Poster

Advancing protein conformation analysis: a comprehensive clustering pipeline for the Protein Data Bank

Joseph I. J. Ellaway[1], PDBe-KB consortium[1]

[1]*Protein Data Bank in Europe, European Bioinformatics Institute, Hinxton, UK*

jellaway@ebi.ac.uk

Proteins, as highly dynamic macromolecules, adopt specific conformations to perform biological functions. Understanding these conformations is vital for elucidating disease mechanisms. Factors influencing protein structural states include experimental techniques, ligand presence, oligomeric state, and experimental conditions. AI-based computational methods, like AlphaFold [1] and ESMFold [2], have enabled the exploration of protein conformation space.

The Protein Data Bank (PDB) [3] offers a sampling of protein conformational space; however, inconsistent annotations limit its usability. To address this, the PDB in Europe (PDBe) [4] team developed a standardised clustering pipeline to capture structural variability and enable the accurate identification of distinct conformational states. This open-source pipeline releases updated clustering data weekly for the entire PDB.

Our method calculates pairwise C-alpha distances between peptides to generate a global conformation difference score (GLOCON score), which is then used in the UPGMA agglomerative clustering algorithm. Representative chains are selected for each cluster based on model quality. Users can structurally superpose predicted models from the AlphaFold Protein Structure Database [5] onto PDB structural clusters, allowing identification of the model's conformational state.

At the PDBe, we are now working on the analysis of these conformational states across the PDB archive. Although models captured in the PDB are static representations of dynamic molecules, these structures often depict biologically important states that (when appropriately annotated) could act as way-points for the validation of molecular dynamics simulations.

In conclusion, this clustering pipeline advances protein conformation analysis within the PDB archive by providing a standardised, comprehensive approach. Access to up-to-date structure clusters will aid in understanding the relationship between sequence, structure, and dynamics and improve researchers' ability to study molecular mechanisms in health and disease.

[1] Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021).

[2] Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379, 1123–1130 (2023).

[3] wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res. 47, D520–D528 (2019).

[4] Varadi, M. et al. PDBe and PDBe-KB: Providing high-quality, up-to-date and integrated resources of macromolecular structures to support basic and applied research and education. Protein Sci. 31, e4439 (2022).

[5] Varadi, M. et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 50, D439–D444 (2022).