Structural Bioinformatics 2020

Last document update 22 May 2020

Project no. I - Identification of structure clusters in Molecular Dynamics trajectories from Residue Interaction Networks

Introduction

Residue Interaction Networks are derived from proteins structures estimating contacts from distance measures. RING is a command line tool implemented in C++ which takes in input a PDB file and returns the list of contacts in a protein complex. RING is able to classify different types of contacts based on geometrical and physico-chemical properties of the amino acids. Types of contacts are:

- Hydrogen bonds (HBOND)
- Van der Waals interactions (VDW)
- Disulfide bridges (SBOND)
- Salt bridges (IONIC)
- π-π stacking (PIPISTACK)
- π-cation (PICATION)
- Inter-Atomic Contact (IAC), generic contact simply based on distance

RING generates two files. An “edge” file containing the contacts and a “node” file containing the list of residues which are in contacts and their attributes. A node is identified by a string like this “A:159:_:PRO”, in which the chain, residue index, insertion code and residue name are column (“:”) separated. An insertion code equal to “_” indicates there is no insertion code for that residue.

Project goals

The output of a Molecular Dynamics (MD) simulation is a trajectory file which describes the change of atomic coordinates from an initial state to a final state (after a certain amount of time) when a forcefield is applied. The full trajectory can be described by a subset of intermediate conformations / atomic coordinates / snapshots. Each team of students is requested to develop a software that identifies clusters of related (similar) structural conformations (snapshots) in a MD simulation. Additionally, the software should be able to identify those residues and contacts (pairs of residues) which are relevant to describe the transition between clusters.

- The clustering has to be generated by comparing the contact maps (not the structures) calculated by RING on the MD snapshots. The number of snapshots is ca. 100-1000 for each trajectory. RING contacts maps and the corresponding PDB files for each snapshot are provided by the teacher (see Dataset)

- The software should be able to both estimate an “optimal” number of clusters and provide a hierarchy of clusters (dendrogram). The distance metric used to compare contact maps is chosen by the team
- RING generates different types of contacts. They can be considered separately or weighted differently
- Consider that in molecular dynamics simulations the protein changes conformation very frequently and the same conformation can appear in different moments during the trajectory. As a consequence clusters may not correlate linearly with time
- In order to evaluate the clustering you can compare it with the clustering generated by measuring the distance between structures (instead of contact maps). For example by calculating all-against-all pairwise RMSD (RMSD can be calculated with the TM-score program available here: https://zhanglab.ccmb.med.umich.edu/TM-score/)
- Also, to evaluate the MD you can calculate and plot the pairwise RMSD against the first structure of the MD (this corresponds to the first row of the matrix generated above)

Project requirements

- The software has to be implemented in Python3. The use of the BioPython.PDB, Argparse and Logging modules is strongly encouraged
- The software has to be documented extensively in English 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)
- The output generated for all input examples provided by the teacher (see below output specifications)

Command line interface

Input

Mandatory

- A file with the path to RING contact map files (edge files). One file per line

Optional arguments

- Configuration file with algorithm parameters
- Output directory
- Temporary file directory

Example usage

```
$ python3 ring-sclusters.py contact_maps_paths.txt -conf params.ini -out_dir results/ -tmp_dir /tmp
```

Output

Mandatory:

- The distance matrix representing the distance between contact maps
- The clusters dendrogram

For the “optimal” clusterization (when the number of cluster is estimated automatically):

- The cluster assignment (a number representing the cluster ID) for each contact map in the input file and possibly a score representing the distance to the cluster center (or any other score which indicates something similar)
- A representative contact map for each cluster
- The list of relevant contacts/residues which contribute more to the transition between clusters. One list of contacts/residues for each possible cluster transition

Project evaluation

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

Dataset

Example datasets is available for download from this link:
https://drive.google.com/drive/folders/1lpL-wRGl78VtWCrfDy5reTXPZduZN9yj?usp=sharing
Each TAR archive corresponds to a different protein (frataxin, antibody) and contains a “pdb/” and an “edges/” folder which contains the conformation (PDB coordinates) and RING contact maps for each snapshot of the MD simulations. Files inside the “pdb/” and “edges/” with the same name correspond to the same snapshot.