the proposed method model in this paper was used to compare with other already existing method models. The comparison methods include the following: CDA, CDA and PseAA , Fourier-spectrum, PseAA, Wavelet, Dipeptide, CPSR, and Two-stage SVM. The overall accuracies of these methods on dataset 4 were 79.4%, 87.5%, 87.0%, 90.3%, 91.4%, 90.1%, 95.2%, and 96.7%, respectively. Compared with weighted SVM using PseAA (90.3%) and Two-stage SVM (96.7%), our proposed method (99.0%) was more effective, obtaining gains of 8.7% and 2.3%, respectively.

4 Conclusions

We used five methods, AvBlock, DCT, DWT, HOG, and PsePSSM, to extract the protein features. The constructed hypergraph neural network model achieved better results on different datasets. The fusion of different features, driven by multimodal data, further improved the accuracy of membrane protein identification. Therefore, HGNN has the advantages of strong scalability for multimodal features and flexibility of hyper-edge generation.