Semester Report

Anirudh Tiwari, 201164104

Problem Statement

To develop an automated tool for identification of domains(number as well as their boundaries) in protein contact networks.

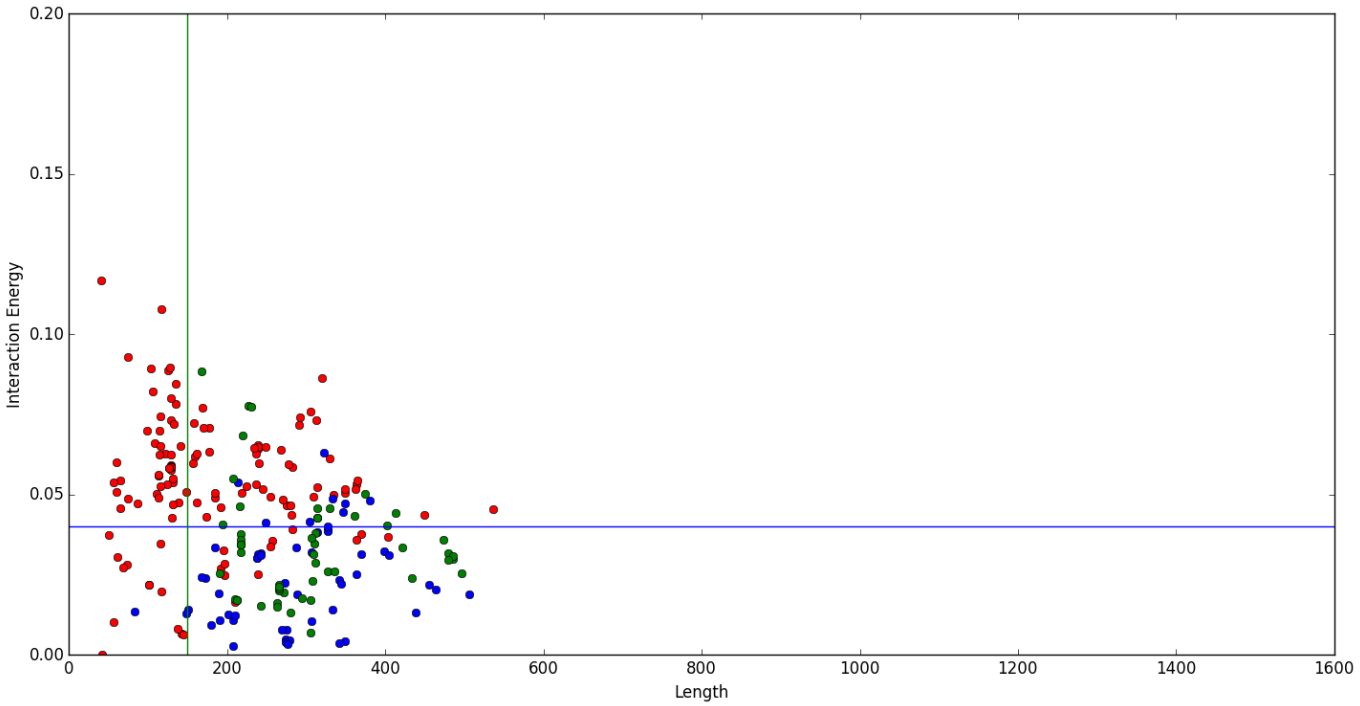
Work Done

1. To establish the importance of length, radius of gyration and interaction energy in identifying the number of domains in a given protein. I plotted graphs(length, radius of gyration and interaction energy) for contiguous and non-contiguous two domain proteins first separately then together, comparing them with single domain proteins. I repeated the process for comparing two domain proteins with three domain proteins and so on and so forth.

Single Domain vs Two Domain Proteins

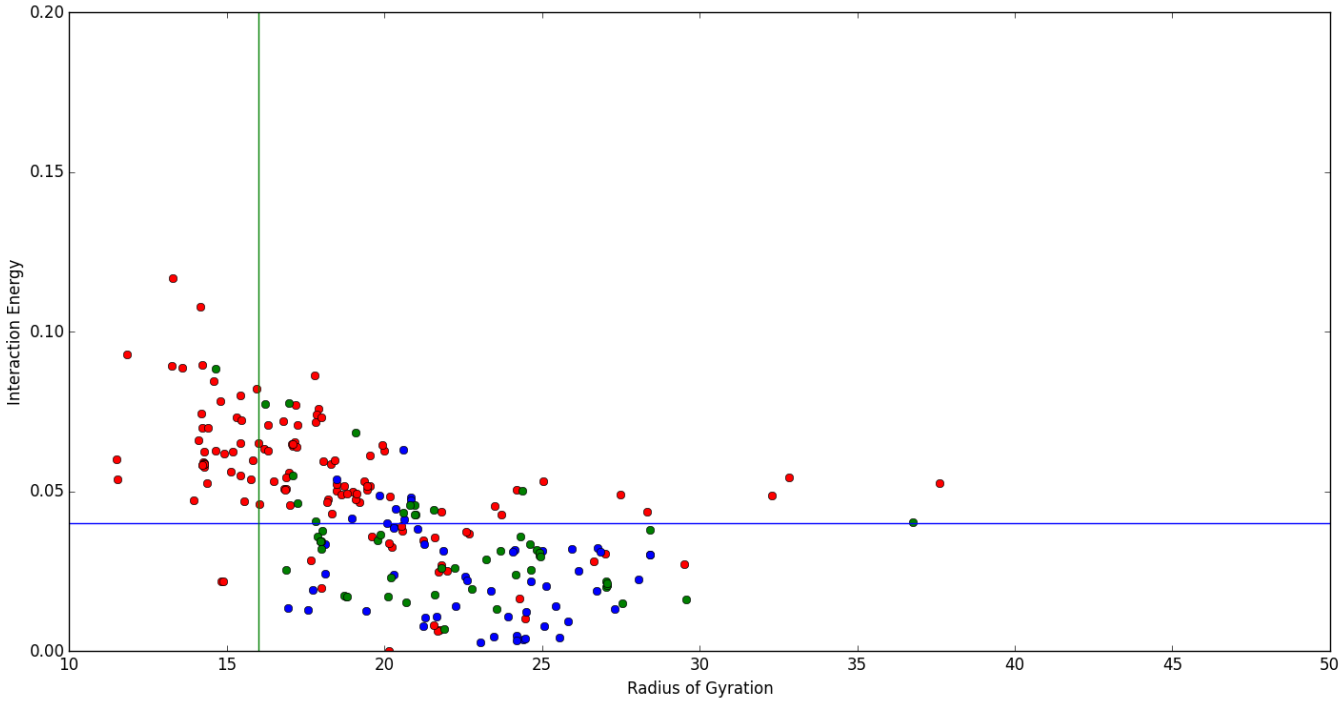
Length vs Interaction Energy

(red dots→ single domain, blue dots→two domain contiguous, green dots→ two domain non-contiguous)



Radius of Gyration vs Interaction Energy

(red dots→ single domain, blue dots→two domain contiguous, green dots→ two domain non-contiguous)

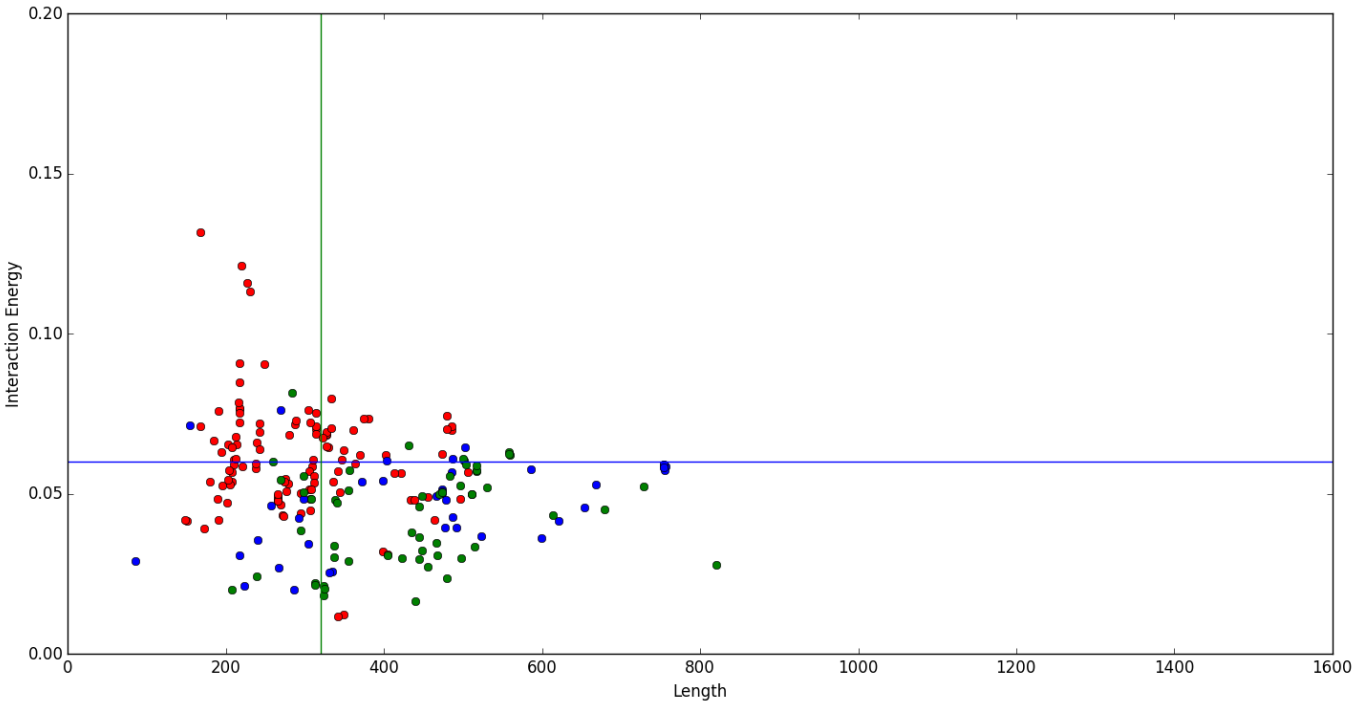


	Length	Interaction Energy	Radius of Gyration	True Positives	False Positives
Single vs Two	>=150	<=0.04	>=16.0	82/108	12/120

Two Domain vs Three Domain Proteins

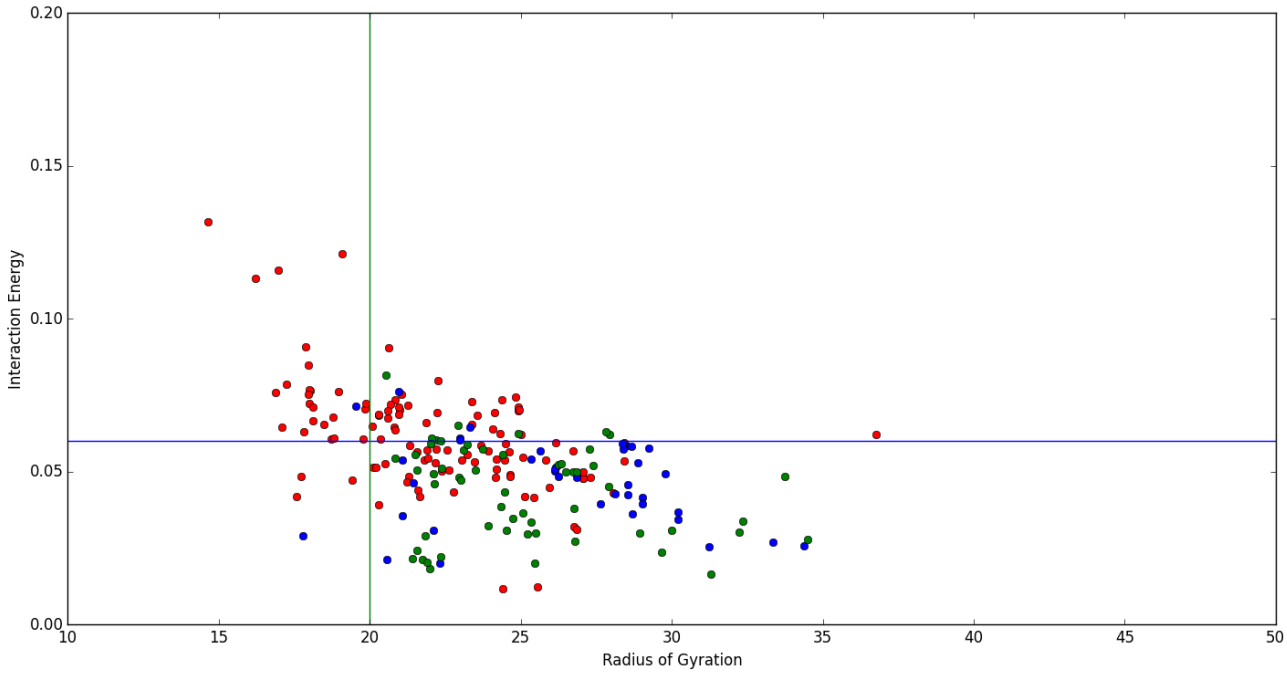
Length vs Interaction Energy

(red dots→ two domain, blue dots→three domain contiguous, green dots→ three domain non-contiguous)



Radius of Gyration vs Interaction Energy

(red dots→ two domain, blue dots→three domain contiguous, green dots→ three domain non-contiguous)

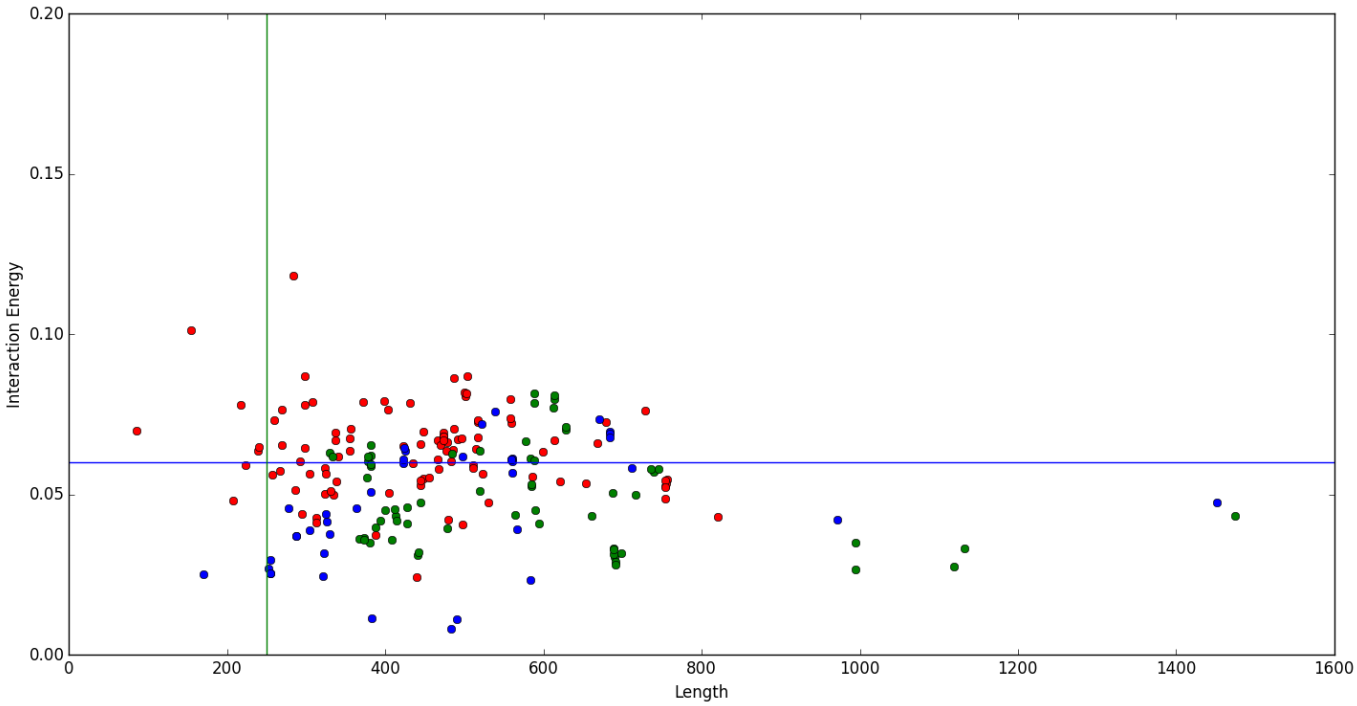


	Length	Interaction Energy	Radius of Gyration	True Positives	False Positives
Two vs Three	>=320	<=0.06	>=20.0	65/97	16/112

Three Domain vs Four Domain Proteins

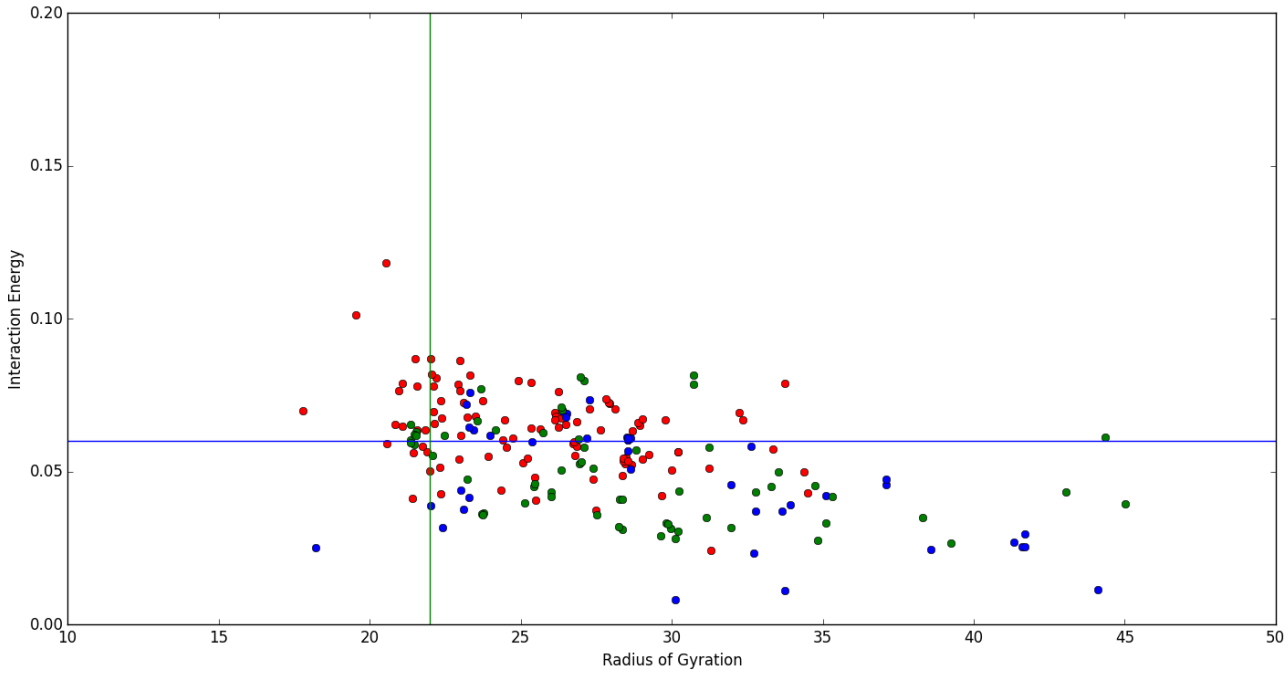
Length vs Interaction Energy

(red dots→ three domain, blue dots→four domain contiguous, green dots→ four domain non-contiguous)



Radius of Gyration vs Interaction Energy

(red dots→ three domain, blue dots→four domain contiguous, green dots→ four domain non-contiguous)

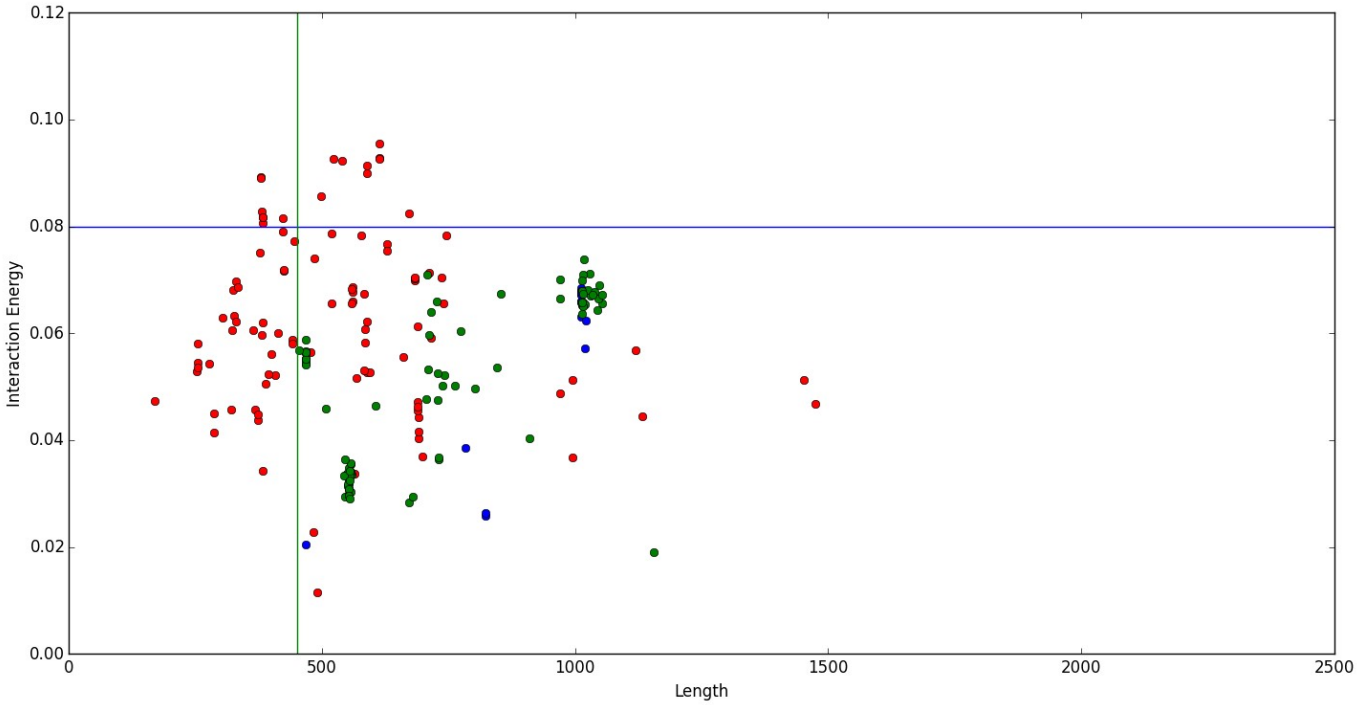


	Length	Interaction Energy	Radius of Gyration	True Positives	False Positives
Three vs Four	>=250	<=0.06	>=22.0	67/102	33/98

Four Domain vs Five Domain Proteins

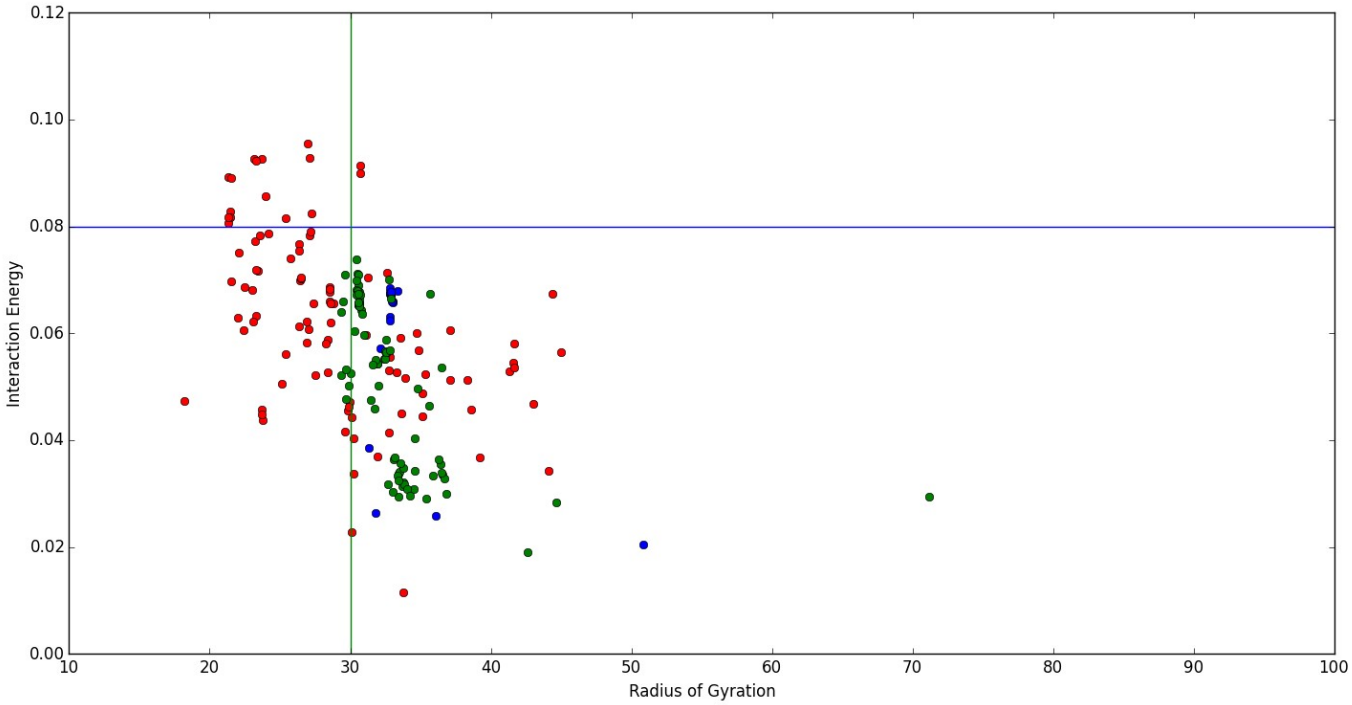
Length vs Interaction Energy

(red dots→ four domain, blue dots→five domain contiguous, green dots→ five domain non-contiguous)



Radius Of Gyration vs Interaction Energy

(red dots→ four domain, blue dots→five domain contiguous, green dots→ five domain non-contiguous)



	Length	Interaction Energy	Radius of Gyration	True Positives	False Positives
Four vs Five	>=450	<=0.08	>=30.0	93/100	22/98

Summary Of Results

	Length	Interaction Energy	Radius of Gyration	True Positives	False Positives
Single vs Two	>=150	<=0.04	>=16.0	82/108	12/120
Two vs Three	>=320	<=0.06	>=20.0	65/97	16/112
Three vs Four	>=250	<=0.06	>=22.0	67/102	33/98
Four vs Five	>=450	<=0.08	>=30.0	93/100	22/98

As it can be observed from the above table, the analysis gives a high number of false positives in comparing three vs four domain proteins, otherwise it is working fine.

2. Once the number of domains in a protein were identified, I compared various clustering techniques which could be used to identify the domain boundaries. The below table gives the summary of results

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 it can be seen from the tables above, K-Means stand out as the most accurate clustering algorithm.

Next Step

Now with a way to find the number of domains and using k-means to find the domain boundaries, the next task is to test these on some benchmark data sets and the entire pdb data to validate the correctness of our method. One of the benchmark data set to be constructed is as per the one suggested by Holland et al[1] and I am currently working to reconstruct it.

References

[1] Partitioning Protein Structures into Domains: Why Is it so Difficult? Timothy A. Holland, Stella Veretnik, Ilya N. Shindyalov and Philip E. Bourne.