Refinement of Macromolecular Crystal Structures

Garib N Murshudov
MRC-LMB
Cambridge
Contents

1) Purpose of refinement
2) Model and Data
3) Model parameters
4) Bulk solvent modelling
Purpose of refinement

Crystallographic refinement has two purposes:

1) Fit chemically and structurally sensible atomic model into observed–X-ray crystallographic data

2) To calculate best possible (electron density) map so that atomic model can be rebuild
We have observed amplitudes: $|F_{\text{obs}}|$

But we don’t have phases: $\varphi$
Suppose we have a starting model:

| H, K, L | |F_{calc}| | Φ_{calc} |
|---------|--------------|--------|----------------|
| ...     |             |        |                |
| 5, 5, 5 | 355         |        | 27°            |
| 5, 5, 6 | 387         |        | 8°             |
| 5, 5, 7 | 146         |        | 75°            |
| 5, 5, 8 | 340         |        | 31°            |
| ...     |             |        |                |
Model Refinement

Idea:
Iteratively improve the model, optimising the agreement between $|F_{\text{obs}}|$ and $|F_{\text{calc}}|$.

Purpose: improve phase estimates: $\phi_{\text{calc}}$.
Knowledge about macromolecules used in refinement

1. Macromolecules consist of atoms bonded to each other in a specific way
   - Standard restraints: bonds, angles etc
2. Oscillation of atoms close to each other in 3D cannot be dramatically different
   - B-factor restraints, TLS restraints
3. If there are two copies of the same molecule present then they will likely be similar to each other
   - NCS/local symmetry restraints
4. If there are two molecules with sufficiently high sequence identity then it is likely that they will be structurally similar
   - External restraints to homologous structures – ProSMART
5. Proteins tend to form secondary structures
   - Generic H-bonding restraints – ProSMART
6. DNA/RNA tend to form base-pairs, stacked bases tend to be parallel
   - Generic base-pair and stacking restraints – LibG
Restraints

Standard restraints (used by default) include:

• Bond lengths
• Angles
• Chirals
• Planes
• Some torsion angles
• B–values
• VDW repulsions

These help to ensure that the model is chemically sensible

Note – we generally deal with restraints, not constraints
NCS
Three ways of dealing with NCS

1) NCS constraints: copies of molecules are considered to be exactly same. Only one set of atomic parameters per molecule is refined, other copies are kept to be exactly same.

2) NCS restraints: Molecules are superimposed and difference between corresponding atoms after superposition minimised.

3) NCS local restraints: Molecules are assumed to be locally similar, globally they may be different.
Auto NCS: local and global

1. Align all chains with all chains using Needleman-Wunsh method
2. If alignment score is higher than predefined (e.g. 80%) value then consider them as similar
3. Find local RMS and if average local RMS is less than predefined value then consider them aligned
4. Find correspondence between atoms
5. If global restraints (i.e. restraints based on RMS between atoms of aligned chains) then identify domains
6. For local NCS make the list of corresponding interatomic distances (remove bond and angle related atom pairs)
7. Design weights

The list of interatomic distance pairs is calculated at every cycle
In many cases it could be expected that two or more copies of the same molecule will have (slightly) different conformation. For example if there is a domain movement then internal structures of domains will be same but between domains distances will be different in two copies of a molecule.
External (reference) structure restraints

Restrains to external structures are generated by the program ProSmart:
1) Aligns structure in the presence of conformational changes. Sequence is not used
2) Generates restraints for aligned atoms
3) Identifies secondary structures (at the moment helix and strand, but the approach is general and can be extended to any motif).
4) Generates restraints for secondary structures

Note 1: ProSmart has been written by Rob Nicholls and available from CCP4.

Note 2: Robust estimator functions are used for restraints. I.e. if differences between target and model is very large then their contributions are down-weighted
External reference restraints

ProSMART: restraints from homologous structures
Basepair restraints

monomers C G  label C:G
bond atom N4 C atom O6 G value 2.91 sigma 0.15
bond atom N3 C atom N1 G value 2.95 sigma 0.1
bond atom O2 C atom N2 G value 2.86 sigma 0.15
chiral atom N1 G atom C6 G atom C2 G atom N3 C value 0.0 sigma 0.5
chiral atom N3 C atom C4 C atom C2 C atom N1 G value 0.0 sigma 0.5
torsion atom C4 C atom N3 C atom N1 G atom C2 G value 180.0 sigma 10.0
Tools for generation of extra DNA/RNA restraints

- Generate extra restraints on sub-structures such as base-pairs, backbone conformation, stacking planes in DNA, RNA and their protein complex for refinement

- Data from single-particle electron cryo-microscopy
- Resolution 3.2Å
- G:U base pair. Purple: before refinement. Yellow: after refinement. The base-pair is put to the same plane via refinement
Tools for generation of extra DNA/RNA restraints

- Generate extra restraints on sub-structures such as base-pairs, backbone conformation, stacking planes in DNA, RNA and their protein complex for refinement

- Data from single-particle electron cryo-microscopy
- Resolution 3.2Å
- Green: before refinement. Blue: after refinement without stacking. Yellow: after refinement with stacking
Restraints to current distances (jelly-body)

The term is added to the target function:

$$\sum_{\text{pairs}} w(|d| - |d_{\text{current}}|)^2$$

Summation is over all pairs in the same chain and within given distance (default 4.2A). $d_{\text{current}}$ is recalculated at every cycle. This function does not contribute to gradients. It only contributes to the second derivative matrix.

It is equivalent to adding plastic springs between atom pairs. During refinement inter-atomic distances are not changed very much. If all pairs would be used and weights would be very large then it would be equivalent to rigid body refinement.

It could be called “implicit normal modes”, “soft” body or “jelly” body refinement.
Usher complex structure solution

Jelly body refinement (Refmac)
Data and their property

1) Amplitudes of structure factors from single crystals - $|F_o|$, $\sigma_o$. It is the most common case. Usually structure factors are reliable, uncertainties are not so.

2) Intensities or amplitudes are from “twinned” crystals

3) Amplitudes of structure factors are available for $|F+|$ and $|F-|$ - SAD case

4) Amplitudes of structure factors are available from multiple crystal forms
Crystallographic refinement

The function in crystallographic refinement has a form:

\[ L(p) = wL_X(p) + L_G(p) \]

Where \( L_X(p) \) is -loglikelihood and \( L_G(p) \) is -log of prior probability distribution – restraints: bond lengths, angles etc.

It is one of many possible formulations. It uses Bayesian formulation. Other formulation is also possible.
TWIN
**merohedral and pseudo-merohedral twinning**

<table>
<thead>
<tr>
<th>Crystal symmetry:</th>
<th>P3</th>
<th>P2</th>
<th>P2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Constrain:</strong></td>
<td>-</td>
<td>$\beta = 90^\circ$</td>
<td>-</td>
</tr>
<tr>
<td>*<em>Lattice symmetry <em>:</em></em></td>
<td>P622</td>
<td>P222</td>
<td>P2</td>
</tr>
<tr>
<td>(rotations only)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Possible twinning:</strong></td>
<td>merohedral</td>
<td>pseudo-merohedral</td>
<td>-</td>
</tr>
</tbody>
</table>

**Domain 1**

**Domain 2**

Crystal lattice is invariant with respect to twinning operator.

The crystal is NOT invariant with respect to twinning operator.
Effect on intensity statistics

Take a simple case. We have two intensities: weak and strong. When we sum them we will have four options w+w, w+s, s+w, s+s. So we will have one weak, two medium and one strong reflection.

As a results of twinning, proportion of weak and strong reflections becomes small and the number of medium reflections increases. It has effect on intensity statistics

In probabilistic terms: without twinning distribution of intensities is $\chi^2$ with degree of freedom 2 and after perfect twinning degree of freedom increases and becomes 4. $\chi^2$ distributions with higher degree of freedom behave like normal distribution
# Twin: Few warnings about R values

R values for random structures (no other peculiarities)

<table>
<thead>
<tr>
<th>Twin</th>
<th>Modeled</th>
<th>Not modeled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>0.41</td>
<td>0.49</td>
</tr>
<tr>
<td>No</td>
<td>0.52</td>
<td>0.58</td>
</tr>
</tbody>
</table>

R-value for structures with different model errors: Combination of real and modeled perfect twin fractions
Where’s the density for my ligand (2.15Å)?

R-factor (R-free) 25.5% (26.9%) – after initial rigid body and restrained refinement.
Fo-Fc – 3 sigma

R-factor (R-free) 15.9% (16.3%) – re-run restrained ref. with twin on (refined twin fractions 0.6043/0.3957).
Fo-Fc – 3 sigma

Borrowed from B. Bax, GSK, Stevenage, UK
Model Parameterisation

Usual parameters (if programs allow it)
1) Positions x,y,z
2) B values – isotropic or anisotropic
3) Occupancy

Derived parameters
4) Rigid body positional
   • After molecular replacement
   • Isomorphous crystal (liganded, unliganded, different data)
5) Rigid body of B values – TLS
   – Useful at the medium and final stages
   – At low resolution when full anisotropy is impossible
6) Torsion angles
Model Parameterisation

Standard refinable parameters

Atomic model:
• Position – (x,y,z) coordinates
• Uncertainty – B–factors
• (Occupancies)

Overall parameters (scaling)
• Overall B–factor (and anisotropic U)
• Solvent treatment

<table>
<thead>
<tr>
<th>ATOM</th>
<th>CB</th>
<th>ASP A</th>
<th>8</th>
<th>-30.909</th>
<th>9.723</th>
<th>18.264</th>
<th>1.00</th>
<th>33.70</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATOM</td>
<td>CG</td>
<td>ASP A</td>
<td>8</td>
<td>-31.252</td>
<td>9.345</td>
<td>16.825</td>
<td>1.00</td>
<td>41.96</td>
<td>C</td>
</tr>
<tr>
<td>ATOM</td>
<td>OD1</td>
<td>ASP A</td>
<td>8</td>
<td>-31.072</td>
<td>10.248</td>
<td>15.981</td>
<td>1.00</td>
<td>46.18</td>
<td>O</td>
</tr>
</tbody>
</table>
Model Parameterisation

Standard refinable parameters

Atomic model:
• Position – \((x, y, z)\) coordinates
• Uncertainty – B–factors
• (Occupancies)

Overall parameters (scaling)
• Overall B–factor (and anisotropic U)
• Solvent treatment

<table>
<thead>
<tr>
<th>ATOM</th>
<th>CB</th>
<th>ASP A</th>
<th>8</th>
<th>(x)</th>
<th>(y)</th>
<th>(z)</th>
<th>(B)</th>
<th>C (U)</th>
<th>Solvent</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td></td>
<td></td>
<td>8</td>
<td>-30.909</td>
<td>9.723</td>
<td>18.264</td>
<td>1.00</td>
<td>33.70</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td>8</td>
<td>-31.252</td>
<td>9.345</td>
<td>16.825</td>
<td>1.00</td>
<td>41.96</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td>8</td>
<td>-31.072</td>
<td>10.248</td>
<td>15.981</td>
<td>1.00</td>
<td>46.18</td>
<td></td>
</tr>
</tbody>
</table>
Standard refinable parameters

Atomic model:
- Position – \((x,y,z)\) coordinates
- **Uncertainty** – B–factors
- (Occupancies)

Overall parameters (scaling)
- Overall B–factor (and anisotropic U)
- Solvent treatment

<table>
<thead>
<tr>
<th>ATOM</th>
<th>CB</th>
<th>ASP A</th>
<th>5</th>
<th>-30.909</th>
<th>9.723</th>
<th>18.264</th>
<th>1.00</th>
<th>33.70</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATOM</td>
<td>CG</td>
<td>ASP A</td>
<td>6</td>
<td>-31.252</td>
<td>9.345</td>
<td>16.825</td>
<td>1.00</td>
<td>41.96</td>
<td>C</td>
</tr>
<tr>
<td>ATOM</td>
<td>OD1</td>
<td>ASP A</td>
<td>7</td>
<td>-31.072</td>
<td>10.248</td>
<td>15.981</td>
<td>1.00</td>
<td>46.18</td>
<td>O</td>
</tr>
</tbody>
</table>
Model Parameterisation

Standard refinable parameters

Atomic model:
- Position – (x,y,z) coordinates
- Uncertainty – B-factors
- (Occupancies)

Overall parameters (scaling)
- Overall B-factor (and anisotropic U)
- Solvent treatment
Model Parameterisation

Standard refinable parameters

Atomic model:
- Position – (x,y,z) coordinates
- Uncertainty – B–factors
- (Occupancies)

Overall parameters (scaling)
- Overall B–factor (and anisotropic U)
- Solvent treatment

| ATOM | 5 | CB  | ASP A | 8 | -30.909 | 9.723 | 18.264 | 1.00 | 33.70 | C  |
| ATOM | 6 | CG  | ASP A | 8 | -31.252 | 9.345 | 16.825 | 1.00 | 41.96 | C  |
| ATOM | 7 | OD1 | ASP A | 8 | -31.072 | 10.248 | 15.981 | 1.00 | 46.18 | O  |
Notes on B–factors

- Overall B–factor and atomic B–factor are different!
- Atomic B–factors can be modelled in a different way according to the quality of the data: **isotropic** (1 parameter per atom), **anisotropic** (6 parameters per atom), **TLS** (20 parameters per group of atoms)
- B–factors are sometimes also referred to as atomic displacement parameters (ADPs) or thermal/temperature factors
- B–factors describe relative positional uncertainty
- Should not compare atomic B–factors between different models
TLS Groups

Describe rigid body motion – e.g. for chains/domains/subunits

Suitable for medium resolution, when full anisotropy is impossible

Per group (20 parameters):
- **Translation** – 6 parameters
- **Libration** – 6 parameters
- **Screw rotation** – 8 parameters

Define groups using CCP4i
or TLSMD webserver:
http://skuld.bmsc.washington.edu/~tlsmd/
Bulk solvent
Method 1: Babinet’s bulk solvent correction

At low resolution electron density is flat. Only difference between solvent and protein regions is that solvent has lower density than protein. If we would increase solvent just enough to make its density equal to that of protein then we would have flat density (constant). Fourier transformation of constant is zero (apart from F000). So contribution from solvent can be calculated using that of protein. And it means that total structure factor can calculated using contribution from protein only

\[
\rho_s + \rho_p = \rho_T \quad \leftrightarrow \quad F_s + F_p = F_T \\
\rho_s + k\rho_p = c \quad \leftrightarrow \quad F_s + kF_p = 0 \\
F_s = -kF_p \quad \Rightarrow \quad F_T = F_p - kF_p = (1-k)F_p
\]

k is usually taken as \( k_b \exp(-B_b s^2) \). \( k_b \) must be less than 1. \( k_b \) and \( B_b \) are adjustable parameters.
Bulk solvent
Method 2: Mask based bulk solvent correction

Total structure factor is the sum of protein contribution and solvent contribution. Solvent region is flat. Protein contribution is calculated as usual. The region occupied by protein atoms is masked out. The remaining part of the cell is filled with constant values and corresponding structure factors are calculated. Finally total structure factor is calculated using

\[ F_T = F_p + k_s F_s \]

\( k_s \) is adjustable parameter.

Mask based bulk solvent is a standard in all refinement programs. In refmac it is default.
Map calculation

- After refinement programs give coefficients for two type of maps: 1) 2Fo-Fc type maps. 2) Fo-Fc type of maps. Both maps should be inspected and model should be corrected if necessary.

- Refmac gives coefficients:

\[ 2m F_o - D F_c \] – to represent contents of the crystal

\[ m F_o - D F_c \] - to represent differences

\( m \) is the figure of merit (reliability) of the phase of the current reflection and \( D \) is related to model error. \( m \) depends on each reflection and \( D \) depends on resolution. Unobserved reflections are replaced by DFc.

If phase information is available then map coefficients correspond to the combined phases.
Available refinement programs

• SHELXL
• CNS
• **REFMAC5**
• TNT
• BUSTER/TNT
• Phenix.refine
• RESTRAINT
• MOPRO
• XD
• MAIN
What can REFMAC do?

- Simple maximum likelihood restrained refinement
- Twin refinement
- Phased refinement (with Hendrickson-Lattmann coefficients)
- SAD/SIRAS refinement
- Structure idealisation
- Library for more than 10000 ligands (from the next version)
- Covalent links between ligands and ligand-protein
- Rigid body refinement
- NCS local, restraints to external structures
- Helical, point group NCS constraints
- TLS refinement
- Fit into EM map
- Map sharpening
- etc
What and when

- Rigid body: At early stages - after molecular replacement or when refining against data from isomorphous crystals
- “Jelly” body – At early stages and may be at low resolution
- TLS - at medium and end stages of refinement at resolutions up to 1.7-1.6Å (roughly)
- Anisotropic - At higher resolution towards the end of refinement
- Adding hydrogens - they could be added always
- Phased refinement - at early and medium stages of refinement
- SAD – at the early stages
- Twin – when you are sure that crystal is twinned
- NCS local – always?
- Ligands - as soon as you see them
- What else?
Conclusion

- Information about ligands and their chemistry should be used in refinement.
- NCS restraints are useful tools.
- External restraints can be a powerful tool for reliable atomic model derivation.
- “Jelly” body can be very powerful at the early stages of refinement.
- Twin refinement improves statistics and occasionally electron density: Rfactors may be misleading.
- Refinement is just one step in X-ray structure analysis – it is often used as part of phase improvement and model building.
Acknowledgment

Rob Nicholls          Pavol Skubak          Alexei Vagin
Fei Long              Raj Pannu              Andrey Lebedev
Paul Emsley           Lamzin group, Hamburg
Andrea thorn          Perrakis group, Amsterdam
Oleg Kovalevskiy

CCP4, LMB people

REFMAC is available from CCP4 and from our website:
http://www2.mrc-lmb.cam.ac.uk/groups/murshudov/