Validating Protein Structure Models Using Internal Energy

By: Branden Lee and Kim Kwan

ECS 129 Option 5

Github Repository: <https://github.com/GoodProtein/ProScore>

Abstract

This project seeks to determine which structure of a protein is more valid by comparing free energy. Free energy scores are calculated by making approximations of the summations that include Van der Waals forces, Coulomb forces, and solvation energy. Our program is run in standard python 3 with preprocessed protein data files provided.

Introduction

The human body requires proteins to carry out structural, enzymatic, and transport functions. Since the start of molecular biology research, determining protein structure from primary structure has been a priority of scientists worldwide. Currently, there are methods of sequencing the amino acids of proteins such as mass spectometry. There are also protein gels and column chromatography techniques that allow us to analyze the polarity or size of the protein. Methods used to study secondary and tertiary structure include circular dichroism, X-ray crystallography, and NMR spectrometry. However, given a polypeptide sequence and two possible structures, we want to identify the most likely native state the sequence will form.

The 1972 Nobel Prize winner, Christian Anfinsen, hypothesized a protein’s structure is a unique, stable and kinetically accessible minimum of the free energy (Anfinsen) in a normal physiological environment. Also known as the thermodynamic hypothesis, Anfinsen’s dogma is the basis for many protein folding computations. This also ties with the folding funnel hypothesis which states that the protein’s natural state is one where its free energy is minimum within the environment of a cell.

One such way of calculating free energy of a protein is using experimental-based approximations with OPLS force fields. OPLS force field parameters for amino acids are used to predict the free energy inside a protein. Intermolecular forces such as Van der Waals forces, Coulomb, and solvation energy must be factored into the calculation.

The primary estimation of free energy used in this program is:

Equation 1.1. Total energy of a protein structure. (Koehl)

|  |  |
| --- | --- |
|  | Equation 1.2 |

The approximation for Van der Waals energy is done using the equation for Lennard-Jones-Potential as shown in figure 1.2. The Lennard-Jones potential is a simple mathematical 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. This equation accounts for the attraction and repulsive forces that an atom may experience depending on its distance relative to other atoms within the peptide.

|  |  |
| --- | --- |
|  | Equation 1.3 |

Electrostatic potential energy results from conservative Coulomb forces and is associated with the configuration of a particular set of point charges within a defined system (Wikipedia). In the case of an amino acid, the partial charges of each atom are experimentally derived. In this equation, qi and qj stand for the two charges that interact with each other with rij representing the distance between interacting particles. ε0 and εr are electric constants.

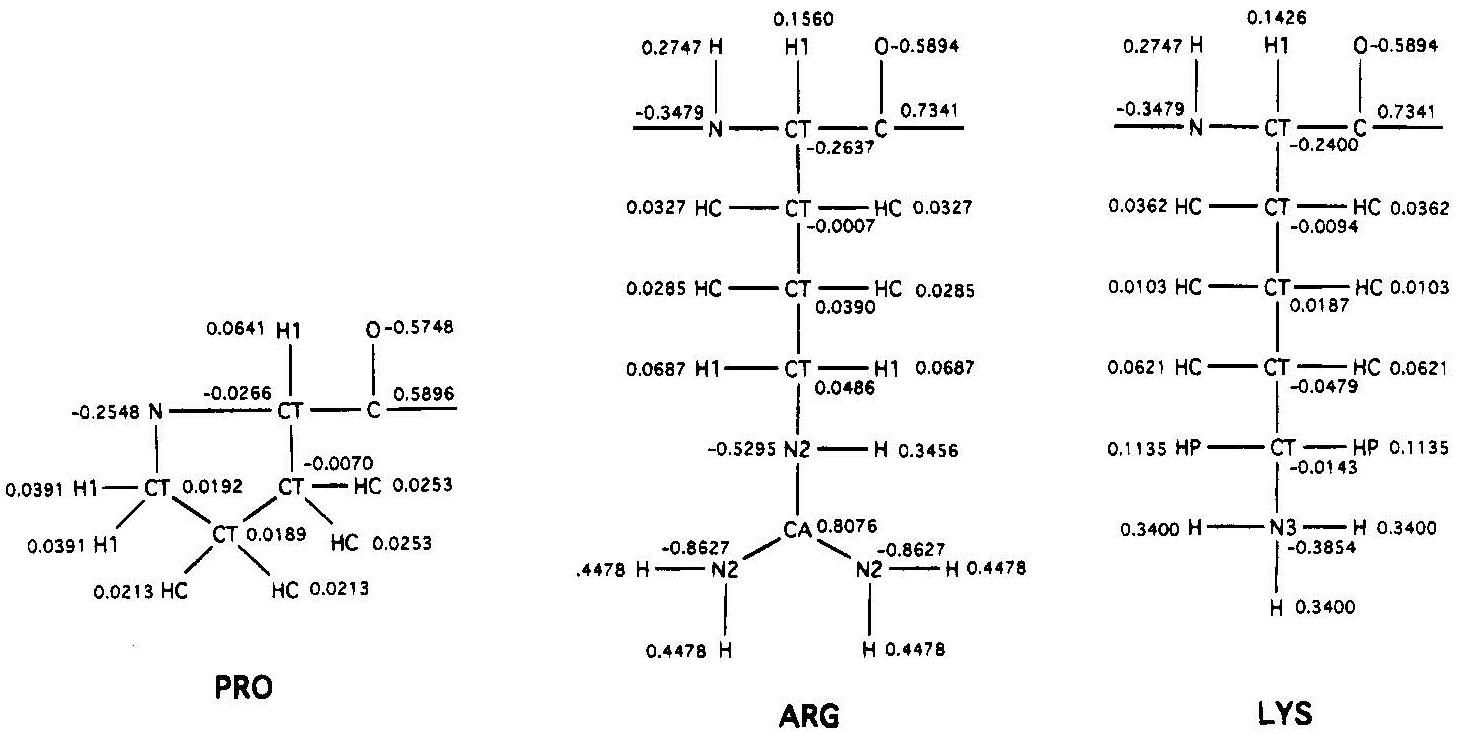


Fig 2. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. (Cornell)

|  |  |
| --- | --- |
|  | Equation 1.4 |

While van der Waals and Coulomb act as repulsion terms between non-bonded atoms, solvation energy is also a very important component of protein free energy.

Implicit solvation (sometimes termed continuum solvation) is a method to represent solvent as a continuous medium instead of individual “explicit” solvent molecules, most often used in molecular dynamics simulations and in other applications of molecular mechanics (Wikipedia). The free energy of solvation of a solute molecule in the simplest ASA-based method is given by figure 1.4. ASP(i) represents the atomic solvation parameter for atom i, which was provided in the data we used.

|  |  |
| --- | --- |
|  | Equation 1.5 |

ASA(i) stands for the accessible surface area for atom i where ri is the van der Waals radius for atom i and RH2O is the radius of a water molecule. ASA(i) is calculated from the rough approximation in figure 5.

It is important to note that because the implicit solvation equation 1.5 is a poor estimation for actual solvation energy, the calculated solvation energy from this program is a very inaccurate and is more of a metric or score to compare the relative total energy in a protein structure. A more accurate way to calculate the ASA based on Gromiha’s method is:

|  |  |
| --- | --- |
|  | Equation 1.6 |

R is the radius. Li is the length of the arc for the atom i, Zi is the distance from the center of the sphere to the atom i, Z is the spacing between two different sections, and ′Z is ΔZ/2 or R − Zi, whichever is smaller.

Methods

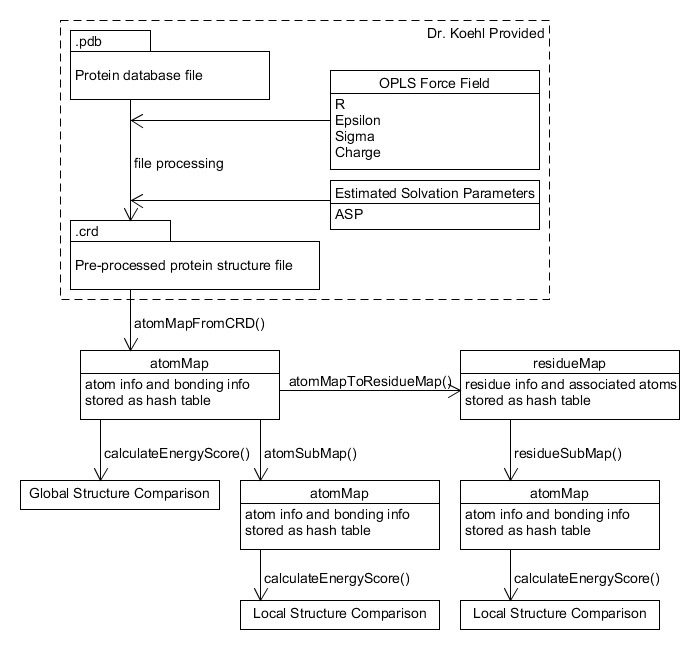
An internal energy calculator was designed with python. The script opens a preprocessed protein file that contains a tabularized list of atoms in the protein with their associated numerically defined properties. The atoms are stored as a python dictionary and are looped through to calculate internal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

Fig 3. Created with UMLet.

Global structure comparison has results output to the console. Local Structure comparison is output to a comma-separated values (.csv) file.

To run the a global structure comparison of two protein molecules with the same amino acid sequence. Use the following command in shell.

python mainEnergyScore.py

To run local sequence comparison based on sampling ranges of atoms:

python mainAtomScoreCompare.py

To run local sequence comparison based on sampling ranges of residues:

python mainResidueScoreCompare.py

The pre-processed protein structure file must be in the following format:

Line 1: number of atoms or

Lines with leading pound (#) character will be ignored and not interfere with atom count. Leading pound is used for in-file annotations and comments such as column labels.

The next lines contain rows of atom data with columns delimited by whitespace. Columns are not fixed width. Column width is determined by data type size. Atom numbers go from 1 to inclusive without sequence skipping.

Atom data columns:

|  |  |  |
| --- | --- | --- |
| Column | Data Type | Description |
| 1 | Integer | Atom number |
| 2 | Real(10.4) | X |
| 3 | Real(10.4) | Y |
| 4 | Real(10.4) | Z |
| 5 | Real(10.4) | R |
| 6 | Real(10.4) | Epsilon |
| 7 | Real(10.4) | Sigma |
| 8 | Real(10.4) | Charge |
| 9 | Real(10.4) | ASP |
| 10 | Char(6) | Atom name |
| 11 | Char(6) | Residue name |
| 12 | Integer | Residue number |

The next lines contain rows of atom bonding data with columns delimited by whitespace.

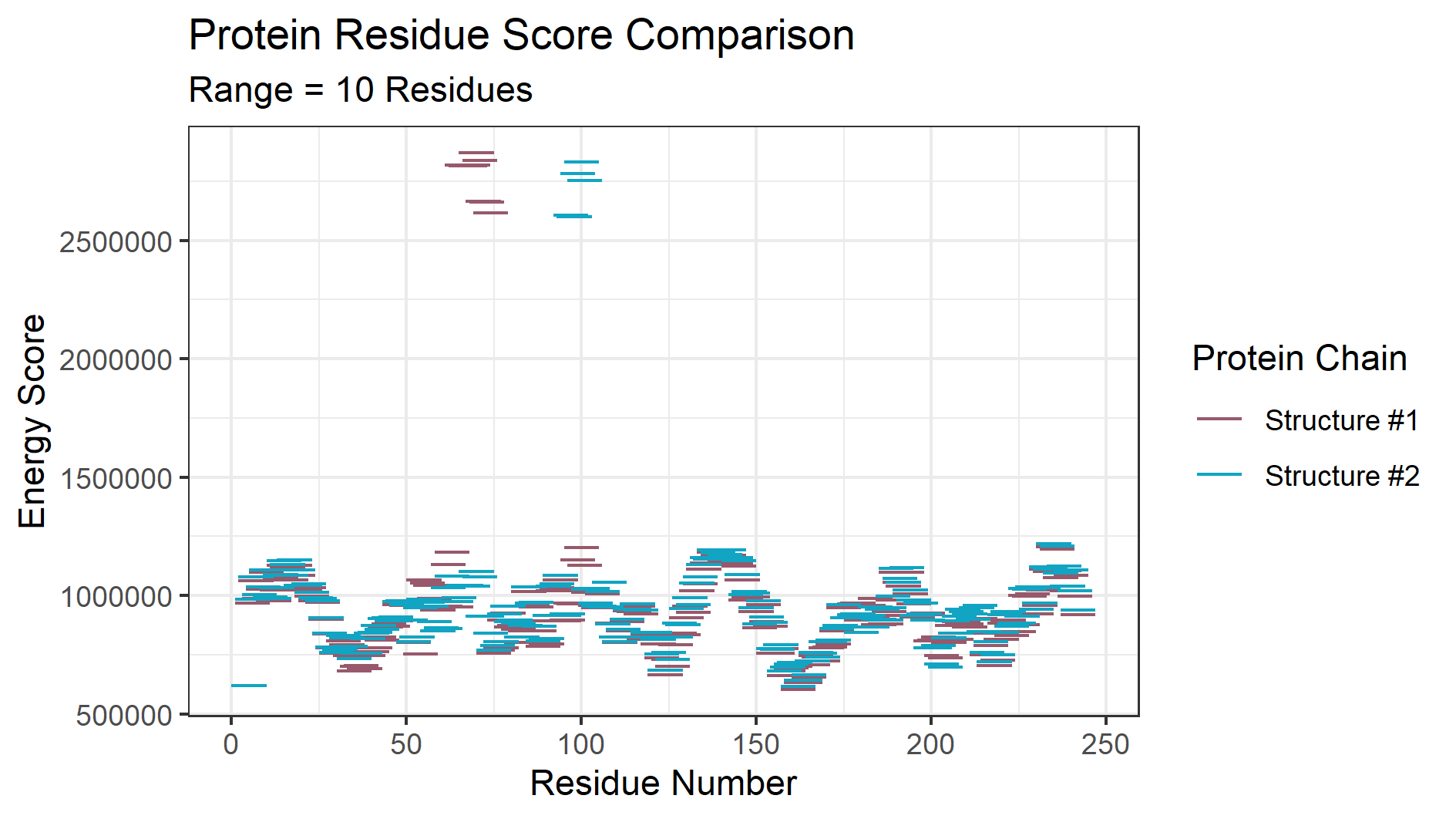
Atom bonding data columns:

|  |  |  |
| --- | --- | --- |
| Column | Data Type | Description |
| 1 | Integer | Atom number |
| 2 | Integer | Size of subsequent integer array |
| 3 | Integer Array | Bonded atom number |

The calculateEnergyScore() method is an implementation of the total energy equation in equation 1.1. This method has a time complexity of due to the nested loop.

Results

Using the global comparison program, the energy score of structure #1 is , while conformation #2 is. We concluded that there is a significant difference in the energy scores between both protein conformations. The energy score of Structure #2 is lower and is more likely to be the native state using the lowest free energy noted by Anfinsen's dogma.



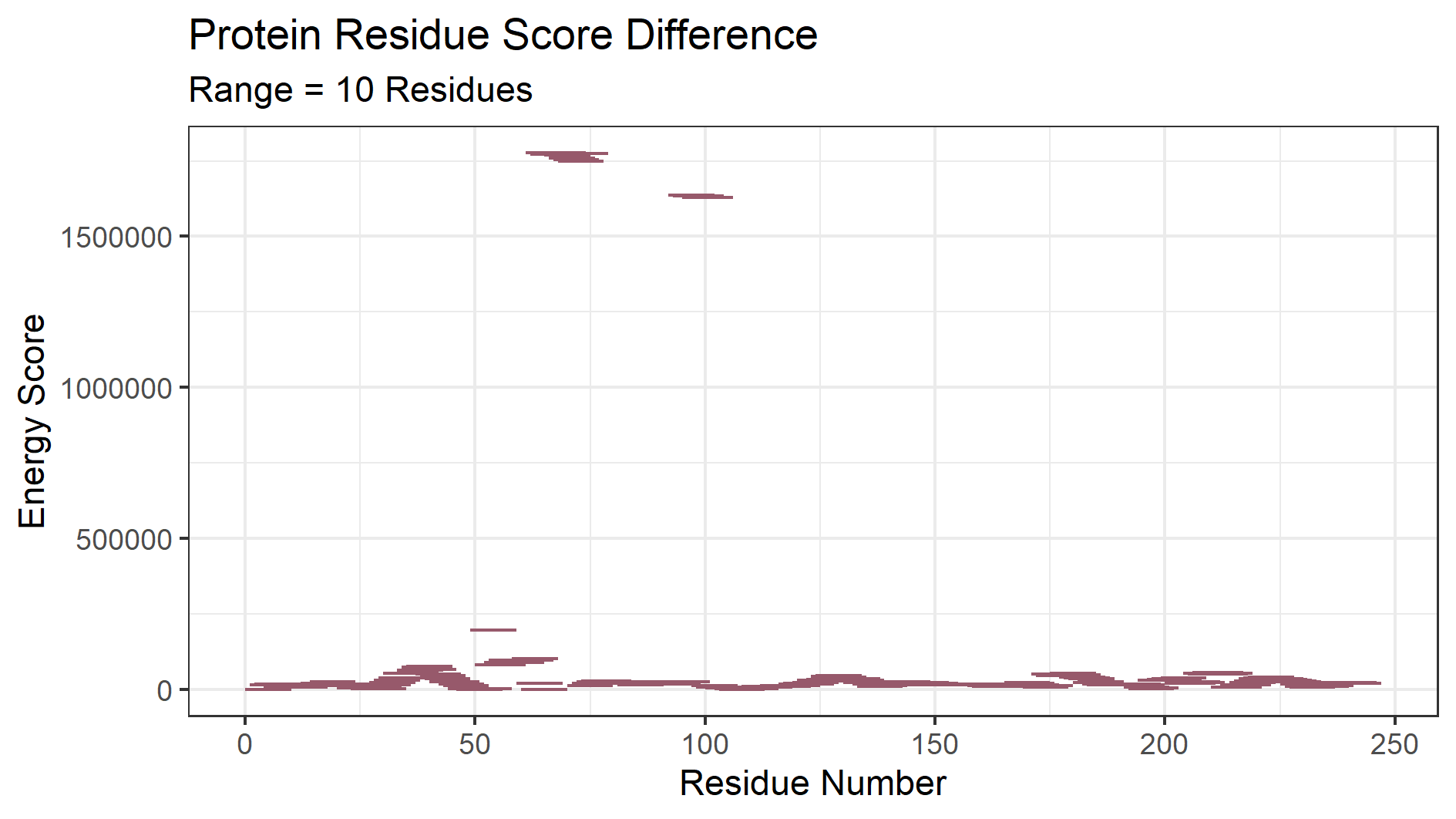


Fig 4.1 (top) and 4.2 (bottom) were created with R Studio ggplot2 package.

Both structures of the protein had a global energy score difference of with a magnitude of ; however, from amino acid 66 to 69, there was a significant difference. A subarray range of ten residues was used to discover the local energy scores. Ten was arbitrarily chosen, because it gave the most clear energy score levels.

A further investigation was performed to discover the structure differences owing to the high difference in energy score at the local residue sequence. Both protein structures were superimposed using UCSF Chimera’s Matchmaker Algorithm for structure comparison using best-alignment of chains between structures. Residues #60 to 78 with residue seqence AAALVPWKNENAGIDGTKA were selected and focused on to view structure differences that would cause such a large energy score difference. The range was chosen by identifying the region of the highest local energy score difference between both protein structures and extending the ranges by an arbitrary amount.

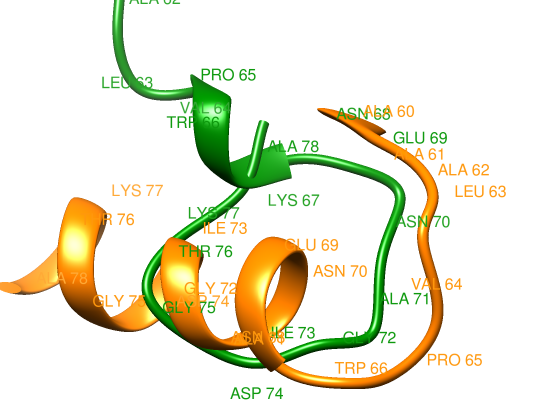


Fig 5. Green chain is structure #1. Orange chain is structure #2. Created using UCSF Chimera.

The main structural differences of this region seems to be that structure #2 forms an alpha helix after TRP 66, but structure #1 starts a long turn back into the chain instead.

**Time Complexity:**

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

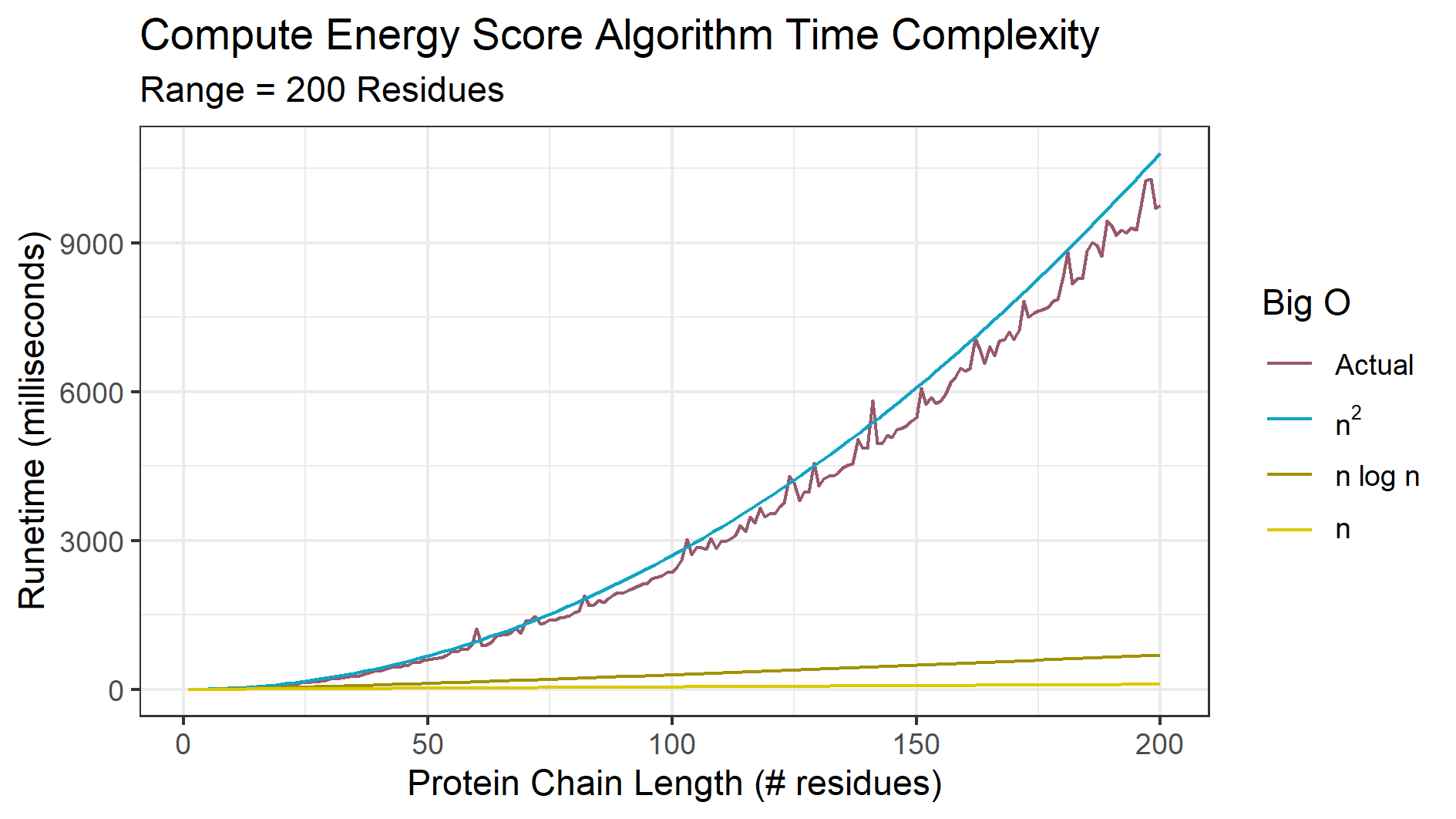


Fig 6. Computing Energy Score algorithm time. Created with R Studio ggplot2 package.

Discussion

It is important to reiterate this program does not calculate a free energy with a specific physical meaning, but rather acts as a score or metric to validate protein structures. A lower score is better.

The protein structure local residue energy score comparison in Figure 4.1 revealed that the difference in structure energy is localized to two specific regions. The structural analysis done in Figure 5 gave further insight into structural differences at one of the regions of highest energy score difference. It is hypothesized that LYS 67 turning back into ALA 78 creates large unfavorable interactions between atoms, leading to an increase in energy score.

Structure #2 is the more valid structure between the two structures compared. However, establishing that structure #2 is the native structure of the protein in vivo would require validation of all possible protein conformations. Cyrus Levinthal, attributed with Levinthal's Paradox, noted that each protein molecule has an astronomical number of possible conformations. Validating each conformation would be time prohibitive, thus better methods should be used. One such method is using artificial intelligence.

The program takes over 9 seconds to run on a randomly generated protein chain with 200 lysine residues. With a sufficiently large amount of possible protein structures for a single protein that need to be validated, the script would take a large amount of time to finish. The free energy calculation was also written with C++ with no thread or GPU enhancement, and ran in 1 second for 200 lysine residues (or 2600 atoms). Lysine residues were arbitrarily chosen for the benchmark. Residues were not randomized because residues do not all contain equal numbers of atoms.

In the future, we would like to determine the presence and location of misalignments and incorrect folding patterns that may have a small or huge impact on the protein in question. 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 protein folding 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 viruses 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.

Works Cited

Anfinsen, Christian B. “Principles That Govern the Folding of Protein Chains.” Science, vol. 181, no. 4096, July 1973, pp. 223–30. science.sciencemag.org, doi:10.1126/science.181.4096.223.

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/

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

Cornell, Wendy D., et al. “A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules.” Journal of the American Chemical Society, vol. 117, no. 19, May 1995, pp. 5179–97. ACS Publications, doi:10.1021/ja00124a002.

Gromiha, M. Michael. “Protein Bioinformatics” Chapter 3 - Protein Structure Analysis,

Academic Press, 2010, pages 63-105

Koehl, P.. ECS 129: Validating protein structure models. University of California, Davis. dscdvd2020.

Levinthal, Cyrus. How to Fold Graciously. Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois.

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.

Wikipedia. Lennard-Jones potential. https://en.wikipedia.org/wiki/Lennard-Jones\_potential

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