Structural Bioinformatics 2020-21

Last document update 1 May 2021

Conformational analysis of protein structural ensembles

Introduction

Intrinsically disordered proteins (IDPs) lack a fixed 3D structure and instead exhibit extreme conformational dynamics in the free state. Similar to the unfolded state, IDPs and **intrinsically disordered regions (IDRs)** must be described as ensembles of heterogeneous, rapidly interconverting conformations. **Conformational ensembles** are representative sets of conformers reflecting on the structural dynamics of IDPs sampling the space. Ensemble modeling usually relies on experimental data. These measurements are then used to define local or nonlocal structural constraints for the computational modeling of the conformational ensemble. Solving structural ensembles, however, is fraught with uncertainties, because the number of degrees of freedom is inherently much larger than the number of experimentally determined structural restraints. We don't yet know how to select the 'best' ensemble from multiple alternatives, neither can we be sure if an actual ensemble is a faithful representation of the real physical state of the IDP/IDR, nor is only a reasonable fit to experiment observations. To help address these issues, solved IDP/IDR ensembles are collected and made available in the dedicated **Protein Ensemble Database** (**PED**, https://proteinensemble.org/) [1].

Comparison of alternative ensembles

Structural comparisons rely on quantitative similarity measures. The most common measure is the **root mean square deviation (RMSD)** of the atomic positions between two structures, which is minimized upon rigid-body superimposition of these structures. But, the RMSD is often not very informative because it averages out differences across regions of the structures with varying similarity levels. Characterizing and comparing IDP ensembles is therefore particularly challenging. First, their extreme conformational heterogeneity makes it **difficult to evaluate the degree of global similarity between two ensembles** by any measure, let alone by RMSD-based metrics. Second, the function of disordered proteins is often mediated by short, **sequentially contiguous binding motifs adopting locally relevant conformations**. The latter are interconnected through more structurally variable linkers that determine the relative overall configuration of these important motifs. Therefore, the similarity of the IDP and IDR ensembles must be evaluated at both the local and global levels in a statistically meaningful approach [2].

Project requirements

1. **Implement a software with two components**. The first component processes a single PED ensemble. The second component compares two PED ensembles of the same protein (same PED entry) based on the output of the first software component.
2. **Write a report** describing main findings. Includes tables and figures as generated by the software. The text should **not exceed 1500 words**.

Project goals

The main goal of the project is to **implement a software** (two components) to identify conformational relationships **1) within a single ensemble** and **2) between different ensembles,** and visualize them.

Task 1

Relationships within an ensemble will be identified considering the structural features of single conformations (see single conformation features). The first software component is required to:

1. Input
    a. A file (or folder) containing the PDB structures (conformations) of **one single PED ensemble.**
2. Output
    a. **Features** files (see Features). To be used as input in the second component.
    b. **A graph** (text and figure) where nodes are a subset of representative conformations and edges represent their similarity (or distance). The similarity is calculated combining all feature values (see below). Representative conformations are found by unsupervised clustering and the number of clusters is identified automatically.
    c. **An Pymol image** including the PDB structures corresponding to the graph nodes. Centered (translate) the selected structures on the position(s) with the lowest feature variance. Residues in each structure will be displayed (color, size, ...) based on their feature variability within the ensemble. Alternative and more effective visualizations are welcomed.

Task 2

Relationships between different ensembles of the same protein will be identified considering ensemble features (see ensemble features) calculated from the output of the first software component (Task 1). The students are asked to identify a measure (**global score**) to quantify global differences between ensembles pairs and another measure (**local score**) to identify low/high variance positions along the sequence.

1. Input
    a. The features files as generated by the first component of two or more ensembles.
2. Output
    a. Feature files (see Features).
    b. A **dendrogram/heatmap** representing the distance (global score) between ensembles.
    c. A plot showing **features values along sequence positions** (local score).

Features

Single conformation features

N = number of residues in one conformation

1. Radius of gyration of the structure. *Single value*
2. Relative accessible surface area (ASA) for each residue. *Vector of of size N*
3. Secondary structure (SS) for each residue, mapped into four classes based on phi/psi angles: α-helix, left-handed helix, β-strand, polyproline I and II. *Vector of size N*
4. Residue distance matrix considering Cα atoms. *Matrix of shape N x N (symmetric)*

Ensembles features (multiple conformations)

N = number of residues in one conformation

M = number of conformations in one ensemble

1. Radius of gyration for each conformation in the ensemble. *Vector of size M*
2. Secondary structure entropy for each position across ensemble conformations. *Vector of size N*
3. Median solvent accessibility for each position across ensemble conformations. *Vector of size N*
4. Median RMSD for each position across ensemble conformations[1]. *Vector of size N*
5. Median distance of each pair of equivalent positions across ensemble conformations. *Matrix of shape N x N (symmetric)*
6. Standard deviation of the distance of each pair of equivalent positions across ensemble conformations. *Matrix of shape N x N (symmetric)*

Software specifications

1. The software has to be written in Python3.
2. The use of the BioPython.PDB, Argparse and Logging modules is strongly encouraged
3. The software has to be documented by providing the algorithm description, user guide and usage. The documentation has to be provided in a README-like file using Markdown notation (like in GitHub)

Project evaluation

The algorithm will be blindly tested against other ensembles (not available to the students group) and compared with the analysis provided by experts in the field. Clarity of the documentation and software usability will be evaluated.

Dataset

PED (https://proteinensemble.org) entries are assigned based on the students group number.

1. PED00017 - Structural ensemble of isoform Tau-f of microtubule-associated protein tau
2. PED00020 - Structural ensemble of measles virus nucleoprotein
3. PED00154 - Structure and dynamics of the MKK4
4. PED00153 - Structure and dynamics of the MKK7
5. PED00142 - Structural ensemble of KISS-1 - *consider only the first 5 ensembles!*
6. PED00022 - Structural ensemble of Protein enhancer of sevenless 2B - *consider only the first 5 ensembles!*

References

1. Lazar et al. *PED in 2021: a major update of the Protein Ensemble Database for intrinsically disordered proteins*. (2021) Nucleic Acids Research
2. Lazar et al. *Distance-Based Metrics for Comparing Conformational Ensembles of Intrinsically Disordered Proteins*. (2020) Biophysical journal

[1] Split the structure in fragments of 5 residues and perform a structural alignment against the first structure, then measure the RMSD. See eq. 5 in [2].