Weekly Report(Up until 13thth August, 2015)

Anirudh Tiwari

Work Done

Analyzed the results of k-means on two domain proteins and compared it with the CATH database.

Results

1. Most of the contiguous proteins matched exactly with the CATH database, while some of them showed an error of 5-10 residues at the boundary points.
2. Similarly, the non-contiguous showed some error with 10-15 residues being present in the wrong domain(at the boundary).
3. Both, the contiguous as well as the non-contiguous proteins had some scattered patches(non-boundary) of residues(5-20) in the wrong domain.
4. Also, there were some proteins which didn't match with CATH output at all.

Next Steps

1. Improve the k-means algorithm by making slight adjustments to absorb scattered patches of residues into their corresponding domains, also figure out if something can be done to ensure that the boundary points are identified correctly.
2. Once the validity of k-means is established, figure out a way to find the k to be given as an input by establishing some correlation with energy, length and radius of gyration of a protein.