Data statistics

Table 1. Data Collection Statistics for Native Structure Determination

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Native</th>
<th>EMTS</th>
<th>K₃PtCl₄</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preparation of heavy-atom derivative</td>
<td>-</td>
<td>0.1 mM, 2 hr +</td>
<td>0.2 mM, 4 hr +</td>
</tr>
<tr>
<td>Wavelength (Å)</td>
<td>1.0000</td>
<td>1.0063</td>
<td>1.0715</td>
</tr>
<tr>
<td>Number of reflections²</td>
<td>337,229 (52,430)</td>
<td>247,466 (38,398)</td>
<td>162,593 (30,588)</td>
</tr>
<tr>
<td>Completeness (%)²</td>
<td>97.9 (95.0)</td>
<td>93.2 (61.8)</td>
<td>97.4 (95.0)</td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>30–2.5</td>
<td>30–2.8</td>
<td>30–3.0</td>
</tr>
<tr>
<td>Rmerge (%)bc</td>
<td>6.9 (26.7)</td>
<td>5.4 (23.0)</td>
<td>7.6 (33.7)</td>
</tr>
<tr>
<td>Rₜₒ (15.0–3.0 Å)</td>
<td>13.6 (17.4)</td>
<td>20.0 (25.0)</td>
<td></td>
</tr>
<tr>
<td>Rₜₒ (13.5)</td>
<td>2.5 (13.5)</td>
<td>4.2 (17.4)</td>
<td></td>
</tr>
<tr>
<td>Riso (centric/a-centric)</td>
<td>0.82/0.93 (MIR)</td>
<td>0.93/0.86 (MIR)</td>
<td></td>
</tr>
<tr>
<td>Phasing power (centric/a-centric)</td>
<td>1.10/0.73 (MIR)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Figure of merit</td>
<td>30–2.5 Å</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Refinement statistics</td>
<td>21.3% for 49,367 reflections</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rₜₒ</td>
<td>24.5% for 2,767 reflections</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rms bond length and bond angles</td>
<td>0.008 Å, 1.470°</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

² Numbers in parentheses are the number of unique reflections.

Numbers in parentheses are for the highest-resolution bins.

Rmerge = \[\sum_{h} \sum_{i} \frac{|I_h| - \langle I_h \rangle}{\sum_{h} |I_h|} \sum_{h} I_h\]

Riso = \[\sum_{h} |F_{h} - F_{h}|/\sum_{h} F_{h}\]

Rapo = \[\sum_{h} |I_{h} - I_{h}^{obs}|/\sum_{h} |I_{h}^{obs}| + |I_{h}^{obs}|\]

Rcrss = \[\sum_{h} |F_{h} - F_{h}^{calc}| - |F_{h}^{calc}|/\sum_{h} |F_{h} - F_{h}^{calc}|\] where \(F_{h}, F_{h}^{calc}, F_{h}^{calc}\) are the structure amplitudes.

Phasing power = \[\text{rms}(F_{h}/E)\] where \(E\) is the lack of closure error.
Phase methods:
Molecular replacement

Lecture 9
2-13-06
Methods for phase determination

**Molecular replacement (MR)**
- Only one native crystal is needed. One homologous structure must be available. Quick and simple.

**Multiple isomorphous replacement (MIR)**
- At least one native crystal and two crystals soaked in two heavy atom solutions must be available. Give good quality experimental maps. Need no information about the unknown structure.

**Multiwavelength anomalous dispersion (MAD)**
- Only one crystal is needed but multiple data sets must be collected at three different wavelengths. Se-Met protein is usually needed. Excellent electron density map.
Molecular replacement and Michael Rossmann

Molecular replacement method was initially developed by Michael Rossmann (MR). He used the structure of Tomato Bushy Stunt Virus, a plant virus, to determine the crystal structure of the human rhinovirus 14 (the common cold virus) in the early 80’s.

Basic principle of molecular replacement

- The goal in MR is to orient and position the search model, such that it coincides with the position of the unknown protein in the crystal. The model can then provide phase information for the unknown structure.
- A total of six parameters need to be determined – 3 angles and 3 translational elements.

Given a crystal with proteins "P" of unknown structure and the known structure of a related protein "P"

The major challenge of MR is to find out how to orient and position the known structure such that it corresponds with the unknown structure in the crystal.
Basic principle of molecular replacement, cont’d

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

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

– $F_{\text{calc}}$ 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:

$$\rho_{(x,y,z)} = \frac{1}{V} \sum_{h} \sum_{k} \sum_{l} F_{(h,k,l)} \exp[-2\pi i (hx + kx + lx)]$$

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

Molecular replacement provides $\alpha_{\text{calc}(h,k,l)}$ for electron density calculation
Molecular replacement – an example

- **Unknown**
  - Observed intensities only

- **Known**
  - Calculated intensities and phases

*Cat and cat diffraction*
*Manx and Manx FT*

Cat diffraction intensity
Manx phases

\(X\)
Several scenarios suitable for the molecular replacement method

- The structure of a homologous molecule is known, e.g. bacterial polymerase vs. polymerase II
- The structure of a large fraction of the unknown molecule is known, e.g. LacR vs. LacR+DNA
- The structure of the same molecule in another functional state is known, e.g. $K^+$ channel in the open vs. closed state
- Same molecule solved in different space groups
- NMR or theoretical model is available
Four Major Steps in Molecular Replacement (MR)

Commonly used to solve the phases for an unknown structure that is closely related to a known structure.

1. Find a homologous structure.
2. Rotation - determination of relative orientation of the same (or similar) molecule in different unit cells.
3. Translation - defines the unknown molecule’s position in the unit cell with respect to the origin.
4. Phase determination - An initial set of phases can be determined by placing the known structure into the unit cell of the unknown structure using the rotation and translation determined in steps (2) and (3).
Step 1 - Are there a homologous structure available

- All solved structures are deposited in a database called Protein Data Bank (http://www.rcsb.org/pdb/)

- BLAST search against the PDB structure database
  - Go to http://us.expasy.org/tools/blast/
  - Input either your protein sequence or a Swiss-Prot accession number
  - >20% sequence identity usually means similar 3-D structures.
The number of entries in PDB has increased exponentially in recent years.
Structure is better conserved than sequence during evolution, especially the hydrophobic core and the enzymatic active site.

- Structure is better conserved than sequence during evolution, especially the hydrophobic core and the enzymatic active site.
- Human contains ~24,500 protein-coding genes but only ~100 structure folds in total.
Step 2 – the rotation function

• Determination of relative orientation of the same (or similar) molecule(s) in different unit cells.
• In rotation function, the known molecule is used to search the unit cell of the unknown structure for a similar pattern. The search is commonly carried out by comparing Patterson maps.

Review - Patterson Map

\[
\rho(u, v, w) = \frac{1}{V_{\text{unit cell}}} \sum \sum \sum |F|^2 \cos 2\pi(hu + kv + lw),
\]

The Patterson map is a representation of the interatomic vectors in the unit cell. The Patterson space unit cell will contain \(N^2\) peaks, corresponding to the \(N\) vectors that exist for each of the \(N\) atoms.

The Patterson map for the structure for which you are trying to compute phases is a stationary map, the Patterson map for the search model is rotated. Only a subset of Patterson vectors are chosen for comparison – the so called self Patterson peaks.
Patterson maps contain both intramolecular and intermolecular vectors

- Peaks in a Patterson map are caused either by intramolecular or intermolecular vectors
- Intramolecular vectors are:
  - Located near origin
  - Independent of position
  - Dependent on orientation
Rotation function matches experimental and model Pattersons

Rotating the search model to the correct orientation will give a calculated Patterson that matches the observed Patterson, in region close to the origin (intramolecular vectors).
Rotation function

- Mathematically, rotation function can be quantified as:

\[
R(\kappa, \phi, \psi) = \int_{r_{\text{min}}}^{r_{\text{max}}} P_{\text{nat}}(u) P_{\text{mod}}(\kappa, \phi, \psi, u) du
\]

\[
= \frac{U}{V^3} \sum_h \sum_{h'} |F(h)|^2 |F([C]h')|^2 \times G[-(h + h')]
\]

- A product function
- The maximum value is obtained when two Patterson maps can be well superimposed
- Integration is calculated in volume from \( r_{\text{min}} \) to \( r_{\text{max}} \) from the Patterson origin
Rotation function – important parameters

• Resolution
  – Often uses 10 to 3Å
  • High resolution cutoff: to minimize model differences
  • Low resolution cutoff: to minimize solvent effects

• Integration radius
  – To maximize the number of intramolecular vectors, but minimize intermolecular vectors
  – Patterson origin should be removed

• Search model box size
  – Large enough to avoid overlapping of intramolecular vectors

• Search interval size
  – Larger step sizes save computation time
  – Should be fine enough so the correct solution will not missed

• Search range
  – Depending on the symmetry of both Patterson maps

\[
R(\kappa, \phi, \psi) = \int_{r_{\text{min}}}^{r_{\text{max}}} P_{\text{nat}}(u) P_{\text{mod}}(\kappa, \phi, \psi, u) du
\]
The effect of resolution

Advanced Methods in Modern Biomolecular Crystallography

Importance of resolution

Reduced disorder at low temperature

Dramatic improvements in the overall structure are likely to result from better definition of disordered regions regardless of resolution.
The unit cell for the search model should be large enough to allow good separation of self and cross Patterson peaks.

Rule of thumb: \( a = \text{diameter of the molecule} + \text{radius of integration} + \text{low resolution cutoff} \)
Self rotation function

• In self rotation function, a Patterson map is rotated and then superimposed onto itself.
• Self rotation function detects non-crystallographic symmetry (NCS) and should be used in conjunction with the Matthew’s coefficient.
• Should be calculated before MR, MIR or MAD
  – NCS averaging is a powerful method in improving the quality of electron density maps.
Definition of rotation angles

Three consecutive rotations:
Rotate around the Z axis by $\alpha$
Rotate around the new position of the Y axis for $\beta$
Rotate around the newest Z axis position for $\gamma$

In this definition, $\phi$ and $\psi$ define the direction of the rotation axis, and $\kappa$ defines the rotation angle.
Correct solutions of rotation function should give $R$ values significantly above the noise level.

$$R(\kappa, \phi, \psi) = \int_{r_{\text{min}}}^{r_{\text{max}}} P_{\text{nat}}(u) P_{\text{mod}}(\kappa, \phi, \psi, u) du$$
Step 3 – Translation function

Translation function defines the unknown molecule’s position in the unit cell with respect to the origin.

The actual translation function again compares calculated and observed Pattersons. The information about the position is present in the intermolecular vectors so this time we use the full Patterson, minus the origin, and we again use a product function. In this case the product function can be evaluated by a Fourier transform using the correlation theorem.

\[
T(t) = \int_{cell} P_{2 \rightarrow 1}(u - t) P_{nat}(u) du
\]

\[
= \frac{1}{V} \sum_{h} \left( F_1(h) F_2^*(h) \right)^* \left| F_O(h) \right|^2 \exp(-2\pi i h \cdot t)
\]

\[
= \frac{1}{V} \sum_{h} F_1^*(h) F_2(h) \left| F_O(h) \right|^2 \exp(-2\pi i h \cdot t)
\]

When the model is correctly positioned, the maximum of T is obtained at t
Translation function and Patterson maps

For translation function, the integration is calculated over the whole unit cell. Both intermolecular and intramolecular vectors are included for comparison.

\[ T(t) = \int_{\text{cell}} P_{2 \rightarrow 1}(u - t) P_{\text{nat}}(u) d\mathbf{u} \]

\[ = \frac{1}{V} \sum_{\mathbf{h}} \left( \mathbf{F}_1(\mathbf{h}) \mathbf{F}_2^*(\mathbf{h}) \right)^* |\mathbf{F}_O(\mathbf{h})|^2 \exp(-2\pi i \mathbf{h} \cdot \mathbf{t}) \]

\[ = \frac{1}{V} \sum_{\mathbf{h}} \mathbf{F}_1^*(\mathbf{h}) \mathbf{F}_2(\mathbf{h}) |\mathbf{F}_O(\mathbf{h})|^2 \exp(-2\pi i \mathbf{h} \cdot \mathbf{t}) \]
Practical considerations for translation function

- **Origin**
  - The choice of origin is dependent on space group symmetry.

- **Screw axis**
  - For space group with 3, 4, and 6-fold screw axes, diffraction intensities do not provide sufficient information to differentiate 3(1)/3(2), 4(1)/4(3), 6(1)/6(5), or 6(2)/6(4) axes. Therefore, we have to try different possibilities.

- **Search volume**
  - Depends on the space group symmetry

- **Resolution**
  - Often 10 to 3Å
• Correct solutions of the translation function should give:
  – small R factors when $F_{\text{calc}}$ is compared to $F_{\text{obs}}$.  
    $$R_{\text{diff}} = \Sigma\{|F_{\text{calc}} - F_{\text{obs}}|/F_{\text{obs}}\}$$
  – Good crystal packing arrangements with very few intermolecular clashes
Step 4 – Phase determination

An initial set of phases can be determined by placing the known structure into the unit cell of the unknown structure using the rotation and translation.

The search model is used as a rough “guesstimate” of the unknown molecule.
Phasing model

Observed Diffraction intensity

Intensity

Phases

Calculated map

The tail is back!!! But if we use a dog as the phasing model, then ....

- The electron density map is calculated using observed structure factor amplitude and theoretical phases calculated from molecular replacement solution.
- Phases usually contain more structural information than the diffraction intensity
- The calculated structure may be biased towards the phasing model
- The more similar the phasing model is to the unknown, the less biased the calculated structure will be.
How to treat the model bias problem

• Use a poly-alanine model for molecular replacement. Side chain should emerge if phases are sufficiently accurate.

• Calculate a composite omit map and see if omitted features can be recovered.
  – A small fraction of the model, i.e. ~5%, is systematically omitted for phasing. The recovered densities in the omitted region are pasted together to form the so-called composite omit map.