Research Proposal

The identification of the interfacial residues involved in protein-protein interactions is particularly important in the field of immunology. Antigens induce immune responses to generate humoral and cell-mediated immunity. The humoral and cell-mediated responses are targeted against extracellular and intracellular pathogens, respectively. Humoral immunity involves plasma B cells which secrete antibodies that bind to the antigenic determinant regions, or epitopes, of antigens. The binding of an antibody to its cognate antigen is highly specific. This is attributed to the variable domains located at the upper tip of the Fab regions of immunoglobulin G (Figure 1).

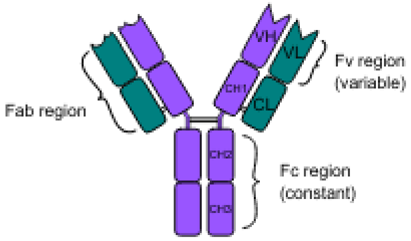

Figure 1. Structure of Immunoglobulin G [1]

These domains feature a high degree of complementarity, including size, shape, charge, and polarity, with the epitope region. A variety of experimental techniques are employed to map epitope regions, such as NMR spectroscopy and X-ray crystallography. While NMR does not require samples to be crystallized, its capability to determine the structures of smaller proteins with molecular weights around 20 kDa is limited. X-ray crystallography is generally regarded as the benchmark standard to determine interfacial residues as it provides a high atomic resolution of the three-dimensional image of the antibody-antigen complex [2]. However, this method is largely low-throughput, costly, and ultimately relies on the availability of the crystallized complex. Additionally, antigens could have several epitopes which each bind to different antibodies. Since X-ray crystallography requires crystallization of each antibody-antigen complex, only a single epitope may be identified in a given complex. For example, there are three experimentally determined epitopes for human interleukin 6 (PDB ID: 1ALU) when complexed with olokizumab, camelid, and Llama Fab fragment 68F2 antibodies. The interfacial residues are indicated by the pink spheres in Figure 2.

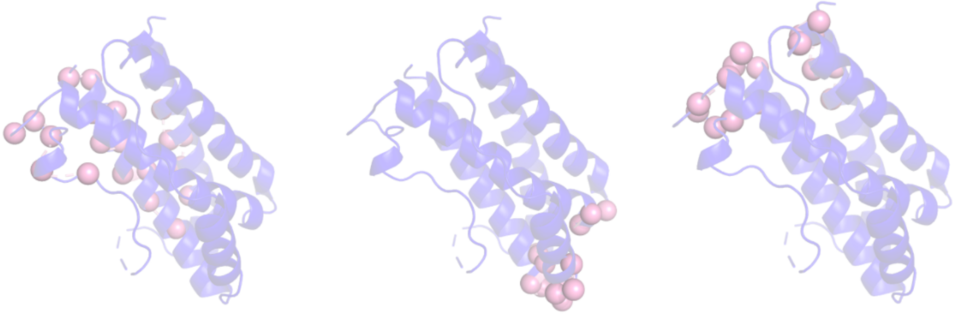

Figure 2. Experimentally Determined Epitopes of Human Interleukin 6

Computational methods can be developed for the identification of the epitope regions even in the absence of the crystallized antibody-antigen complex. Furthermore, computational methods may reveal multiple distinct epitopes, generating additional targets in the development of monoclonal antibodies and vaccines [3]. Several of these 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 [4].

Recently, Jespersen et al. have curated a data set of 335 antibody-antigen complexes [5]. This set incorporates both pathogenic and non-pathogenic antigens and includes multiple complexes of a single antigen bound to different antibodies, thereby revealing multiple epitopes. Because the antibody-antigen complexes are expected to undergo large conformational changes to exhibit tighter binding, epitope regions are more readily identified by analyzing the complexed antigen structures [6]. Since we are developing a method that does not depend on the availability of the complexed structure, we will find uncomplexed antigens with 95% sequence similarity to those in the complexed antigen test set using the RCSB protein similarity searching tool. We will incorporate both the complexed and uncomplexed antigen structures into distinct test sets to make predictions of the epitopes.

Computational predictions of the epitopes will be performed using at least three previously published individual classifiers, ISPRED4, DOCKPRED, and SPPIDER II, which assign an epitope probability score to each residue in the antigen based on sequence and structure-specific features. In our work, we describe 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 to identify antigen epitopes in both the complexed and uncomplexed conformations. The performance of ISPIP will be compared with the individual classifiers described above, 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, we will use a hierarchical clustering approach to identify potential epitope regions. Hierarchical clustering can separate the predicted residues into distinct patches based on their relative distances in 3D space [7]. 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 four statistical measures. Two of these measures, F1-score and Matthews correlation coefficient (MCC) score, are based on establishing a threshold value, N, for the number of top-ranking residues that comprise the protein epitopes. To determine the threshold value, a dynamic cutoff for each query protein will be calculated according to the following equation: $N = 6.1\, R^{0.3}$ where R is the number of surface-exposed residues for each query protein [8]. Using this threshold value, the elements of the confusion matrix, True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) can be determined. F1 and MCC scores will then be calculated to determine the prediction accuracies. The other two measures, Receiver Operator Characteristic (ROC) curve and precision-recall (PR) curve use the area under their respective curves as a measure of success. ROC and 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 [9]. Collectively, these statistical measures will serve as comprehensive indicators of the relative performances of each prediction model and will guide our analysis of the protein folds, families, and superfamilies in which the predictions both succeed and fail.

My role in this project consists of several steps: (I) curate the dataset of uncomplexed antigens. (II) Identify the training and test sets for the complexed and uncomplexed datasets to incorporate into the machine learning algorithm. (III) Receive data from group members on the individual methods and integrate these results using ISPIP. (IV) Develop codes to perform the clustering of the predicted residues. (V) Develop codes to perform the four statistical measures. (VI) Compare the performance of ISPIP with other computational methods. My participation in this project will contribute to my growth at Yeshiva University by expanding beyond the undergraduate science curriculum and pursuing my research interests. As I develop and advance this project, I will learn more about the process of research including manuscript preparation. I hope that my efforts will contribute to advancements, albeit in a small way, to a greater understanding of antibody-antigen interactions.

References

1. Kramer, David. “Anti-Immunoglobulin G (IgG) Secondary Antibodies.” *Antibodies*, https://www.antibodies-online.com/resources/16/675/anti-immunoglobulin-g-igg-secondary-antibodies/.
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.
3. 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.
4. 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.
5. 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.
6. Abbott, W Mark et al. “Current approaches to fine mapping of antigen-antibody interactions.” *Immunology* vol. 142,4 (2014): 526-35.
7. Data Mining and Knowledge Discovery Handbook. United Kingdom, Springer US, 2005.
8. 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
9. 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.

Timeline

Summer 2022:

- Review literature pertaining to antigen epitope predictions
- Curation of Jespersen bound-conformation antigen dataset
- RCSB protein similarity searching to curate the uncomplexed antigen dataset
- Map experimental interfacial residues on the bound-conformation antigens to the corresponding unbound-conformation antigens
- Perform residue epitope probability calculations with ISPRED4, DOCKPRED, and SPPIDER II, ElliPro, DiscoTope, meta-PPISP, and VORFFIP on both datasets

Fall 2022:

- Curation of test and training sets for ISPIP cross-validation
- Perform ISPIP calculations and cross-validation
- Analysis and comparison of the performance of all the prediction methods using the four statistical measures
- Investigate the most appropriate and effective parameters to design the hierarchical clustering script
- Perform hierarchical clustering using the epitope predictions of each computational method

Spring 2023:

- Analyze and compare the performance of all the prediction methods after the clustering of distinct epitopes
- Enhance the performance of ISPIP
- Attend ACS conference

Budget

- ACS conference registration and travel (Mar. 26-30, 2023): $1000
- Manuscript publication cost: $500