Order in Disorder: Clustering to identify conformational ensembles in IDP using Node2Vec embedding & GMMs.

By- Aryan Laroia

Under Professor Jitin Singla, CSE Department, IITR

May 2024

Abstract

Intrinsically disordered proteins (IDPs) are crucial biomolecules with diverse functions and connections to various diseases. Their intrinsic flexibility and dynamic behaviour challenging traditional protein structure-function relationships. Advanced computational methods are necessary to understand their dynamic behaviour and unravel their roles in biology. Our focus is on employing advanced, unconventional computational approaches to tackle this challenge. Specifically, we target the Aβ42 peptide, central to Alzheimer's disease pathogenesis, whose intrinsic disorder complicates traditional structure-function analyses. By integrating molecular dynamics simulations and machine learning techniques, we aim to map the extensive conformational space of IDPs, identifying crucial ensembles for functional understanding and potential therapeutic interventions. A key breakthrough lies in reducing vast conformational datasets to manageable representative structures, enabling efficient analysis and enhancing our grasp of disorder-function relationships. These insights pave the way for innovative drug design strategies targeting IDPs, promising advancements in therapeutic interventions.

1 Introduction

Intrinsic disorder in proteins signifies the absence of a stable three-dimensional structure under physiological conditions, contrasting with traditional structured proteins. Intrinsically disordered proteins (IDPs) exhibit significant structural plasticity, transiently adopting secondary structure elements without dominance. IDP regions feature low sequence complexity and a high proportion of disorder-promoting amino acids like proline, glycine, and charged residues. These flexible regions enable IDPs to adopt multiple conformations and interact with diverse binding partners.

Intrinsically disordered proteins (IDPs) are integral to diverse cellular processes, acting as molecular switches, participating in signalling pathways, and facilitating protein-protein interactions. Their association with diseases like cancer, viral infections, and neurodegenerative disorders highlights their multifaceted roles. In neurodegenerative diseases, IDPs like tau and alpha-synuclein aggregate into amyloid fibrils, contributing to disease pathology. Despite challenges in clinical management, understanding IDPs' dynamic nature is crucial, as they lack stable structures and contribute to various cellular functions and disease mechanisms. Conventional structure-based drug discovery (SBDD) relies on stable binding sites, presenting challenges with intrinsically disordered proteins (IDPs) due to their flexible conformational states. IDPs' transient binding sites lack well-defined geometric characteristics, hampering small-molecule

interactions and leading to a scarcity of effective therapeutic agents against them. Traditional SBDD methods struggle to accommodate the

multiplicity and plasticity of IDPs, highlighting the need for innovative approaches in drug discovery.

Recent advances in Molecular Dynamics (MD) simulations have revolutionized the study of Intrinsically Disordered Proteins (IDPs), offering detailed insights into their dynamic conformational landscapes. Enhanced computational power and IDP-specific force fields enable the generation of realistic simulations, producing extensive datasets of potential 3D conformations. However, processing and interpreting this vast data present challenges in identifying function- ally relevant conformational ensembles. To tackle this, the integration of advanced Artificial Intelligence/Machine Learning (AI/ML) algorithms is performed. AI/ML techniques, particularly clustering algorithms, offer powerful tools for analysing complex datasets, classifying protein conformations, and identifying structural similarities. This innovative approach not only enhances our understanding of IDP conformational diversity but also identifies potential therapeutic targets to modulate their pathological roles.

2 Methodology

2.1 Clustering and its challenges

Clustering is a vital technique in data analysis that seeks to arrange a set of items so that those within the same group, or cluster, share greater similarities with each other than with those in other clusters. Clustering analysis of intrinsically disordered proteins (IDPs) poses significant methodological challenges due to their dynamic conformational behaviour. Unlike structured proteins, IDPs exhibit a wide range of fluctuating conformations, rendering traditional clustering algorithms less effective as they typically rely on the presence of stable structural features. Moreover, the conformational space of IDPs is inherently high-dimensional, owing to the myriad degrees of freedom within their structure. This complexity often leads to sparse data distributions, exacerbating the ”curse of dimensionality” problem and diminishing the efficacy of conventional clustering techniques. Compounding these challenges is the lack of clear boundaries between conformational clusters in IDPs, as their inherent flexibility results in overlapping or continuous distributions of structural states. Consequently, the application of standard clustering algorithms becomes challenging, as they typically assume distinct separations between clusters. Addressing these methodological hurdles requires the development of novel clustering approaches tailored to accommodate the dynamic and high-dimensional nature of IDP data, thereby facilitating more accurate categorization and analysis of these biologically important molecules.

2.2 Generating Contact Maps from MD Trajectories:

In the analysis of Molecular Dynamics (MD) simulations of IDPs, Python libraries including NumPy, SciPy, and Matplotlib were utilized for numerical operations, scientific computations, and visualization of contact maps, respectively. The

methodology involves selecting Cα atoms to represent amino acid residue position. Distance matrices are then calculated to depict pair- wise distances between Cα atoms in each trajectory frame. These matrices undergo transformation to emphasize closer interactions, followed by contact map computation using a defined threshold (e.g., 10 angstroms) to identify significant contacts. Modifications, such as adjusting diagonal elements to prevent division by zero and normalization, are applied to enhance close contacts. Finally, the contact maps are visualized using colormap representations, where colour intensity reflects atom proximity. Normalization of contact maps is also crucial to reduce bias due to varied numbers of contacts per residue, enhancing the interpretation of interaction strengths and frequencies across the protein.

2.3 Node2Vec Embedding:

In the context of analysing contact maps of intrinsically disordered proteins (IDPs), an essential step involves leveraging techniques like Node2Vec for effective data representation and analysis. Node2vec is an algorithm used for generating embeddings for nodes in a graph. It is based on the skip-gram model, which is commonly used in word embeddings like Word2Vec. In skip-gram, the objective is to predict the context words (neighbours) given a target word. Similarly, in node2vec, the algorithm aims to learn embeddings for nodes in such a way that nodes with similar neighbourhood structures have similar embeddings.Node2vec achieves this by performing random walks on the graph. These random walks are biased to explore both local neighbourhoods (breadth-first) and global structures (depth-first). By sampling these walks, node2vec captures both local and global graph structures, allowing it to generate meaningful embeddings for nodes. The project's pipeline involved embedding graphs (or contact maps) into finite-dimensional vectors, followed by dimensional reduction and clustering. To accomplish this, node2vec embeddings were utilized to individually embed all nodes into finite-dimensional vectors. To aggregate these individual embeddings, the mean and standard deviation of the individual embeddings were concatenated, resulting in the aggregate graph embedding. The Node2Vec code was implemented from scratch.

2.4 K-PCA Dimensional Reduction:

The aggregated embeddings obtained from Node2Vec were characterized by high dimensions, presenting a challenge for clustering the graphs effectively. Consequently, Kernel Principal Component Analysis (KPCA) was implemented to address this issue. KPCA serves as an extension of traditional Principal Component Analysis (PCA), facilitating nonlinear dimensionality reduction by implicitly mapping data into a higher-dimensional space through a kernel function prior to PCA application. This approach enables KPCA to discern and capture intricate patterns and structures within the data, which may elude efficient capture by linear techniques such as PCA.

2.5 Clustering : Gaussian Mixture Model:

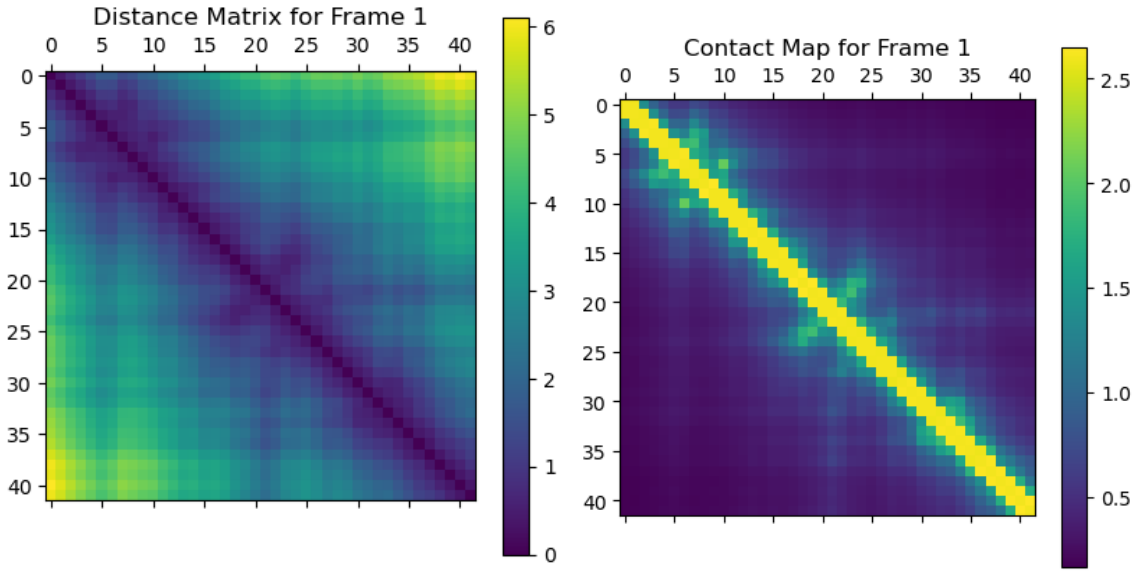
Clustering methods like k-means and t-SNE are commonly used for data analysis, but when dealing with Intrinsically Disordered Protein (IDP) structures, a unique challenge arises. The continuous nature of conformations in IDPs makes it difficult to distinctly classify them into single clusters. To address this challenge, we opted for Gaussian

Mixture Models (GMM) due to their ability to accommodate overlapping clusters using a probabilistic approach. In GMM, the assumption is that data points are generated from a mixture of several Gaussian distributions, each characterized by its mean and covariance. The primary objective of GMM is to uncover these underlying Gaussian distributions that best describe the data. This is achieved through iterative adjustments of parameters such as means, covariances, and mixture coefficients, aiming to maximize the likelihood of observing the data. The code of GMM was written from scratch.

3 Results and Conclusion

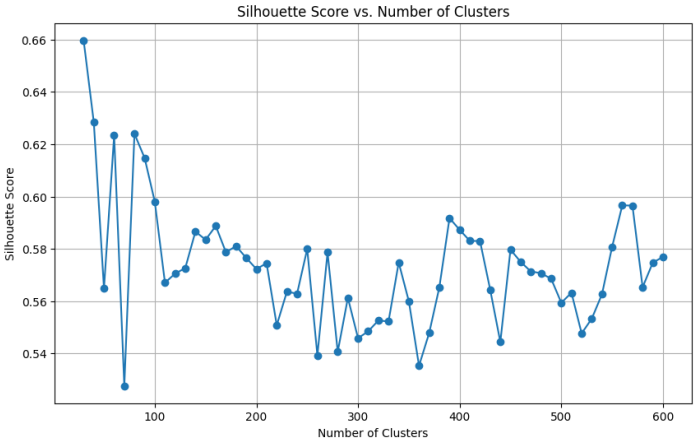
3.1 Contact Maps Analysis

The contact maps generated from the simulation offer a detailed view of the interactions between individual residues within the protein. Each contact map consist of 42X42 matrix, as Aβ42 has 42 Cα atoms. Distance Map is a representation of spacial distance between atoms and contact map is a binary matrix based on a particular threshold (e.g., 10 angstroms) value to consider it as a contact or not. By analysing each frame of the simulation, we gain insight into how the protein's shape and structure change dynamically over time. This approach helps in understanding the protein's conformational landscape and how it evolves throughout the simulation.



3.2 GMM Clustering

As a result of the above steps, in total 80 clusters were obtained, with silhouette score 0.623, and DBI(davies bouldin score) score 0.622 which indicated excellent clustering results.


Silhouette Score vs. Number of Clusters
Silhouette Score
Number of Clusters
0.66
0.64
0.62
0.60
0.58
0.56
0.54
100
200
300
400
500
600