Introduction to Protein Structure

Function, evolution & experimental methods

Thomas Holberg Blicher
Novozymes A/S

Who’s the Teacher?

- Thomas Holberg Blicher
- Ph.D. in protein crystallography and immunology.
- Areas of interest:
  - Protein structure and function relationships
  - Drug design
  - Immunology
- Previously associate professor at CPR-KU/CBS-DTU.
- At Novozymes A/S since June 2014.
Learning Objectives

- Outline the basic levels of protein structure.
- Outline key differences between X-ray crystallography and NMR spectroscopy.
- Identify relevant parameters for evaluating the quality of protein structures determined by X-ray crystallography and NMR spectroscopy.
- Navigate a protein structure using the program PyMOL.

Outline

- Protein structure evolution and function
  - Inferring function from structure.
  - Modifying function
- Experimental techniques
  - X-ray crystallography
  - NMR spectroscopy
  - Structure validation
"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."

Once Upon a Time…

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

John Kendrew, 1959, with myoglobin model

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.

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

- Myoglobin
  - Surface
  - Interior
### Hydrophobic vs. Hydrophilic

<table>
<thead>
<tr>
<th>Globular protein (in solution)</th>
<th>Membrane protein (in membrane)</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Myoglobin" /></td>
<td><img src="image2" alt="Aquaporin" /></td>
</tr>
</tbody>
</table>

### Hydrophobic vs. Hydrophilic

<table>
<thead>
<tr>
<th>Globular protein (in solution)</th>
<th>Membrane protein (in membrane)</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image3" alt="Cross-section" /></td>
<td><img src="image4" alt="Cross-section" /></td>
</tr>
<tr>
<td>Myoglobin</td>
<td>Aquaporin</td>
</tr>
</tbody>
</table>
Backbone Problems?

- Polar backbone groups form regular secondary structure to satisfy hydrogen bonding donors and acceptors.

Turns, Loops & Bends

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

Quaternary Structure

- **Assembly of monomers/subunits into protein complex**
  - Backbone-backbone, backbone-side-chain & side-chain-side-chain interactions:
    - Intramolecular vs. intermolecular contacts.
    - Mutations have (relatively) little effect on individual subunits but potentially large effect on complex formation.
- **Myoglobin**
- **Hemoglobin**
The Peptide Bond

- The peptide bond
- A polypeptide chain

Ramachandran Plot

- Allowed backbone torsion angles in proteins

The Ramachandran Ploot.
Ramachandran Plots

X-ray structure – good data.  NMR structure – low quality data...

Small Exercise (5 minutes)

- For the polypeptide on the right discuss the following with your neighbour:
  - Why is the lower right quadrant a "forbidden" region in the Ramachandran plot?
  - What makes Gly a special amino acid when it comes to Ramachandran plots?
  - What about Pro?
Hunting for Function in Structures

- Look what nature has done.
- Analyse and compare!
- Find more examples.

Comparing Proteins

- Analysis = 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
Structure & Evolution

- 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

A Rare Case

1 sequence + 1 mutation = 2 folds

A minimal sequence code for switching protein structure and function.
Grouping Amino Acids

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

http://palaeo.gly.bris.ac.uk/palaeofiles/marsupials/index.htm
Convergent Evolution

- Trypsin
- Subtilisin

Convergent evolution of function

Structure & Evolution

- Rhamnogalacturonan acetylesterase (A. aculeatus) (1K7C)
- Platelet activating factor acetylhydrolase (B. taurus) (1WAB)
- Serine esterase (S. scabies) (1ESC)
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.

- Especially suited for studying structural details.

Lattice and unit cell

X-rays

Fourier transform
The Importance of Resolution

Model vs. Data Agreement

- Individual reflections
  \[ I_{hkl} \propto |F_{obs}(hkl)|^2 \]

- R-factor:
  \[ R = \frac{\sum |F_{obs} - |F_{calc}|}{\sum |F_{obs}|} \]

- \( R_{free} \):
  - Like R-factor
  - Unbiased measure.
  - Calculated on 5-10% of data not included in refinement.
Key Parameters

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

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
NMR Spectroscopy

Well-defined structures
RMSDs < 0.6 Å

Less well-defined structures
RMSDs > 0.6 Å

Evaluation of NMR Structures

- Atomic backbone RMSD:

  1T1H, Andersen et al. JBC, 2004

  3GF1, Cooke et al. Biochemistry, 1991
RMSD

- Root mean square deviation (or distance).
  (Sometimes just RMS)
- Pairwise comparison of structures.

\[ RMSD = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_{1,i} - x_{2,i})^2} \]

**RMSD Example**

\[ x_1-y_1 = 0.8\text{Å}, \quad x_2-y_2 = 1\text{Å}, \quad x_3-y_3 = 1.2\text{Å}, \quad x_4-y_4 = 2\text{Å} \]

\[ RMSD = 1.33\text{Å} \]
Evaluation of NMR Structures

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.
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
PDB
The Protein Structure Database

Protein Data Bank

- [http://www.rcsb.org/](http://www.rcsb.org/)
- Contents
- File structure
  - Types of structures
- Structure reports & summaries
- Quality check
- Searching
- Molecule of the Month

Free app!
The PDB also contains nucleotide and nucleotide analogue structures.

## PDB Growth 1971-2013

<table>
<thead>
<tr>
<th>Year</th>
<th>X-ray</th>
<th>NMR</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sep. 2001</td>
<td>13116</td>
<td>2451</td>
<td>338</td>
<td>15905</td>
</tr>
<tr>
<td>May 2006</td>
<td>30860</td>
<td>5368</td>
<td>200</td>
<td>36428</td>
</tr>
<tr>
<td>Mar. 2014</td>
<td>86817</td>
<td>10324</td>
<td>976</td>
<td>98117</td>
</tr>
</tbody>
</table>

## PDB File Header

```
HEADER    IMMUNE SYSTEM                           16-AUG-04   1X7Q
TITLE     CRYSTAL STRUCTURE OF HLA-A*1101 WITH SARS NUCLEOCAPSID
TITLE    2 PEPTIDE
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: HLA CLASS I HISTOCOMPATIBILITY ANTIGEN, A-11 ...
REMARK   3 DATA USED IN REFINEMENT.
REMARK   3 RESOLUTION RANGE HIGH (ANGSTROMS) : 1.45
REMARK   3 RESOLUTION RANGE LOW (ANGSTROMS) : 30.00 ...
SEQRES   1 C   9 LYS THR PHE PRO PRO THR GLU PRO LYS
HET      SO4  B4001       5 ...
HETNAM    SO4 SULFATE ION ...
HELIX    1 1 ALA A   49 GLU A   53   5 5 ...
CRYST1   58.465  80.623   56.570  90.00 116.79  90.00 P 1 21 1 2 ...
```
### PDB File Atom Coordinates

<table>
<thead>
<tr>
<th>ATOM</th>
<th>N</th>
<th>THR</th>
<th>1</th>
<th>25.200</th>
<th>26.068</th>
<th>37.670</th>
<th>1.00</th>
<th>25.43</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATOM</td>
<td>CA</td>
<td>THR</td>
<td>1</td>
<td>26.443</td>
<td>26.547</td>
<td>37.135</td>
<td>1.00</td>
<td>16.70</td>
<td>C</td>
</tr>
<tr>
<td>ATOM</td>
<td>C</td>
<td>THR</td>
<td>1</td>
<td>27.568</td>
<td>25.589</td>
<td>37.431</td>
<td>1.00</td>
<td>13.12</td>
<td>C</td>
</tr>
<tr>
<td>ATOM</td>
<td>O</td>
<td>THR</td>
<td>1</td>
<td>27.577</td>
<td>25.073</td>
<td>38.554</td>
<td>1.00</td>
<td>15.92</td>
<td>O</td>
</tr>
<tr>
<td>ATOM</td>
<td>CB</td>
<td>THR</td>
<td>1</td>
<td>26.745</td>
<td>27.891</td>
<td>37.843</td>
<td>1.00</td>
<td>20.41</td>
<td>C</td>
</tr>
<tr>
<td>ATOM</td>
<td>OG1</td>
<td>THR</td>
<td>1</td>
<td>25.564</td>
<td>28.674</td>
<td>37.550</td>
<td>1.00</td>
<td>26.40</td>
<td>O</td>
</tr>
<tr>
<td>ATOM</td>
<td>CG2</td>
<td>THR</td>
<td>1</td>
<td>27.995</td>
<td>28.594</td>
<td>37.359</td>
<td>1.00</td>
<td>22.25</td>
<td>C</td>
</tr>
<tr>
<td>HETATM</td>
<td>O3</td>
<td>GOL</td>
<td>B5007</td>
<td>23.356</td>
<td>27.459</td>
<td>10.492</td>
<td>1.00</td>
<td>29.71</td>
<td>O</td>
</tr>
<tr>
<td>HETATM</td>
<td>O</td>
<td>HOH</td>
<td>A5007</td>
<td>19.520</td>
<td>54.389</td>
<td>40.841</td>
<td>1.00</td>
<td>10.97</td>
<td>O</td>
</tr>
<tr>
<td>ATOM</td>
<td>N</td>
<td>GLY</td>
<td>1</td>
<td>9.360</td>
<td>23.828</td>
<td>42.512</td>
<td>1.00</td>
<td>45.34</td>
<td>N</td>
</tr>
<tr>
<td>ANISOU</td>
<td>N</td>
<td>GLY</td>
<td>1</td>
<td>6582</td>
<td>3670</td>
<td>6975</td>
<td>2086</td>
<td>1854</td>
<td>1681</td>
</tr>
<tr>
<td>ATOM</td>
<td>CA</td>
<td>GLY</td>
<td>1</td>
<td>8.104</td>
<td>24.285</td>
<td>41.984</td>
<td>1.00</td>
<td>33.31</td>
<td>C</td>
</tr>
<tr>
<td>ANISOU</td>
<td>CA</td>
<td>GLY</td>
<td>1</td>
<td>5450</td>
<td>1898</td>
<td>5309</td>
<td>-137</td>
<td>1446</td>
<td>394</td>
</tr>
<tr>
<td>ATOM</td>
<td>C</td>
<td>GLY</td>
<td>1</td>
<td>7.740</td>
<td>25.713</td>
<td>42.380</td>
<td>1.00</td>
<td>28.54</td>
<td>C</td>
</tr>
<tr>
<td>ANISOU</td>
<td>C</td>
<td>GLY</td>
<td>1</td>
<td>4079</td>
<td>1849</td>
<td>4915</td>
<td>-6</td>
<td>728</td>
<td>561</td>
</tr>
</tbody>
</table>

### Protein Structure Visualisation

Introduction to PyMOL
Example 1

- **mRNA**
- **Large subunit:** Catalytic mechanism
- **Small subunit:** Specificity

**Protein**

Example 2

- "Leci n'est pas une pipe.

René Magritte
What is PyMOL?

- Open-source molecular viewing program

http://www.pymol.org

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.

PyMOL

- Representations
  - Lines, sticks, ribbon, spheres, cartoon(s)
- Surfaces
  - Transparency, quality
- Ray-tracing (rendering)
  - Modes
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