Weekly Report(Up until 8th October, 2015)

Anirudh Tiwari

Work Done

I constructed a data set of contiguous and non-contiguous proteins to perform a comparison of various clustering techniques. Up until last time I only compared them on the Jones Dataset which had a lot of single domain proteins. Thus, I created my own data set which had a lot of multi-domain proteins.

Also, I did another analysis of comparing interaction energies when a single domain protein was split into two and when a two domain protein was split into two.

Results

- Comparing various clustering techniques

Contiguous Multi-Domain Proteins

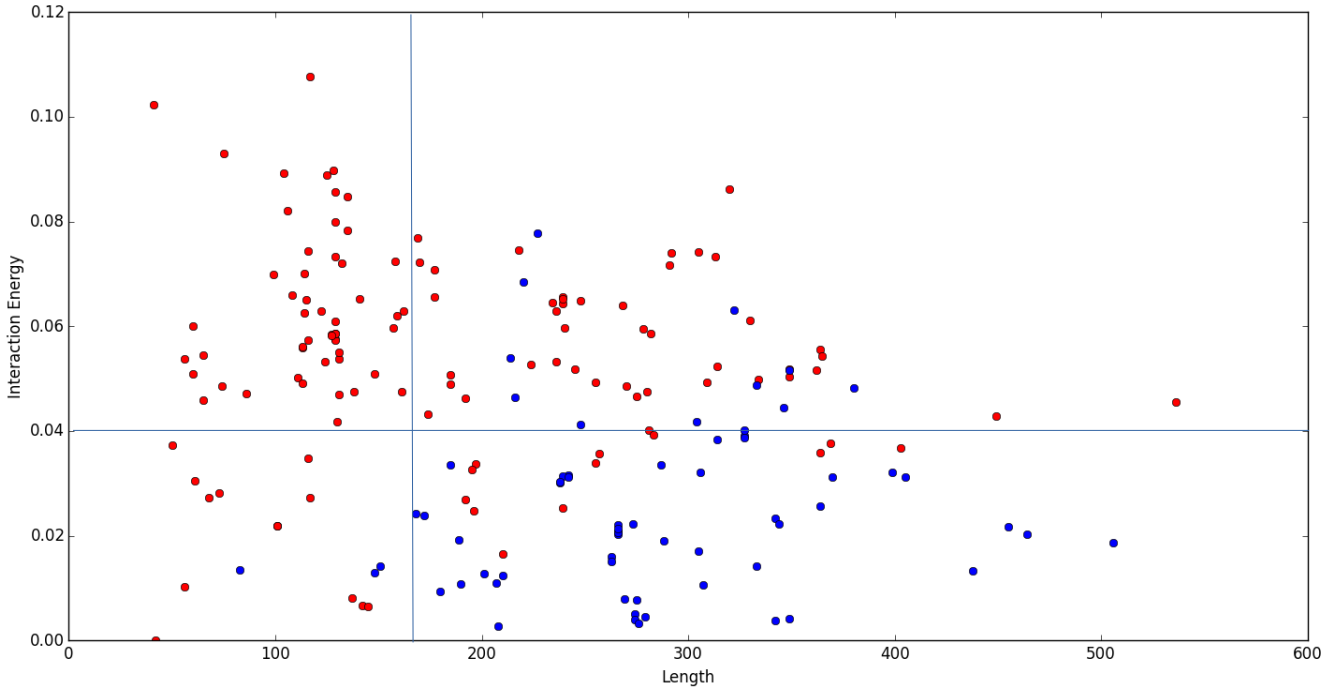
Clustering Technique	K-Means	Birch	Mean Shift	Agglomerative	DBSCAN
Average Overlap	82.01%	76.38%	47.12%	79.36%	37.04%

Non-Contiguous Multi-Domain Proteins

Clustering Technique	K-Means	Birch	Mean Shift	Agglomerative	DBSCAN
Average Overlap	81.47%	73.97%	39.53%	73.97%	36.92%

As can be seen from the above results, K-Means stands out as the best clustering technique for our problem. And this overlap of 81-82% is increased to about 87-88% after putting mismatched fragments into their correct domain.

- Comparing interaction energies when a single domain protein is split into two(red dots) and when a two domain protein is split into two(blue dots).



It can be seen that, generally the red dots(signifying the splitting of a single domain protein into two) are above the blue dots(splitting a two domain protein into two). Thus, a horizontal line can be drawn from y=0.04 and a vertical line can be drawn somewhere in between 180-200, signifying that the enclosed area is of two domain proteins(x>=180 and y<=0.04) and the rest is of one domain proteins, with some false positives and false negatives.

Next Step

- Use the information of radius of gyration also to reduce the number of false positives and false negatives as much as possible in classifying two domain proteins from single domain ones.
- Start reading about other graph spectral approaches employed in domain identification.