Protein Simulation II (April 7th):

Objectives of the Lecture

By the end of this lecture, students should be able to:

- Understand the concepts behind protein core packing.
- Explain how hard-sphere models are used to study protein structures.
- Identify key methods for predicting protein side-chain conformations and their packing in protein cores.

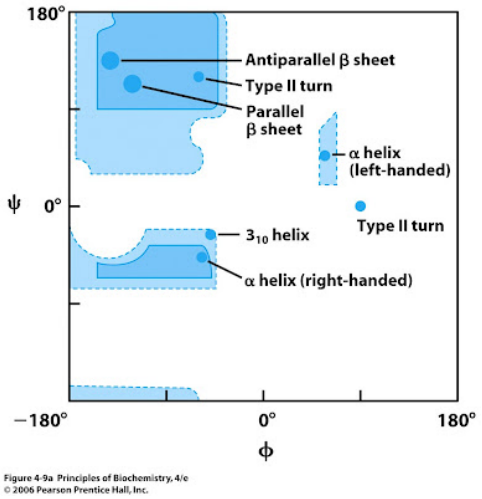
Key Concepts and Definitions

- Dihedral Angle: The angle between two planes formed by four atoms in a molecule, often used to describe the orientation of atoms in side-chain rotations of proteins.
- Ramachandran Plot: A graphical representation of the φ (phi) and ψ (psi) dihedral angles in proteins, showing allowed regions for these angles in protein structures.
- Rotamers: Different spatial conformations of a side-chain around its rotatable bonds.
- Hard Sphere Interaction: A model in which atoms or molecules are treated as spheres that cannot overlap, used to represent steric interactions in protein packing.
- Contact Distance: The minimum distance at which atoms in a protein come into close contact.
- Packing Fraction: The ratio of the volume occupied by atoms in a protein structure to the total volume available, which reflects how efficiently the protein core is packed.
- Voronoi Tessellation: A method for partitioning space into regions based on the proximity to a set of points, applied to study the distribution of atoms in a protein core and its packing efficiency.

Main Content / Topics

Overview of Protein Structure:
Proteins are made of amino acids, with the same backbone but different side chains. Each residue has three backbone torsions (φ, ψ, ω) and additional χ angles for side chains, which can be considered the degrees of freedom of a protein. Proteins generally have a central core region and a surface which interacts with water. Side chains on the surface are rotameric but not fixed, while in the core they are constrained and have definite values.

Ramachandran Plots:
These provide a graphical representation for formalising valid conformations of protein structure. The plot is a probability density function used to determine steric clashes. By plotting the φ (phi) and ψ (psi) dihedral angles on the horizontal and vertical axes respectively, regions that are populated represent common secondary structures. For example, regions for right- and left-handed alpha helices and beta structures (either single-stranded or as a linked beta sheet) can be observed. The plot can be theoretical based on 'toy' models or can be what actually happens with analysis of real protein data. Note that variation between these can be explained from the size of atoms used in modelling. For example, if the atoms were considered to be points, the whole diagram would be open due to no steric clashes.



A model Ramachandran plot

Rotamers:
These are different spatial arrangements of specific side chains based on rotation about their internal bonds. The smallest side chains have only one dihedral angle (χ_1), but the largest chains could have up to five. There are three frequently occurring possibilities for side chain angles. The number of possible rotamers of a side chain is equal to the square of its number of dihedral angles. For example, isoleucine (χ_2) has 3^2=9 rotamers, and their frequency can be plotted in a 2D grid for proteins of known structure.

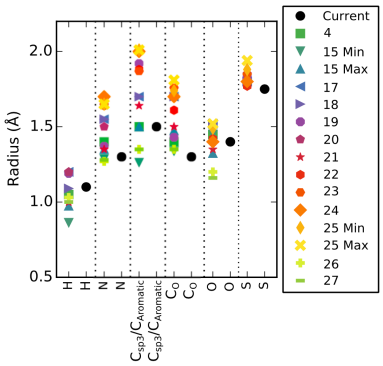
Side-chain recovery:
This is a method to determine the rotameric form of a particular side chain. It involves removing side chains from the backbone and guessing where they go based on surrounding interactions. The lowest energy form is your rotameric prediction. Note that there are different levels of 'forgetting' when trying to insert the side chain - for example, forgetting the specific amino acid makes reinserting the 'puzzle piece' much more difficult because structural conformations must also be tried. In general, methionine is difficult to predict because it is χ_3 and has low electron density, while serine is challenging because it hydrogen bonds with adjacent amino acids to localise conformation.

Energy Functions:
These are key to computational protein design software, which sample all possible conformations for dihedral angles, calculate the energy then find the lowest value. They include many terms with differing importance, but often the best method is to simply look at the experimental Ramachandran plot. Terms of an energy function include:

- Stereochemistry: Potentials that enforce equilibrium bond lengths and angles derived from small molecule crystal functions (very important)
- Repulsive / attractive Van der Waals atomic interactions (useful)
- Hydrogen bonding (generally good)
- Electrostatics and desolvation energies (more important for surface interactions)
- Disulfide bond energies (not very useful)

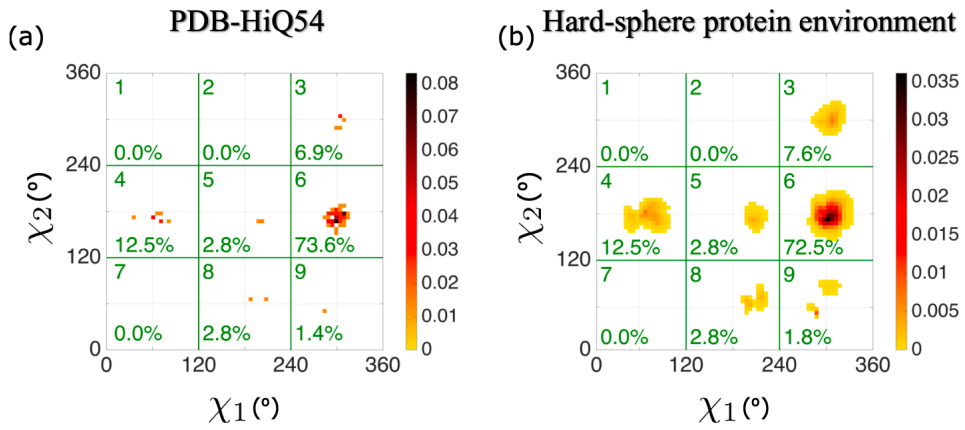
Hard Sphere Interactions:
This is an assumption for atomic interaction that assumes atoms are hard (ie non-overlapping) spheres. The contact distance is the sum of the radii of the interacting atoms. If the separation between centres is greater than the contact distance, the energy is zero, otherwise it can be calculated using the hard sphere potential equation. This can be turned into a probability using a Boltzmann weight. Note that this yields a good approximation but does not account for surrounding (inter) contacts between amino acids.

Atom Size Assumptions:
In order to have accurate modelling, we need a good approximation of atomic size. From experimental data, we have ballpark values but need to consider deviations in our simulations. Thus, the best option is to try many different values within a reasonable range. For side chain analysis, use the atomic sizes that show rotamers.



The range of experimental values for atomic radii

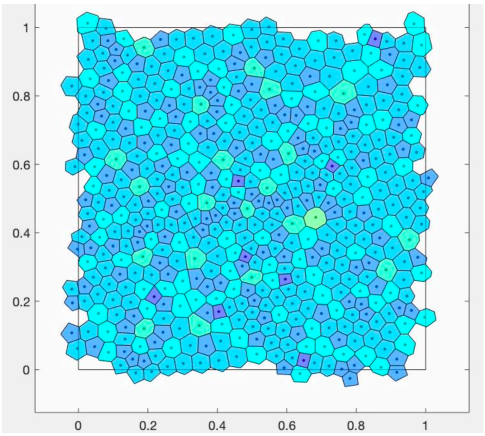
Hard Sphere Protein Environment:
This is a technique that applies the hard sphere interaction assumption to the side chain recovery problem. It gives a more precise answer than hard sphere dipeptide modelling, and gives agreement with real data to within a percentage. This can be used to optimise energy functions using machine learning methods by finding the deviation.



Experimental versus predicted rotameric probability for isoleucine

Packing Fraction:
This is the proportion of space 'occupied' inside a protein (assuming each atom is sized consistently with observed dihedral angles). In general, ordered arrangements fill better than disordered ones, and the most optimal spherical packing arrangement (face-centred cubic lattice) fills ~74% of 3D space. Protein cores have high packing densities and low compressibility.

Voronoi Partition and Tessellation:
These are methods to calculate global packing fractions. For a Voronoi partition, shape is ignored and only centres are considered. The intersections of perpendicular bisectors between all pairwise points partition space, meaning that points inside polygons are closer to the centre of their own polygon than any other. Local packing fractions can be calculated from the area of the disc divided by the area of the polygon, and their average gives the global packing fraction. A Voronoi tessellation extends this to 3D using an agglomeration of overlapping spheres.



An example of a 2D Voronoi partition

Discussion / Comments

This lecture focused on understanding how protein cores are packed and how physical models, especially hard-sphere models, can help explain and predict the structure of these densely packed regions. Protein cores are made up of tightly packed side chains, and unlike the more flexible surface regions, they are highly constrained. This makes them a good system for applying simplified physical models.

One key approach discussed was modeling atoms as hard spheres that can't overlap, which allows researchers to study side-chain conformations and predict rotamer states based on steric hindrance. Tools like Ramachandran plots help visualize backbone angles, while rotamer grids show the preferred side-chain orientations.

The lecture also covered how side-chain recovery tests how well different methods can predict where side chains belong after being removed. Energy functions used in protein design software were introduced, showing how different terms—like bond lengths, van der Waals interactions, and hydrogen bonding—are weighted to find the lowest energy conformation.

A major theme was packing efficiency, measured by the packing fraction, which tells us how much of the available space is actually filled by atoms. While perfect packing (like in crystals) reaches about 74%, protein cores usually pack closer to 56%, indicating a balance between tight packing and necessary flexibility.

Lastly, Voronoi tessellation was introduced as a method to calculate local and global packing based on how space is divided between atoms. Overall, the lecture showed how combining structural data and physical models can help us better understand the organization of protein interiors.

Suggested readings:

2021 "Artificial intelligence powers protein-folding predictions"

- This Nature article is accessible and does a good job of summarizing the significance of the AlphaFold2 for a general scientific audience.
- This article discusses the transformative impact of AI-based tools like AlphaFold2 and RoseTTAFold in predicting protein structures. It highlights how these deep-learning algorithms can accurately determine a protein's 3D shape from its amino acid sequence, a significant advancement for structural biology. Understanding these tools provides context for how computational methods are revolutionizing our ability to model protein structures, which is directly relevant to the lecture's focus on protein core packing.
- Useful subsections:
    - The discussion on AlphaFold2's performance and its impact on protein modeling is especially relevant. However, since it's more of a high-level overview, it could

be complemented with more technical content for students with a stronger background in structural biology.

2021 “Computed structures of core eukaryotic protein complexes”

- This Science article dives a little deeper into methodology and results, showing how deep learning was used to model thousands of protein-protein interactions. This actually aligned nicely with the lecturer's focus on protein packing and interaction predictions
- Useful subsections:
    - The Methods and Results sections are especially relevant. The way the authors use coevolutionary signals and integrate AlphaFold for interaction prediction connects well with the lecture’s emphasis on side-chain packing, rotamers, and computational modeling accuracy.

References ISL/ESL

There are not a lot of direct references to ISL/ESL in the content of the lecture. However, some concepts are related to those presented in the textbooks. Methods like side-chain recovery and rotamer prediction are related to supervised learning approaches (ESL and ISL Ch 2.) Unsupervised learning approaches from ISL Ch. 12 are also applicable. Protein modeling uses energy minimization in a way that is analogous to regression analysis (ESL Ch. 3 and ISL Ch. 6). Energy functions are parametrized and constrained, which mirrors the concept of regularization (ESL Ch. 5 and ISL Ch. 6). In addition, principles for model assessment and selection from ESL Ch. 7 are applied to protein structure prediction models. Protein structure prediction problems are generally high-dimensional, where the number of parameters exceeds the number of data points. Dealing with this issue is discussed in ESL Ch. 18.

Other Suggest references

1. Gaines, Jennifer C., et al. "Packing in protein cores." *Journal of Physics: Condensed Matter* 29.29 (2017): 293001.
2. Carugo, Oliviero, and Kristina Djinović-Carugo. "Half a century of Ramachandran plots." *Biological Crystallography* 69.8 (2013): 1333-1341.
3. Haddad, Yazan, Vojtech Adam, and Zbynek Heger. "Rotamer dynamics: analysis of rotamers in molecular dynamics simulations of proteins." *Biophysical journal* 116.11 (2019): 2062-2072.
4. Igashov, Ilia, et al. "VoroCNN: deep convolutional neural network built on 3D Voronoi tessellation of protein structures." *Bioinformatics* 37.16 (2021): 2332-2339.
5. Michael, Eleni, and Thomas Simonson. "How much can physics do for protein design?." *Current Opinion in Structural Biology* 72 (2022): 46-54.