Validating Protein Structure Models Using Internal Energy

By: Branden Lee and Kim Kwan

ECS 129 Option 5

<https://github.com/UC-Davis-ECS-129-Project/Protein-Internal-Energy>

Brad:

We should write this project as if it was an actual paper we wanted to publish, so we can get writing experifence.

Background questions:

1. Why are we answering this? What is the scientific question we are trying to answer?

2. What is currently known about this topic? Who has already worked on this?

3. What have you done?

Abstract

This project seeks to determine the conformation a protein is more likely to be found in by minimizing the internal energy of proposed topologies. Internal energy is calculated by making approximations of Van der Waals forces, electrostatic energy, and solvation energy. The project is run in standard python 3 with preprocessed protein data files that were preprocessed.

Background

Proteins are extremely important to help the human body function. Right now, we have a general idea of the general structures that a protein may have, such as a variety of secondary and tertiary structures a protein may be made of. In addition, many quaternary structures have been determined for a wide variety of proteins with functions ranging from regulatory to structural purposes. However, given a polypeptide sequence, we would like to accurately identify the resulting structure, which includes widespread alpha helices and beta sheets in addition to more rare structures like the beta sheet hairpins.

Intermolecular forces such as Van der Waals, electrostatic, and solvation energy have a profound effect on protein folding. Gibbs free energy and spontaneity dictates that a molecule wants to reduce its internal energy, therefore staying in a lower energy conformation. By predicting internal energy of multiple possibilities for 3D protein structures, one can identify the most probable folded structure by finding the structure that minimizes internal energy.

The approximation for Van der Waals energy is done using the equation for Lennard-Jones-Potential as shown in figure 1.



The Lennard-Jones potential is a mathematically simple model that approximates the interaction between a pair of neutral atoms or molecules. A form of this interatomic potential was first proposed in 1924 by John Lennard-Jones. ε is the depth of the potential well, sij is the distance at which the potential reaches its minimum, and r is the distance between the particles.

TALK ABOUT THE OTHER PARTS OF THE EQUATION THAT ARE MISSING HERE:

We would also like to determine the presence and location of misalignments and incorrect folding patterns that may have a small or huge impact on the protein affecting a person’s daily life. While computer models have aided researchers to predict protein models, algorithms built to analyze sequences are not perfect and constantly change over time.

Currently, scientists in this field have been able to detect variations of the same protein to determine the most likely structure that the protein may have with very little variability. However, not much research has been done about finding accurate sequences from two different proteins and observing how they aligned with each other. We are not yet able to observe the variability from the sequence and the model compared to the protein in vivo. We plan to take a sequence that one may have to code for amino acids in a protein and accurately determine which of the two forms given is more energetically favorable.

By predicting accurate models, researchers would be able to identify minute differences between distantly related models. We could use this data to observe how the structure of a protein changes due to mutations in the sequence. Once a library of models is made, professionals in the medical field can access this database and see which specific variations of protein they may need to target. Medical professionals may also use this data to determine if a person is able to react to certain virus or bacteria that may enter the immune system. They may also use this information to target enzymes that regulate or are a part of many metabolic pathways.

SOLVATION ENERGY NEEDS TO BE DISCUSSED!!

Methods

An internal energy calculator was designed with python. The script opens a preprocessed protein file that contains a tabularized list of atoms in the protein with their associated numerically defined properties. The atoms are stored as a python dictionary and are looped through to calculate internal energy based on the atomic interactions.

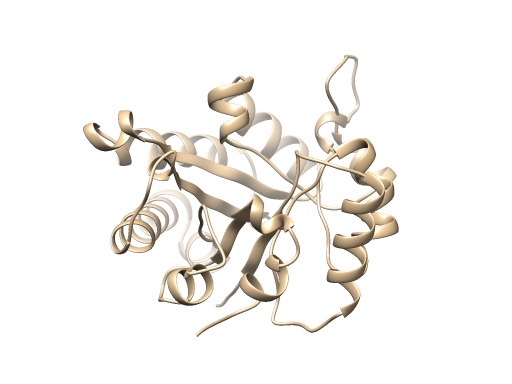
The Tabularized list contains properties for each atom. The properties are available are:

|  |  |  |
| --- | --- | --- |
| Property | Unit | Description |
| X | Angstroms (Å) | 3D space coordinate |
| Y | Angstroms (Å) | 3D space coordinate |
| Z | Angstroms (Å) | 3D space coordinate |
| R | Angstroms (Å) | Van der Waals radius |
| Epsilon | kcal/mol | Depth of the potential well (Wikipedia) |
| Sigma | Angstroms (Å) | Distance at which the potential reaches its minimum (Wikipedia) |
| Charge | Coulomb Fraction |  |
| ASP | Å2 | Atomic solvation parameter |
| ASA | kcal/mol/Å2 | accessible surface area |

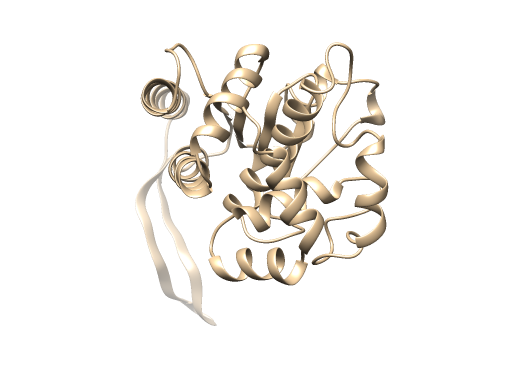
(Page Break)

Results

Protein 1:

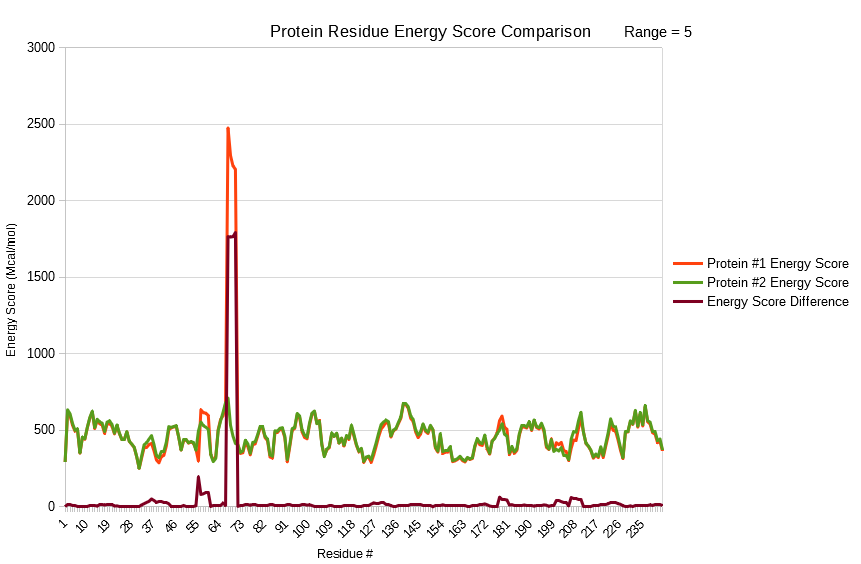


Protein 2:

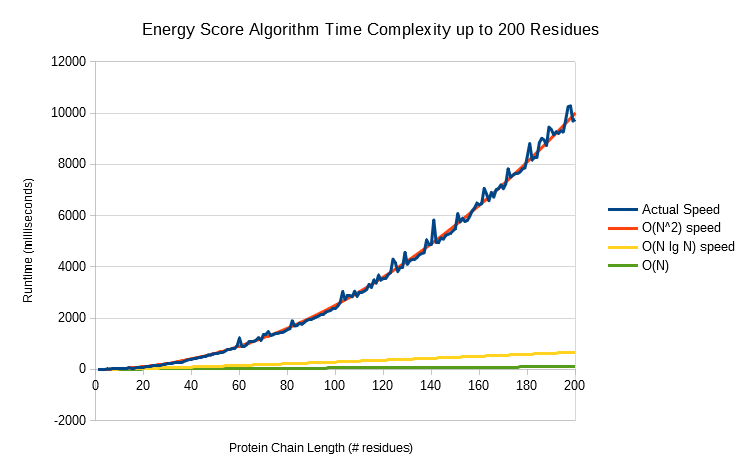


The energy score of conformation #1 was 8.1e9 kcal/mol, while conformation #2 was 1.7e6. There was a significant difference in the energy scores between both protein conformations. Conformation #2 energy score is lower and is the more likely conformation because of thermodynamic reasons.

Both conformations of the protein had an average energy score difference ~10 Mcal/mol; however, from amino acid 66 to 69, there was a significant difference. A subarray range of five residues was used to discover the local energy scores.



The Lennard-Jones potential and electrostatic energy calculations are in a nested loop, thus the time complexity of the algorithm is theorized as O(N^2). Running the protein energy scoring algorithm on randomly generated protein chains of N length up to N=200 confirms that the algorithm runs at a O(N^2) speed.



Discussion

The program takes around 7 seconds which is quite lengthy. With a sufficiently large amount of possible protein topologies for a single protein that need to be validated, the script could take a large amount of time to finish. A compiled language such as C would be a better fit.

Works Cited

Breda A, Valadares NF, Norberto de Souza O, et al. Protein Structure, Modelling and Applications. 2006 May 1 [Updated 2007 Sep 14]. In: Gruber A, Durham AM, Huynh C, et al., editors. Bioinformatics in Tropical Disease Research: A Practical and Case-Study Approach [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2008. Chapter A06. Available from: https://www.ncbi.nlm.nih.gov/books/NBK6824/

Maiorov and Abagyan, 1998 V. Maiorov, R. Abagyan Energy strain in three-dimensional protein structures Fold. Des., 3 (1998), pp. 259-269 <https://www.sciencedirect.com/science/article/pii/S1359027898000376>

UCSF Chimera--a visualization system for exploratory research and analysis. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. J Comput Chem. 2004 Oct;25(13):1605-12.

Chang, Raymond. *Physical Chemistry for the Biosciences.* Sausalito, CA. University Science Books, 2005. (498-500)

Wikipedia. Lennard-Jones potential. <https://en.wikipedia.org/wiki/Lennard-Jones_potential>

Wikipedia. Implicit solvation. https://en.wikipedia.org/wiki/Implicit\_solvation