Domain Identification in Proteins

- **Problem Statement:** To develop an automated method to identify all the domains in a protein.
- Advancements up until now.
- Thus, there are two sub problems which are to be solved:
  - Identify the number of domains.
  - Identify the domain boundaries.

Identifying Domain Boundaries

- Compared various clustering algorithms.

Contiguous Multi-Domain Proteins

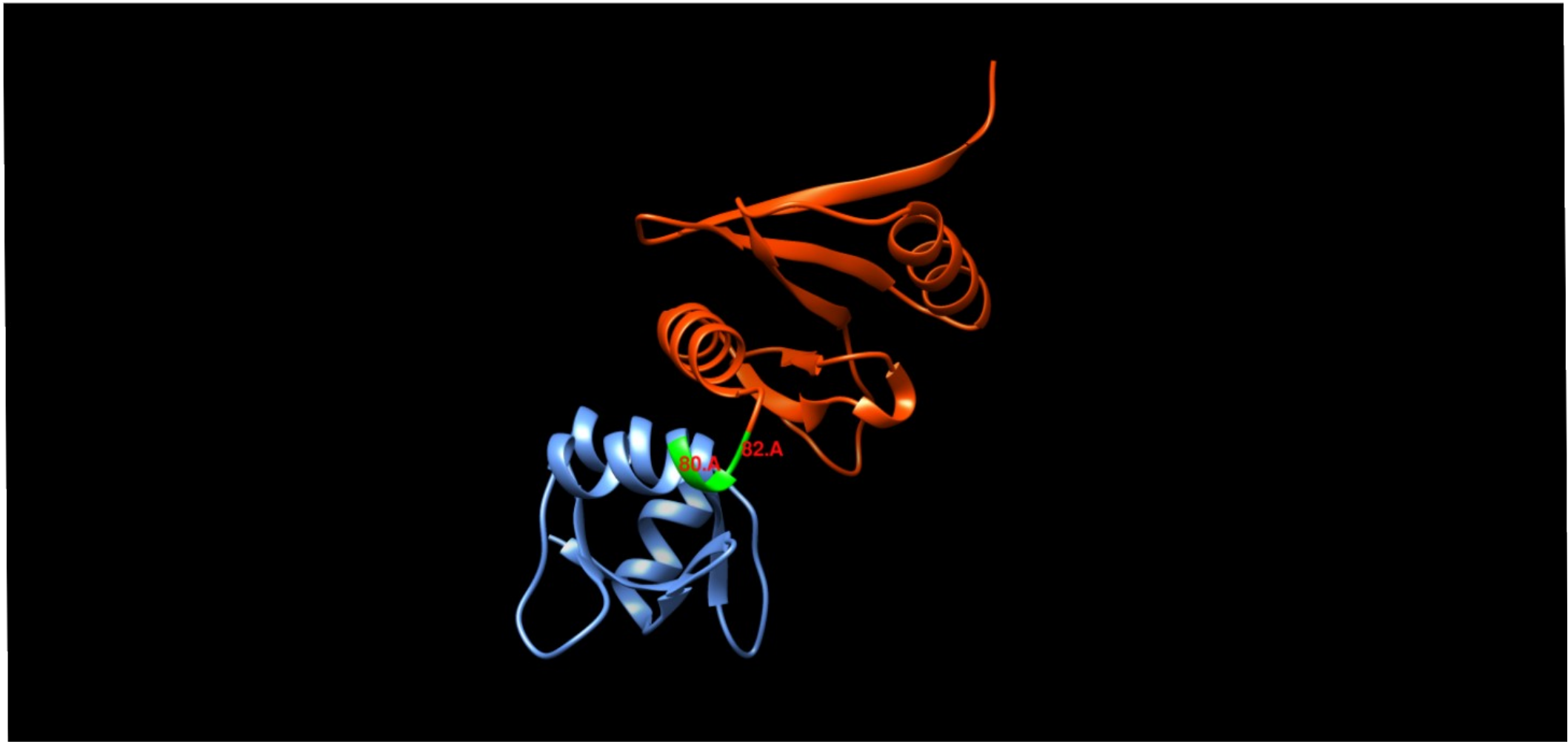
Clustering Technique	K-Means	Birch	Mean Shift	Agglomerative	DBSCAN
Average Overlap	82.01%	76.38%	47.12%	79.36%	37.04%

Non-Contiguous Multi-Domain Proteins

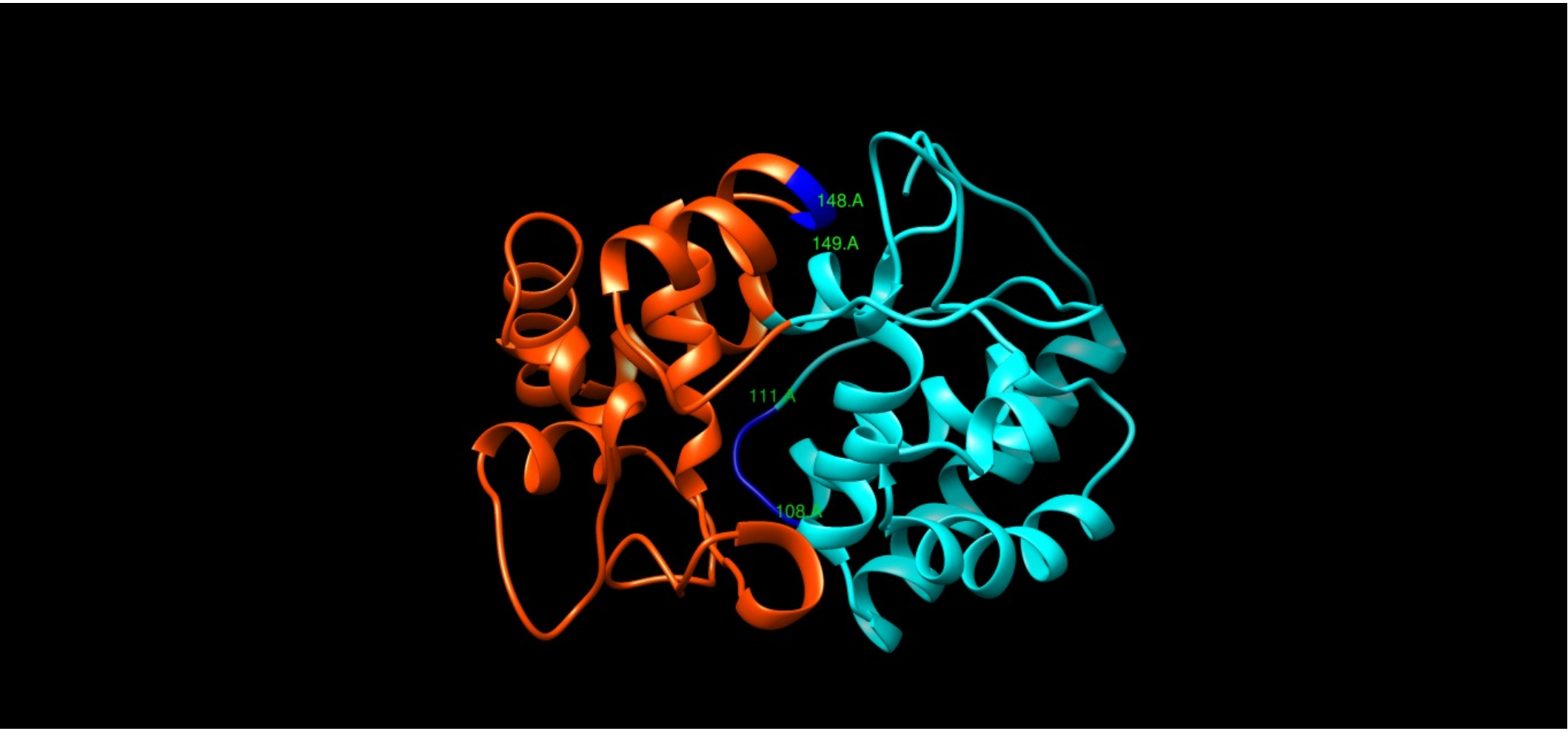
Clustering Technique	K-Means	Birch	Mean Shift	Agglomerative	DBSCAN
Average Overlap	81.47%	73.97%	39.53%	73.97%	36.92%

Problems with K-Means

1. Incorrect Boundary prediction



2. Mismatched residues



3. It requires K(the number of clusters) as an input

Finding The Number Of Domains

Tried to establish a correlation of the number of domains with the following attributes of a protein:

- Length
- Radius of Gyration
- Interaction Energy