NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE
School of Computer Science and Engineering
College of Engineering

Prediction of Neutralising Antibodies for Novel Coronavirus with Machine Learning

Student: Jordon Kho Junyang Supervisor: Assoc Prof Kwoh Chee Keong

Project Motivation

Since the beginning of the 21st century, coronaviruses were responsible for 3 major viral outbreaks – SARS, MERS, and COVID-19. Coronavirus infections can cause serious respiratory disease and even death. While neutralising antibodies have potential to prevent future infections, conventional lab-based processes are often too time-consuming and expensive. Therefore, machine learning approaches have been gaining popularity in expediting the search for potential antibody candidates.

Project Aims

- Investigate the utility of graph features for discovering neutralising SARS-CoV-2 antibodies as compared to molecular fingerprints
- Evaluate mean and max pooling methods used for preparing the graph features
- Determine which oversampling technique – Synthetic Minority Oversampling Technique (SMOTE) or SMOTE-Nominal (SMOTE-N) – is more suitable for resolving class imbalance in the data set

Data Set

- Taken from the Coronavirus Antibody Database
- 225 neutralising antibodies, 82 non-neutralising antibodies
- A pair of epitope and paratope sequences were obtained from each antibody
- SMOTE and SMOTE-N were applied on the data set separately

Feature Set

- Mean Pooling Graph Features (k = 74)
- Max Pooling Graph Features (k = 74)
- Combined Pooling Graph Features (k = 148)
- Molecular Fingerprints (Baseline)

Models

- Random Forest
- Decision Tree
- Extreme Gradient Boosting
- Light Gradient Boosting Machine
- Logistic Regression
- Support Vector Machine
- Multilayer Perceptron

Results

Best Performing Feature Set	Best Performing Oversampling Technique	Accuracy	F1 Score
Mean Pooling Graph Features	SMOTE-N	72% - 82%	77% - 84%

Mean pooling features was found to capture sequence information more accurately than max pooling features and SMOTE-N was evaluated to be more compatible with graph features than SMOTE as the latter was susceptible to noise generation. However, the models had high false positive rates of up to 41% and thus, other oversampling techniques in combination with undersampling techniques could be explored in future work.