Lecture Title and Date:

SIMULATION - Protein Simulation I (April 2nd, 2025)

Objectives of the Lecture

By the end of the lecture, students should:

1. Understand what proteins are, amino acids, synthesis, and protein structure
2. Understanding the driving forces behind protein folding and their dynamics via energy landscapes, Ramachandran plots
3. Have a strong foundation of proteins and folding energetics to build onto for learning about computational modeling of proteins and protein structure

Key Concepts and Definitions

Protein Folding: The transition that a protein takes from a linear chain of polypeptides to its native folded state. A protein must be in its folded state to perform its biological function. Many proteins have their folded states solved in crystal structures. Unfolding can occur due to high temperature or denaturants in solution.

Amino Acids: The individual units that make up a peptide chain. They contain a carbonyl group and an amine group connected by a single alpha carbon. The alpha carbon can be attached to a side chain, which determines its identity, and there are 20 typical amino acids present in human proteins, characterized by polarity and charge. They can chain together by peptide bonds into longer polymers which form proteins.

Energy Landscape: The free energy available associated with a given conformation of an amino acid chain. The landscape potential is determined by the forces acting on the atoms in the given conformation, and a minimum occurs when the gradient is zero in a convex area. The landscapes for wild-type proteins under evolutionary pressure are thought to be generally smooth with few local minima and a steep global minimum representing the folded state.

CASP: The Critical Assessment of Structure Prediction competition in which different models compete to most accurately predict the unreleased structure of a provided amino acid sequence.

Folding Driving Forces: The forces that act on a peptide chain that define the energy landscape and direct folding into the native state. Forces that contribute to folding include the hydrophobic effect, van der Waals interactions and hydrogen bonding.

Solvent Accessible Surface Area (SASA): The surface area of a given amino acid that could potentially be exposed to solvent, measured in square angstroms. The relative SASA (rSASA) normalizes this value relative to the total surface area of the amino acid to give a number in the

range [0, 1]. The rSASA can be used to quantify if an amino acid is in the core; values below 10^{-2} to 10^{-3} usually indicate it is in the core, otherwise it is a surface residue.

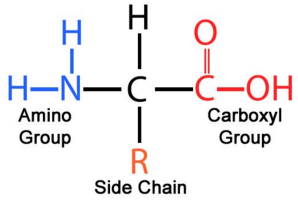
Secondary Structure: The local structure of a protein, categorized into several different motifs. These include alpha helices and beta strands, both of which are heavily stabilized by hydrogen bonds. The amino acids in secondary structures often have strict bond angle requirements compared to more disordered regions; alpha helices often have phi and psi angles of -60 and -45 respectively while the same angles for a beta sheet are closer to -135, 135.

Bond Angles: The angle made by three adjacent covalently bonded atoms which can be plotted as a histogram or probability distribution. These distributions made for proteins and their side chains have relatively small variances (RMSD of 5-10 angstroms), since such angles are determined by fixed stereochemistry.

Dihedral Angles: The angle made by the two normal vectors of the adjacent planes that are made by four covalently bonded atoms in sequence. The most relevant ones in protiens are the phi angle, which rotates around the alpha carbon-amino nitrogen bond, and the psi angle, which rotates around the alpha carbon-carbonyl carbon bond.

Ramachandran Plot: A graph showing the occurrence of pairwise phi and psi angles for a given protein or group of amino acids, typically with phi angles on the x-axis and psi angles on the y-axis. The distribution typically reveals certain clusters associated with secondary structures such as alpha helices and beta sheets.

Main Content/Topics

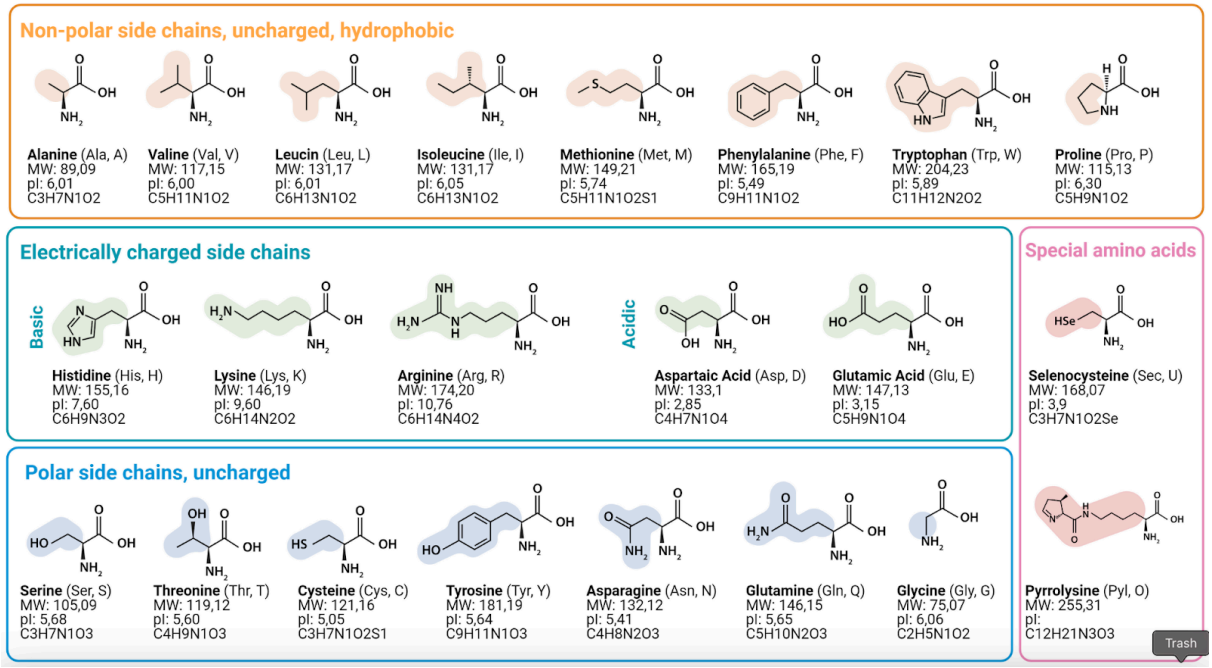


What is a protein?

They catalyze and regulate reactions in cells, give structure to cells, etc.
They are composed of amino acids (of which there are 20 naturally) and formed by the linking of these amino acids into long chains. Different amino acids have different amino acid sequences, and the sequence determines the biochemical capabilities of the protein, giving rise to sequence-dependent structure and functions. They are synthesized in an unfolded manner, so a long chain, and then fold into a structured protein. They can also be unfolded, or denatured, in adverse conditions such as high temperatures or improper pH environments.

Amino Acids (AAs):

Building blocks of proteins. There are 20 naturally occurring amino acids, but we can synthesize additional ones.All AAs have the same backbone (an amino group bonded to a central carbon (alpha carbon) bonded to a carboxyl group) but variable side chains. The different side chains are what determines the brunt of the AA's biochemistry.

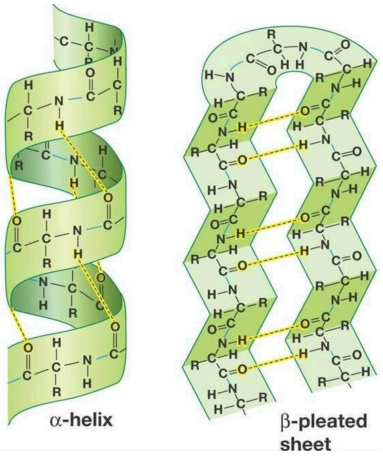


How is a protein able to fold from a long chain of AAs into a functional protein? The AAs' variable side chains give rise to disparate interactions between them and an aqueous environment, allowing the sequence of these AAs to direct how a protein is folded. Additionally, chaperone proteins and other effectors can assist in the folding process of more complicated or challenging proteins.

Levinthal's Paradox - there are too many possible folded states for the protein to fold so rapidly, and yet it folds so rapidly. The number of possibilities is ~ to the number of allowable dihedral angles raised to the number of AA's multiplied by 2.

What drives folding?

Hydrophobicity appears to drive folding, with hydrogen bonds, electrostatic forces, and van der waals forces between AAs being used to stabilize the interaction. Hydrophobicity can be measured as 0 - 1 where 0 is hydrophobic and 1 is hydrophilic. Hydrophobic AAs typically occur in the core of the protein, while hydrophilic ones arise on the surface. This leads to the idea of Solvent Accessible Surface Area, which is a measure of how much of the protein's surface area is accessible to the solvent.



Two main secondary structures arise in aqueous environments of proteins.

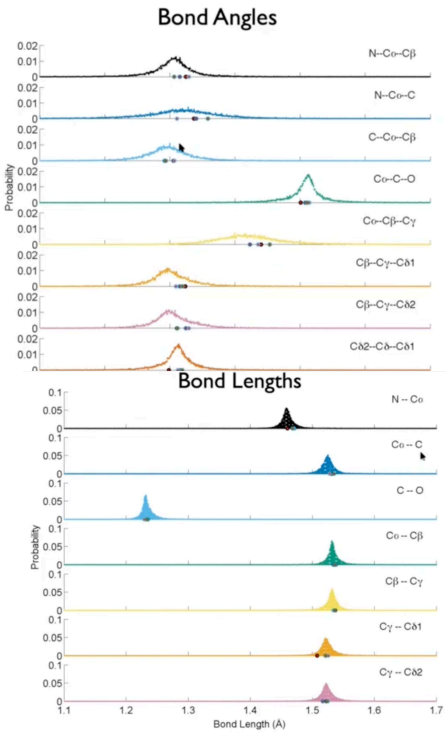
Alpha helix - right handed, three turns. There are ~3.6 AA/turn, which corresponds to roughly 100 degrees per AA. They're stabilized by hydrogen bonding. Side chains are found on the outside of the helix, typically pointing toward the N-terminus. Commonly formed by Met,

Ala, Leu, and Glu. The phi, psi angles for this structure are about (-60, -45).

Bet sheets - peptide backbone is fully extended, can run parallel or antiparallel where one strand is hydrogen bonded to the next. Commonly formed by Val, The, Tyr, Trp, Phe, Ile. The phi, psi angles for this structure are about (-135, 135).

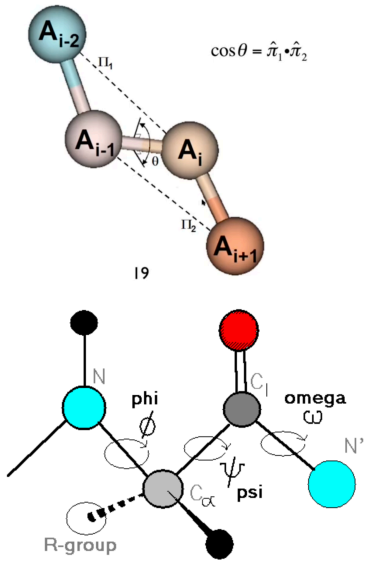
How well do we know structures?

We're approaching 1 million crystal structures in the PDB. From these datasets, we can measure the lengths of bond angles between different types of atoms. In amino acids. This information will be useful when we try to model protein folding and protein-protein interactions. The bond angles are fairly well defined around 110 degrees. They are not fixed but flexible, and dictated by the stereochemistry of the AA. The RMSD of these measurements is about 5 degrees. Similarly, we can measure the bond lengths as well from crystal structures and you can see most types of bonds are about 1.5Å in length, with some exceptions. They generally differ due to the nature of the covalent bond (single, double, etc) and are typically more fixed than bond angles.



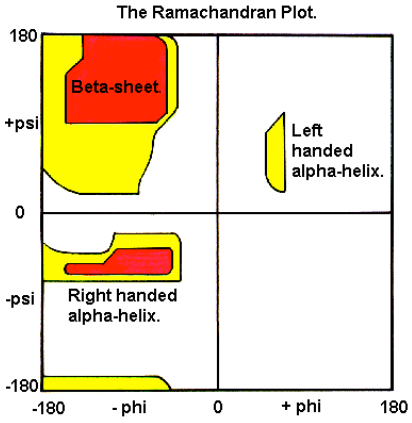
Dihedral angles

Whereas the bond angle is defined by two atoms and the bond length three, these are defined by the position of four atoms. Three atoms can define a plane, two groups of three atoms defines two planes, and the two planes generate an angle between them, which is defined as a dihedral angle. They can be calculated with backbone or side chain atoms. The *phi* angle is the angle around the -N-CA- bond (where 'CA' is the alpha-carbon). The *psi* angle is the angle around the -CA-C- bond. The *omega* angle is the angle around the -C-N- bond (i.e. the peptide bond).



Ramachandran plots

These graphs plot the phi and psi angles of atoms. Invented by Ramachandra through his use of atom modeling, they are useful in determining the structure of a protein and estimating how much it is able to move. The yellow regions correspond to greater movement (generated by using smaller atoms) while the red regions correspond to lesser movement (generated by using larger atoms). Beta sheets and alpha helices have been found to correspond to specific areas of the plot.



Discussion/Comments

Protein Folding Problem (Levinthal's Paradox): How do proteins spontaneously fold from a linear chain of amino acids into their correct structures? Theoretically, if a protein tried every possible conformation to find its most stable state, it would take longer than the age of the universe. However, in reality, proteins fold within seconds. This contradiction is known as Levinthal's Paradox. Rather than relying on random sampling, it is now believed that proteins follow multiple guided folding pathways (rough landscapes), forming intermediate structures within energy minimums that help lead to the final folded form.

Quite astonishingly, Google's AlphaFold has managed to tackle this problem quite successfully, being able to accurately predict protein folding and structure based off of amino acid sequence alone by utilizing deep learning methods, using established protein databases to recognize protein patterns and to conducting multiple sequence alignments (MSA) to better hypothesize structure.

Degrees of Freedom (DoF): The concept of DoF in molecules has many layers that work as limitations governing molecular behavior. When you start, theoretically you could have 3N possible movements (N is the number of atoms and this describes x,y,z coordinates), molecules then lose degrees of freedom through various constraints. First, bond lengths are fixed, removing N-1 degrees of freedom. Then, bond angles introduce further restrictions, reducing mobility by another N-2 degrees. Finally, dihedral angles - the rotational possibilities around bonds - further constrain the molecule, ultimately leaving just N-3 degrees of freedom. 3N−6 of those account for overall translation and rotation and would be the result of the summation of all the previous limitations. Therefore, the number of DoF would be 3N-6 and would account for all the constraints.

Suggested readings:

Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

Baldwin, R. L. Energetics of Protein Folding. *Journal of Molecular Biology* **2007**, *371* (2), 283–301. https://doi.org/10.1016/j.jmb.2007.05.078.