Molecular Replacement

Andrey Lebedev

CCP4
http://www.ccp4.ac.uk/tutorials/

Basic phasing tutorials

(Includes both MR and Experimental Phasing)
MR Problem

Known crystal structure

New crystal structure

Given:  
- Crystal structure of a homologue  
- New X-ray data

Find:  
- The new crystal structure
MR Technique

Method:

- 6-dimensional global optimisation
  - one 6-d search for each molecule in the AU
    >> split further to orientation + translation searches = 3 + 3?

Required:

- Scoring
  - the match between the data and an (incomplete) crystal model
  - ideally: the highest score = the correct model
Two reasons to discuss the real and reciprocal spaces

Perception of crystallographic methods

• Many concepts formulated in one space have their counterparts in another space
• The concepts formulated in real space are sometimes more intuitive

Terminology vs. methods

• Sometimes names are taken from real space but assume calculations in the reciprocal space:
  – "Search in the electron density"
  – "Patterson search"
Functions in Real and Reciprocal spaces

Real space $\leftrightarrow \mathcal{F} \rightarrow$ Reciprocal space

Map $\rho(r)$ $\leftrightarrow \mathcal{F} \rightarrow$ F(s) $\rightarrow$ structure factors

$\downarrow$ $\downarrow$

convolution $\leftrightarrow \mathcal{F} \rightarrow$ product

$\downarrow$ $\downarrow$

Patterson $P(r)$ $\leftrightarrow \mathcal{F} \rightarrow$ F(s) $F^*(s) = I(s)$ intensities

Next few slides: $F$ and $I$ as experimental data
Structure factors $F(h,k,l)$
- A discrete complex function in the reciprocal space

Each $F$:
- Complex number: $F = A + iB$
- Can be expressed via structure amplitude and phase $F = |F|\exp(i\phi)$

Electron density map
- periodic 3-d function in real space

is directly interpretable
- model building
- real-space fitting of fragments
**Intensities**  \[ I(h,k,l) \]  
- 3-d discrete real function in the reciprocal space

**Patterson map:**  
- 3-d function in real\(^{(*)}\) space

- Nothing reminds a protein molecule  
- **Model building**, residue by residue, is impossible
MR: Two distinct cases dependent on availability of phases

- **Data = structure factors** *(include phases)*
  - "Search in the electron density"
  - Electron density maps are compared: calculated vs. observed
  - Can be treated in a more straightforward way (model building)
    » Useful in special cases

- **Data = observed intensities** *(no phases)*
  - "Patterson search"
  - Patterson maps are compared: calculated vs. observed
  - Direct model building is impossible in the absence of phases
    » The most common case of MR

As a rule, all computations are in the reciprocal space
Self and cross vectors

Electron density map = peaks from all atoms
Patterson map = peaks from all interatomic vectors

- **self-vectors**: vectors between atoms belonging to the same molecule
- **cross-vectors**: vectors between atoms belonging to different molecules

Electron density map

Patterson map

One molecule

\[ r_1 - r_1 \]
\[ r_2 - r_2 \]

Two molecules

\[ r_1 - r_2 \]
\[ r_2 - r_1 \]

Contribution from **self-vectors**
- is centred at the origin
- dominates in a vicinity of the origin
What can be seen to be separated?

Electron density maps:

• Peaks from atoms or larger fragments are separated in space
  – Model building is possible

Patterson map:

• Contribution from self-vectors is centred at the origin
• Self-vectors are, in average, shorter than cross-vectors
  – Peaks from self-vectors dominates in a vicinity of the origin
  – Peaks from cross-vector dominates away from the origin

• One 6-dimensional search splits into
  – Rotation Function: 3-dimensional search (using self-vectors)
  – Translation Function: 3-dimensional search (using cross-vectors)
Rotation Function

\[ RF(\alpha, \beta, \gamma) = P_{\text{obs}} \times P_{\text{self}}^{\text{calc}}(\alpha, \beta, \gamma) \]
Rotation Function

$$RF(\ ,\ ,\ ) = \int P^{obs}\ P^{calc(\ ,\ ,\ )}d\mathbf{r}^3$$

$P^{calc(\ ,\ ,\ )}$ contains only

- **self-vectors**

$P^{obs}$ contains

- **self-vectors (mostly signal)**
- **cross-vectors (noise)**
Analytical steps

- The centre of molecule 1
  - Parameter $t$

- Centres of molecules 2, 3 and 4
  - Form symmetry operations

- $F_h^{\text{calc}}(t), \ I_h^{\text{calc}}(t)$

- $TF(t) = \sum_h I_h^{\text{obs}} I_h^{\text{calc}}(t)$

- Can be converted to a form:
  
  $TF(t) = \sum_h G_h \exp(2 \pi i h \times t)$

  - FFT techniques can be applied

Numerical calculations

- $G_h$

- FFT: $G_h \rightarrow TF(t)$

- Peak search in $TF(t)$: best $t$
Almost the same equation as for a single molecule search,

\[ TF(t) = \sum_h I_h^{\text{obs}} \left| F_h^{\text{fixed}} + F_h^{\text{calc}}(t) \right|^2 \]

Again, can be converted to a form:

\[ TF(t) = \sum_h G_h \exp(2i\ h \times t) \]

and FFT technique can be used
Translation Function

\[ TF(t) = P_{\text{obs}} \int P_{\text{cross}}(t) \, dr \]

\( P_{\text{calc}}(t) \) contains only

- cross-vectors

\( P_{\text{obs}} \) contains

- self-vectors (noise)
- cross-vectors (mostly signal)
Molecules in the crystal **do not overlap**

**How can we use this information?**

» Patterson map does not explicitly reveal molecular packing

**Reject MR solutions**

• Restrict distance between centres of molecules
• Count close interatomic contacts

**Modify TF**

• Divide by Overlap Function
• Multiply by Packing function
Estimation of overlap
• Using mask from search model
• FFT

Packing Function
\[ PF = 1 - \text{overlap} \]

Modified Translation Function
\[ TF' = TF \times PF \]

Peak search
• Using \( TF' \)
• No irrelevant peaks are passed to rescoring step

Implemented in \textit{MOLREP}
As implemented in *MOLREP*

\[ RF = \sum_{hkl} w^* I_o^* I_c(\alpha\beta\gamma) \]

\[ TF = \sum_{hkl} w^* I_o^* I_c(xyz) \]

Rescoring: Correlation Coefficient* PF

**Diagram:**
- **Search model** → **RF** → **TF * PF** → **Rescoring** → **Partial structure**
- **TF * PF** → **Rescoring** → **Possible solution**
Molrep

Alexey Vagin

YSBL University of York
molrep -f data.mtz -m model.pdb -mx fixed.pdb -s target.seq
Molrep default protocol

molrep -f data.mtz -m model.pdb -mx fixed.pdb -s target.seq

- model correction if sequence provided
- defines the number of molecules per AU
- modification of the model surface
- anisotropic correction of the data
- weighting the data according to model completeness and similarity
- check for pseudotranslation and use it if present
- 30+ peaks in Cross RF for use in TF (accounts for close peaks)
- applied packing function
- make use of partial structure (fixed model)
Molecular Replacement

Gábor Bunkóczi
Model error

Position errors

Model incompleteness
Model error

Position errors

Calculate from sequence identity (Chothia & Lesk)

Model incompleteness

Calculate from unit cell composition
Likelihood of $|F_{\text{obs}}|$ (given $R_1, r_1, D, \sigma_\Delta$)

$DF_c(R_1, r_1)$
Likelihood of $|F_{\text{obs}}|$ 

$(\text{given } R_2, r_2, D, \sigma_\Delta)$
Likelihood of $|F_{\text{obs}}|$ (given $R_3, r_3, D, \sigma_\Delta$)

$DF_c(R_3, r_3)$

\((R_1, r_1)\) \hspace{1cm} \((R_2, r_2)\) \hspace{1cm} \((R_3, r_3)\)
Likelihood usage

Calculate approximation by FFT

Peak search

Rescore with likelihood
The rotation and translation functions were performed on a (not very fine) grid.

The solution can be improved if the grid is taken away and the rotational and translational parameters optimized.
Fast search

Back to classical MR? Not quite

Advantages of using Likelihood approach
• Final scoring is the most robust
• Maximum consistency between preliminary and final scoring

A way to push boundaries of the method
Log Likelihood Gain (LLG)

- $p_1 = p(E_{OBS} \mid E_{CALC} = 1.3, A = 0.8)$
  The model is assumed to be good

- $p_0 = p(E_{OBS} \mid E_{CALC} = \text{Any}, A = 0)$
  The model is assumed to be random

- $\text{LLG}_1 = \log p_1 - \log p_0$
  Log Likelihood Gain (LLG)

- An empirical criterion $\text{LLG} > 70$:
  - identification of the MR successful run
  - prediction of sufficient amount of data (resolution cut-off)
Identifying solutions

TFZ: 2 3 4 5 6 7 8 9 10

06/11/2014 KEK-CCP4 Workshop
A small number of clashes is acceptable
Tree search with pruning

1\textsuperscript{st} component

Pruned
Propagated

Perform search

2\textsuperscript{nd} component
Model search order

65% of structure
35% identical model

Better model if data resolution is low

35% of structure
50% identical model

Better model if data resolution is high

Determined automatically!
If automated MR does not work
• Automated MR is quite advanced, is convenient, but
  – still may fail to produce complete model
  – may be very slow if there are many components to position
  – limited feedback from user

In general, automated MR is good for finding partial model, but it is not the most efficient way to generate a nearly complete model.

• Combine MR with
  – refinement of partial model
  – density improvement
  – manual or automated model building

MR pipelines: combinations with other automated methods

• Specialised MR techniques can be useful
• Handling Translational Non-Crystallographic symmetry
  – Non-origin peaks in the Patterson map indicate the presence of TNCS
  – Molecules related by TNCS can be found in one go as they have the same orientation

• Locked RF and TF
  – Using point symmetry of oligomers

• Exhaustive and stochastic searches

• Using phase information in MR

• Self Rotation Function
Refinement refinement + Phased MR

Partial structure

\[ \text{Refmac} \rightarrow \text{Map coefficients} \]

Search model

\[ \text{Molrep} \rightarrow \text{Extended model} \]

Search in the map

- Calculate 2-1 or 1-1 maps after restrained refinement of partial structure
- Flatten the map corresponding to the known substructure
- Calculate structure amplitudes from this map
- Use these modified amplitudes in Rotation Function

  – Theoretically quite similar to RF with fixed model in Phaser

- And finally – Phased TF
Spherically Averaged Phased Translation Function
(FFT based algorithm)

\[ \text{SAPTF}(s) = \int_{\text{Map}(s,r)} - \int_{\text{Model}(r)} r^2 \, dr \]
1. Find approximate position:
   Spherically Averaged \textit{Phased} Translation Function

2. Find orientation:
   \textbf{Local} \textit{Phased} Rotation Function
   – Local search of the orientation in the density

3. Verify and adjust position:
   \textit{Phased} Translation Function
1. Find approximate position:
   Spherically Averaged Phased Translation Function

2. Find orientation:
   Local Rotation Function
   – Structure amplitudes from the density within the SAPTF sphere

3. Verify and adjust position:
   Phased Translation Function

   • Local RF is less sensitive than Phased RF to inaccuracy of the model position
• Asymmetric unit: two copies
• Resolution: 2.8 Å

Phane et al. (2011) Nature, 474, 50-53
Usher complex structure solution

1. Conventional MR
   - FimC-N + FimC-C
   - FimH-L + FimH-P
   - FimD-Pore

2. Jelly body refinement (Refmac)
   - FimD-Pore

3. Fitting into the electron density
   - FimD-Plug
   - FimD-NTD
   - FimD-CTD-2

4. Manual building
   - FimD-CTD-1
**Performance of fitting methods**

<table>
<thead>
<tr>
<th>search model</th>
<th>sequence identity</th>
<th>&quot;Masked&quot; RF PTF</th>
<th>SAPTF PRF PTF</th>
<th>SAPTF Local RF PTF</th>
</tr>
</thead>
<tbody>
<tr>
<td>FimD-Plug</td>
<td>3fip_A</td>
<td>prf n</td>
<td>– (–)</td>
<td>1 (2)</td>
</tr>
<tr>
<td>FimD-NTD</td>
<td>1ze3_D</td>
<td>2 (2)</td>
<td>1 (2)</td>
<td>2 (2)</td>
</tr>
<tr>
<td>FimD-CTD-2</td>
<td>3l48_A</td>
<td>– (–)</td>
<td>2 (2)</td>
<td>– (–)</td>
</tr>
</tbody>
</table>

Trying several methods is a good practice (also because of cross-validation)
Twinned crystal with pseudo-symmetric substructure

Human macrophage receptor CLEC5A for dengue virus

- Substructure (A): Patterson search (4 copies), search in density (2 copies)
- Substructure (B): Search in the density (3 copies)

3-fold axes with respect to the true structure:
- ▲ crystallographic
- ▲ ▲ pseudosymmetry for (A)
Fitting into EM maps

*SPP1 portal protein*
Self Rotation Function (SRF)

Preliminary analysis of X-ray data
- Oligomeric state of the protein in crystal
- Selection of oligomeric search model

Limited use
- No clear interpretation or even artifact peaks in high symmetry point groups (e.g. 622)
- Different oligomers with the same symmetry

Example of SRF
- Space group P21
- One 222-tetramer in the AU
Locked Rotation Function

- Uses SRF to derive NCS operations
- Averages RF over NCS operations
- In favorable cases Improves signal to noise ratio in RF

Automatic mode:

```
molrep -f s100.mtz -m monomer.pdb -s s100.seq -i <<+ lock y +
```

There is an option of selecting specific SRF peaks
One-dimensional exhaustive search (exotic case)

\[ \chi_{SRF} = 120^\circ \]

\[ \chi_{SRF} = 180^\circ \]

SRF helps restrict dimensionality in an exhaustive search

- Orientation of the ring is known from the analysis of SRF
- Unknown parameter: rotation about 3-fold axis
- One-parametric exhaustive search using TF as score function
MR substructure solution (exotic case)

• Select isomorphous derivative
  – by comparing native SRF and SRF from D-iso

  \[ F_{\text{obs}} \quad \text{(native)} \quad D_{\text{iso}} \quad \text{(Hg-1 – native)} \quad D_{\text{iso}} \quad \text{(Hg-2 – native)} \]

  \[ x = 180^\circ \]

• Hg-substructure is a 13-atom ring (from native SRF analysis)
  – Orientation of the ring is known from the analysis of SRF
  – Unknown parameters: radius of the ring, rotation about 13-fold axis

• Two-parametric exhaustive search
Which direction does MR go?

Automation. Therefore:

❌ Collection of tricks

✔ Improvement of "standard" methods
✔ Better scoring system

✔✔ Models
Model improvement

✓ Structural information in MR models
✓ Extra information in sequence alignment
✓ Many homologous structures

• Remove residues that do not align
• Remove "excessive" atoms from aligned residues
• Ensemble models (several superposed models)
CCCP4 programs for model preparation

Single model correction:
- Chainsaw
- Molrep
- Sculptor

Preparation of ensemble models – fitting models:
- Lsqkab
- ProSMART
- SSM (also in Coot)
- Gesamt

Automated preparation of ensemble models:
- Ensembler
Model modification in MOLREP

- Performs model correction:
  - Identifies secondary structure in the model
  - Aligns target and model sequences
    » no deletions or insertions in α-helixes or β-strands
  - Retains aligned residues
  - Retains "aligned" atoms in aligned residues

- Adds B-factor to residues exposed to solvent

- Uses sequence identity to down-weight high resolution data

```
molrep -f data.mtz -m model.pdb -s target.seq
```
Molecular Replacement in CCP4

MR Programs

• AMoRe
• Molrep
• Phaser

MR Pipelines

• MrBUMP
• Balbes
• AMPLE
Acknowledgements

**Molrep**

Alexey Vagin, University of York, UK
Garib Murshudov, MRC LMB, UK

**Phaser**

Randy Read, University of Cambridge
Airlie McCoy, University of Cambridge
Gabor Bunkozci, University of Cambridge

**FimD-FimC-FimH**

Gilles Phan, ISMB, London, UK
Gabriel Waksman, ISMB, London, UK