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.

- 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.
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.
Protein Synthesis

- newly born protein
- amino acids
- large subunit
- tRNA
- mRNA
- small subunit
- P site
- A site
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)
  - Myoglobin

- Membrane protein (in membrane)
  - Aquaporin

Cross-section
Backbone Problems?

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

- Aligned peptide units → Dipolar moment
- Ion/ligand binding
- Secondary and quaternary structure packing
- Capping residues
- The α helix (i→i +4)
- Other helix types! (3_{10}, π)
β-Sheets

- Multiple strands → sheet
  - Parallel vs. antiparallel
  - Twist

- Flexibility
  - Vs. helices
  - Folding
  - Structure propagation (amyloids)
  - Other…

Thioredoxin
β-Sheets

- Multiple strands → sheet
  - Parallel vs. antiparallel
  - Twist

- Flexibility
  - Vs. helices
  - Folding
  - Structure propagation (amyloids)
  - Other…
β-Sheets

- Multiple strands → sheet
  - Parallel vs. antiparallel
  - Twist

- Flexibility
  - Vs. helices
  - Folding
  - Structure propagation (amyloids)
  - Other…
**β-Sheets**

- Multiple strands $\rightarrow$ sheet
  - Parallel vs. antiparallel
  - Twist

- Flexibility
  - Vs. helices
  - Folding
  - Structure propagation (amyloids)
  - Other…
**β-Sheets**

- Multiple strands → sheet
  - Parallel vs. antiparallel
  - Twist

- Strand interactions are **non-local**

- Flexibility
  - Vs. helices
  - Folding
Turns, Loops & Bends

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

MSSVLLGH I KKLEMGS...
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
Structures of Unstructured Regions

- Estimate: 20% of all proteins contain unstructured regions.
  - 1% of structures in PDB contain unstructured regions.

- Structural genomics
  - Special structural genomics projects
  - Selection and modification of targets
  - Prediction of crystallisable domains

Protein disorder publications in PubMed

Iakoucheva & Dunker, *Structure* 2003
Degrees of Structure

- **Unstructured** (conformational ensemble): For example, ACTR (no NCBD)
- **Molten globule** (conformational ensemble): For example, NCBD (no ACTR)
- **Linked folded domains** (based on a string): For example, zinc fingers (no DNA)
- **Mostly folded, local disorder**: For example, eIF4E (N terminus is unfolded)

**Folding on target binding**

- ACTR-NCBD complex
- Zinc-finger-1-3-DNA complex
- eIF4E-eIF4G complex
What’s the Fuss About?

- Properties of Disordered Regions
  - Flexible, i.e. adaptable
  - Accessible
    - Contain Extended Linear Motifs (ELM)
  - Different behaviour in interaction interfaces
    - Very adaptable
    - Many hydrophobic interactions (close packing)
  - No fixed structure without interaction partner
  - Folding upon binding
Proteins Are Polypeptides

- The peptide bond

- A polypeptide chain
Ramachandran Plot

- Allowed backbone torsion angles in proteins

The Ramachandran Plot.

Beta-sheet.

Left handed alpha-helix.

Right handed alpha-helix.

Peptide torsion angles.
Torsion Angles

\[ \phi = 180^\circ \]  \[ \phi = 0^\circ \]  
\[ \phi = +90^\circ \]  \[ \phi = -90^\circ \]

© O. S. Smart, 1995
Ramachandran Plots

X-ray structure – good data.

NMR structure – low quality data…
The Amino Acids

Acidic and amide side chains
- Aspartate
- Asparagine
- Glutamate
- Glutamine

Basic side chains
- Lysine
- Histidine
- Arginine

Aliphatic side chains
- Valine
- Isoleucine
- Glycine
- Alanine
- Leucine

Aromatic side chains
- Tryptophan
- Phenylalanine
- Tyrosine

Hydroxyl or sulfur-containing side chains
- Serine
- Methionine
- Threonine
- Cysteine

Cyclic side chain
- Proline

http://www.ch.cam.ac.uk/magnus/molecules/amingo/
Grouping Amino Acids

Amino Acids

A alanine (ala)
R arginine (arg)
N asparagine (asn)
D aspartic acid (asp)
C cysteine (cys)
Q glutamine (gln)
E glutamic acid (glu)
G glycine (gly)
H histidine (his)
I isoleucine (ile)
L leucine (leu)
K lysine (lys)
M methionine (met)
P phenylalanine (phe)
P proline (pro)
S serine (ser)
T threonine (thr)
W tryptophan (trp)
Y tyrosine (tyr)

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

- Based on Blosum62 matrix
- Measure of evolutionary substitution probability
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
Form vs. Function

- Divergent evolution
  - Common ancestor
  - New function

- Convergent evolution
  - Different ancestor
  - Same function
Sequence vs. Function – I

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

- Divergent evolution
  - Same fold
  - Different specificities
  - Small changes in binding pocket
Sequence vs. Function – II

- Trypsin
- Subtilisin

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

- HIV protease
  

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

1 Å
2 Å
3 Å
4 Å

low

high

0.5 Å
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 Å², intermediate $\sim$20-30 Å², flexible 30-50 Å², invisible $>60$ Å².
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 (0.5-1 mM) @ room temperature
NMR Spectroscopy
Evalutation of NMR Structures

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

Well-defined structures
RMSDs < 0.6 Å

1T1H, Andersen et al. JBC, 2004

Less well-defined structures
RMSDs > 0.6 Å

3GF1, Cooke et al. Biochemistry, 1991
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.

Image adapted from: National Human Genome Research Institute.
Summary

- 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
PDB
The Protein Structure Database
Protein Data Bank

- http://www.rcsb.org/
- Contents
- File structure
  - Types of structures
- Structure reports & summaries
- Quality check
- Searching
- Molecule of the Month
Holdings of the Protein Data Bank (PDB):

<table>
<thead>
<tr>
<th></th>
<th>Sep. 2001</th>
<th>May 2006</th>
<th>Nov. 2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-ray</td>
<td>13116</td>
<td>30860</td>
<td>60342</td>
</tr>
<tr>
<td>NMR</td>
<td>2451</td>
<td>5368</td>
<td>8666</td>
</tr>
<tr>
<td>Other</td>
<td>338</td>
<td>200</td>
<td>502</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>15905</strong></td>
<td><strong>36428</strong></td>
<td><strong>69510</strong></td>
</tr>
</tbody>
</table>

The PDB also contains nucleotide and nucleotide analogue structures.
<table>
<thead>
<tr>
<th>ATOM</th>
<th></th>
<th>N</th>
<th>THR</th>
<th>A</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>2</td>
<td>CA</td>
<td>THR</td>
<td>A</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>3</td>
<td>C</td>
<td>THR</td>
<td>A</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>4</td>
<td>O</td>
<td>THR</td>
<td>A</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>5</td>
<td>CB</td>
<td>THR</td>
<td>A</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>6</td>
<td>OG1</td>
<td>THR</td>
<td>A</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>7</td>
<td>CG2</td>
<td>THR</td>
<td>A</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>
</tbody>
</table>
### PDB File Fields

<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>
Worldwide Structural Genomics

- "Fold space coverage"
- Complete genomes
  - Disease-causing organisms
  - Model organisms
- Membrane proteins
- Protein-ligand interactions

Hou et al., PNAS 2003, 100: 2386-2390
Protein Structure and Visualisation

Introduction to PyMOL
Overview

- A brief introduction
  - About PyMOL
  - Objects vs. selections

- Installation
  - Mac OS X, Linux, Windows
  - Download and install

- A few comments about molecular graphics...
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.

- But you should contribute!

- 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.
PyMOL

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

- **PDB (protein structure database)**
  - [www.pdb.org/](http://www.pdb.org/)

- **PyMOL home:**
  - [http://www.pymol.org/](http://www.pymol.org/)

- **PyMOL manual:**
  - [http://pymol.sourceforge.net/newman/user/toc.html](http://pymol.sourceforge.net/newman/user/toc.html)

- **PyMOL Wiki:**
  - [http://www.pymolwiki.org/index.php/Main_Page](http://www.pymolwiki.org/index.php/Main_Page)
Molecular Graphics

Visualizing protein structures
Example 1

- **mRNA**
- **rRNA**
- **Protein**

- **Large subunit:**
  - Catalytic mechanism

- **Small subunit:**
  - Specificity
Example 2

René Magritte

Leci n’est pas une pipe.
Aesthetics of Molecular Images

- Personal taste

- General guidelines:
  - Focus on relevant parts
  - Strip away unnecessary information
  - Find good viewing angle!
  - Choice of representation
  - No need for excessive graphics

- Good figures are (almost) self-explanatory.
Molecular Images

- Molecular graphics is an example of selective reduction of complexity.
Molecular graphics is an example of selective reduction of complexity.
Molecular Images

- Molecular graphics is an example of selective reduction of complexity.
Molecular Images

- Molecular graphics is an example of selective reduction of complexity.
Molecular Images

- Molecular graphics is an example of selective reduction of complexity.
Molecular Images

- Molecular graphics is an example of selective reduction of complexity.
Molecular Images

- Molecular graphics is an example of selective reduction of complexity.
Molecular graphics is an example of selective reduction of complexity.
Molecular Images

- Molecular graphics is an example of selective reduction of complexity.
Figures in PyMOL
Representations
Representations
Representations
Representations
Representations
Representations
Focus on the Relevant Parts

- Homologous peptides in different proteins
- Intuitive colouring
Residue Conservation

- http://www.mcgnmr.ca/ProtSkin/
Cross-Sections

- Proteasome
Electrostatic Surface Potentials
Composite Images

- MHC-I binding groove
- Images combined in Adobe Illustrator.
Simplified Composite Images