Validating Protein Structure Models Using Internal Energy

By: Branden Lee and Kimberly Kwan

ECS 129 Option 5
Github Repository: https://github.com/bradosia/Validating-Protein-Structure-Models

Abstract

This project implements a simplified method, originally devised by Koehl, for scoring the quality of a protein structure using an internal energy calculation that includes Van der Waals, electrostatic, and solvation energy. Two protein structures with accompanying pre-processed atom data files are compared using our method. Structure #2 is found to have a higher structure quality because of its lower internal energy score.

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 worldwide. Currently, researchers sequence proteins using mass spectrometry as well as column chromatography to analyze the polarity or size of the protein. Methods used to study secondary and tertiary structure include circular dichroism and NMR spectrometry. Protein structure verification required vast amounts of data to calculate to ultimately build an acceptable model of what a protein may look like. It is crucial to have the ability to create a high quality model since scientists need it to identify how proteins function and interact with other substrates. According to Benkert et al., researchers need to be able to identify accurate protein foldings to be able to treat various diseases by targeting active sites for fields in drug design.

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 since the dogma states that the amino acid sequence dictates the most natural conformation the protein will form. Dill's described funnel-shaped energy landscape, also known as the folding funnel hypothesis, 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.

Methods

The implementation of the internal energy score is a truncated force field equation without bond, angle, and rotation energy. Additionally, an implicit solvation energy is added.

$$U = \sum_{i=1}^{N} \sum_{j=i+1}^{N} NoBond(i,j) \left(\epsilon_{ij} \left(\left(\frac{s_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{s_{ij}}{r_{ij}} \right)^{6} \right) + \frac{1}{4\pi\epsilon_0\epsilon_r} \frac{q_i q_j}{r_{ij}} \right) + \sum_{i=1}^{N} ASP(i) ASA(i)$$

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

$$\epsilon_{ij} \left(\left(\frac{s_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{s_{ij}}{r_{ij}} \right)^{6} \right)$$

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. According to Chang, the Lennard-Jones potential is a simple mathematical model that approximates the interaction between a pair of neutral atoms or molecules. ε is the depth of the potential well, s_{ij} 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.

$$\frac{1}{4\pi\epsilon_0\epsilon_r} \frac{q_i q_j}{r_{ij}}$$

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, q_i and q_j stand for the two charges that interact with each other with r_{ij} representing the distance between interacting particles. ε_0 and ε_r are electric constants with ε_r representing the dielectric constant of water with a value of 4.

$$\sum_{i=1}^{N} ASP(i) ASA(i)$$

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.

$$ASA_i = 0.2 * 4 * \pi * (r_i + R_{H2O})^2 \qquad \text{Equation 1.5}$$

ASA(i) stands for the accessible surface area for atom i where r_i is the van der Waals radius for atom i and R_{H2O} is the radius of a water molecule.

It is important to note that because the implicit solvation equation 1.4 is a poor estimation for actual solvation energy, the calculated solvation energy from this program is more of a metric or score to compare the relative total energy in a protein structure. Equation 1.5 only factors in the radius of water. In reality, physiological conditions include a variety of molecules besides water, which may interact with the protein differently. In addition, implicit solvation does not directly factor in which part of the protein is in contact with water. Hydrophobic parts of proteins conglomerate toward the center of the protein while the hydrophilic portions form the surface to interact with the hydrophilic solvent molecules such as water.

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.

This project introduces three python scripts that are outlined in figure 2 and described as follows. 1. *mainEnergyScore.py* is a basic global comparison of two protein structures. 2. *mainAtomScoreCompare.py* performs a local comparison of two structures using sequential atoms in a user-defined range. The range acts as a sliding window with internal energy score calculated at each frame. Output file and usage is outlined in the readme file distributed with the source code. 3. *mainResidueScoreCompare.py* is a local comparison with residues the same same principle as described previously.

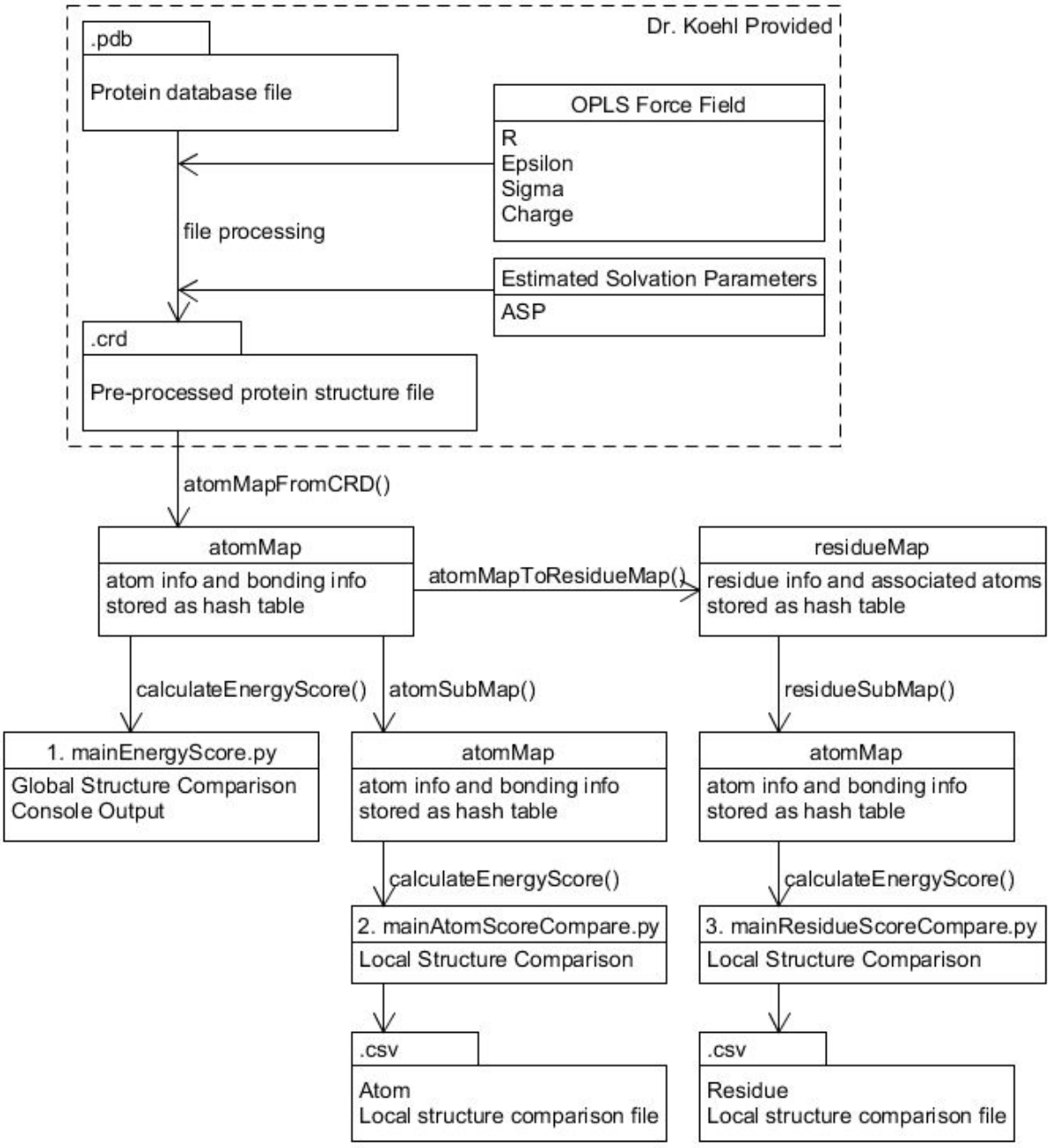

Fig 2. Created with UMLet.

The calculateEnergyScore() method is an implementation of the total energy equation in equation 1.1. This method has a time complexity of $O(N^2)$ due to the nested loop. Global structure comparison has results output to the console. Local Structure comparison is output to a comma-separated values (.csv) file.

Input

The pre-processed protein structure file must be in the following format:
Line 1: number of atoms or N
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 N 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 N 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 N 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

Results

Using the global comparison program, the energy score of structure #1 is $8.1e^9\ kcal/mol$, while conformation #2 is $1.7e^6\ kcal/mol$. We concluded that a significant difference exists 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.

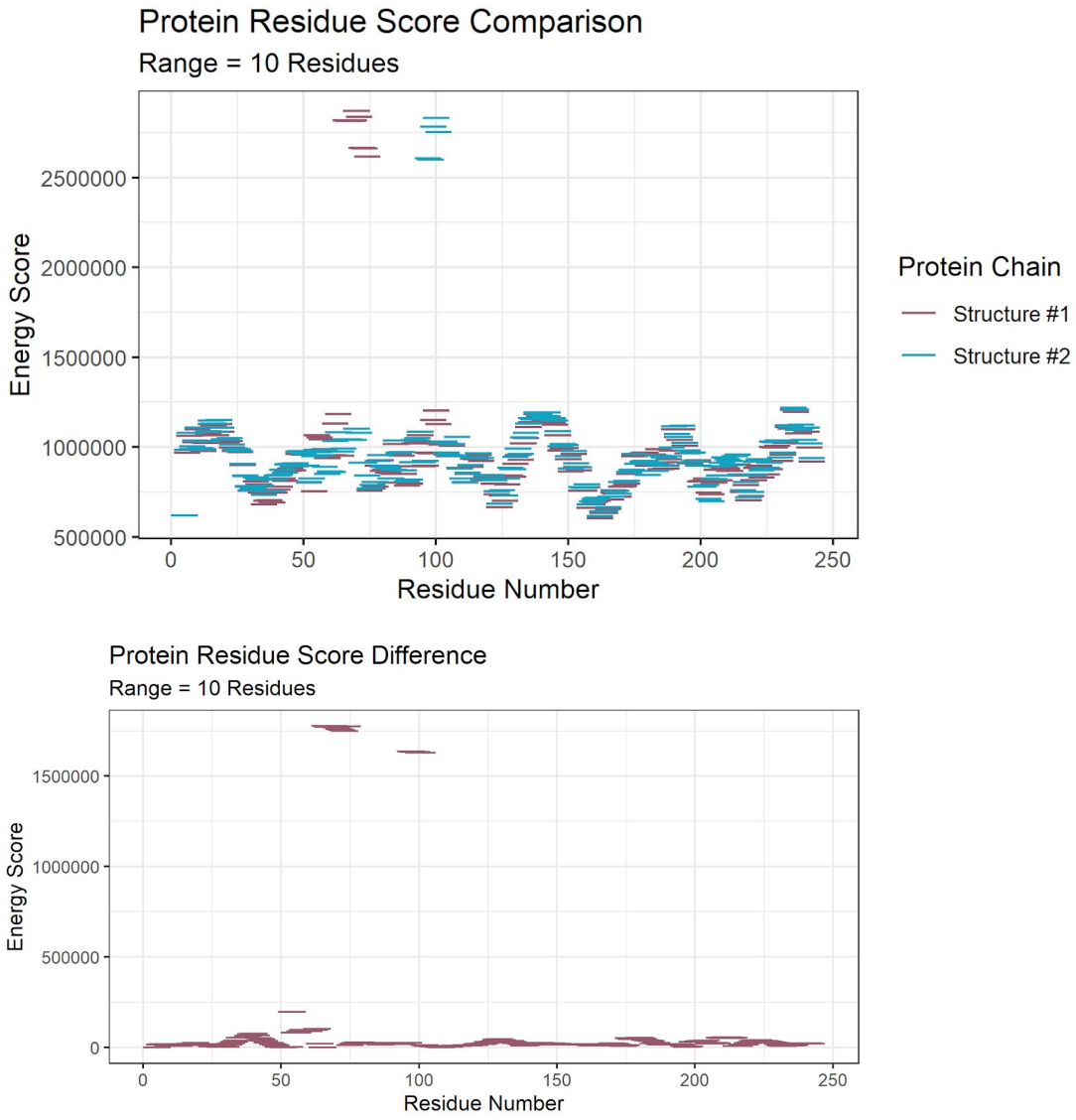

Fig 3.1 (top) and 3.2 (bottom) were created with R Studio ggplot2 package. Data calculated from mainResidueScoreCompare.py script by methods outlined in Figure 2.

Both structures of the protein had a global energy score difference of with a magnitude of 10^6; 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.

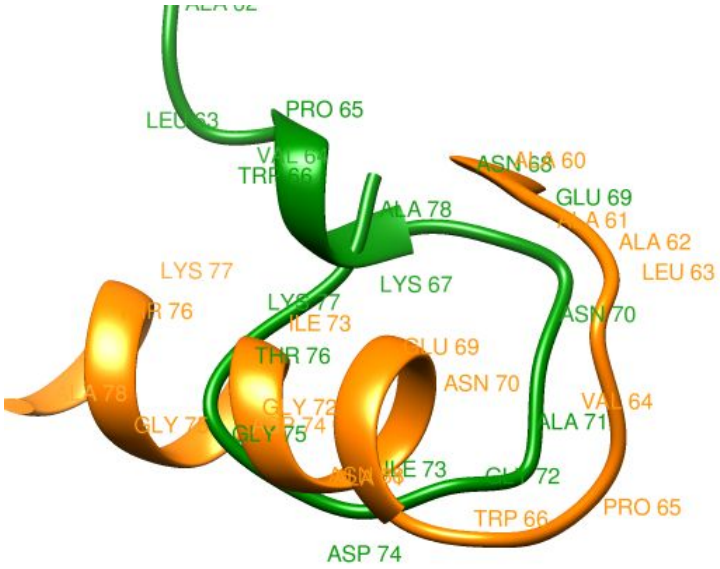

Fig 4. Green chain is structure #1. Orange chain is structure #2. Created using UCSF Chimera.

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 $O(N^2)$. Running the protein energy scoring algorithm on randomly generated protein chains of N length lysine residues up to $N = 200$ confirms that the algorithm runs at a $O(N^2)$ speed.

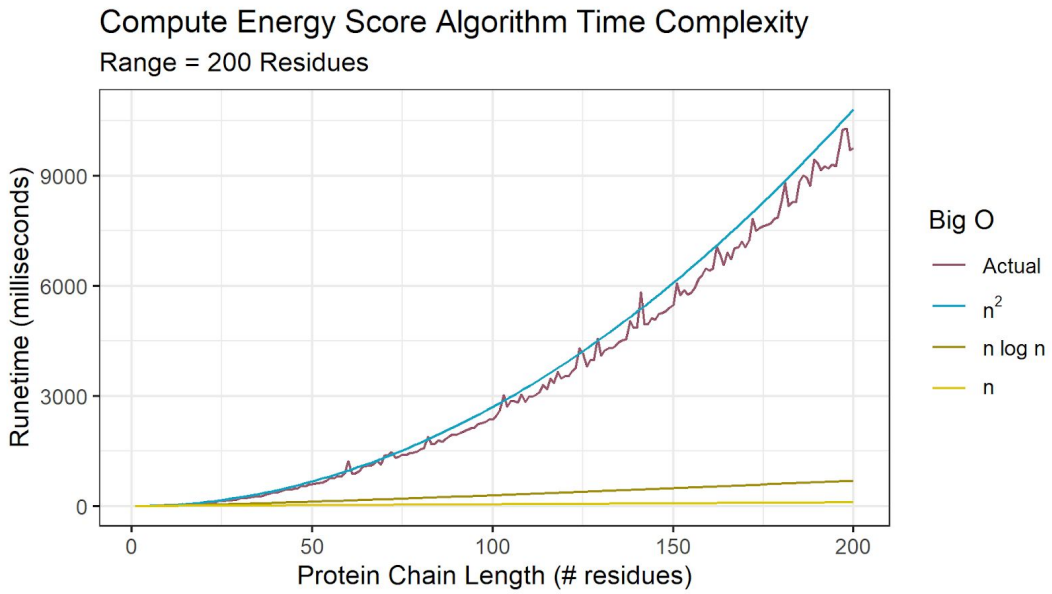

Fig 6. Computing Energy Score algorithm time. Created with R Studio ggplot2 package.

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. According to Dill's Funnel-Shaped Energy Landscapes a lower score is the more likely conformation.

The protein structure local residue energy score comparison in Figure 3.1 revealed that the difference in structure energy is localized to two specific regions. The structural analysis done in Figure 4 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.

Discussion

A more accurate way to calculate the ASA based on Gromiha's method is:

$$\mathrm{ASA} = \sum \left[R/\left(R^2 - Z_i^2\right)^{1/2} \right] L_i \cdot D; D = \Delta Z/2 + \Delta' Z,$$ 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. This equation involves more careful calculation of the exact surface area that is in contact with the solvent than our given equation.

Another group of researchers who have worked on a similar problem to our project is Maiorov and Abagyan. However, they used the interaction energy between a single amino acid and the entire molecule along with the solvent that the protein is in to be able to identify a stable protein. They also focused more on identifying the total strain found within an individual protein instead of comparing different sequences. We focused on identifying which of the two structures matched better with a given sequence.

Benkert et al. used X-ray crystallography data to estimate the quality of a protein structure. They developed the Qmean score. This score can be used on both short and long peptides. His QMEAN Z-score is a metric m of nativeness within a specific solvent useful for structure comparison.

Using the Qualitative Model Energy ANalysis (QMEAN) tool by Benkert et al., Structure #1 had a QMEAN4 value of -11.89 and Structure #2 has a QMEAN4 value of -10.94. The QMEAN4 value is transformed into a Z-score to relate it with what one would expect from high resolution X-ray structures. Both structures have scores far from zero meaning both structures #1 and #2 would be unlikely to be observed from a high resolution X-ray. However, the QMEAN4 score of structure #2 is closer to zero than structure #1, thus structure #2 is comparatively better.

The local quality estimates of the structure were not analyzed because preprocessing of the QMEAN4 removed 209 atoms in structure #1 and 111 atoms in structure #2. Most of the atoms removed in structure #1 were from residues #35 to 67, which was identified in figure 3.1 as a region of a large score difference.

Compared to both of these researchers, our method takes a known sequence and identifies a structure that best matches the native conformation by calculating energies. While Maiorov and Abagyan did include solvation energies, we decided to have a complete analysis by adding energies caused by charges between individual amino acids and Lennard Jones potentials caused by distances between atoms.

Benkert takes out project one step further by analyzing aspects of protein including solvent accessibility and backbone geometry when building an accurate protein model.

References

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.

Benkert, P., Biasini, M., Schwede, T. Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics 27, 343-350 (2011). https://doi.org/10.1093/bioinformatics/btq662

Chang, Raymond. Physical Chemistry for the Chemical and Biological Sciences. University Science Books, 2000.

Dill, Ken A., et al. “The Protein Folding Problem.” Annual Review of Biophysics, vol. 37, June 2008, pp. 289–316. PubMed Central, doi:10.1146/annurev.biophys.37.092707.153558.

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. 2020.

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

Maiorov, Vladimir, and Ruben Abagyan. “Energy Strain in Three-Dimensional Protein Structures.” Folding and Design, vol. 3, no. 4, Aug. 1998, pp. 259–69. ScienceDirect, doi:10.1016/S1359-0278(98)00037-6. 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.

“Implicit Solvation.” Wikipedia, 5 Mar. 2020. Wikipedia, https://en.wikipedia.org/w/index.php?title=Implicit_solvation&oldid=944132814.