Weekly Report(Up until 10th September, 2015)

Anirudh Tiwari

Work Done

Up until last time, I was looking at various means to increase the overlap of k-means. Meanwhile, I did a comparative analysis of K-means with various other clustering algorithms on the Jones et al. data set. Detailed analysis file is attached with the report.

Results

Clustering Methods	**K-Means**	**Affinity Propagation**	**Mean Shift**	**Hierarchical Ward**	**Complete Agglomerative**	**Average Agglomerative**	**DBSCAN**	**Birch**
Average Overlap	97.90%	12.37%	87.83 %	97.25%	96.10%	96.25%	86.62%	97.25%

This is the average overlap table of various clustering methods on the Jones et al. Data set. The following should be noted on the obtained results:

1. K-Means, Hierarchical Ward, Complete Agglomerative, Average Agglomerative & Birch takes the number of domains as an input. Hence, they have clearly performed well in the comparison.

2. Majority of the proteins in the data set were single domain, thus all those clustering methods which took the number of clusters as an input had a clear advantage over other methods.

3. Out of all the clustering approaches, K-Means performed the best even while clustering multi-domain proteins.

4. Mean Shift and DBSCAN were the two standout clustering methods which gave good results even without giving number of clusters as an input. But these overlap scores only suggest that they can fairly accurately predict the number of domains as one in single domain proteins. Otherwise they performed pretty bad on multi-domain proteins with their average overlap lying in the range of 15-50% only.

Next Steps

1. Report the analysis of change in interaction energy with splitting of domains.
2. Report the analysis of variation in Radius of gyration and interaction energy with varying length of proteins(250 & 300).
3. Read about graph based and machine learning approach for domain prediction.
4. Incorporate information of secondary structures in your clustering methods.