**Preliminary Analysis of Epitope Identification**

We have identified the epitopes of ~274 antigens using four predictors – ISPRED, DOCKPRED, VORFFIP, Meta-PPISP. The global F-scores, MCC, and areas under the ROC curve and PR curve are given in the following table for bound antigens (using the structure of known bound antigens based on the crystal structure of the antigen – antibody complex).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | ISPRED | DOCKPRED | VORFFIP | Meta-PPISP |
| F-score | 0.256 | 0.222 | 0.243 | 0.186 |
| MCC | 0.180 | 0.135 | 0.165 | 0.099 |
| Area-ROC | 0.685 | 0.677 | 0.714 | 0.668 |
| Area-PR | 0.159 | 0.154 | 0.161 | 0.135 |

The global F-scores are low compared to the results for other proteins. We show the distribution of F-scores for each of the four methods below:

In order to assess the suitability of the methods to predict epitopes of antigens when no complex structure is experimentally determined, we identified antigen structures (with only one chain) that have a >95% similarity to that of the bound antigen in the complex using ……….. We then mapped the known epitopes (from the complex structure) to the residue numbers in the unbound antigen. We were able to find 189 antigens that had a high similarity score to the bound antigens. The statistical measures for the effective identification of epitopes for these antigens by the four different prediction methods is given in the following table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | ISPRED | DOCKPRED | VORFFIP | Meta-PPISP |
| F-score | 0.194 | 0.18 | 0.173 | 0.154 |
| MCC | 0.126 | 0.104 | 0.103 | 0.078 |
| Area-ROC | 0.657 | 0.669 | 0.662 | 0.653 |
| Area-PR | 0.115 | 0.130 | 0.101 | 0.110 |

We show the distribution of F-scores for each of these methods below:

We hypothesized that we maybe identifying the region of the epitopes by these methods even if we do not identify the exact epitope residues identified from the X-ray crystal structure of the complex. To test this hypothesis, we used a clustering method.

We first needed to identify the optimum number of clusters to be used in our analysis. We clustered the epitopes determined from the experimental structure of the antigen-antibody complex. We used K-clustering with different values of k and used both Elbow method and Silhouette Method to identify the optimal number of clusters. The elbow method looks at the average distance of every point in the cluster to its centroid and chooses the optimal size that leads to a steep decrease in this distance. The Silhouette method on the other hand computes a parameter for each point, i, based on its average distance to every point in its, ai, and the average distance of this point to every other point in all other clusters, bi. It then computes a Silhouette parameter, Si, for each point and then an average value for all points.

Si =

The Silhouette parameter, Si, ranges from -1 to +1. The optimal cluster size can be determined from the cluster size that yields a maximum value for Si.

We show the distribution of proteins as a function of the optimal number of clusters determined by both the Elbow and Silhouette methods below:

We used the centroid of all the atomic coordinates of each epitope residue in the clustering method. The Elbow method predicts too large an optimal cluster size to be physically meaningful for our purposes. The Silhouette method predicts an optimal cluster size of 2 and hence we decided to group the epitope residues into two clusters.

We used hierarchical clustering with number of clusters = 2 to determine how the known epitopes and predicted epitopes cluster together. (**Question: Should we really have used K-clustering with k =2?)**. We combined the centroids of all the annotated epitope residues and the predicted epitope residues by each predictor method for the entire antigen set. We then compared the % annotated and % predicted that falls within each cluster.

We analyzed the data as follows:

Number of epitopes predicted by experiment = annotated residues = Nannoted

% Annotated residue in clusteri = X 100

Similarly, if the Number of predicted epitopes = Npredicted

% predicted epitopes in clusteri = X 100

To test our hypothesis, we first looked at cases where >90% annotated epitope residues fall within a single cluster. Depending upon the prediction method used, this subset of antigens ranges from 95 to 140 out of the total 270 antigens. For this set of antigens, we looked at the performance of the different prediction methods. As seen from the graph below, the % antigens that have > 50% predicted epitopes within this cluster range from 52% for meta-PPISP to 76% for DOCKPRED to 81% for VORFFIP and 91% for ISPRED.

The two clusters are very far apart in about 70% of the antigens indicating that the predicted epitopes are likely to represent two distinct binding sites for the antibodies. The prediction methods are able to identify two distinct patches as represented by the two clusters. This explains why the F-scores are low as these are based on the experimental identification of a single binding site in the antigen-antibody complex.

In the cases where the annotated residues are distributed between the two clusters, we find the following distribution of epitopes between the clusters.

We look at the distance distribution between the centroids of the two clusters in these cases and it is observed that for 74% to 93% of the antigens in this set, the distance between the clusters is < 25 A. This is a very different distribution of the distance between clusters compared to when >90% of the annotated residues are found in a single cluster. It is likely that in these cases the epitopes form a large single patch rather than two isolated patches.