YESHIVA COLLEGE
THE JAY AND JEANNIE SCHOTTENSTEIN HONORS PROGRAM

THESIS PROPOSAL COVER

This form must be accompanied by a proposal, written in consultation with the student's mentor. For guidelines on the format of the thesis proposal, consult the Honors Program website.

STUDENT NAME: Moshe Carroll **YUID:** 800609823

MENTOR: Raji Viswanathan

STUDENT'S MAJOR: Biochemistry

DATE OF PLANNED GRADUATION (Check and complete one):

___ **January, 20__** ✓ **June, 2023** __**September, 20__**

THESIS TITLE: Hierarchical Clustering to Identify Multiple Epitopes in Antigens

COURSEWORK RELEVANT TO THESIS:

Principles of Biology 1/2
General Chemistry 1/2
Organic Chemistry 1/2
Biochemistry
Bioinformatics
Statistics

STUDENT SIGNATURE: Moshe Carroll **DATE:** 11/7/22

MENTOR SIGNATURE: Raji Viswanathan **DATE:** 11/7/22

Hierarchical Clustering to Identify Multiple Epitopes in Antigens

Moshe Carroll

The identification of the interfacial residues involved in protein-protein interactions is particularly important in the field of immunology. The binding region of an antibody to its cognate antigen is highly specific, incorporating a high degree of complementarity with respect to its size, shape, charge, and polarity, with the antigen epitope region. Experimental techniques, including NMR spectroscopy and X-ray crystallography, may be used to map epitope regions but each method has its own limitations and are largely costly and have low-throughput. Additionally, antigens could have multiple epitopes which each bind to different antibodies. Experimental techniques, however, may only identify a single epitope in a given crystallized complex with an antibody. In turn, less targets are available for monoclonal antibody and vaccine development. Computational methods can be developed for the identification of the epitope regions even in the absence of the crystallized antibody-antigen complex and can also reveal multiple distinct epitopes, revealing new targets against a query antigen. Several methods to identify B cell epitopes have already been developed, such as ElliPro and DiscoTope. These current models largely base their predictions on either the sequential or structural information of a query antigen. Collectively, there is significant opportunity in the field of computational biology for the development of methods that combine both structure and sequence-based approaches to improve epitope predictions which are notoriously difficult to accurately determine.

Our laboratory recently described in *BMC Bioinformatics* the Integrated Structure-based Protein Interface Prediction (ISPIP) method that improves the predictive performance of interfacial residues on a query protein by combining methods that rely on orthogonal structure-based properties by linear or logistic regression, random forest, or gradient boosted tree models. The objective of this project is to test and enhance the performance of ISPIP in identifying multiple antigen epitopes in both the complexed and uncomplexed conformations using a data set of 335 antibody-antigen complexes curated by Jespersen et al. ISPIP's performance will also be compared to those of state-of-the-art individual classifiers — such as docking, template, and structural-based models — as well as other combined prediction methods, such as VORFFIP and meta-PPISP, and the antigen-specific methods ElliPro and DiscoTope. Since antigens could have multiple epitopes, I will develop and implement a hierarchical clustering methodology to identify potential epitope regions in an antigen. Hierarchical clustering can separate the predicted residues into distinct patches based on their relative distances in 3D space. Clustering will be applied to compare the predicted and experimentally determined epitopes and elucidate the presence of additional epitopes for further experimental investigation. The performance of each prediction model will be assessed using the F1-score and MCC score. Receiver Operator Characteristic (ROC) curve and precision-recall (PR) curves will be generated, and the ROC-AUC and PR-AUC will be calculated and compared among the various models to determine their specificities and sensitivities. These statistical measures will serve as an indicator of each prediction model's performance both before and after applying hierarchical clustering and will guide our analysis of the protein folds, families, and

superfamilies in which the predictions both succeed and fail. I hope that my efforts will contribute to a greater understanding of antibody-antigen interactions.

Bibliography.

Walder, M., Edelstein, E., Carroll, M. *et al.* Integrated structure-based protein interface prediction. *BMC Bioinformatics* 23, 301 (2022). https://doi.org/10.1186/s12859-022-04852-2

Toride King, Moeko, and Cory L Brooks. "Epitope Mapping of Antibody-Antigen Interactions with X-Ray Crystallography." *Methods in molecular biology (Clifton, N.J.)* vol. 1785 (2018): 13-27.

Norman, Richard A et al. "Computational approaches to therapeutic antibody design: established methods and emerging trends." *Briefings in bioinformatics* vol. 21,5 (2020): 1549-1567.

Soria-Guerra, Ruth E et al. "An overview of bioinformatics tools for epitope prediction: implications on vaccine development." *Journal of biomedical informatics* vol. 53 (2015): 405-14.

Jespersen, Martin Closter et al. "Antibody Specific B-Cell Epitope Predictions: Leveraging Information From Antibody-Antigen Protein Complexes." *Frontiers in immunology* vol. 10 298. 26 Feb. 2019.

Abbott, W Mark et al. "Current approaches to fine mapping of antigen-antibody interactions." *Immunology* vol. 142,4 (2014): 526-35.

Data Mining and Knowledge Discovery Handbook. United Kingdom, Springer US, 2005.

Zhang, Q.C., et al., *PredUs: a web server for predicting protein interfaces using structural neighbors.* Nucleic Acids Res, 2011. 39(Web Server issue): p. W283-7

Saito, T. and M. Rehmsmeier, *The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets.* PLoS One, 2015. 10(3): p. e0118432.