Introduction to Protein Structure

Function, evolution & experimental methods

Thomas Holberg Blicher
NNF Center for Protein Research
University of Copenhagen
Learning Objectives

- Outline the basic levels of protein structure.
- Identify relevant parameters for evaluating the quality of protein structures determined by X-ray crystallography and NMR spectroscopy.
- Navigate protein structures and perform simple analyses using PyMOL.
Outline

- Protein structure evolution and function
  - Inferring function from structure.
  - Modifying function

- Experimental techniques
  - X-ray crystallography
  - NMR spectroscopy

- Structure validation
Watson, Crick & DNA, 1952
"We wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest....

...It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material."

“Could the search for ultimate truth really have revealed so hideous and visceral-looking an object?” Max Perutz, 1964, on protein structure
Why are Protein Structures so Interesting?

- They provide a detailed picture of interesting biological features, such as active site, substrate specificity, allosteric regulation etc.

- They aid in rational drug design and protein engineering.

- They can elucidate evolutionary relationships undetectable by sequence comparisons.

- They can be used to put mutations in the proper structural context.
Protein Synthesis
Proteins Are Polypeptides

- A polypeptide chain
- Hydrophobic collapse
Protein Folding

- Initially formed structure is in molten globule state (ensemble).

- Molten globule condenses to native fold via transition state.
Protein Folding

- Hydrophobic collapse
  - Hydrophobic residues cluster to “escape” interactions with water.

Myoglobin

Surface

Interior
Hydrophobic vs. Hydrophilic

- Globular protein (in solution)
  - Myoglobin

- Membrane protein (in membrane)
  - Aquaporin
Hydrophobic vs. Hydrophilic

- Globular protein (in solution)
- Membrane protein (in membrane)

Cross-section

Myoglobin

Aquaporin
Polar backbone groups form regular secondary structure to satisfy hydrogen bonding donors and acceptors.
Structure Levels

- Primary structure = Sequence
- Secondary Structure = Helix, sheets/strands, loops & turns
- Structural Motif = Small, recurrent arrangement of secondary structure, e.g.
  - Helix-loop-helix
  - Beta hairpins
  - EF hand (calcium binding motif)
  - Etc.
- Tertiary structure = Arrangement of Secondary structure elements

MSSVLLGHIKKLEMGHS...
Quaternary Structure

- Assembly of monomers/subunits into protein complex
  - Backbone-backbone, backbone-side-chain & side-chain-side-chain interactions:
    - Intramolecular vs. intermolecular contacts.
    - For ligand binding side chains may or may not contribute. For the latter, mutations have little effect.

- Myoglobin

- Hemoglobin
- Between helices and sheets
- On protein surface
- Intrinsically “unstructured”
Residue Patterns

- Helices
  - Helix capping
  - Amphiphilic residue patterns

- Sheets
  - Amphiphilic residue patterns
  - Residue preferences at edges vs. middle

- Special residues
  - Proline
    - Helix breaker
  - Glycine
    - In turns/loops/bends
The Peptide Bond

- The peptide bond
- A polypeptide chain
Ramachandran Plot

- Allowed backbone torsion angles in proteins
Ramachandran Plots
Comparing Proteins

- Comparison to proteins with known function
  - By structural similarity (convergent evolution)
  - By sequence similarity (divergent evolution)

- **Sequence alignment vs. structure alignment**
  - Evolution vs. physics/chemistry(/topology)
The structure of a protein is uniquely determined by its amino acid sequence (but sequence is sometimes not enough):
- prions
- pH, ions, cofactors, chaperones

In evolution structure is conserved much longer than both function and sequence.
- Structure > Function > Sequence
Grouping Amino Acids

http://www.dreamingintechnicolor.com/InfoAndIdeas/AminoAcids.gif
The Evolution Way

- Based on Blosum62 matrix (sequence alignments)
- Measure of evolutionary substitution probability
Form vs. Function

- Divergent evolution
  - Common ancestor
  - New function

- Convergent evolution
  - Different ancestor
  - Same function
Divergent Evolution

- Trypsin
  - positive
- Chymotrypsin
  - large hydrophobic
- Elastase
  - Small hydrophobic

- Divergent evolution
  - Same fold
  - Different specificities
  - Small changes in binding pocket
Convergent Evolution

- Trypsin
- Subtilisin

Convergent evolution of function
Rhamnogalacturonan acetylesterase (A. aculeatus) (1K7C)

Platelet activating factor acetylhydrolase (B. taurus) (1WAB)

Serine esterase (S. scabies) (1ESC)
Rhamnogalacturonan acetyesterase

Serine esterase

Platelet activating factor acetylhydrolase

Inferring biological features from the structure

Topological switchpoint
Engineering & Design

- **Protein engineering**
  - Overpacking
  - Buried polar groups
  - Cavities

- **Drug design**
  - Target specificity/selectivity
  - Function
  - Mutations

**COX-1/COX-2**
- Arthritis
- Designed to prevent drug side effects

http://publications.nigms.nih.gov/structlife/chapter4.html

**HIV protease**

Blundell et al. (2002), *High-throughput crystallography for lead discovery in drug design*, Nature Reviews Drug Discovery 1, 45-54.

---

A Rare Case

A minimal sequence code for switching protein structure and function.

1 sequence + 1 mutation = 2 folds
Experimental Methods

Crystallography & NMR spectroscopy
Methods for Structure Determination

- X-ray crystallography
- Nuclear Magnetic Resonance (NMR)
- Modelling techniques

- More exotic techniques
  - Cryo electron microscopy (Cryo EM)
  - Small angle X-ray scattering (SAXS)
  - Neutron scattering
X-ray Crystallography

- No size limitation.
- Protein molecules are "stuck" in a crystal lattice.
- Some proteins seem to be uncrystallizable.
- Slow.

- Lattice and unit cell

- Especially suited for studying structural details.
X-rays → Fourier transform
The Importance of Resolution

The resolution of a structure is a measure of the level of detail that can be discerned. A higher resolution indicates a more detailed and accurate structure. Here are examples of structures at different resolutions:

- **4 Å**: Low resolution, showing a general shape with less detail.
- **3 Å**: Slightly higher resolution, providing more detail than 4 Å.
- **2 Å**: Moderate resolution, offering a good balance between detail and clarity.
- **1 Å**: High resolution, providing the most detailed view, allowing for precise structural analysis.

The rightmost image illustrates a highly detailed structure at 0.5 Å resolution, showcasing the pinnacle of structural clarity possible through advanced imaging techniques.
Key Parameters

- **Resolution**
- **R values**
  - Agreement between data and model.
  - Usually between 0.15 and 0.25, should not exceed 0.30.
    - $R + 0.05 > R_{\text{free}} > R$.
- **Ramachandran plot**
- **B factors**
  - Contributions from static and dynamic disorder
    - Well determined $\sim$10-20 Å$^2$, intermediate $\sim$20-30 Å$^2$, flexible 30-50 Å$^2$, invisible $>60$ Å$^2$. 
NMR Basics

- NMR is
  - nuclear magnetic resonance
  - done on proteins **IN SOLUTION**
  - especially suited for studies of protein **dynamics** and **folding**
  - slow!

- Only certain atoms can be detected: $^1\text{H}$, $^{13}\text{C}$, $^{15}\text{N}$

- Proteins must be
  - below **50 kDa**
  - **stable** at high concentration (0.5-1mM) @ room temperature
Well-defined structures
RMSDs < 0.6 Å

Less well-defined structures
RMSDs > 0.6 Å

Atomic backbone RMSD:

$$RMSD = \sqrt{\frac{\sum_{1}^{n}(x_i - \langle x'_i \rangle)^2}{n}}$$

1T1H, Andersen et al. JBC, 2004

3GF1, Cooke et al. Biochemistry, 1991
What regions in the structure are most well-defined?

Look at the pdb ensembles to see which regions are well-defined

1RJH

Nielbo et al, Biochemistry, 2003
Summary I – Protein Structure

- Proteins consist of amino acids.
- Polypeptide chains fold into specific 3D structures.
- Function is performed by the folded protein.
- Proteins are dynamic and only marginally stable.

Image adapted from: National Human Genome Research Institute.
Summary – Part I

- In evolution **structure** is conserved longer than both **function** and **sequence**.

- **X-ray crystallography**
  - Proteins of any size
  - Proteins in crystal
  - Complete data/total map of structure
  - Many details – one model
  - Resolution, R-values, Ramachandran plot

- **NMR spectroscopy**
  - Proteins below 50 kDa
  - Proteins in solution
  - Incomplete data
  - Fewer details – many models
  - Restraint violations, RMSD, Ramachandran plot
Summary – Part II

- Analysis aims at
  - Understanding function
    - Modifying function
  - Finding relevant examples

- Can use
  - Structure
  - Sequence predictions
  - Evolutionary information

- Can be qualitative or quantitative
The Protein Data Bank

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>X-ray</td>
<td>13116</td>
<td>30860</td>
<td>74567</td>
</tr>
<tr>
<td>NMR</td>
<td>2451</td>
<td>5368</td>
<td>9605</td>
</tr>
<tr>
<td>Other</td>
<td>338</td>
<td>200</td>
<td>674</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>15905</strong></td>
<td><strong>36428</strong></td>
<td><strong>84846</strong></td>
</tr>
</tbody>
</table>

The PDB also contains nucleotide and nucleotide analogue structures.
ATOM  1    N    THR A 1  25.200  26.068  37.670  1.00  25.43 N
ATOM  2   CA   THR A 1  26.443  26.547  37.135  1.00  16.70 C
ATOM  3    C    THR A 1  27.568  25.589  37.431  1.00  13.12 C
ATOM  4    O    THR A 1  27.577  25.073  38.554  1.00  15.92 O
ATOM  5   CB   THR A 1  26.745  27.891  37.843  1.00  20.41 C
ATOM  6   OG1   THR A 1  25.564  28.674  37.550  1.00  26.40 O
ATOM  7  CG2   THR A 1  27.995  28.594  37.359  1.00  22.25 C
<table>
<thead>
<tr>
<th>COLUMNS</th>
<th>DATA TYPE</th>
<th>FIELD</th>
<th>DEFINITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - 6</td>
<td>Record name</td>
<td>&quot;ATOM&quot;</td>
<td></td>
</tr>
<tr>
<td>7 - 11</td>
<td>Integer</td>
<td>serial</td>
<td>Atom serial number.</td>
</tr>
<tr>
<td>13 - 16</td>
<td>Atom</td>
<td>name</td>
<td>Atom name.</td>
</tr>
<tr>
<td>17</td>
<td>Character</td>
<td>altLoc</td>
<td>Alternate location indicator.</td>
</tr>
<tr>
<td>18 - 20</td>
<td>Residue name</td>
<td>resName</td>
<td>Residue name.</td>
</tr>
<tr>
<td>22</td>
<td>Character</td>
<td>chainid</td>
<td>Chain identifier.</td>
</tr>
<tr>
<td>23 - 26</td>
<td>Integer</td>
<td>resSeq</td>
<td>Residue sequence number.</td>
</tr>
<tr>
<td>27</td>
<td>AChar</td>
<td>iCode</td>
<td>Code for insertion of residues.</td>
</tr>
<tr>
<td>31 - 38</td>
<td>Real(8.3)</td>
<td>x</td>
<td>Orthogonal coordinates for X in Angstroms</td>
</tr>
<tr>
<td>39 - 46</td>
<td>Real(8.3)</td>
<td>y</td>
<td>Orthogonal coordinates for Y in Angstroms</td>
</tr>
<tr>
<td>47 - 54</td>
<td>Real(8.3)</td>
<td>z</td>
<td>Orthogonal coordinates for Z in Angstroms</td>
</tr>
<tr>
<td>55 - 60</td>
<td>Real(6.2)</td>
<td>occupancy</td>
<td>Occupancy.</td>
</tr>
<tr>
<td>61 - 66</td>
<td>Real(6.2)</td>
<td>tempFactor</td>
<td>Temperature factor.</td>
</tr>
<tr>
<td>77 - 78</td>
<td>LString(2)</td>
<td>element</td>
<td>Element symbol, right-justified.</td>
</tr>
<tr>
<td>79 - 80</td>
<td>LString(2)</td>
<td>charge</td>
<td>Charge on the atom.</td>
</tr>
</tbody>
</table>
Protein Structure Visualisation

Introduction to PyMOL
What is PyMOL?

- Open-source molecular viewing program

http://www.pymol.org
Benefits

- It’s free!
  - For academia…
  - …not for industry.
  - Version 0.99rc6
  - Users can always compile the latest version.

- Pay to get support, manual, latest version etc.
Potential Weaknesses

- Few!
- Not a fully integrated modelling environment.
- Not fully developed for experimental structure determination/fitting.
- Mostly for qualitative analyses.
- No undo function…
Selections & Objects

- Every molecule (pdb file) is an object.
- Selections refer to objects
  - Make smaller or composite objects
- Changes in representation can affect objects or selections.
Links

- PDB (protein structure database)
  www.pdb.org/

- PyMOL home:
  http://www.pymol.org/

- PyMOL manual:
  http://pymol.sourceforge.net/newman/user/toc.html

- PyMOL Wiki:
  http://www.pymolwiki.org/index.php/Main_Page
Installation

- Linux
- Windows
- Mac OS X
  - Special behaviour
  - Make a copy of the executable and rename to get plugin menu!
    - MacPyMOL → MacPyMOL copy
      → MacPyMOLX11hybrid