Prof. Dr. Oliver Kohlbacher
Fabian Aicheler, Lukas Zimmermann
Mathematisch-Naturwissenschaftliche Fakultät
Abteilung für Angewandte Bioinformatik
Zentrum für Bioinformatik Tübingen

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Project 2
Implementation of a residue-based interaction potential for conformational energy differences

Supervisor: Fabian Aicheler (aicheler@informatik.uni-tuebingen.de)
Students: Dominic Boceck, Jonas Kübler, Lakshmipriya Sivaprakash Thrukonda

Predicting the stability of different protein conformations is a common task in structural bioinformatics. Force-field-based calculations are still quite compute intense and often have problems correctly predicting folding free energies. Statistical potentials have been developed to address some of these issues. A simple method of approximating a conformation's effective free energy is reducing it to contributing interaction energies between its amino acid residues. This can be done by summing pairwise interaction energies of amino acids that are in contact (i.e. physically close to each other) in a given conformation. We say two amino acids are in contact if their side chains' distance is below a certain threshold. The side chains are hereby typically represented by the position of their C_β atoms or the geometric center of their heavy atoms.

In such models the interaction potential between two amino acids can be as simple as a function of what amino acids they are and their physical distance. Moreover, distance dependence in the most basic case can be just a threshold value (Miyazawa & Jernigan [1]).

Knowledge-based interaction potentials are derived from existing structural data and rely on the assumption that the native conformations of proteins correspond to structures of lowest free energy. By observing contact frequencies of different amino acid pairs in a large set of structures, one can estimate how energetically favorable they are, assuming interaction energies adhere to Boltzmann distribution in ensembles of pair contacts.

Your task is to implement a residue-based interaction potential for conformational energy approximation and apply them to the provided sets of query structures. The sets originate from the Critical Assessment of Techniques for Protein Structure Prediction (CASP) project, containing different conformations of the same protein obtained from different tertiary structure predictors, and a crystal structure of the reference protein.

Using BioPython's PDB processing module, write a program that computes a structure's total interaction potential based on the Miyazawa & Jernigan potential.

- Read the provided PDB structures for every set, and strip down each of them to a polypeptide chain, removing water, possible ligands if present, etc.
- Compute the geometric centers of residue side-chains, where applicable. For glycine, reconstruct the position where the C_β atom would be if it existed by using 1.53 Å C_α-C_β bond length and tetrahedral geometry for C_α's bond angles.
- Determine contact pairs using a 6.5 Å distance threshold. To account for uncertainties in the boundary regions, use a sigmoidal transition with a width of 1 Å following the authors' proposition.
- Determine the total folding free energy of the system using the contact energy table presented in Miyazawa & Jernigan 1993.
- Evaluate all conformations of the same protein and draw conclusions as to which one of them should be the most stable structure.

- Contacts arising from chain connectivity (i.e. residue pairs which are connected via a peptide bond) should be ignored. They are likely forced by the sequence rather than interresidue interactions. How does considering for this in your algorithm influence your rankings?
- Align the structures to the native structure. Superimpose the C_α atoms of the two structures in a way that their root mean square deviation (RMSD) is minimal, and discuss your findings in the light of the query structures' folding free energy values. Visualize the alignments of top scoring structures and the crystal structure.
- Compute the RMSD between every pair of conformations in the set. Perform hierarchical clustering to identify sets of similar conformations among them.
- Include the run times of your calculations in the report for different proteins and discuss them in the light of the algorithmic complexity of the method.

In your final report, please summarize everything you did for your project. Your final report should be 5 pages long including figures, tables, and references.

References

[1] Miyazawa, S., & Jernigan, R. L. (1996). Residueresidue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. Journal of molecular biology, 256(3), 623-644.