Title: Graph Convolutional Network with Self-Attention Pooling for the Prediction of Neutralizing Paratope Sequences of SARS-CoV2 Antibodies

Authors: Shamima Rashid[1], Ng Yue Hao Shaun[1,2], Ng Teng Ann[1], Kwoh Chee Keong[1]

Affiliation: 1. School of Computer Science & Engineering, Nanyang Technological University
2. School of Chemical & Biomolecular Engineering, Nanyang Technological University

Background

The COVID-19 pandemic caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV2) pathogen has resulted in a great loss to human lives and economic disruption. Although the severity of the disease outbreak has been overcome and normal operations have resumed in many countries, therapeutics to treat COVID-19 still remain necessary as many in the population continue to get re-infected with circulating variants of the SARS-CoV2 pathogen. It would be ideal to have a repertoire of suitable antibody or paratope sequences which can be rapidly designed for therapeutic needs, based on emergent strains.

In-silico models provided by deep graph networks are an avenue for high-throughput discoveries of neutralizing antibody sequences. Graph neural networks have emerged as promising architectures in several aspects of health and molecular medicine, such as in adaptive graph relations for antibody prediction, [1] models of drug-target interactions [2] and to aggregate spatially related cellular data [3]. Here, a deep graph neural network employing graph convolution with self-attention pooling was trained to detect pairs of neutralizing paratopes and epitopes from sequence data alone.

Methods

A graph network was trained on a dataset of approximately 300 pairs of SARS-CoV2 epitope paratope peptide sequences. Their corresponding neutralization data were used as labels for classification (1 indicates neutralization and 0 indicates non-neutralization). The sequence-based input data were converted to graph-based features using RDKit library [4], where an atom corresponds to a node in the graph and a bond represents the edge between two connected atoms. The network consists of 2 stacked layers of graph convolutional network followed by self-attention graph pooling. The output consists of 2 linear layers and the network was trained for 65 epochs using a binary cross-entropy loss function. Approximately two-thirds of the data were used in training and the remainder one-third was used in testing.

Results

The model obtained a test accuracy of approximately 71% and a high F1 measure of 81.5% using the train-test split method. Although the sensitivity achieved was high (greater than 90%), the specificity needs to be improved owing to the large number of false positives. The ROC-AUC and Matthew's Correlation Coefficient (MCC) scores were 57.4% and 0.201, respectively. We also verified our results with 5-fold cross validation and obtained similar

performance across all metrics tested. Further settings such as a top k pooling layer and Laplacian eigenvector based positional encodings were also explored, but the performance remained similar or showed a slight degradation.

Conclusions

The results on the SARS-CoV2 dataset indicate that it is possible to develop a graph convolutional classifier using only the corresponding epitope and paratope sub-sequences from their respective antigen and antibody molecules. In comparison, many algorithms require knowledge of the VH or VL regions (sequences or structures) to make the prediction. More pairs of data and advanced graph featurization and representation approaches will assist the development of good models to rapidly detect neutralizing epitope-paratope pairs.

References

[1] J. Zhang *et al.*, "Predicting unseen antibodies' neutralizability via adaptive graph neural networks," *Nat Mach Intell*, vol. 4, no. 11, Art. no. 11, Nov. 2022, doi: 10.1038/s42256-022-00553-w.

[2] H. Wang, G. Zhou, S. Liu, J.-Y. Jiang, and W. Wang, "Drug-Target Interaction Prediction with Graph Attention networks." arXiv, Jul. 10, 2021. Accessed: May 29, 2023. [Online]. Available: http://arxiv.org/abs/2107.06099

[3] J. Li, S. Chen, X. Pan, Y. Yuan, and H.-B. Shen, "Cell clustering for spatial transcriptomics data with graph neural networks," *Nat Comput Sci*, vol. 2, no. 6, Art. no. 6, Jun. 2022, doi: 10.1038/s43588-022-00266-5.

[4] RDKit: Open-source cheminformatics. https://www.rdkit.org.