Introduction to Protein Structure

Function, evolution & experimental methods

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

- Outline the basic properties of protein enzymes.

- 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
Watson, Crick and 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.
Proteins & Cells

David Godsell, http://mgl.scripps.edu/people/goodsell/
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

Cross-section
Aquaporin
Backbone Problems?

- 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
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
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
Turns, Loops & Bends

- Between helices and sheets
- On protein surface
- Intrinsically “unstructured” proteins
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
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/amino/
Grouping Amino Acids

http://www.dreamingintechnicolor.com/InfoAndIdeas/AminoAcids.gif
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

- HIV protease


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

http://publications.nigms.nih.gov/structlife/chapter4.html
The Evolution Way

- Based on Blosum62 matrix
- Measure of evolutionary substitution probability
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
What Is an Enzyme?

![Enzyme Diagram](image)
# Enzyme Types

## Classification of Enzymes

<table>
<thead>
<tr>
<th>Group of Enzyme</th>
<th>Reaction Catalysed</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Oxidoreductases</td>
<td>Transfer of hydrogen and oxygen atoms or electrons from one substrate to another.</td>
<td>Dehydrogenases Oxidases</td>
</tr>
<tr>
<td>2. Transferases</td>
<td>Transfer of a specific group (a phosphate or methyl etc.) from one substrate to another.</td>
<td>Transaminase Kinases</td>
</tr>
<tr>
<td>3. Hydrolases</td>
<td>Hydrolysis of a substrate.</td>
<td>Estrases Digestive enzymes</td>
</tr>
<tr>
<td>4. Isomerases</td>
<td>Change of the molecular form of the substrate.</td>
<td>Phospho hexo isomerase, Fumarase</td>
</tr>
<tr>
<td>5. Lyases</td>
<td>Nonhydrolytic removal of a group or addition of a group to a substrate.</td>
<td>Decarboxylases Aldolases</td>
</tr>
<tr>
<td>6. Ligases (Synthetases)</td>
<td>Joining of two molecules by the formation of new bonds.</td>
<td>Citric acid synthetase</td>
</tr>
</tbody>
</table>
Enzyme Inhibition

Enzyme binds substrate
Enzyme releases products

Enzyme binds inhibitor
Inhibitor competes
Active Site Residues

Diagram a: Asp, Ser, His

Diagram b: Cys, His, Zn²⁺

Diagram c: Asp, Asp

Diagram d: Glu, His, X
Catalytically Active (or Modified)?

- **Acidic and amide side chains**
  - Aspartate
  - Asparagin
  - Glutamate
  - Glutamine

- **Basic side chains**
  - Lysine
  - Histidine
  - Arginine

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

- **Aromatic side chains**
  - Tryptophan
  - Phenylalanine
  - Tyrosine

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

- **Cyclic side chains**
  - Proline

[Link to Molecular Properties](http://www.ch.cam.ac.uk/magnus/molecules/amino/)
Proteases

Aminopeptidase

Carboxypeptidase
Enzymes can be large.

Proteasome

Ribosome
Structure & Evolution

Rhamnogalacturonan acetylesterase (A. aculeatus) (1k7c)

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

Serine esterase (S. scabies) (1ESC)
Inferring biological features from the structure

Topological switchpoint
Amylases

$\alpha$-amylase  $\beta$-amylase
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 \rightarrow \text{Fourier transform}
The Importance of Resolution

low

4 Å

3 Å

2 Å

1 Å

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 \(~10-20\,\text{Å}^2\), intermediate \(~20-30\,\text{Å}^2\), 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: $^1H$, $^{13}C$, $^{15}N$

- Proteins must be
  - below 50 kDa
  - stable at high concentration (0.5-1mM) @ room temperature
NMR Spectroscopy
Evaluation of NMR Structures

- Atomic backbone RMSD:
  - 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
RMSD

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

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

\[ \sum (x_i - y_i)^2 = 1 \AA^2 \]

\[ \sum (x_i - y_i)^2 = 0.8 \AA^2 \]

\[ \sum (x_i - y_i)^2 = 2 \AA^2 \]

\[ \sum (x_i - y_i)^2 = 1.2 \AA^2 \]

\[ \text{RMSD} = 1.33 \AA \]
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
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.
<table>
<thead>
<tr>
<th>ATOM</th>
<th>N</th>
<th>THR A 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 A 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 A 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 A 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 A 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 A 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 A 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>
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
Homology Modelling
Why Do We Need Homology Modelling?

- *Ab Initio* protein folding (random sampling):
  - 100 aa, 3 conf./residue gives approximately $10^{48}$ different overall conformations!

- Random sampling is *NOT feasible*, even if conformations can be sampled at picosecond ($10^{-12}$ sec) rates.
  - Levinthal’s paradox

- Do homology modelling instead.
How Is It Possible?

- The structure of a protein is uniquely determined by its amino acid sequence (but sequence is sometimes not enough):
  - prions
  - pH, ions, cofactors, chaperones

- Structure is conserved much longer than sequence in evolution.
  - Structure > Function > Sequence
How Often Can We Do It?

- There are currently ~50000 structures in the PDB
  - Reduces to ~8000 structures <30% identical (sequence) with a resolution <3.0 Å.

- 25% of all sequences can be modeled.

- 50% can be assigned to a fold class.
Protein Folds (SCOP) in PDB

No new folds!
Protein Folds

http://www.jcsg.org/
How Is It Done?

- Identify template(s)
  - Initial alignment
- Improve alignment
- Backbone generation
- Loop modelling
- Side chains
- Refinement
- Validation ←
Template Quality

- Selecting the best template is crucial!
- The best template may not be the one with the highest % id (best p-value...)
  - Template 1: 93% id, 3.5 Å resolution 😞
  - Template 2: 90% id, 1.5 Å resolution 😊
How Well Can We Do It?

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
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.
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