Evan Edelstein

Mentor: Dr. Raji Viswanathan

Prediction of Antigen Epitopes

Abstract

Proteins have a wide range of biological functions. They facilitate many biochemical reactions via catalysis, enable communication between cells, and are the main component of immunological responses. Proteins are formed from the linkage of Amino Acid monomer subunits. There are 21 unique amino acids, each with their own chemical, physical and electrical properties. The unique combination of amino acids in a protein results in a complex 3-dimensional structure. A protein’s function is determined by how its three-dimensional structure can interact with ligands, or binding partners. Protein-protein interactions occur at specific binding sites, known as the interface. Building computational tools for predicting the amino acid sequence at the interface of a protein has been of interest to researchers. Currently there are many programs that predict protein interfaces, but few are optimized for predicating the more complex epitope sequence. In immunological settings, knowledge of the protein interface is a critical first step in antibody treatment. Specifically, researchers aim to determine the antigen-antibody binding interface, or epitope of pathogens. Pathogens are invading cells, such as viruses and bacteria; antigens are the proteins presented on their surface. Antibodies are produced by the immune system to target the antigens presented by pathogens, in order to trigger a targeted immune response. Antibodies require specific and complementary binding to antigens in order to function properly. For this reason, determining the structure of the epitope of an antigen is needed to build antibodies that target the antigen. Experimental methods for protein-protein interface determination are costly, time consuming and inefficient. Therefore, computational algorithms have been developed to more effectively predict protein interfaces. We have developed Meta-DPI, a protein interface prediction program that combines three prediction algorithms to accurately predict protein interfaces We intend to develop Meta-DPI further in order to effectively predict antigen epitopes. By incorporating more complex machine learning tools such as Random Forest and Decision Tree as well as incorporating more robust decision-making protocols we hope to more accurately and efficiently predict the antigen epitopes. In conjunction with progressing Meta-DPI we will develop automated visualization tools to better understand the interface regions of antigen-antibody complexes. Knowledge of the protein’s epitopes utilized by a novel pathogen, such as Sars-CoV-2, can be analyzed using computational tools, to quickly determine the efficacy of treatment options, or create new drugs that target specific antigens presented by the virus.

All proteins, including antigens and antibodies, are composed of sequentially linked amino acids. These amino acids are often referred to as residues. The amino acid residues of a protein that participate in binding with another protein are said to be at the protein’s interface. Thus, it can be said that the epitope of an antigen is composed of the interface residues of the antigen. Understanding the structure of the antigen’s epitope is the critical initial step that enables researchers to create artificial antibodies against the epitopes, allowing for the production of synthetic vaccines

Background

The human body’s defense against foreign invaders, or pathogens, can be divided into two components: 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 - like macrophages, dendritic cells, and neutrophils - 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 cells and B cells to recognize and kill 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 its neutralization, phagocytosis, antibody-dependent cellular cytotoxicity, or complement mediated lysis. Due to the key defensive functions of antibodies, the specific antigen-antibody binding complex, or the epitope, is of utmost priority to researchers.

All proteins, including antigens and antibodies, are composed of sequentially linked amino acids. These amino acids are often referred to as residues. The amino acid residues of a protein that participate in binding with another protein are said to be at the protein’s interface. Thus, it can be said that the epitope of an antigen is composed of the interface residues of the antigen.

Understanding the amino acid sequence of the antigen’s epitope is the critical initial step that enables researchers to create synthetic antibodies against the epitope, allowing for the production of synthetic vaccines. Being able to accurately and efficiently predict the epitope residues of a query antigen with an unknown complex structure is an important step in developing a pharmaceutical drug to fight a pathogen.

Protein Interface Prediction

There are a variety of experimental methods for predicting the structure and binding interface of proteins. The method with the highest resolution is X-Ray Crystallography, which determines the protein-binding complex based on x-ray diffraction data. In order to perform X-Ray Crystallography a high-resolution crystalized protein complex needs to be prepared. This is a difficult, time consuming and costly procedure. Once a high-resolution crystal is formed, it is analyzed via X-ray diffraction to determine the structure of the protein binding complex. An amino acid is said to be at the interface if it within a certain distance from the binding partner. This method is inconsistent and dependent on choice of threshold size and resolution of synthesized crystals. X-Ray Crystallography also fails to capture the dynamic and flexible binding interactions between proteins. For these reasons, computer algorithms have been developed to more accurately and efficiently predict the interface region of a protein. There are two major categories of computational prediction methods, template-based predictors and intrinsic-based predictors.

Template Based Predictors

Template based programs find structurally similar proteins or evolutionarily related (homologous) proteins, to the query protein. The interface region of the homologous protein is then mapped onto the query region to predict the binding interface. For a given query protein, the efficacy of template-based predictors is based on the number and similarity of homologues with a known complex structure.

Intrinsic Based Predictors

Intrinsic based predictors determine the protein interface by using features of a proteins sequence and structure. Each residue is assigned features such as evolutionary conservation, solvent accessible surface area and protrusion index. A feature classification model is used to predict which residues will be at the interface based on the features of the residue. Through a learning process the model is optimized by assigning weights to each feature. The most robust intrinsic based predictor is ISPRED4 which uses 46 different features to predict protein interface. These methods do not require the experimentally determined interface of homologous proteins, giving them an advantage over template-based predictors. Although they are limited by the amount of number of features that can be used as classifiers in the prediction model.

Docking Based Predictors

A third type of prediction software called DockPred, predicts interface binding complex based on binding energy models. The area of a protein that binds most regularly is known as a “sticky site”. DockPred predicts the most energy favorable conformation of the bound protein complex, in which the query protein is bound to multiple other ligands. The energetically favored sticky site is used to predict the interface region of the protein.

Meta-DPI

Over the last year, we have developed Meta-DPI to incorporate all three prediction programs into a single meta-method capable of outperforming each method individually. PredUS (Template-based), ISPRED4 (Intrinsic-based) and DockPred (Docking-based) were used to develop Meta-DPI. The development of the meta-method justifies our hypothesis that combining orthogonal prediction metrics would produce an overall more accurate prediction software. Meta-DPI utilizes machine learning algorithms to combine and optimize the three prediction scores.

Epitope prediction

There are multiple aspects of antigen epitope binding sites that make antigen epitope prediction more difficult then general protein interface prediction. Experimental research suggests that unlike general protein interfaces, that contain stabilization reactions between amino acids at the interface, amino acids at the epitopes don not interact differently than amino acids elsewhere on the protein. Furthermore, there are usually multiple epitope sites on a single antigen. Even within a single epitope site, antigen epitopes present binding and non-binding patches based on complex folding patterns or the confirmation of the epitope. The conformation of the antigen presents certain amino acids on its surface that participate in antigen-antibody interactions. The antigen folding structure also covers up amino acids, blocking them from participating in binding, resulting in non-sequential patches of interface residues within the epitope. These complications require specific computational algorithms to predict the epitope amino acid sequence. Decision making protocols can allow prediction algorithms to simultaneously predict multiple epitope sites. Advanced sequence-based tools can lower the probability of folded or non-surface amino acids from being predicted as interface. Furthermore, advanced machine learning algorithms, like Random Forest, can train on known antigen epitope amino acid sequences to determine refined epitope classifier prediction classes, increasing the accuracy of Meta-DPI in predicting antigen epitope sequences.

Decision Tree and Random Forest

Machine learning algorithms provide robust tools for prediction optimization. Meta-DPI already incorporates Logistic Regression to compute optimization coefficients. This allows for a weighted combination of template, intrinsic, and docking based predictors. Decision Tree and Random Forest are two other machine learning algorithms that can enhance the accuracy of epitope prediction. Decision Tree is a machine learning method that creates classification rules for a specific data set. The input data is loaded into a “tree” network that procedurally sorts the information at nodes. The data is then sorted into classified classes or “leaves”. By training the decision tree algorithm a more accurate classification system is developed. The benefits of implementing a Decision Tree learning process are the clear visualization of the algorithm’s classification procedure and flexibility in modifying the Decision Tree classification procedure. Random Forest is a machine learning algorithm that generates multiple Decision Trees to create many classification protocols. Through learning the Random Forest is able to tweak the weight each individual decision tree is given, generating a weighted formula for combining each decision tree into a single classification node. Random Forest lack the easy readability and flexibility of standard Decision Tree algorithms but it creates a more accurate classification procedure. The training of the machine learning algorithms will be done on known antigens with known epitope sequence, this will allow the classification process to be optimized for antigen epitope prediction.

Results, Data Analysis and Visualization

Metrics were used to assess the efficacy of Meta-DPI protein interface prediction; similar tools will be utilized to better understand Meta-DPI’s success rate for antigen epitope prediction. The two metrics used were Receiver Operator Characteristic (ROC) curve and Precision-Recall (PR) curve. An Interface-Probability (IP) score was calculated for each residue in the query protein using PredUS a template-based predictor, ISPRED4 an intrinsic-based predictor, DockPred a docking-based predictor and Meta-DPI. For each predictor a ROC curve and PR curve are generated. Both curves sort the residues of the protein by IP score and iterate over an increasing threshold value; a residue is said to be at the interface if its IP score is greater than the threshold. The residues are then sorted into one of four categorizes. The true positives (TP) are residues predicted to be at the interface that in fact are at the interface. False positives (FP) are residues predicted to be at the interface but in reality, are not. True negatives (TN) are residues predicted to be outside the interface that are in fact not a part of the interface. False Negatives (FN) are residues that are predicted to be not in the interface but are really at the interface. These four classes are used in the computation of ROC curves and PR curves. The ROC curve compares the TP rate (TPR) to the FP rate (FPR). The X-axis is the FPR calculating by the ratio between the FP’s and the total number of residues. The Y-axis is the TPR, which the quotient of TP’s and total number of residues. In PR curves, recall and precision are compared. Recall is defined as the number of TP divided by the sum of TP’s and FN. Precision is calculated by dividing the number of TP’s over the sum of TP’s and FP’s. to quantify the success of the prediction methods, the area under the curve (AUC) for both ROC curves and PR curves is calculated. A greater AUC is associated with more successful prediction algorithm. Meta-DPI showed a greater AUC than any individual prediction method or combination of prediction methods. Cross validation between different data sets, Benchmark and NOX, were used to verify the performance of the predictors. Both data sets contain a total 241 proteins with known interface residues. Each predictor was trained on Benchmark set and used then computed an IP score for all the proteins in the NOX data set. This procedure was then repeated, training on NOX and calculating IP scores for Benchmark. A Matthews Correlation Coefficient (MCC) and F1 score were also calculated for each predictor, Meta-DPI outperformed in these metrics as well.

Conclusion

The necessity for rapid, cost effective and accurate prediction of protein interfaces is of utmost priority to researches. Therefore, computational systems have been developed to predict the residue sequence of protein interfaces. It has been shown that Meta-DPI successfully performs general protein interface prediction. We intend to bolster Meta-DPI ability to accurately predict antigen epitopes. To do so robust Machine Learning tools and decision protocols will be developed alongside advancement of Meta-DPI’s prediction algorithm. Being able to accurately predict the epitope of antigens is the first step in targeted protein therapy and artificial vaccine development. Alongside Meta-DPI a protein visualization program will developed to help researches visualize the epitope of antigens.

Statistical tools including ROC curves and PR curves will be used to assess the success of Meta-DPI.

References

Timeline

Summer 2020:

* Study literature regarding current methods for predicting antigen epitopes to set benchmark.
* Study literature on Sars-CoV-2 including its structure, epitope and antibody binding complex.
* Study literature regarding the development of anti-viral medication to better streamline the epitope prediction-drug development process.
* Complete “Final touches” on Meta-DPI to prepare for further development.

Fall 2020:

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

Spring 2021:

* Create presentations for group Einstein meetings.
* 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 prediction program.
* Prepare for ACS (American Chemistry Society) conference
* Continue research on Machine learning and Neural Networks to enhance Meta-Dpi antigen epitope prediction.

Budget

* External Hard-Drive (data storage): 50$
* Traveling, Registration and Hotel for ACS national Meeting and Exposition (Date TBD Because of Covid19): 750$

|  |  |
| --- | --- |
| Evaluation Metric | Formula |
| F1 Score |  |
| Matthews Correlation Coefficient |  |

A screenshot of a cell phone

Description automatically generated

|  |  |
| --- | --- |
| **Method** | **ROC\_AUC** |
| PredUs 2.0 | 0.743 |
| ISPRED4 | 0.792 |
| DockPred | 0.866 |
| Meta-DPI | 0.871 |

|  |  |
| --- | --- |
| **Method** | **ROC\_AUC** |
| Meta-DI | 0.844 |
| Meta-DP | 0.760 |
| Meta-PI | 0.826 |
| Meta-DPI | 0.849 |

|  |  |
| --- | --- |
| **Method** | **PR\_AUC** |
| PredUs 2.0 | 0.260 |
| ISPRED4 | 0.359 |
| DockPred | 0.264 |
| Meta-DPI | 0.389 |

|  |  |
| --- | --- |
| **Method** | **PR\_AUC** |
| Meta-DI | 0.375 |
| Meta-DP | 0.299 |
| Meta-PI | 0.373 |
| Meta-DPI | 0.389 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Method | Benchmark ROC-AUC | NOX ROC-AUC | Benchmark PR-AUC | NOX PR-AUC | Benchmark F1 | NOX F1 | Behcmarck MCC | NOX MCC |
| PredUS | 0.743 | 0.666 | 0.246 | 0.26 | 0.34 | 0.351 | 0.314 | 0.282 |
| ISPRED4 | 0.792 | 0.815 | 0.263 | 0.359 | 0.342 | 0.4 | 0.316 | 0.366 |
| DockPred | 0.866 | 0.761 | 0.288 | 0.264 | 0.375 | 0.361 | 0.355 | 0.293 |
| Meta-DPI | 0.871 | 0.849 | 0.364 | 0.389 | 0.4 | 0.422 | 0.384 | 0.36 |

Proteins have a wide range of biological functions. They facilitate many biochemical reactions via catalysis and immunological signaling. Proteins are polymers of 20 different amino acid monomer subunits, all of which have their own chemical, physical and electrical properties. The unique combination of amino acids in a protein results in a complex 3-dimensional structure. A protein’s function is determined by not only  its 3-dimensional structure assigned by its amino acid sequence, but also by its interactions with ligands, or binding partners. Protein-protein interactions occur at specific binding sites, known as the interface.

Proteins are polymers made from different combinations of the 20 amino acid building blocks. Each combination leads to a unique protein with different chemical, physical, and electrical properties. These properties lead to the protein’s unique shape, or conformation, that allows it to bind to specific binding partners. The unique amino acid sequence that participates in protein binding is called the interface. Proteins are essential for many biochemical reactions throughout the body, like in the immune system.

Proteins are polymers of 20 different amino acid monomer subunits, all of which have their own chemical, physical and electrical properties. The unique combination of amino acids in a protein results in a complex 3-dimensional structure. A protein’s function is

determined by not only  its 3-dimensional structure assigned by its amino acid sequence, but also by its interactions with ligands, or binding partners. Protein-protein interactions occur at specific binding sites, known as the interface.

Proteins have a wide range of biological functions. These functions are determined by the specific binding interactions between proteins. Proteins interact with one another at specific sites, known as the protein interface. The unique shape, chemical and electrical