Weekly Report(Up until 3rd September, 2015)

Anirudh Tiwari

Work Done

Up until last time, my output of k-means had an overlap of around 80%(on an average) with CATH. I made further corrections in my functions which stitched the scattered fragments to their correct domain and also some other bugs were resolved.

Results

1. The average overlap increased from 80 to 90%. This was tested on a sample of 300 2 domain proteins.
2. Earlier, it was observed that for fragments of size <=5, the overlap was maximum. After making the above mentioned corrections, I found out that the maximum overlap(90%) was for fragments of size <=15. Although this has been only tested on 300 proteins and has to be tested on the entire data set to arrive at a fixed number.

Next Steps

1. Try to improve the overlap score as much as possible by tweaking the k-means. Scikit learn provides ways to alter the number of iterations, minimum distances between between two points to be converged etc.
2. Get the right fragment length and the right combination of k-means attributes by running it on the entire database of 2 domain proteins.