IN-791 Report

Computational Investigation of Protein Structures

Supervisor: Dr. Ashutosh Srivastava

Sayantoni Chaudhuri

20310060

Biological Engineering, IIT Gandhinagar

May 08, 2021

Abstract

3D representative structures of proteins give insights into mechanisms of diseases and allow the design of diagnostic agents. So, a computational understanding of protein structures along with experimental techniques is advantageous. Several methods popularly used to determine the structure of a protein are: X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy. The type of data generated in each case is different, and these data, along with known sequence, bond geometry preferences, etc, models are created from scratch. The Protein Data Bank (PDB) is a repository of solved structures of biomolecules using these experimental techniques. It allows for searching and exploring deposited structures under a unique PDB entry; and the information can be downloaded as the PDB file made available in 3 different file formats- the PDB, the mmCIF, and the PDBML fileformats. Typically, the PDB file integrates the following information : (i) HEADER, TITLE, AUTHOR records (ii) REMARK records, (iii) SEQRES records (iv) ATOM records and (v) HETATM records. Often, the most used information in a structure file is the coordinates of each ATOM record in each structure that describe its location in 3D space, an associated occupancy, and a temperature factor. These structure files can be viewed by a simple text editor and visualized with the help of programs like UCSF Chimera that reads and displays interactive 3D representations of the structures. To analyze these structure files, various open-source platforms are available. A widely-used platform is Biopython, a Python-based package that provides a set of libraries or "tools" for computational biology and bioinformatics. Of the many functionalities available, it offers easy parsing options of bioinformatics files into local data structures. To "parse" PDB files, Biopython provides users with a PDB parser. Biomolecules follow a building hierarchy, from atoms, residues, chains, and so on. Similarly, structure files can be navigated and the information contained in them can be extracted following a Structure-Model-Chain-Residue-Atom (SMCRA) hierarchy. In this project, various tools of Biopython have been used to analyze a set of PDB structures using the following workflow: (i) Downloading 20 structures of the human Casein Kinase 2 Alpha subunit (CK2α) from the PDB, solved at different resolutions and differ among each other with respect to bound ligands. (ii) parsing their structure files using Biopython [1], (iii) Perform a multiple sequence alignment of these structures, of ~340 residues and extract regions 30-300 for further analysis.

The Root Mean Square Deviation (RMSD) is a measure of the average distance between ATOMS, usually the CA-backbones. It can be used as a measure of similarity between structures. To calculate RMSD, we perform a geometric translation followed by a rotation of one structure over the target structure. Given a set of n points, the RMSD between two protein structure backbones were calculated using the following formula thereby obtaining a good, minimized RMSD [2]:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_{ci}-x_{di})^2+(y_{ci}-y_{di})^2+(z_{ci}-z_{di})^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{n}{d_x}^2+{d_y}^2+{d_z}^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{n}\vec{d}}$$

where $(x_{ci}\ ,\ y_{ci}\ ,\ z_{ci})$ are the coordinates of Reference Structure CAs and $(x_{di}\ ,\ y_{di}\ ,\ z_{di})$ the coordinates of Reference Structure CAs and, are the coordinates of Sample Structure CAs. A stacked line-plot of the RMSD all protein structures showed some regions of variability across all the structures. Using visualization software UCSF Chimera[3], it was observed that most of these variabilities come from loop regions.

The B-factor describes the displacement of the atomic positions from an average value (mean-square displacement). Residue-wise B-factor plots were generated for the 20 structures using Biopython. It was seen that regions with low local residue B-factors corresponded to the stable regions in the protein structure, like helices and sheets, whereas regions with high local B-factors corresponded to flexible regions in the protein structure, like loops. Comparing the B-factor plots of the 20 structures, it was found that at good resolutions, the per residue B factors were low, ie., uncertainty in atomic positions was low. This is reflective of the fact that at poor resolutions, (low empirical electron density) higher temperature factors and hence disorder is observed.

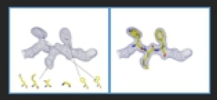
Computational Investigation of Protein Structures

Sayantoni Chaudhuri, IN791, Supervised by: Dr. Ashutosh Srivastava, Department of Biological Engineering, IIT-Gandhinagar

Introduction

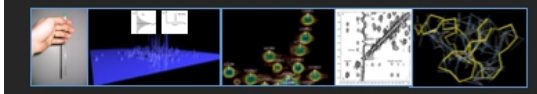
1. X-Ray crystallography (XRC)

- Obtain crystals: "Protein Crystallization"
- Diffraction experiment; obtain a diffraction pattern
- Data fitting to analyze the electron density distribution
- Build a model of atomic arrangement of the crystal



2. Nuclear Magnetic Resonance (NMR)

- Sample preparation
- Data acquisition and Spectral Processing
- Structural analysis



3. Cryo-Electron Microscopy (Cryo-EM)

- Negative-Stain EM Screening
- Biochemical sample preparation
- Sample vitrification, screening and data acquisition
- 3D reconstruction

Each technology has its own advantages.
Protein Data bank which is a repository of solved structures of biomolecules use these experimental techniques.

4. File Formats:

PDB mmCIF PDBML

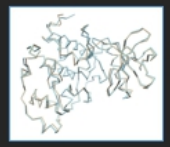
5. Going 3D: Biopython and BioPDB module

Create structure objects from PDB files | Navigate throuh a structure object | Extract information

6. Structural Analysis: Per Residue RMSD

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_{ci}-x_{di})^2 + (y_{ci}-y_{di})^2 + (z_{ci}-z_{di})^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{n} d_x{}^2 + d_y{}^2 + d_z{}^2}$$

Superimposition of backbones | Reference: 3war, Sample: x

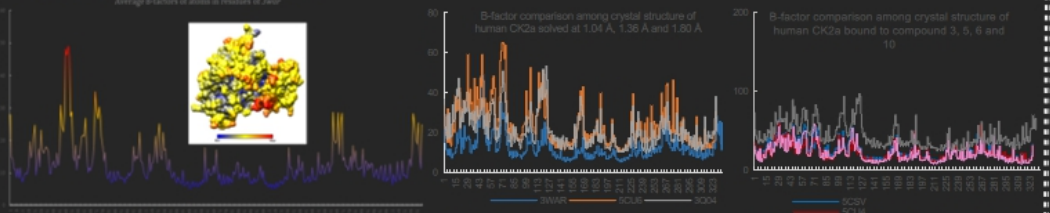


7. Average B-Factor of atoms in each residues

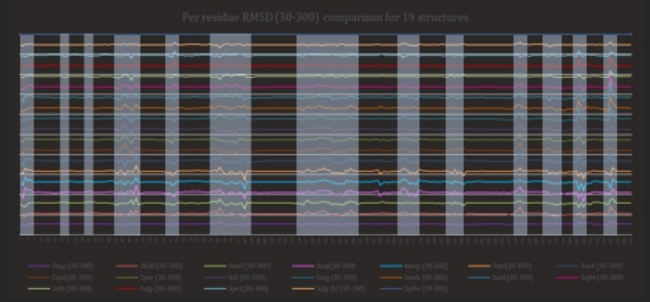
$$\text{B Factor} \propto \frac{1}{\text{Atomic Packing Density}}$$

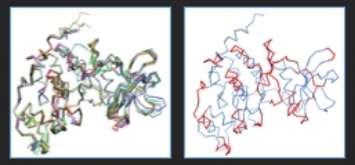
Noises from lattice disorder, crystal packing effects, and type of structure refinementmay lead to a discrepancy between B-factors and the root-mean-squarefluctuations of the atoms (RMSD).

Results & Discussion



From Left: Lineplots of (1) Avg. B-Factors of atoms in residues of 3war (1.04 A) (2) Avg. B-Factor comparison among CK11a solved at diff. resolutions. (3) Avg. B-Factor comparison among CK11a structures bound with different ligands (Compounds 3,5,6,10)





Regions of variability in samples with respect to reference (3war)

Motivation

The three-dimensional structure of proteins and protein complexes provide great insights into the mechanism of diseases, and thereby allowing rational design of novel diagnostic and therapeutic agents.

Objectives

- Protein structure determination: XRD, NMR, Cryo-EM
- The PDB and the PDB file format
- Biopython and the BioPDB module
- Structural analysis of 20 human CK2a proteins

Materials & Methods

PDB Structures used: 3bqc, 3h30, 3mb7, 3nsz, 3pe1, 3war, 4kwp, 5csv, 5cu4, 5cu6, 2pvr, 5clp, 5csp, 5cvg, 3r0t, 3q9w, 3q04, 5cs6, 3owk, 4rll

RCSB PDB PROTEIN DATA BANK
UCSF Chimera
biopython
SPYDER

References

1. Srivastava, Ashutosh, et al. "Conformational dynamics of human protein kinase CK2α and its effect on function and inhibition." Proteins: Structure, Function, and Bioinformatics 86.3 (2018): 344-353.

2. Cock, Peter JA, et al. "Biopython: freely available Python tools for computational molecular biology and bioinformatics." Bioinformatics 25.11 (2009): 1422-1423.

3. Pettersen, Eric F., et al. "UCSF Chimera—a visualization system for exploratory research and analysis." Journal of computational chemistry 25.13 (2004): 1605-1612.

Questions/Suggestions

1. What is 3D protein structure, do a 2D structure exist ?

3D protein structure is the tertiary structure of a protein, defined by it's atomic co-ordinates. These structures can be represented in 2D for visualization and analysis.

References

[1] Cock, Peter JA, et al. "Biopython: freely available Python tools for computational molecular biology and bioinformatics." Bioinformatics 25.11 (2009): 1422-1423.3.

[2] Srivastava, Ashutosh, et al. "Conformational dynamics of human protein kinase CK2α and its effect on function and inhibition." Proteins: Structure, Function, and Bioinformatics 86.3 (2018): 344-353.2.

[3] Pettersen, Eric F., et al. "UCSF Chimera—a visualization system for exploratory research and analysis." Journal of computational chemistry 25.13 (2004): 1605-1612.