Protein (Macromolecular)
Crystallography

Comparison to
Chemical
Crystallography
Differences are all a result of the unique properties of proteins.

Crystallization
Diffraction Properties
Structure Determination

Crystal Quality
X-ray Source
Accuracy of Structure

Exaggerated example of Mosaicity

Moderately diffracting protein crystal
Protein Size

- Proteins are large in comparison to most compounds (heteropolymer of amino acids)
  - average molar mass is 30,000 g/mol per polypeptide chain
  - range between 5000 and 1,000,000 g/mol
- Proteins can be composed of one or more polypeptide chains
  - bacterial ribosome contains over 50 polypeptide chains and 3 RNA molecules (molar mass ~2,500,000 g/mol)

Top Left: rifampcin 880 g/mol
Bottom: RNA polymerase 345,000 g/mol
Protein Structure

- Proteins adopt a single or very few structures – **folded states**
  - Largely determined by **non-covalent interactions**
    - Involves regions both close and widely separated in the covalent structure
  - Dependent upon aqueous phase
    - Non-covalent forces are solution dependent
- Protein structures are typically compact and globular
  - Charged and polar residues are located on the surface of proteins
  - Interior of proteins is (almost) exclusively hydrophobic

Structure determines function at the molecular level
Protein Stability

- Folded state is marginally stable (by design?)
  - Relatively mild conditions disrupt the folded state
    - Temperature above 45°C (mammals)
    - pH < 4 or > 9
    - Low ionic strength (< 50 mM) or low dielectric solvent

- Proteins are susceptible to spontaneous chemical modification
  - Oxidation of sulfur, deamination, hydrolysis

- Proteins have limited solubility
  - Few proteins can be concentrated to 1 mM

Unfolded State

Folded State
Protein Crystallization

**Difficult !!! The “bottleneck” in protein structure determination**

- Must utilize conditions that do not disrupt the folded state
  - aqueous solutions
  - narrow temperature range (0 – 37°C)
  - narrow pH range (4-9)
- Large surface area (protein) and long, slow crystallization process
  - days to months
  - require novel crystallization methods
- Slow crystallization process increases likelihood of chemical modification

- Vapour diffusion crystallization trial

\[
[ppt]_{\text{drop}} = \frac{[ppt]_{\text{reservoir}}}{2}
\]
Protein Crystals

- Crystals are small and have a large unit cell
  - 0.1 x 0.1 x 0.1 mm is a typical crystal size
  - 100 Å per edge is a typical unit cell

- Contain 30-70% solvent (present as channels)
  - Solvent is critical for protein structure and therefore lattice structure
  - Solvent is largely disordered

- Restricted number of Space Groups
  - amino acids are all L-stereoisomer; no inversion or mirror symmetry
  - lower symmetry (¾ of crystals are orthorhombic or lower)
More Protein Crystals

- Limited contacts between symmetry related molecules
  - crystals are mechanically fragile ('crush' as opposed to 'fracture')
  - temperature, pressure, X-ray, etc. sensitive
    - crystals typically nucleate at interface of solvent: support or solvent: air

- High Mosaicity (0.2 - 1.5°)
  - partly due to mechanical fragility?
  - elongated spot shape, anisotropic diffraction
Crystal Packing

YZ plane
Unit Cell: 45.95 x 140.02 x 76.30 Å

XY plane
Space Group: P2₁2₁2₁
Diffraction Properties

Protein crystals diffract poorly (low signal / noise and resolution)

- large unit cell and small crystal size means there are fewer unit cells/crystal and less constructive interference

- proteins are composed of light atoms (H, C, N, O, S) with weak scattering factors

- disordered solvent generates diffuse scattering background (especially between 3.2-4.0 Å)

- solvent and protein atoms have significant thermal motion
  - rapid falloff in diffraction as function of resolution
More on Diffraction

- Long data collection
  - weak diffraction necessitates longer exposures
  - large $d$ spacing requires larger crystal to detector distance
  - low symmetry (more unique reflections/resolution shell)

- Radiation damage
  - crystals are damaged (indirectly) by X-rays and their diffraction changes as a function of time
  - longer Cu $K_{\alpha}$ more damaging than shorter wavelengths
    - X-ray generate free radicals within solvent
  - Freezing minimizes damage BUT requires suitable cryoprotectant
    - Glycerol, glycols, sugars, oils, etc.
X-ray Source

- Cu radiation (not Mo)
  - longer wavelength increases spot separation in reciprocal space (required due to large $d$ spacing)
  - longer wavelength X-rays are diffracted more efficiently
  - protein crystals rarely diffract beyond Cu limit (0.77 Å)

- Rotating (liquid cooled) Anode
  - dissipates heat of incident electron beam allowing greater electron flux
  - produces more X-ray photons

- Longer crystal to detector spacing
  - again due to large crystal $d$ spacing
Synchrotron X-ray Source

- Elliptical or circular particle accelerators
- Particle deflection produces intense electromagnetic radiation (100-1000 fold more intense)
- Emission wavelength can be easily changed (0.1 fm steps)
- Huge advantage for structure determination
  - Multiple Wavelength Anomalous Dispersion

(1) Electron Gun
(2) Linear Accelerator
(3) Booster Ring
(4) Storage Ring
(5) Beamline
(6) X-ray Station
Structure Determination (MR)

- Molecular Replacement
  - requires knowledge of “closely similar” structure
  - based upon overlap of Patterson maps from the experimental data and known structure
    - intensity data is used to calculate experimental Patterson
    - atomic coordinates are used to calculate known Patterson
  - determination of superposition matrix allows placement of known structure in unknown unit cell and provides phases

Molecular Replacement
(Real space example)
Structure Determination (MIR)

Multiple Isomorphous Replacement

• exploits stoichiometric binding of heavy atoms
  – Heavy atom positions are determined from differences in experimental intensities
  – Phases are derived from heavy atom positions
• require at least two unique heavy atom derivatives to solve a novel structure
  – $\alpha_{HH}$ from heavy atom position
  – $F_{HH} = F_{PH} - F_{PH}$
• Weak diffraction, thermal motion and non-isomorphism greatly complicate calculation
  – low information content of intensities is also a problem
Harker Construct - From Patterson to Phase

Knowns (measured):
\[ F_{\text{HHI}} \quad F_{\text{PHI}} \]

Knowns (calculated):
\[ F_{\text{HHI}} = F_{\text{PHI}} - F_{\text{PHI}} \] (algebra)
\[ \alpha_{\text{HHI}} \] (Patterson soln)

STEP 1)

Arbitrary Origin

Radius = \( F_{\text{PHI}} \)

(1) Origin represents tail of vector \( F_{\text{PHI}} \) (structure factor)
(2) Circle represents all possible values of \( \alpha_{\text{HHI}} \)

STEP 2)

\( d/\lambda = \alpha_{\text{HHI}} \)

(1) Adding \( F_{\text{PHI}} \) to \( F_{\text{PHI}} \) yields tail of vector \( F_{\text{PHI}} \)
(2) Circle represents all possible values of \( \alpha_{\text{HHI}} \)
(1) Intersection of circles represents possible solutions for $\alpha_{\text{HH}}$ and $\alpha_{\text{HN}}$ that are consistent with the known value of $\alpha_{\text{HH}}$.

(2) Two solutions are possible when using a single derivative.

(1) Introducing a second derivative resolves the ambiguity and a unique solution for the phase problem is obtained.

*Note: With the errors in real data this is not nearly as straightforward as portrayed.*
Multiple Wavelength Anomalous Dispersion

- conceptually identical to isomorphous replacement
- requires tunable X-ray source (synchrotron)
  - data collected at different wavelengths
    - remote, inflection and peak of atomic absorption edge
    - completely isomorphous**
- molecular biology techniques allow reliable introduction of Se as heavy atom (~2 / 100 residues)

<table>
<thead>
<tr>
<th>λ (pm)</th>
<th>keV</th>
</tr>
</thead>
<tbody>
<tr>
<td>97.937</td>
<td>12.6592</td>
</tr>
<tr>
<td>97.947</td>
<td>12.6578</td>
</tr>
<tr>
<td>105</td>
<td>11.8076</td>
</tr>
</tbody>
</table>
Structural Model

Limited resolution cannot identify individual atoms

- cannot determine structure without external information
  - require knowledge of bond lengths and angles
  - require knowledge of covalent structure of protein

- can determine fit of known fragments of structure to electron density
  - orientation of side groups may be ambiguous (carboxamides, imidazoles)

- cannot locate H atoms

Structural model is built into electron density
Final refined model fit to electron density at 3.5 & 1.8 Å resolution

- Note the difference in the definition of the oxygen atoms (red)
Can we know the structure is accurate?

- In several cases protein structures have been determined at greater than 1.0 Å resolution
  - validate structure of same protein determined at lower resolution
- Successfully explains wide array of biological data
  - existing experimental data can be rationalized using the structure
- Successfully at predicting results of biological and physical experiments
  - repeatedly proven to be model of choice for designing experiments
- Same structure as determined by independent techniques (NMR, cryoEM) at lower resolution
Common Indicators of High Quality Structure

- **R-factors**
  - $R_{merge} < 5\%$ for data with $I/\sigma(I) > 2$
  - $R_{refine} < 20\%$ for data to 2.0 Å resolution (and $R_{free} < 25\%$)

- **Stereochemistry**
  - Bonds (rmsd) $\sim 0.010$ Å, Angles (rmsd) $\sim 1.2^\circ$
  - Ramachandran Plot $\sim 90\%$ favored (main chain torsion angles)

- **Model**
  - $> 95\%$ of protein atoms fit and $\sim 1$ H2O per residue (2.0 Å resolution)

*Must explain functional and experimental data!*
Structural Visualization

- Structures are complex and require simplified representation
  - “Cartoons” of protein provide overall view of structure
  - Electrostatic surface provide visualization of local regions

Substrate binding requires both shape and charge complementarity
Summary

- Proteins have a number of unique properties that affect the ease of production and quality of protein crystals.
- Relatively low quality of protein crystals compromise the quality of intensity measurements.
- Relatively weak and low resolution intensity measurements increase the difficulty of structure determination and decrease the accuracy of the final structure.
- Protein crystallography is not (yet?) a routine technique that can be performed by a qualified technician.