Lecture Title and Date

Protein Simulation III - 04/09

Objectives of the Lecture

- Quantify differences between X-ray crystal structures and NMR structures
- Compare average properties of high-resolution x-ray crystal structures (5621) and high-quality NMR structures (6449)
- Compare NMR and x-ray crystal structures of the same protein (702 pairs).

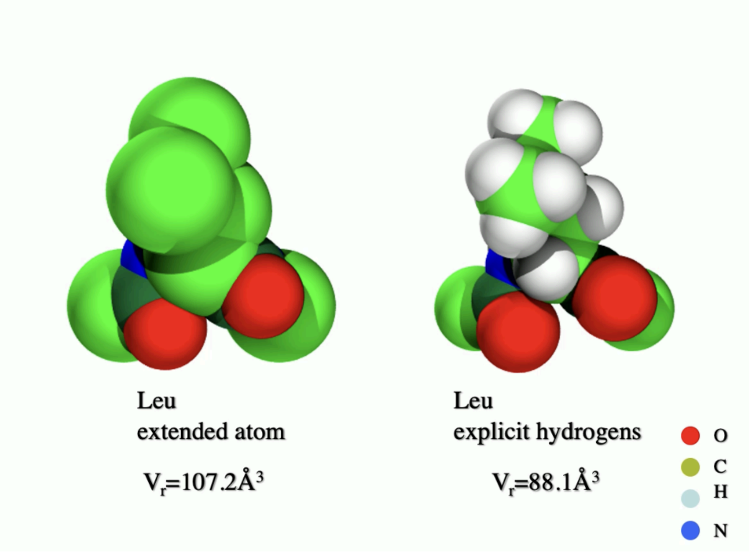
Key Concepts and Definitions

NMR: Nuclear magnetic resonance; allows us to determine the structure of a molecule in solution
X-ray crystallography: Allows us to determine the structure of a molecule; requires the molecule to crystallize to do so
Protein core: the part of a protein completely enveloped (inaccessible from the outside, doesn't touch any solvent)
Packing fraction: How much volume is occupied compared to how much volume could be occupied
Bond-orientational order: the preferred number of "neighbors" of something depends on its shape and dimensions; as order increases, the packing fraction increases

Main Content/Topics

Studies of Packing in Protein Cores

- A *packing fraction* is the ratio of volume occupied to volume possible. When drawing packing fraction boxes, we want to consider what spaces best enclose the object (i.e., define the best container).
- For amino acids, the packing fraction is the volume of the amino acids divided by the container volume we use to enclose the amino acids.
- In an extended atom model, we do not consider the hydrogen atoms—just the heavy atoms (the non-hydrogens). The extended atom model increases the size of a heavy atom until it is beyond the outskirts of the hydrogen atoms.

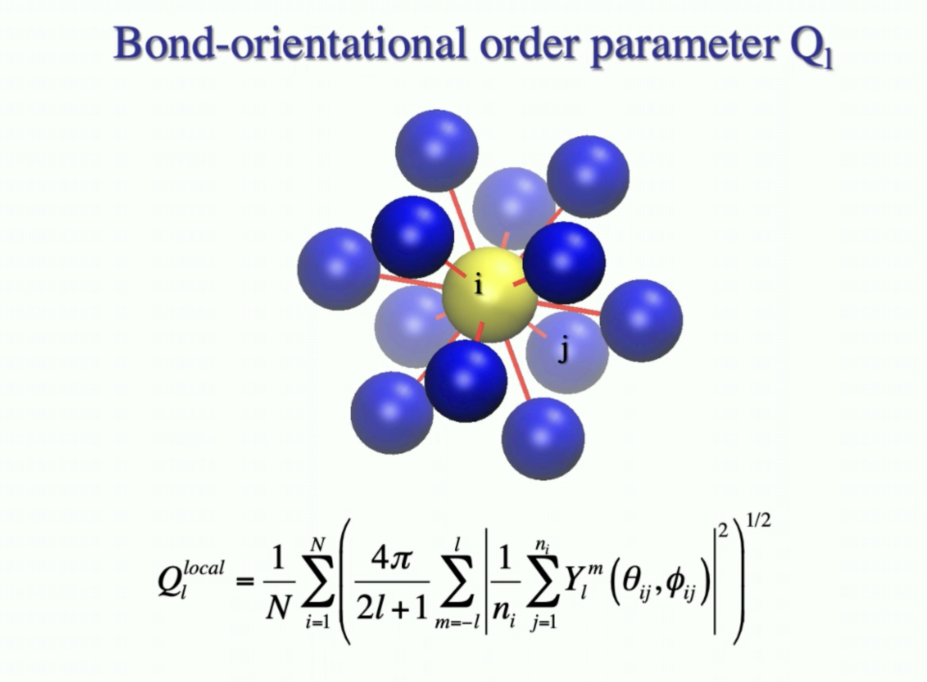
  - 

Protein Cores

- You can calculate the packing fractions of various proteins by downloading the data for them and running the calculations on your computer.
- Until the late 1980s, protein cores were thought to have packing fractions in excess of 70 percent (using extended atom methods). After using more detailed models, such as explicit hydrogen, the packing fractions found were closer to 55 percent.

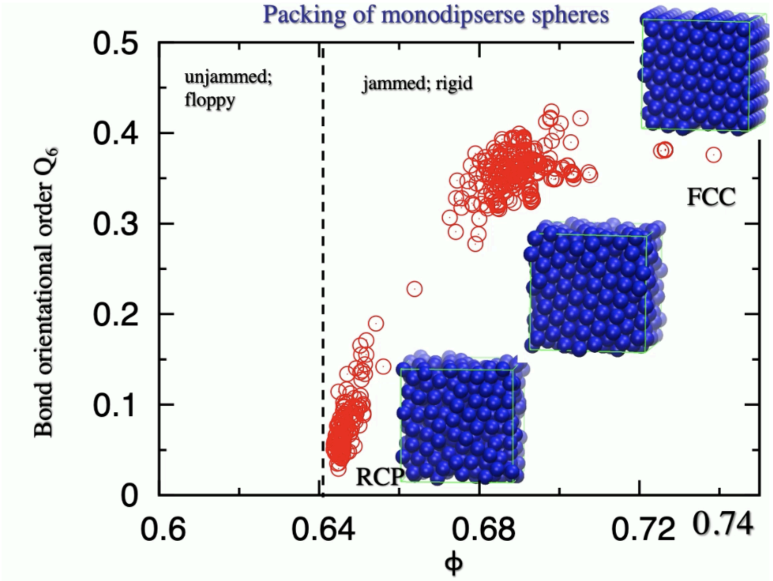
Why should the Packing Fraction of Protein Cores not be 70 percent?

- Think about it like this: packing is related to order. One type of order is bond-orientational order. For two-dimensional circles, it is preferable to have 6 neighbors. For three-dimensional spheres, having 12 neighbors (a 6-fold symmetry) is preferable. Another type of order is positional order, where the literal positions of atoms are defined along some hexagonal lattice.
- The formula below notes that we prefer a bond-orientational order in three dimensions.

- ○ 

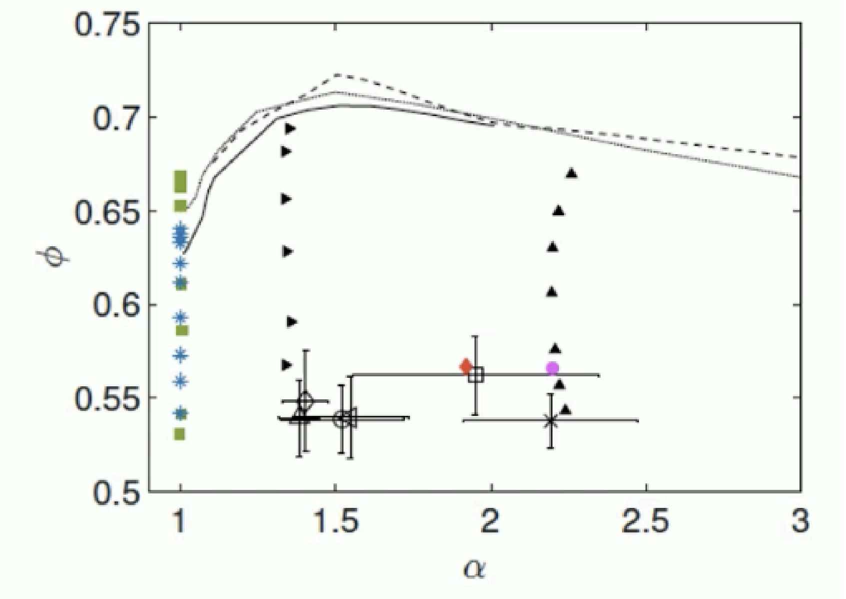
  Bond-orientational order parameter Q_l

  $$Q_l^{local} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{4\pi}{2l+1}\sum_{m=-l}^{l}\left|\frac{1}{n_i}\sum_{j=1}^{n_i}Y_l^m\left(\theta_{ij},\phi_{ij}\right)\right|^2\right)^{1/2}$$

- The general correlation between packing fraction and order is that as the level of order increases, so does the packing fraction.
  - ○ 

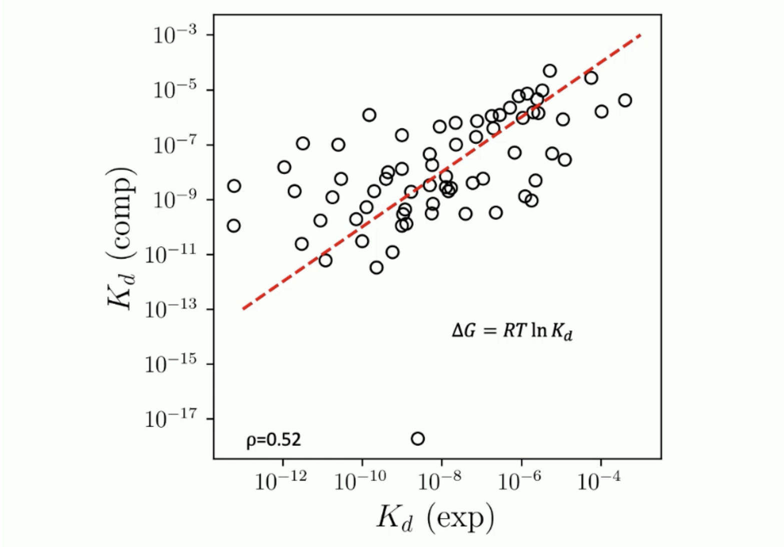
    Packing of monodipserse spheres

- Remember that when proteins crystallize, their center of mass crystallizes, not their internal structure. Individual proteins are not ordered; they have secondary structures (e.g., alpha helices, etc.), but they are not crystals.
- The next things we should consider are bumpiness and aspect ratio—these are two important qualities that impact how something packs.
  - Aspect ratio (dividing the longest and smallest dimensions) affects the packing fraction.

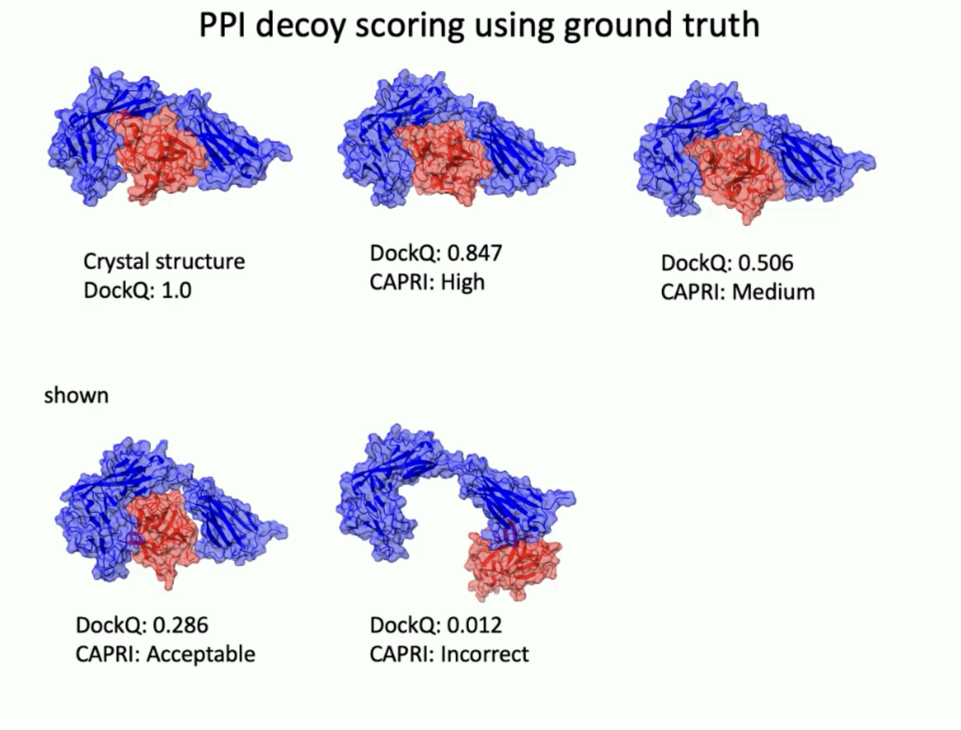
    - For bumpiness (or roughness), remember that frictionless (or smooth) objects can only constrain the contact that is normal to them. This provides fewer constraints than frictional contacts.
    - The plot below shows the packing fraction versus some aspect ratio (alpha) for frictional spheres (the blue asterisks). Amino acids are bumpy and slightly elongated; as such, they pack at roughly 55 percent. This is consistent with the number calculated through the explicit method and not the extended method.
        - 

Protein-Protein Binding & Interactions

- In the human proteome, there are roughly 25,000 distinct proteins. Each of these proteins can bind with itself (homo dimers) or with another, different protein in the proteome (hetero dimers). Thus, the number of dimers we could have is the square of 25,000, or roughly 50 million.
- The plot below shows the dissociation constant, which tells us about inverse binding affinity. Strong binders have low dissociation, so they would be found in the bottom left. The horizontal axis has experimentally measured values; the vertical axis has computationally predicted values. The correlation between these is 0.52; thus, there is some correlation, but nowhere near what would be ideal.

- ○ 

  Vangone A, Bonvin AM. Contacts-based prediction of binding affinity in protein-protein complexes. Elife. 2015 Jul 20;4:e07454. doi: 10.7554/eLife.07454.

- It is currently difficult to predict the exact bounding forms. For rigid body docking, it is possible to try all potential docking points (brute force sampling) to find the best docking. We need a ground truth score to know what the best docking is (for training the model). One metric is DockQ, where 1 means you have the right answer (best docking) and zero means you aren't close.
  - ○ 

    PPI decoy scoring using ground truth

    Crystal structure
    DockQ: 1.0

    DockQ: 0.847
    CAPRI: High

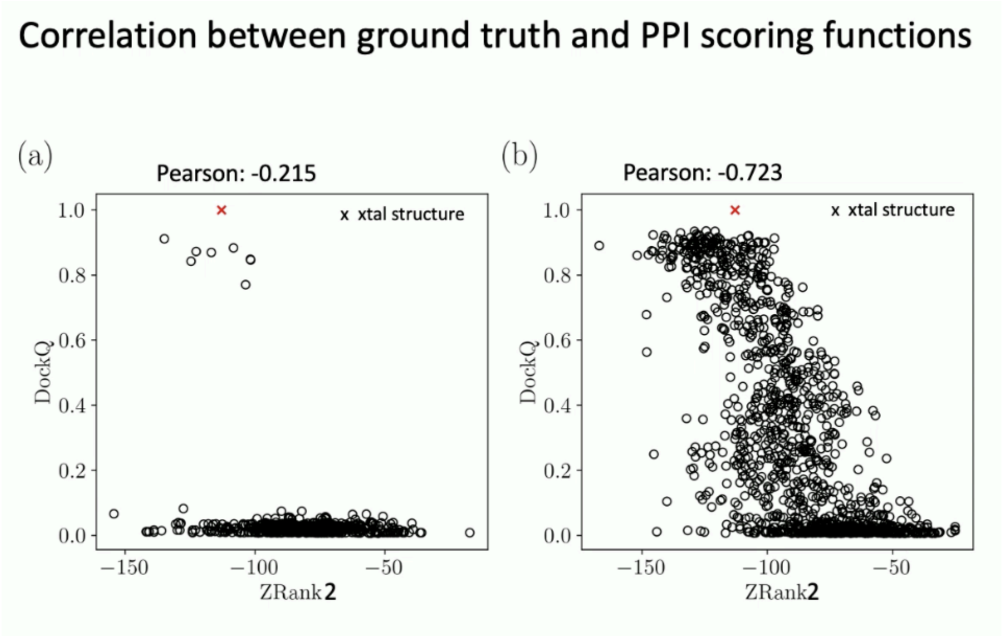
    DockQ: 0.506
    CAPRI: Medium

    shown

    DockQ: 0.286
    CAPRI: Acceptable

    DockQ: 0.012
    CAPRI: Incorrect

What is the Performance of PPI Scoring Functions on Models obtained from Bound Forms?

- In the chart, we see that there are a few good scoring functions and a lot of bad ones. The reasoning for this might be because the scoring functions get tricked (i.e., the model can assign a low score to incorrect dockings). Note that low scores are meant to indicate a more correct answer.
    - 

Correlation between ground truth and PPI scoring functions

Discussion/Comments

The lecture presents the findings that compare X-ray crystallography and NMR structures, particularly on protein packing and protein-protein interactions. The revelation that protein cores pack at approximately 55% rather than the previously assumed 70% demonstrates how methodological refinements (switching from extended atom to explicit hydrogen models) lead to more accurate structural understanding.

This 55% packing fraction aligns with physical principles of disordered systems. This is consistent with what we'd expect from amino acids, which are bumpy and slightly elongated. The lecture connects the concepts of bond-orientational order, aspect ratio, and bumpiness to explain why protein cores exhibit this specific packing density. This represents an important advancement in our structural understanding of proteins.

The protein-protein interaction discussion underlines a critical challenge in computational structural biology - the modest correlation (0.52) between experimentally measured and computationally predicted binding affinities. This limited predictive power becomes especially significant when considering the human proteome's approximately 50 million potential protein-protein interactions. The DockQ metric and visualization of various docking models illustrate why current scoring functions struggle to identify correct binding configurations consistently.

Here are several questions regarding this lecture that connect to broader implications:

1. How do the structural differences between crystal structures (used in X-ray crystallography) and solution structures (used in NMR) impact drug design strategies? The artificial constraints in crystal structures may lead to targeting binding pockets that behave differently in cellular environments.

2. Does the relatively loose packing (55%) in protein cores provide a physical explanation for the conformational flexibility essential for allosteric regulation and protein function? This could explain why proteins can undergo subtle structural changes propagating through their structure.

3. What specific computational improvements might overcome the current limitations in predicting protein-protein interactions? The lecture shows that many scoring functions perform poorly, suggesting fundamental issues in our approach.

4. How does the dynamic nature of proteins, including post-translational modifications, affect packing fraction and protein-protein interactions in ways not captured by static structural determination methods?

5. Given the computational challenges in accurately predicting protein-protein interactions, what integrated experimental-computational approaches might be most effective for mapping the interactome?

The correlation plots between ground truth and PPI scoring functions emphasize the need for improved methods to better distinguish correct binding modes from incorrect ones. The negative Pearson correlation (-0.215) shown in one scoring function indicates it's predicting the opposite of what should be expected, while the stronger negative correlation (-0.723) in another function shows more promise but still requires refinement.

The comparison between extended atom models and explicit hydrogen models is particularly insightful, as illustrated by the leucine example showing volumes of 107.2Å^3 versus 88.1Å^3, respectively. This ~18% difference in volume calculation directly impacts our understanding of packing fractions and explains why earlier studies using extended atom models incorrectly estimated protein core packing at 70%.

Another important consideration is how the relatively loose packing in protein cores might be evolutionarily advantageous. The 55% packing density likely represents an optimal structural stability and functional flexibility balance. It allows proteins to maintain their fold while accommodating the conformational changes necessary for catalysis, binding, and allosteric regulation. This balance may be disrupted in disease-causing mutations that affect core packing, potentially leading to misfolding or altered dynamics.

The challenge in predicting protein-protein interactions suggests that incorporating dynamics rather than solely relying on rigid body docking may be necessary for improving predictive models. Future approaches might benefit from integrating molecular dynamics simulations with machine learning methods to capture the conformational ensembles relevant for binding.

List all suggested reading here and please answer:

Are the readings for the class useful? If so, are the specific subsections useful or would they change?
If not, are there other references you could suggest? Please suggest one.

- There are no suggested readings for this class. I suggest three readings regarding protein packing, NMR vs. X-ray structure comparison, and protein-protein interactions.

1. Protein packing: https://jamming.research.yale.edu/files/papers/Topical_review.pdf
2. NMR vs. X-ray structure comparison: https://www.creative-biostructure.com/comparison-of-crystallography-nmr-and-em_6.htm
3. Protein-protein interaction: https://www.thermofisher.com/us/en/home/life-science/protein-biology/protein-biology-learning-center/protein-biology-resource-library/pierce-protein-methods/overview-protein-protein-interaction-analysis.html

Suggested references for many of the key concepts

1. ScienceSketch - *NMR spectroscopy visualized*
    a. https://www.youtube.com/watch?v=RZLew6Ff-JE
    b. Excellent video to better understand the workings of NMR