Identification of Membrane Protein Types via Deep Residual Hypergraph Neural Network

Jiyun Shen, Zhiqiang Hui, Long Cheng

Suzhou University of Science and Technology

Abstract. Conventional computational methods for identifying the species of membrane proteins tend to ignore two issues: high-order correlation among membrane proteins and the scenarios of multi-modal representations of membrane proteins, which leads to information loss. To tackle those two issues, we use a deep residual hypergraph neural network (DRHGNN) to learn the representations of membrane proteins further and to achieve accurate identification of membrane proteins' types eventually.

1 Methods

In order to extract features from membrane proteins' PSSM, we employ Average Blocks (AvBlock), Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT), Histogram of Oriented Gradient (HOG), and Pseudo-PSSM (PsePSSM). Each type of PSSM-based feature is used to generate a hypergraph G which can be represented by an incidence matrix H. Then, five types of features and corresponding H are concatenated, respectively, and both are fed into a deep residual hypergraph neural network (DRHGNN) to identify the types of membrane proteins. Figure.1. depicts the schematic diagram.

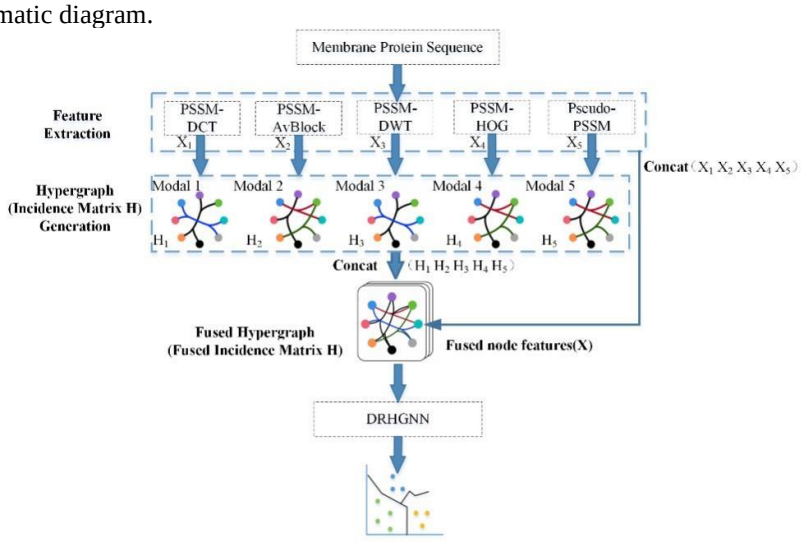

Figure. 1. The schematic diagram of our proposed method.

Figure.2. illustrates the detail of the deep residual hypergraph neural network (DRHGNN). Those multi-types of node features and corresponding incidence matrix H modelling complex high-order correlation are concatenated, respectively, which overcomes the scenarios of multi-modal representations of membrane proteins. Then, concatenated features and incidence matrix are fed into deep residual hypergraph neural network to get nodes output labels and eventually achieve classification task. We build a residual enhanced hypergraph convolution layer. Then we naively stack multiple residual hypergraph convolution blocks to tackle the problem of over-smoothing in HGNN and enjoy an accuracy increase. Additional Linear transforms are incorporated into the model's first and last layer, and the residual hypergraph convolutions are utilized for information propagation. The deep embeddings are finally used for classification tasks.

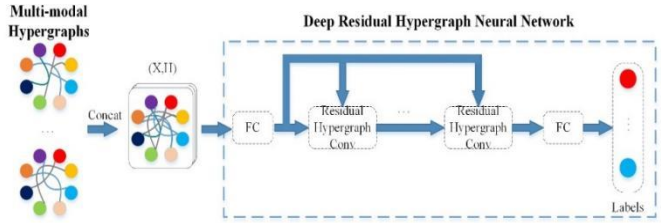

Figure .2. The DRHGNN framework. FC represents a fully connected layer.

2 Datasets

We judge the performance of DRHGNN on the classification of membrane proteins based on four datasets. Table 1. outlines the details of the datasets.

Table 1. The scale of training and testing samples in four different membrane proteins' 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

* - represents not available.

3 Results

As Figure. 3. shows, HGNN using identify mapping can mitigate the problem of over-smoothing a little, and HGNN using initial residual can reduce the over-smoothing problem greatly. Meanwhile, adopting initial residual and identity mapping together can significantly improve performance while effectively reducing the over-smoothing problem. Furthermore, we find that the experimental results of HGNN adopting initial residual and identity mapping together and HGNN using initial residual are very close. However, HGNN adopting both outperforms in terms of accuracy and the macro average of the F1-score and reaches the best result faster than just adopting the initial residual.

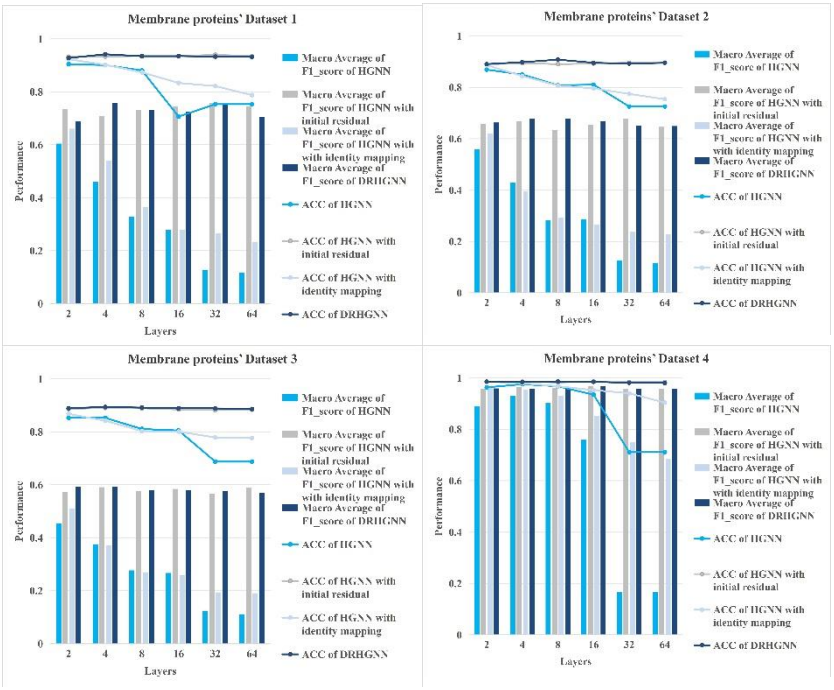

Figure. 3. The performance comparison of DRHGNN, HGNN, HGNN with initial residual, HGNN with identity mapping with different layers on membrane protein classification task.

4 Conclusions

DRHGNN resolves the following issues: the high-order correlation among membrane proteins and the scenarios of multi-modal representations of membrane proteins.

We carry out extensive experiments whose results demonstrate the better performance of DRHGNN on membrane protein classification task. Experiments also show that DRHGNN can handle the over-smoothing issue as the number of model layers increases compared with HGNN.