Evan Edelstein
Mentor: Dr. Raji Viswanathan
Prediction of Antigen Epitopes

Proposal Approved for Submission
Raji Viswanathan

Abstract

Proteins have a wide range of biological functions. Proteins are polymers made from different combinations of the 20 amino acid building blocks. The unique combination of amino acids in a protein results in a unique 3-D structure. A protein's function is determined by not only its sequence and 3-D structure, but also by its interactions with ligands, or binding partners. Protein-protein interactions occur at specific binding sites, known as the interface. Protein interactions are a key component in many biological processes including immunological response pathways. The two main proteins involved in the immune system are antigens and antibodies.

Infection-causing invaders, called pathogens, express certain proteins, known as antigens, on their surface. Antibodies are proteins produced by the immune system to target antigens and trigger an immune response. They require specific and complementary binding to the antigen in order to function effectively. The interface, or binding region of an antigen in the antigen-antibody complex is called the epitope. Knowledge of the structure of the epitope of an antigen is needed to build antibodies that target the antigen. Experimental methods for predicting protein interfaces are costly, time consuming and inefficient. Therefore, computational algorithms have been developed to more effectively predict protein interfaces. However, few programs are currently optimized for predicting the antigen epitopes. We have recently developed Meta-DPI, a protein interface prediction program, that combines three prediction algorithms to more accurately predict protein interfaces.

We intend to further develop Meta-DPI in order to effectively predict antigen epitopes by incorporating more complex machine learning algorithms as well as robust decision-making protocols. Additionally, we will develop automated visualization tools to better understand the interface regions of antigen-antibody complexes. Development of computational methods to identify the epitope region of antigens presented by pathogens, like Sars-CoV-2, will enable the development of treatment options, or create new drugs that target the virus. For this reason, epitope prediction has become increasingly relevant since the Covid19 outbreak.

Background

The human body's defense against foreign invaders, or pathogens, can be divided into two classes: the innate and adaptive immune systems. The innate immune system is the first line of defense. It includes physical and chemical barriers against pathogen entry into the body and specialized cells with nonspecific defense activity for when pathogens do enter. This response is immediate and nonspecific. In contrast, the adaptive immune response takes longer to mount, but is more specific and has a memory component. This response utilizes T and B cells to recognize and kill specific pathogens. B cells produce one of the most important components of the immune system: antibodies. Antibodies are Y-shaped proteins that recognize a specific part of an invading pathogen called the antigen. This binding facilitates the clearing of the pathogen through neutralization, phagocytosis, antibody-dependent cellular cytotoxicity, or complement mediated lysis. Due to the key defensive functions of antibodies, determining their specific target region on the antigen, or the epitope, is of utmost importance to researchers. The ability to accurately and efficiently predict the epitope of a query antigen with an unknown complex structure is an important step in developing a pharmaceutical drug to fight novel pathogens with unknown antigen epitope regions[1].

Experimental Protein Interface Prediction

There are a variety of experimental methods for predicting the Interface, or binding region, of a protein. Of these methods X-Ray Crystallography has the highest resolution; it determines the 3D structure of a protein complex based on X-Ray diffraction data, and requires crystallization of protein complexes, which is often difficult. X-Ray Crystallography also fails to capture the dynamic and flexible binding interactions between proteins. Because of these two drawbacks, computer algorithms have been developed to more accurately and efficiently predict the interface region of a protein. This proposal will develop these methods further in order to predict the more complex interfaces of antigens, the epitopes.

Computational Protein Interface Predictors

There are three major categories of previously established computational prediction methods: template-based predictors,

intrinsic-based predictors and docking-based predictors. Template based programs find proteins which are structurally similar or evolutionarily related to the query protein, known as homologs. The interface of the homologous protein is then mapped onto the query protein to predict the binding interface[2]. For a given query protein, the efficacy of template-based predictors is determined by the number and similarity of homologs with known complex structures. Intrinsic-based predictors determine protein interfaces using features of the protein's own sequence and structure. These features include evolutionary conservation, solvent accessible surface area, protrusion index, to name a few. A feature classification model is then used to predict which residues will be at the interface[3]. These methods do not require the experimentally determined interfaces of homologous proteins, giving them an advantage over template-based predictors. However, they are limited by the number of features that can be used as classifiers in the prediction model. Docking-based predictors determine the interface of proteins based on the energetics of docked complexes. A recently developed method, DockPred, identifies the frequencies with which residues are observed at the docked complexes of the query protein with various binding partners. It was observed that proteins use certain "sticky sites" to form complexes, irrespective of the binding partner[4].

Meta-DPI

After the development of Dockpred, we hypothesized that creating a meta-method by combining other predictors that are orthogonal, or statistically independent, to DockPred would have more success than DockPred or any individual predictor alone. Over the past year, we have developed Meta-DPI, a program that incorporates all three prediction models into a single method capable of outperforming each method individually. PredUS (Template-based), ISPRED4 (Intrinsic-based) and DockPred (Docking-based) were combined to develop Meta-DPI. The improved performance of the meta-method on a large dataset validates our hypothesis.

Epitope Prediction

There are multiple aspects of antigen binding sites that make their predictions more difficult than general protein interface predictions. Other protein interfaces contain stabilization reactions between their

amino acids; whereas, amino acids at the epitope do not interact differently than those in the rest of the protein[5]. Further complicating the prediction is that there are usually multiple epitope sites on a single antigen. Moreover, even within a single epitope site, there are binding, and non-binding patches based on the conformation of the epitope, resulting in non-sequential patches of interface residues within the epitope region. Due to the complexity of antigen epitopes, more robust computational tools are required to accurately predict them. Incorporating decision making protocols into prediction algorithms allows them to predict multiple epitopes. Furthermore, advanced machine learning algorithms can train using known epitopes to determine more refined classifier prediction classes. These methods will be investigated to increase the accuracy of Meta-DPI in predicting antigen epitopes.

Decision Tree and Random Forest

Machine learning algorithms provide robust tools for prediction optimization. Meta-DPI already incorporates Logistic Regression to compute optimization coefficients, which allows for a weighted combination of template, intrinsic, and docking based predictors. Decision Tree and Random Forest are two other machine learning algorithms that could enhance the accuracy of epitope prediction. Decision Tree is a machine learning method that creates classification rules for a data set. The input data is loaded into a "tree" network that procedurally sorts the information into nodes. The data is then sorted into classes or "leaves." By training the Decision Tree algorithm, a more accurate classification system can be developed. The benefits of implementing a Decision Tree learning process include the clear visualization of the algorithm's classification procedure and flexibility in modifying the classification procedure[6]. Random Forest is another machine learning algorithm that generates multiple Decision Trees to create multiple classification protocols. Through learning, the Random Forest is able to change the weight each individual Decision Tree is given, generating a formula for combining each Decision Tree's output into a single prediction class[7]. Random Forest lacks the readability and flexibility of standard Decision Tree algorithms, but it generates more accurate classification procedures. The training of the machine learning

algorithms can be performed on antigens with experimentally determined epitope sequences, which will allow the classification process to be optimized for antigen epitope prediction.

Data Analysis and Visualization

Statistical metrics were used to assess the performance of Meta-DPI protein interface prediction. Similar tools will be utilized to assess Meta-DPI's success rate for antigen epitope prediction. Receiver Operator Characteristic (ROC) curve and Precision-Recall (PR) curve were used as graphical metrics. An Interface-Probability (IP) score was calculated for each residue in the query protein using a PredUS, ISPRED4, DockPred and Meta-DPI. The IP score measures the probability that the residue is at the protein's interface. Using the IP scores, and varying threshold values for IP, a ROC and PR curve was generated for each individual predictor and combination of predictor methods. The area under the curve (AUC) for both the ROC curves and PR curves was calculated. A greater AUC indicates a more successful prediction algorithm. Cross validation between different data sets, Docking Benchmark and NOX, were used to verify the performance of the predictors. Each predictor was trained on the Benchmark data set and used to compute an IP score for all proteins in the NOX data set. This procedure was then repeated, training on NOX and calculating IP scores for the Docking Benchmark. Meta-DPI showed a greater AUC than any single prediction method. Two other statistical measures were also used to further analyze the performance of Meta-DPI: A Matthews Correlation Coefficient (MCC) and F_1, and Meta-DPI outperformed in these metrics as well. Each of the four metrics used have independent drawbacks and biases but by using a consensus of the four, Meta-DPI's performance can be analyzed objectively.

Conclusion

The necessity for rapid, cost effective, and accurate prediction of protein interfaces, especially the antigen epitopes, has become increasingly apparent during the COVID19 pandemic. Therefore, computational algorithms continue to be developed to predict the residue sequence of protein interfaces. It has been shown that Meta-DPI successfully predicts general protein interfaces. We intend to improve Meta-DPI's ability to accurately predict antigen epitopes.

This will be accomplished by developing robust Machine Learning tools and decision protocols alongside the advancement of Meta-DPI's prediction algorithm. Ability to accurately predict the antigen epitopes is the first step in targeted protein therapy and artificial vaccine development. We will also develop visualization tools to help us understand the 3-D structure of the antigen epitopes. Statistical measures including ROC curves, PR curves, MCC and F_1 score will be used to assess the success of the method.

References

1. Grifoni, A., Sidney, J., Zhang, Y., Scheuermann, R. H., Peters, B., and Sette, A. (2020, April 8) A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2. Cell host & microbe.
2. 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.
3. Savojardo, C., et al., ISPRED4: interaction sites PREDiction in protein structures with a refining grammar model.Bioinformatics, 2017. 33(11): p. 1656-1663.
4. Viswanathan, R., et al., Protein-protein binding supersites. PLoS Comput Biol, 2019. 15(1): p. E1006704.
5. Sun, J., Xu, T., Wang, S., Li, G., Wu, D., and Cao, Z. (2019, February 13) Does difference exist between epitope and non-epitope residues? Analysis of the physicochemical and structural properties on conformational epitopes from B-cell protein antigens. Immunome Research. Longdom Publishing SL.
6. Darnell, S. J., Page, D., & Mitchell, J. C. (2007). An automated decision-tree approach to predicting protein interaction hot spots. Proteins: Structure, Function, and Bioinformatics, 68(4), 813–823.doi:10.1002/prot.21474
7. Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

Timeline

Summer 2020:

- Perform literature search to learn about the current computational methods for predicting antigen epitopes.
- Perform literature search on Sars-CoV-2 including its structure, epitope and antibody binding complex.
- Identify benchmark antigen datasets with known epitopes to measure the prediction ability of our method.
- Begin improving Meta-DPI to prepare for further development.

Fall 2020:

- Continue research on Machine Learning algorithms.
- Develop framework and write scripts for Decision Tree and Random Forest implementation.
- Debug and optimize Meta-DPI and machine learning algorithms for antigen epitope prediction.
- Apply statistical methods for performance analyses of machine learning protocols.
- Develop Protein Interface Visualization software.
- Attend New York Structural Biology discussion group

Spring 2021:

- Prepare and present at research group meetings at YC and Einstein.
- Further optimize machine learning algorithms.
- Validate Meta-DPI's prediction performance and visualize its results.
- Compile Meta-DPI and Visualization toolkits into an easy to use online protein interface/antigen epitope prediction program.
- Prepare for ACS (American Chemistry Society) conference
- Begin a rough draft of the manuscript.

Budget

- External Hard-Drive (data storage): $50
- Registration to attend a meeting at the NY academy of science: $300
- Traveling, Registration and Hotel for ACS national Meeting and Exposition (Date TBD): $750