Biochemistry 412: Proteomics and Functional Genomics

X-ray Crystallography

Rongjin Guan
March 7 2008
Purpose of this class

• Introduce the basic principle of X-ray crystallography
  Symmetry
  X-ray diffraction
  Phase problem
  Methods to solve the phase problem
  Model building and refinement

• Give an overview of all steps in structure determination by X-ray crystallography

• Help students understand crystal structure papers
Overview of X-ray crystallography

1. Grow crystals

2. Measure diffraction

3. Solve phases and refine structure
Why X-rays?

Right wavelength to resolve atoms: C-C bond is 1.54 Å
How are X-rays produced?

Wave-lengths

Cu(\(\kappa_1\)) = 1.54015 Å; Cu(\(\kappa_2\)) = 1.54433 Å
Cu(\(\kappa_\alpha\)) = 1.54015 Å
Cu(\(\kappa_\beta\)) = 1.39317 Å
Synchrotron X-rays

Electron/positron injection

Storage Ring

X-ray

Magnetic Fields

X-rays

Electron/positron beam
An example of home X-ray source (Rigaku)
A synchrotron X-ray source at Argonne, IL
Symmetry in crystals

Why do we need to know the symmetry of the crystals?

1) Reduce the size of the asymmetric unit (a unique part of the crystal)
2) Reduces data collection by reducing the number of unique reflections to collect
3) Symmetry may be useful for phase determination

International Table (volume 1) is a “Bible” for crystal symmetry.
Periodicity and Symmetry in a Crystal

• A crystal has long range ordering of building blocks that are arranged in an conceptual 3-D lattice.

• A building block of minimum volume defines unit cell

• The repeating units (protein molecule) are in symmetry in an unit cell

• The repeating unit is called asymmetric unit – A crystal is a repeat of an asymmetric unit
• Arrangement of asymmetric unit in a lattice defines the crystal symmetry.

• The allowed symmetries are 2-, 3, 4, 6-fold rotational, mirror (m), and inversion (i) symmetry (+/-) translation.

• Rotation + translation = screw

• Rotation + mirror = glide

⇒ 230 space groups, 32 point groups, 14 Bravais lattice, and 7 crystal systems
Only 65 space groups for chiral molecules

<table>
<thead>
<tr>
<th>System</th>
<th>Class</th>
<th>Space group symbols</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triclinic</td>
<td>1</td>
<td>( P_1 )</td>
</tr>
<tr>
<td>Monoclinic</td>
<td>2</td>
<td>( P_2, P_{21}, C_2 )</td>
</tr>
<tr>
<td>Orthorhombic</td>
<td>222</td>
<td>( C_{222}, P_{222}, P_{212121}, P_{21212}, P_{2221}, C_{2221}, F_{222}, I_{222}, I_{212121} )</td>
</tr>
<tr>
<td>Tetragonal</td>
<td>4</td>
<td>( P_4, P_{41}, P_{42}, P_{43}, I_4, I_{41} )</td>
</tr>
<tr>
<td></td>
<td>422</td>
<td>( P_{422}, P_{4212}, P_{4122}, P_{41212}, P_{4222}, P_{4212}, P_{43212}, I_{422}, I_{4122} )</td>
</tr>
<tr>
<td>Trigonal</td>
<td>3</td>
<td>( P_3, P_{31}, P_{32}, R_3 )</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td>( P_{312}, P_{321}, P_{3121}, P_{3112}, P_{3121}, P_{321}, R_{32} )</td>
</tr>
<tr>
<td>Hexagonal</td>
<td>6</td>
<td>( P_6, P_{61}, P_{65}, P_{64}, P_{63}, P_{62}, P_6 )</td>
</tr>
<tr>
<td></td>
<td>622</td>
<td>( P_{622}, P_{6122}, P_{6222}, P_{6322}, P_{6422}, P_{6522} )</td>
</tr>
<tr>
<td>Cubic</td>
<td>23</td>
<td>( P_{23}, F_{23}, I_{23}, P_{213}, I_{213} )</td>
</tr>
<tr>
<td></td>
<td>432</td>
<td>( P_{432}, P_{4132}, P_{4232}, P_{4332}, F_{432}, F_{4132}, I_{432}, I_{4132} )</td>
</tr>
</tbody>
</table>
Steps in structure determination by X-ray crystallography

- Sample preparation
  chemical, conformational and aggregational homogeneity
- Crystallization
- Data collection and process
- Phase determination
  Isomorphous replacement
  Anomalous scattering
  molecular replacement, etc
- Model building and refinement
- Structural analysis
Prepare samples for crystallization

Homogeneous protein samples have the best chance to crystallize. (not only pure in SDS PAGE!!!)

<table>
<thead>
<tr>
<th>Homogeneity Type</th>
<th>Method of detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemical homogeneity</td>
<td>SDS PAGE, mass spectrometry</td>
</tr>
<tr>
<td>Conformational homogeneity</td>
<td>Native PAGE, protease sensitivity</td>
</tr>
<tr>
<td>Aggregational homogeneity</td>
<td>Gel filtration, Native PAGE, light scattering, analytical ultra-centrifugation</td>
</tr>
</tbody>
</table>
How to grow crystals?

By driving solutions to super-saturation; Crystallization is entropically unfavorable, but may be energetically compensated by favorable packing forces; Incomplete factorial search of crystallization conditions combinations of buffers, precipitants (salts, polyethylene glycols- PEGs, organic solvents, etc) and additives.

Methods for crystallization
vapor (phase) diffusion
hanging drop
sitting drop
batch
liquid (phase) diffusion
Phase diagram in crystallization
Factors that affect crystallization

- pH (buffer)
- Protein Concentration
- Salt (NaCl, NH₄Cl, etc.)
- Precipitant
- Detergent (e.g. n-Octyl-β-D-glucoside)
- Metal ions and/or small molecules
- Rate of diffusion
- Temperature
- Size and shape of the drops
- Pressure (e.g. micro-gravity)
## Screening for Crystallization

<table>
<thead>
<tr>
<th>Precipitant Concentration</th>
<th>pH gradient</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 %</td>
<td>4</td>
</tr>
<tr>
<td>15 %</td>
<td>5</td>
</tr>
<tr>
<td>20 %</td>
<td>6</td>
</tr>
<tr>
<td>30 %</td>
<td>7</td>
</tr>
</tbody>
</table>

- Precipitate
- Crystalline precipitate
- Fiber like Micro-crystals
- Small crystals
- Ideal crystal
Monodisperse SeMet Samples
Shipped to both Columbia and Hauptman-Woodward Institute (at HWI) for Crystallization Screening

HTP Robotic Screening (1356 conditions) under oil provides initial hits. DeTitta, Luft, et al

Robotic setup of hanging drop crystallization optimization at Columbia is informed by HWI results. Hunt, Tong, et al

Production crystallization conditions include potential ligands assessed by literature search
A protein crystal gallery
Steps in structure determination by X-ray crystallography

- Sample preparation
  chemical, conformational and aggregational homogeneity
- Crystallization
- Data collection and process
- Phase determination
  Isomorphous replacement
  Anomalous scattering
  molecular replacement, etc
- Model building and refinement
- Structural analysis
Cryo-loop

Crystal

Goniometer

Detector

4-Circle Goniometer (Eulerian or Kappa Geometry)
Bragg’s law: $2d\sin\theta = \lambda$

When this condition is satisfied, a plane with a spacing of $d$ will ‘reflect’ X-rays with an incident angle of $\theta$, due to constructive interference.

Path difference: $2d\sin\theta$

$d$: resolution

$\theta$: scattering angle
Data Collection: measuring positions and intensities of reflections in the diffraction pattern produced by the macromolecular crystal.

- Capillary mounting for crystal characterization and data collection above 0°C.

- Crystal ‘freezing’
  Soak the crystal in a cryo-protectant solution, then collect data around 100K.
  Commonly used cryo-protectant: glycerol, ethylene glycol, PEG400, sugars, MPD, etc

- Each diffraction image contains diffraction from a small angular sweep (oscillation). Rotate the crystal until a complete 3-D diffraction is completed and collected.

- These diffraction images are recorded by imaging plates or CCD detectors as digital images.
Crystal characterizations

- **Diffraction resolution limit.**
  
  \[ d = \frac{\lambda}{2 \sin \theta} \]
  
  will determine what kind of detail you may see from the final structure. Usually 3.0Å or beyond is OK, and 2.0Å or beyond is good. Note that the resolution limit of most macromolecular crystals does not give “atomic” resolution, i.e. electron density does not resolve into atoms; but the fitting of “atomic models” into an electron density leads to “atomic” resolution.

- **Cell dimensions.**

- **Mosaicity**: usually 0.3-1.0°. Indication of variability between different unit cells in the crystal. High mosaicity increases the width of reflections, increases the number of reflections on each image and may cause reflection overlap, leading to problems in data collection.

- **Space group.**

- **Data quality: R factor.**
An example of diffraction image

Resolution
Ice ring
Mosaicity
Spots resolved

e.g. 2.3 Å
Sometimes, there may be problems

- One cell parameter is too long – spots overlaps
- Twinning
- High mosaics
- Water rings
- Radiation damage
- Anisotropic diffraction
A typical data processing output

Space group p212121

<table>
<thead>
<tr>
<th>Shell</th>
<th>I/Sigma in resolution shells:</th>
<th>% of reflections with I / Sigma less than</th>
<th>limit limit</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>&gt;20</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>99.00</td>
<td>4.09</td>
<td>1.0</td>
<td>2.7</td>
<td>4.1</td>
<td>5.1</td>
<td>6.2</td>
<td>9.6</td>
<td>19.2</td>
<td>80.4</td>
<td>99.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.09</td>
<td>3.25</td>
<td>1.0</td>
<td>2.8</td>
<td>4.1</td>
<td>6.0</td>
<td>8.3</td>
<td>14.4</td>
<td>28.0</td>
<td>71.9</td>
<td>99.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.25</td>
<td>2.84</td>
<td>1.2</td>
<td>4.3</td>
<td>6.4</td>
<td>8.8</td>
<td>13.7</td>
<td>24.3</td>
<td>45.3</td>
<td>54.7</td>
<td>100.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.84</td>
<td>2.58</td>
<td>1.4</td>
<td>7.0</td>
<td>12.1</td>
<td>16.0</td>
<td>24.0</td>
<td>39.7</td>
<td>64.8</td>
<td>35.0</td>
<td>99.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.58</td>
<td>2.39</td>
<td>1.5</td>
<td>6.0</td>
<td>12.8</td>
<td>19.4</td>
<td>29.9</td>
<td>48.2</td>
<td>76.7</td>
<td>23.1</td>
<td>99.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.39</td>
<td>2.25</td>
<td>2.8</td>
<td>10.9</td>
<td>18.4</td>
<td>26.7</td>
<td>37.6</td>
<td>57.4</td>
<td>82.1</td>
<td>17.5</td>
<td>99.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.25</td>
<td>2.14</td>
<td>2.4</td>
<td>11.0</td>
<td>20.5</td>
<td>27.9</td>
<td>40.5</td>
<td>62.7</td>
<td>87.4</td>
<td>12.3</td>
<td>99.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.14</td>
<td>2.05</td>
<td>5.0</td>
<td>15.6</td>
<td>27.1</td>
<td>37.8</td>
<td>51.5</td>
<td>73.7</td>
<td>92.3</td>
<td>6.6</td>
<td>98.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.05</td>
<td>1.97</td>
<td>5.2</td>
<td>19.9</td>
<td>32.9</td>
<td>45.2</td>
<td>60.4</td>
<td>83.0</td>
<td>96.5</td>
<td>1.8</td>
<td>98.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.97</td>
<td>1.90</td>
<td>6.3</td>
<td>25.7</td>
<td>43.6</td>
<td>56.0</td>
<td>70.7</td>
<td>88.4</td>
<td>94.6</td>
<td>0.4</td>
<td>95.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All hkl</td>
<td></td>
<td>2.7</td>
<td>10.4</td>
<td>17.9</td>
<td>24.5</td>
<td>33.8</td>
<td>49.4</td>
<td>67.9</td>
<td>31.2</td>
<td>99.1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Summary of reflections intensities and R-factors by shells

- R linear = SUM (ABS(I - <I>)) / SUM (I)
- R square = SUM ( (I - <I>) ** 2 ) / SUM ( I ** 2 )
- Chi**2 = SUM ( (I - <I>) ** 2 ) / (Error ** 2 * N / (N-1) )

In all sums single measurements are excluded

<table>
<thead>
<tr>
<th>Shell</th>
<th>Lower</th>
<th>Upper</th>
<th>Average</th>
<th>Average</th>
<th>Norm.</th>
<th>Linear Square</th>
<th>R-fac</th>
<th>R-fac</th>
</tr>
</thead>
<tbody>
<tr>
<td>limit</td>
<td>limit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>99.00</td>
<td>4.09</td>
<td>5589.5</td>
<td>156.6</td>
<td>143.1</td>
<td>0.734</td>
<td>0.038</td>
<td>0.064</td>
<td></td>
</tr>
<tr>
<td>4.09</td>
<td>3.25</td>
<td>4391.0</td>
<td>131.0</td>
<td>109.4</td>
<td>0.693</td>
<td>0.040</td>
<td>0.053</td>
<td></td>
</tr>
<tr>
<td>3.25</td>
<td>2.84</td>
<td>2181.8</td>
<td>82.2</td>
<td>73.3</td>
<td>0.607</td>
<td>0.050</td>
<td>0.166</td>
<td></td>
</tr>
<tr>
<td>2.84</td>
<td>2.58</td>
<td>1164.9</td>
<td>64.3</td>
<td>60.1</td>
<td>0.583</td>
<td>0.065</td>
<td>0.073</td>
<td></td>
</tr>
<tr>
<td>2.58</td>
<td>2.39</td>
<td>839.3</td>
<td>61.9</td>
<td>59.5</td>
<td>0.572</td>
<td>0.083</td>
<td>0.089</td>
<td></td>
</tr>
<tr>
<td>2.39</td>
<td>2.25</td>
<td>695.7</td>
<td>64.8</td>
<td>61.7</td>
<td>0.524</td>
<td>0.100</td>
<td>0.109</td>
<td></td>
</tr>
<tr>
<td>2.25</td>
<td>2.14</td>
<td>643.9</td>
<td>68.2</td>
<td>65.5</td>
<td>0.519</td>
<td>0.115</td>
<td>0.178</td>
<td></td>
</tr>
<tr>
<td>2.14</td>
<td>2.05</td>
<td>496.0</td>
<td>71.9</td>
<td>70.0</td>
<td>0.436</td>
<td>0.135</td>
<td>0.149</td>
<td></td>
</tr>
<tr>
<td>2.05</td>
<td>1.97</td>
<td>394.6</td>
<td>79.6</td>
<td>78.3</td>
<td>0.421</td>
<td>0.170</td>
<td>0.182</td>
<td></td>
</tr>
<tr>
<td>1.97</td>
<td>1.90</td>
<td>296.9</td>
<td>91.6</td>
<td>90.8</td>
<td>0.343</td>
<td>0.202</td>
<td>0.232</td>
<td></td>
</tr>
<tr>
<td>All reflections</td>
<td>1729.2</td>
<td>88.1</td>
<td>81.9</td>
<td>0.562</td>
<td>0.056</td>
<td>0.077</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Intensities of systematic absences

<table>
<thead>
<tr>
<th>h</th>
<th>k</th>
<th>l</th>
<th>Intensity</th>
<th>Sigma</th>
<th>I/Sigma</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>3</td>
<td>-2.9</td>
<td>7.5</td>
<td>-0.4</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>0</td>
<td>3.9</td>
<td>17.4</td>
<td>0.2</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>8.7</td>
<td>9.1</td>
<td>1.0</td>
</tr>
</tbody>
</table>

completeness

Rmerge

Systematic absences
Steps in structure determination by X-ray crystallography

- Sample preparation
  chemical, conformational and aggregational homogeneity
- Crystallization
- Data collection and process
- Phase determination
  Isomorphous replacement
  Anomalous scattering
  molecular replacement, etc
- Model building and refinement
- Structural analysis
Phase problem in X-ray Crystallography

To calculate an electron density map:

\[ \rho(xyz) = \frac{1}{V} \sum_{h} \sum_{k} \sum_{l} |F(hkl)| \exp\left[(-2\pi i (hx + ky + lz)) + i \alpha_{hkl}\right] \]

Can measure  
|F(hkl)|=I^{1/2}  
Can’t measure  
\( \alpha_{hkl} \)

The phases are lost during measurement, so electron density cannot be directly calculated. We have to estimate them.

This lack of knowledge of the phases is termed phase problem. 
Importance of Phases

Phases dominate the image!
Phase estimates need to be accurate
How to Solve the Phase Problem

• Experimental methods:
  
  * Multiple/Single Isomorphous Replacement (MIR/SIR)
    Heavy atoms (Heavy metals, Xe, Br, I, etc)
  
  * Multi/single Wavelength Anomalous Dispersion (MAD/SAD)
    Heavy atoms (Se-Met, etc),

• Non-experimental methods:
  
  * Molecular Replacement
    Homologous templates (similar structure fold)
  
  * Direct Methods (special cases)
    high resolution data, small proteins (<2000 atoms)
Isomorhous Replacement (MIR)

- Heavy atom derivatives are prepared by soaking or co-crystallizing.
- Diffraction data for heavy atom derivatives are collected along with the native data.

\[ F_{PH} = F_P + F_H \]

- Patterson function \( P(u) = \frac{1}{V} \sum |F(h)|^2 \cos(2\pi u \cdot h) \]
  \[ = \int_{r}^{h} \rho(r) \times \rho(r') \, dv \]

\[ \Rightarrow \] strong peaks for in Patterson map when \( r \) and \( r' \) are two heavy atom positions.
Phase determination by MIR

protein crystal → \( F_P \) → \( F_{H1} \) estimate → \( F_{PH1} \) → \( F_{H2} \) estimate → \( F_{PH2} \)

\( \alpha_P \)

Patterson method

Heavy atom structures

\( F_P \): protein
\( F_{PH1} \): protein + H.A.1
\( F_{PH2} \): protein + H.A.2
\( F_{H1} \): H.A.1
\( F_{H2} \): H.A.2
The Patterson Function

- Obtained by setting $F(h, k, l) = |F(h, k, l)|^2$ in the electron density Fourier equation,

$$P(u, v, w) = \frac{1}{V} \sum_{hkl} |F(h, k, l)|^2 \cos[2\pi(hu + kv + lw)]$$

where $u$, $v$, and $w$ are fractional coordinates in the unit cell. It is a Fourier summation with intensities as coefficients and phase angles equal to zero.

- Patterson peaks corresponding to vectors between atoms, i.e. inter-atomic distances.
Obtaining $F_H$ estimate: the isomorphous difference

When diffraction data for protein and a derivative crystal are available, one can obtain the amplitude difference between the protein and the derivative:

$$\Delta |F|_{iso} = |F_{PH}| - |F_{P}|$$

This isomorphous difference between a heavy atom containing crystal and a native crystal, is an estimate of $|F_H|$. It can be used in the Patterson Function for determining the heavy atom structure.
Phase determination by MIR once the heavy atom structures are known

Because the amplitudes and phases of $F_{H1}$ and $F_{H2}$ are known, the three origins can be defined, which are then used to draw the three circles with radii that correspond to the amplitudes of $F_P$, $F_{PH1}$ and $F_{PH2}$. 
Multiple Anomalous Dispersion (MAD)

At the absorption edge of an atom, its scattering factor

\[ f_{\text{ano}} = f + f' + if'' \]

<table>
<thead>
<tr>
<th>Atom</th>
<th>( f )</th>
<th>( f' )</th>
<th>( f'' )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hg</td>
<td>80</td>
<td>-5.0</td>
<td>7.7</td>
</tr>
<tr>
<td>Se</td>
<td>34</td>
<td>-0.9</td>
<td>1.1</td>
</tr>
</tbody>
</table>

\[ F(h,k,l) = F(-h,-k,-l) \rightarrow \text{anomalous differences} \rightarrow \text{positions of anomalous scatterers} \rightarrow \text{Protein Phasing} \]
Anomalous Scattering breaks the Friedel's Law

No anomalous scattering

\[ |F_p(+)| = |F_p(-)| \]

with anomalous scattering

\[ |F_{PH}(+)| \text{ and } |F_{PH}(-)| \text{ are now different} \]
Obtaining $F_H$ estimate: the anomalous difference

\[ \Delta |F|_{\text{ano}} = [|F_{PH}(+)| - |F_{PH}(-)|] \]

Similarly as for $\Delta |F|_{\text{iso}}$, $\Delta |F|_{\text{ano}}$ is an estimate of $|F_H|$.

It can be used in the Patterson Function method for determining the heavy atom structure.
Se-Met MAD

• Most common method of \textit{ab initio} macromolecule structure determination

• A protein sample is grown in Se-Met instead of Met.

• Minimum 1 well-ordered Se-position/75 amino acids

• Anomalous data are collected from 1 crystal at Se K-edge (12.578 keV).

• MAD data are collected at Edge, Inflection, and remote wavelengths
Resolve the SIR or SAS phase ambiguity (Handedness) by Solvent Flattening

The ISAS process is carried twice, once with heavy atom site(s) at refined locations (+++), and once in their inverted locations (---).

<table>
<thead>
<tr>
<th>Data</th>
<th>FOM&lt;sup&gt;1&lt;/sup&gt;</th>
<th>Handedness</th>
<th>FOM&lt;sup&gt;2&lt;/sup&gt;</th>
<th>R-Factor</th>
<th>Corr. Coef</th>
</tr>
</thead>
<tbody>
<tr>
<td>RHE</td>
<td>0.54</td>
<td>Correct</td>
<td>0.82</td>
<td>0.26</td>
<td>0.958</td>
</tr>
<tr>
<td>0.54</td>
<td>Incorrect</td>
<td>0.80</td>
<td>0.30</td>
<td></td>
<td>0.940</td>
</tr>
<tr>
<td>NP With I&lt;sup&gt;3&lt;/sup&gt;</td>
<td>0.54</td>
<td>Correct</td>
<td>0.80</td>
<td>0.27</td>
<td>0.955</td>
</tr>
<tr>
<td>0.54</td>
<td>Incorrect</td>
<td>0.76</td>
<td>0.36</td>
<td></td>
<td>0.919</td>
</tr>
<tr>
<td>NP With I &amp; S&lt;sup&gt;4&lt;/sup&gt;</td>
<td>0.56</td>
<td>Correct</td>
<td>0.82</td>
<td>0.24</td>
<td>0.964</td>
</tr>
<tr>
<td>0.56</td>
<td>Incorrect</td>
<td>0.78</td>
<td>0.35</td>
<td></td>
<td>0.926</td>
</tr>
</tbody>
</table>

<sup>1</sup>: Figure of merit before solvent flattening  
<sup>2</sup>: Figure of merit after one filter and four cycles of solvent flattening  
<sup>3</sup>: Four Iodine were used for phasing  
<sup>4</sup>: Four Iodine and 56 Sulfur atoms were used for phasing

Molecular Replacement

• If the structure of a similar molecule is known, structure factors of this known model may be calculated, allowing placement of the model in the unit cell by minimizing the R factor as a function of the orientation and position of the model within the unit cell:

\[
R\text{-factor} = \sum_{\text{hkl}} \frac{|F_{\text{obs}}| - |F_{\text{calc}}|}{\sum_{\text{hkl}} |F_{\text{obs}}|}
\]

The templates for a successful molecular replacement must be structural homologues (Cα rmsd < 1.5Å), not necessarily, though often, homologous in sequence.
How phases are calculated in MR

- By artificially packing the model into the crystal unit cell of the unknown molecule, we can calculate the structure factor $F$:

$$F_{\text{calc}}(hkl) = \sum f_j \exp[2\pi i(xh + yk + lz)]$$

- $F_{\text{calc}}$ contains information about both structure factor amplitude and phase $\alpha_{\text{calc}}$

- Using $\alpha_{\text{calc}}$, the unknown structure can then be computed using the following expression:

$$\rho(x,y,z) = \frac{1}{V} \sum_h \sum_k \sum_l F_{(h,k,l)} \exp[-2\pi i(hx + ky + lz)]$$

$$F_{\text{obs}(h,k,l)} \exp(i\alpha_{\text{calc}(h,k,l)})$$

Molecular replacement provides $\alpha_{\text{calc}(h,k,l)}$ for electron density calculation
Two step process in Molecular Replacement

1. find orientation of model (red ==> black)
   Rotation Function
2. find location of orientated model (black ==> blue)
   Translation Function

\[ x_1 = R x_2 + t \]

X1: unknown
X2: known model
Once phases are determined, an electron density map can be calculated

- From diffraction $F(h,k,l)$ to electron density by Fourier Transform:

$$
\rho(x, y, z) = \frac{1}{V} \sum_h \sum_k \sum_l F(h, k, l) \exp(-2\pi i (hx + ky + lz))
$$

$$
= \frac{1}{V} \sum_h \sum_k \sum_l |F(h, k, l)| \exp(i \alpha(h, k, l) - 2\pi i (hx + ky + lz))
$$

- Amplitudes measured from a native diffraction data: $|F| = I^{1/2}$.
- Phases computed from MIR, SIR, SIRAS, MAD, SAD or MR.
- For each reflection, a phase reliability coefficient or ‘figure of merit’ (between 0 and 1) may be multiplied to reduce the contribution from less reliable reflections.
Maps at different resolution show different levels of details

Maps are contoured on a certain sigma value, such as 1.0.
Phase refinement by solvent flattening

E. coli soluble lytic transglycosylase at 3.3Å resolution. Top: MIR map; bottom: solvent flattened map (60% solvent)
Steps in structure determination by X-ray crystallography

- Sample preparation
  - chemical, conformational and aggregational homogeneity
- Crystallization
- Data collection and process
- Phase determination
  - Isomorphous replacement
  - Anomalous scattering
  - molecular replacement, etc
- Model building and refinement
- Structural analysis
Model building is the fitting of atomic models into electronic density.
Least-Squares Refinement

List-squares refinement of atoms (x, y, z, and B) against observed $|F(h,k,l)|$

Target function that is minimized

$$Q = \sum w(h,k,l)(|F_{\text{obs}}(h,k,l)| - |F_{\text{cal}}(h,k,l)|)^2$$

$$\frac{dQ}{du_j} = 0; \; u_j - \text{all atomic parameters}$$
What to refine?

- Atom coordinates: \((x, y, z)\)s
- B factors and occupancy

Steps in refinement

1. Rigid body refinement
2. Energy minimization
3. Simulated annealing refinement
4. Model building in graphics by hand
5. Add waters and other molecules
6. B factor refinement
Geometric Restraints in Refinement

Each atom has 4 (x,y,z,B) parameters and each parameter requires minimum 3 observations for a free-atom least-squares refinement. → A protein of N atoms requires 12N observations.

For proteins diffracting < 2.0 Å resolution observation to parameter ratio is considerable less.

Protein Restraints (bond lengths, bond angles, planarity of an aromatic ring etc.) are used as restraints to reduce the number of parameters.
**R-factor**

\[
R_{\text{cryst}} = \frac{\sum_{hkl} |F_{\text{obs}}(hkl) - kF_{\text{cal}}(hkl)|}{\sum_{hkl} |F_{\text{obs}}(hkl)|}
\]

**Free-R**

R-factor calculated for a test-set of reflections that is never included in refinement.

R-free is always higher than R.

Difference between R and R-free is smaller for higher resolution and well-refined structures.
Free R Factor

Validation of Protein Refinement Models

effect of overfitting the diffraction data

Free R, R

2365 random scatterers
Unrestrianted Protein atoms (2365 atoms)
Protein alone
Protein + Ordered Water Molecules
Prot. + Ord. H2O + 1850 random Water molecules

Free R value

Reasonable Models

poor model of a protein
Echem = 0 bad stereochemistry
When you believe you have done the best model building and Refinement, you can deposit your structural factors and atomic Coordinates to PDB: www.pdb.org
The famous “Table 1”: crystallographic refinement statistics

<table>
<thead>
<tr>
<th>Crystal</th>
<th>hTRANCE-R complex</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRAF6</td>
<td>Residues 346-504</td>
</tr>
<tr>
<td>Peptide</td>
<td>342-QMPTDELEY</td>
</tr>
<tr>
<td>Space group</td>
<td>P2₁2₁2₁</td>
</tr>
<tr>
<td>Cell dimensions</td>
<td>a=38.0Å, b=45.0Å, c=106.5Å</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Diffraction Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resolution</td>
</tr>
<tr>
<td>$R_{sym}$ (last shell)</td>
</tr>
<tr>
<td>Completeness (last shell)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Refinement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resolution</td>
</tr>
<tr>
<td>Sigma cutoff</td>
</tr>
<tr>
<td>Number of protein residues</td>
</tr>
<tr>
<td>Number of protein atoms</td>
</tr>
<tr>
<td>Number of solvent atoms</td>
</tr>
<tr>
<td>Number of reflections used</td>
</tr>
<tr>
<td>$R$ ($R_{free}$)</td>
</tr>
<tr>
<td>RMSD bond length</td>
</tr>
<tr>
<td>RMSD bond angle</td>
</tr>
</tbody>
</table>

In this case, even at 2.0 Å resolution, reflection/parameters ratio $\frac{12396}{4*1331+80} = 2.2$, very under-determined.
Structural analysis

• Can you explain biological data from this structure?
• What is the structural basis of the biological function?
• What implication can you get from the structure? Can you design better inhibitor or substrate (drug design, protein engineering…)
• Are what you observed from this crystal structure believable (or artifact from crystallization?)

• Paper writing
Further reading if you are interested

*Crystallography Made Crystal Clear*, by Gale Rhodes  
*Principles of Protein X-ray Crystallography*, by Jan Drenth  
*International Tables for Crystallography*, vol F, edited by M. G. Rossmann & Eddy Arnold

http://www.ruppweb.org/Xray/101index.html  
http://www.iucr.org/cww-top/edu.index.html

Crystallography history (fun reading)  
Biophysical Features of 230 NESG Structures in PDB

M. Baran, J. Locke, D. Snyder, J. Huang, J. Hunt  Poster
Other possible topics

- Time-resolved X-ray crystallography
- Direct method to solve phase problem
- Small angle X-ray scattering
- X-ray powder diffraction to solve protein structures
- Difference between NMR and X-ray crystallography
- Neutron diffraction
- Drug design
- Virus/supercomplex crystallography