Restraints in Cryo-EM Refinement

CCP-EM Icknield Model Building Workshop
15-18 April 2019
RAL/DLS Oxfordshire UK

Rob Nicholls
nicholls@mrc-lmb.cam.ac.uk
Purpose of Refinement

To fit an atomic model into observed data

- *Model should agree with the observed data*
- *Model must be chemically and structurally sensible*

Data + Atomic model

Fit and refine

Structure factors (reconstruction/map) are treated as observations...
Model Refinement

REFMAC5 uses a Maximum Likelihood approach

Target functions have two components:

\[ f_{\text{tot}} = w f_{\text{data}} + f_{\text{geom}} \]

\[ f_{\text{data}} = -\log[P(\text{obs}; \text{model})] \]
\[ f_{\text{geom}} = -\log[P(\text{model})] \]
\[ w : \text{relative weighting} \]

We have:

• Data – to refine our model against (SFs corresponding to a reconstruction)
• Parameters to refine - describing the model

We also need prior knowledge (restraints)

*These help ensure chemical and structural integrity*
Why Restraints?

Example: two-atom ideal case

Distance between atoms 1.3Å. B-factors 20 and 50

Thin lines – single atoms

Bold line - sum of the two atoms
Why Restraints?

Example: two-atom ideal case

Distance between atoms 1.3Å. B-factors 20 and 50

Thin lines – single atoms

Bold line - sum of the two atoms
Why Restraints?

Example: two-atom ideal case

Distance between atoms 1.3Å. B-factors 20 and 50

Thin lines – single atoms

Bold line - sum of the two atoms
Why Restraints?

Example: Phe at two different resolutions

0.88 Å

2 Å and high mobility
Restraints

Standard restraints (used by default) include:

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

Note – REFMAC5 generally deals with restraints, not constraints (mostly…)
Restained refinement can only be used for high-resolution cryo-EM
What do we know about Macromolecules?

Introduce additional restraints based on our knowledge:

1. Macromolecules consist of atoms bonded to each other in a specific way
   - **Specified by current state of the model**

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 restraints – local/global depending on resolution**

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**
NCS
(Non-Crystallographic Symmetry Restraints)

Three ways of dealing with NCS:

1. NCS constraints
   • NCS-related copies are considered to be exactly the same
   • Only one set of atomic parameters per molecule is refined

2. Global NCS restraints

3. Local NCS restraints
NCS
(Non-Crystallographic Symmetry Restraints)

Three ways of dealing with NCS:

1. NCS constraints
2. Global NCS restraints
   • Molecules are superimposed
   • Differences between corresponding atoms are minimised
3. Local NCS restraints
NCS
(Non-Crystallographic Symmetry Restraints)

Three ways of dealing with NCS:

1. NCS constraints
2. Global NCS restraints
3. Local NCS restraints
   • Molecules are assumed to be locally similar
   • However, they may adopt (slightly) different global conformations
   • Restrain differences between local interatomic distances
Restraints

Why introduce so many restraints?

Answer: to improve the observation:parameter ratio.
Restraints

Why introduce so many restraints?

Answer: to improve the observation:parameter ratio.

Example: Fitting a line $y = a + bx$ (2 parameters)
Restraints

Why introduce so many restraints?

Answer: to improve the observation:parameter ratio.

Example: Fitting a line $y = a + bx$ (2 parameters)
Restraints

Why introduce so many restraints?

Answer: to improve the observation:parameter ratio.

Example: Fitting a line $y = a + bx$ (2 parameters)
Why introduce so many restraints?

Answer: to improve the observation:parameter ratio.

Can fit a line

Line is unreliable

Example: Fitting a line

\[ y = a + bx \]

(2 parameters)
Restraints

Why introduce so many restraints?

Answer: to improve the observation:parameter ratio.

Example: Fitting a line $y = a + bx$ (2 parameters)
Restraints

Why introduce so many restraints?

Answer: to improve the observation:parameter ratio.

Insufficient observations!

Unstable refinement

Example: Fitting a line

\[ y = a + bx \]

(2 parameters)

Ill-posed problem
Restraints

How can we improve the observation:parameter ratio?

1. Reduce number of parameters

Typically 4 parameters per atom:

- Coordinates – 3 parameters
- B-factor – 1 parameter

Alternatively:

- Rigid body refinement – 9 parameters per body
Restraints

How can we improve the observation:parameter ratio?

1. Reduce number of parameters

2. Increase number (and strength) of restraints
   - B-value restraints
   - NCS restraints
   - H-bond and secondary-structure restraints
   - Restraints to homologous known structures
   - Nucleic acid base-pair and base-stacking restraints
   - Jelly-body restraints

Use of prior knowledge to stabilize refinement – Regularisation
Regularisation
Regularisation

Example:

\[ z = (x + y)^2 \]
Regularisation

Example:

\[ z = (x + y)^2 + (|x - y| - 4)^2 \]

Regularise using prior information:

\[ |x - y| = 4 \]
Regularisation

Use of available information (prior knowledge):

Regularisers with a target value:
- Geometry restraints (chemical information)
- B-value restraints
- Local NCS restraints – where applicable
- External restraints – where available

Regularisers without a target value:
- Jelly-body restraints
Jelly Body Restraints

Regularisers without a target:

\[ f = \frac{1}{2} (d - d_{\text{current}})^2 \]

- close atom pairs
- \( d \): interatomic distance
- \( d_{\text{current}} \): current interatomic distance
- \( \sigma \): restraint standard deviation

Does not change likelihood function.
Does not change derivative.
Does change 2\textsuperscript{nd} derivative - curvature.

Model should be less prone to fitting into noise

Typical: \( \sigma = 0.01\text{–}0.02 \)
Distance threshold: 4.2\text{Å}
Jelly Body Restraints
ProSMART

Injection of prior knowledge to aid new structure determination

• External Restraints from homologous structures
  ➢ Protein or nucleic acid chains

• Hydrogen bond restraints
  ➢ Protein backbone

• Generic self-restraints
  ➢ Everything – protein, nucleic acids, ligands, waters, metals

• Structure analysis
  ➢ Alignment & comparison - helps analyse differences between models

Independent of global conformation
ProSMART External Restraints

Prior information:

Stabilises structural features

3g4w – 3.7 Å
ProSMART H-Bond Restraints
External Restraint Generation

structure to be refined

known similar structure (prior)

(abstract representation of an atomic model; circles = atoms)
External Restraint Generation

structure to be refined

known similar structure (prior)

(abstract representation of an atomic model; circles = atoms)
External Restraint Generation

structure to be refined

\[ d \sim N(r, \sigma^2) \]

(abstract representation of an atomic model; circles = atoms)
ProSMART Restraints
ProSMART Restraints
ProSMART Restraints

Red: long
Grey: similar
Blue: short
ProSMART Restraints

Red: long
Grey: similar
Blue: short
ProSMART Restraints
ProSMART Restraints in Coot

Red: long
Grey: similar
Blue: short

Full chain refine
Multi-threaded
ProSMART Restraints in Coot

Red: long
Grey: similar
Blue: short

Full chain refine
Multi-threaded
Robust Estimation
Robust Estimation

Cyan: original
Robust Estimation

Cyan: original
Green: homolog
Robust Estimation

Cyan: original
Green: homolog
Yellow: refined

Red: long
Grey: similar
Blue: short
Robust Estimation

Cyan: original
Green: homolog
Yellow: refined
Red: long
Grey: similar
Blue: short
Robust Estimation

Green: original
Brown: homolog
Robust Estimation

Green:  original
Brown:  homolog
Blue:  refined
Motivational Example from MX

Ovotransferrin

Low-resolution refinement:
Weak signal
Noisy data

Unstable refinement

Result:
Poor quality model

1ryx – 3.5Å
Motivational Example from MX

Ovotransferrin

High-resolution homologue

1ryx – 3.5Å

2d3i – 2.15Å
Motivational Example from MX

Ovo transferrin

Models don’t superpose well
Example: Ovotransferrin

1ryx (3.5Å)
Example: Ovotransferrin

Restraints:
Backbone

1ryx (3.5Å) restrained to 2d3i (2.15Å)

Red: long
Grey: similar
Blue: short
Example: Ovotransferrin

Restraints:
Backbone
Side chains

1ryx (3.5Å) restrained to 2d3i (2.15Å)

Red: long
Grey: similar
Blue: short
Example: Ovotransferrin

Restraints:
- Backbone
- Side chains

After re-refinement

1ryx (3.5Å) restrained to 2d3i (2.15Å)

Red: long
Grey: similar
Blue: short
External Restraints

Ovotransferrin

Refinement Cycles

$R / R_{\text{free}}$

$0 / 10 / 20 / 30 / 40$

$0.20 / 0.25 / 0.30 / 0.35 / 0.40$

$36.5 \%$  $R_{\text{free}}$

$23.9 \%$  $R$-factor
External Restraints

Ovotransferrin

Refinement Cycles

\[ \frac{R}{R_{\text{free}}} \]

- **R** / **R**\text{free}
- **Ovotransferrin**
- **R**free
- **36.5 %**
- **32.4 %**
- **27.3 %**
- **23.9 %**
- **Jelly-body restraints**
- **R-factor**
External Restraints

Ovotransferrin

R / R_{free}

Refinement Cycles

R_{free}:
- 36.5%
- 32.4%
- 30.7%
- 27.3%
- 26.3%
- 23.9%

R-factor:
- Jelly-body restraints
- External restraints
External Restraints

Ovotransferrin

Original Structure

$R/R_{\text{free}} : 0.286/0.330$

In Preferred Regions: 372 (54.39%)
In Allowed Regions: 145 (21.20%)
Outliers: 167 (24.42%)
External Restraints

Ovotransferrin

Original Structure

$R/\bar{R}_{\text{free}} : 0.286/0.330$

Re-refined with External Restraints

$R/\bar{R}_{\text{free}} : 0.263/0.307$

In Preferred Regions: 372 (54.39%)
In Allowed Regions: 145 (21.20%)
Outliers: 167 (24.42%)

In Preferred Regions: 630 (92.11%)
In Allowed Regions: 29 (4.24%)
Outliers: 25 (3.65%)
External Restraints

Original Structure
R/R_{free} : 0.286/0.330

↓

External restraints
(40 cycles)
R/R_{free} : 0.263/0.307
External Restraints

Original Structure
R/R_{free} : 0.286/0.330

↓
External restraints
(40 cycles)
R/R_{free} : 0.263/0.307
External Restraints

Original Structure
R/R_{free} : 0.286/0.330

↓

External restraints
(40 cycles)
R/R_{free} : 0.263/0.307

↓

Build TYR92
Modify LYS209

↓

Jelly body
(40 cycles)
R/R_{free} : 0.252/0.307
External Restraints

When refining MX models at low-resolution, check:

- Refinement statistics - *Not always conclusive*
- Geometry - *Not always conclusive*
- Electron density - *Not always reliable*

Quality/suitability of prior information is important – consider manual re-refinement

- PDB_REDO is useful

External restraints from homologous structures can be useful

Can have confidence to use them in cryo-EM refinement
What if there are no high-resolution homologues?

We still need to stabilise refinement…

• Jelly-body restraints

• Generic external restraints:
  - ProSMART - protein (secondary structure h-bonds)
  - LibG - DNA/RNA (base-pair, base-stacking)
LibG Base-Pair Restraints

Wobble

Reverse wobble
LibG Base Stacking Restraints

Parallel plane restraints

Shift to origin – minimise – transform back to molecule
LibG Nucleic Acid Restraints

Visualise & edit in Coot

Red: long
Grey: similar
Blue: short
LibG Nucleic Acid Restraints
LibG Nucleic Acid Restraints
Ligand Refinement

Geometric restraints for protein / nucleic acids are pre-tabulated

Ligands are more complicated

Need a source of prior information

• Common/known structures are dealt with automatically
  ➢ CCP4/REFMAC Monomer Library has pre-computed descriptions

• New ligands require description (CIF file)
  ➢ AceDRG
Motivation:

High-resolution prior information is required
• MX and cryo-EM data do not inform about bonds, atom assignment
• Must be complemented by prior knowledge:
  o Atomic composition, connectivity
  o Hydrogen bonding capability, protonation of functional groups
    Used to decide the nature of non-bonding interactions
  o Torsion angles, for fine-tuning conformations
  o Chiralities for decisions about stereochemistry

We want to improve the quality of prior information used in refinement
• Restraints for ligands in macromolecular models
• Restraints for macromolecules
How Does ACEDRG Work?

**New atom types**

Full 2\textsuperscript{nd} order neighbour-based atom description

In some cases, 3\textsuperscript{rd} order information is encapsulated also…

H1B: \( \text{H}(\text{CHHO}) \)

C9: \( \text{C}[5,5,6](\text{C}[5,5]\text{CHH})(\text{C}[5,6]\text{CHH})(\text{C}[5,6]\text{CHO})(\text{H}) \)

- Hierarchical organisation of bonded atom-pairs (and angles, etc.)
- Restraint values taken from inspecting small molecule databases
- These are tabulated, distributed as part of CCP4 for quick lookup
What does AceDRG Do?

Functionalities:

(1) Restraint Dictionary Generator
   • Uses restraint tables to generate restraints for given molecule
   • Inputs – mmCIF, SMILES, MDL/SDF, SYBIL/MOL2
   • Output – CIF - bond lengths, angles, torsions, planes, chiralities

(2) Conformer Generator
   • Generates coordinates from graph-based molecule description
   • Generates one of the lower-energy conformations
   • Refines conformation using Refmac
   • Output – PDB

(3) Link Creation
   • Creates link between two components
   • Creates modifications to components where necessary
   • May be from the CCP4 monomer library, or custom CIFs
   • Separate operational mode – requires separate execution
REFMAC Anisotropic Map Sharpening

Idea – remove an overall B value

Original Map

Sharpened map from REFMAC

**Green:** original structure

**Blue:** homologous structure

2r6c (4.0Å) – helix unmodelled

2r6a (2.9Å)
Map Blurring

Idea - apply an overall B value

Blue - electron density (1σ)
Orange - blurred map (B=200)
Map Sharpening/Blurring

Default map

5a1a (2.2Å)
Map Sharpening/Blurring

Blur 20 Å²

5a1a (2.2Å)
Map Sharpening/Blurring

Blur 40 Å²

5a1a (2.2Å)
Map Sharpening/Blurring

Blur 60 Å²

5a1a (2.2Å)
Map Sharpening/Blurring

Blur 80 Å²

5a1a (2.2Å)
Map Sharpening/Blurring

Blur 100 Å² 5a1a (2.2Å)
Map Sharpening/Blurring

Blur 80 Å²

5a1a (2.2Å)
Map Sharpening/Blurring

Blur 60 Å²

5a1a (2.2 Å)
Map Sharpening/Blurring

Blur 40 Å²

5a1a (2.2Å)
Map Sharpening/Blurring

Blur 20 Å²

5a1a (2.2Å)
Map Sharpening/Blurring

Default map

5a1a (2.2Å)
Map Sharpening/Blurring

Sharpen 20 Å²

5a1a (2.2Å)
Map Sharpening/Blurring

Sharpen 40 Å²

5a1a (2.2Å)
Map Sharpening/Blurring

Sharpen 60 Å²

5a1a (2.2Å)
Map Sharpening/Blurring

Sharpen 80 Å²

5a1a (2.2Å)
Map Sharpening/Blurring

Sharpen 100 Å²

5a1a (2.2Å)
Map Sharpening/Blurring
Map Sharpening/Blurring

<table>
<thead>
<tr>
<th>Job title</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input map</strong></td>
<td><strong>Select</strong></td>
</tr>
<tr>
<td><strong>Input PDB</strong></td>
<td><strong>Select</strong></td>
</tr>
<tr>
<td><strong>Resolution</strong></td>
<td>None</td>
</tr>
<tr>
<td><strong>Sharp</strong></td>
<td>50, 100, 150, 200,</td>
</tr>
<tr>
<td><strong>Blur</strong></td>
<td>50, 100, 150, 200,</td>
</tr>
</tbody>
</table>
Map Sharpening/Blurring
Map Sharpening/Blurring

Blur 40 Å²

5a1a (2.2Å)
Blurring / Sharpening

Default map

5a1a (2.2Å)
Blurring / Sharpening

Default map

5a1a (2.2Å)
Blurring / Sharpening

Blur 20 Å²

5a1a (2.2Å)
Blurring / Sharpening

Blur 40 Å²

5a1a (2.2Å)
Blurring / Sharpening

Blur 60 Å²

5a1a (2.2Å)
Blurring / Sharpening

Blur 80 Å²

5a1a (2.2Å)
Blurring / Sharpening

Blur 100 Å²

5a1a (2.2Å)
Blurring / Sharpening

Blur 0-100 Å²  5a1a (2.2Å)
Blurring / Sharpening

Blurring/Sharpening is useful for visual interpretation

- In MX, map blurring/sharpening does not affect refinement

- In cryo-EM, map blurring/sharpening does affect refinement
Blurring / Sharpening

```
job title: None
Multi PDBs/Maps: False
Find in map: False
Input PDB: Select None
Input map: Select None
Resolution: None

weight Auto

Map sharpen: None
```

True
Blurring / Sharpening

Before refinement

After 10 cycles
Blurring / Sharpening

![Graph showing the relationship between FSC and Blurring B-factor. The graph indicates that after 10 cycles, the FSC increases, while before refinement, the FSC remains relatively constant.](image-url)
Auto Box Size – “local” refinement

Helical filament 5jzc (4.2 Å)

Box size is a choice, unlike in MX
Auto Box Size – “local” refinement
Auto Box Size – “local” refinement

Helical filament 5jzc (4.2 Å)
Divide & Conquer Pipeline

Rotavirus 4v7q (3.8 Å)

Oleg Kovalevskiy
Divide & Conquer Pipeline

Rotavirus 4v7q (3.8 Å)
Divide & Conquer Pipeline

Rotavirus 4v7q (3.8 Å)
Divide & Conquer Pipeline

Rotavirus 4v7q (3.8 Å)
Divide & Conquer Pipeline

Rotavirus 4v7q (3.8 Å)
Divide & Conquer Pipeline

Parallel refinement

Rotavirus 4v7q (3.8 Å)
Divide & Conquer Pipeline

Parallel refinement

Rotavirus 4v7q (3.8 Å)
Divide & Conquer Pipeline

Parallel refinement

Available in CCP-EM soon...

Rotavirus 4v7q (3.8 Å)
Summary

• **Restrained refinement can be used for cryo-EM**
  - Need lots of “extra” restraints to regularise refinement
  - Jelly-body restraints are almost always needed

• **External restraints to homologous structures can be useful**
  - Used by REFMAC5 for refinement
  - Used by Coot for refinement; visualised in Coot
  - If homologs are not available, use:
    - Generic h-bond restraints for protein
    - Generic base-pair/stacking restraints for DNA/RNA

• **Other things to think about:**
  - Multiple levels of blurring/sharpening helps, but care needed
  - Box size should be selected appropriately
  - Divide & conquer pipeline for large models… available soon
  - Features in COOT – jiggle fit, model morphing, etc.
  - Validation!
Tools for macromolecular model building and refinement into electron cryo-microscopy reconstructions

The recent rapid development of single-particle electron cryo-microscopy (cryo-EM) now allows structures to be solved by this method at resolutions down to 3 Å. Here, a number of tools to facilitate the interpretation of EM reconstructions with stereochemically reasonable all-atom models are described. The BALBES database has been repurposed as a tool for identifying protein folds from density maps. Modifications to Coot, including new Jiggle Fit and morphing tools and improved handling of nucleic acids, enhance its functionality for interpreting EM maps. REFMAC has been modified for optimal fitting of atomic models into EM maps. As external structural information can enhance the reliability of the derived atomic models, stabilize refinement and reduce overfitting, ProsSMART has been extended to generate interatomic distance restraints from nucleic acid reference structures, and a new tool, LJBG, has been developed to generate nucleic acid base-pair and parallel-plane restraints. Furthermore, restraint generation has been integrated with visualization and editing in Coot, and these restraints have been applied to both real-space refinement in Coot and reciprocal-space refinement in REFMAC.

Current approaches for the fitting and refinement of atomic models into cryo-EM maps using CCP-EM

Recent advances in instrumentation and software have resulted in cryo-EM rapidly becoming the method of choice for structural biologists, especially for those studying the three-dimensional structures of very large macromolecular complexes. In this contribution, the tools available for macromolecular structure refinement into cryo-EM reconstructions that are available via CCP-EM are reviewed, specifically focusing on REFMAC and related tools. Whilst originally designed with a view to refinement against X-ray diffraction data, some of these tools have been able to be repurposed for cryo-EM owing to the same principles being applicable to refinement against cryo-EM maps. Since both techniques are used to elucidate macromolecular structures, tools encapsulating prior knowledge about macromolecules can easily be transferred. However, there are some significant qualitative differences that must be acknowledged and accounted for; relevant differences between these techniques are highlighted. The importance of selecting appropriate levels of map blurring/sharpening is emphasized, which may be facilitated by considering the behaviour of the average map amplitude at different resolutions, as well as the utility of simultaneously viewing multiple blurred/sharpened maps. Features that are important for the purposes of computational efficiency are discussed, notably the Divide and Conquer pipeline for the parallel refinement of large macromolecular complexes. Techniques that have recently been developed or improved in Coot to facilitate and expedite the building, fitting and refinement of atomic models into cryo-EM maps are summarized. Finally, a tool for symmetry identification from a given map or coordinate set, ProSHADE, which can identify the point group of a map and thus may be used during deposition as well as during molecular visualization, is introduced.
Acknowledgements

MRC-LMB Computational Structural Biology Group

Left to right:
Rob Nicholls
Fei Long
Oleg Kovalevskiy
Garib Murshudov
Michal Tykac
Rangana Warshamanage
James Parkhurst
Paul Emsley

CCP-EM:

Tools for cryo-EM model fitting & refinement:

Refinement with REFMAC, ProSMART, LibG, LORESTR:

Contact: nicholls@mrc-lmb.cam.ac.uk

Tom Burnley
Colin Palmer
Agnel Joseph
Martyn Winn

Alan Brown
Jude Short
Ana Casañal

Robbie Joosten
Roberto Steiner
Marcus Fischer
Stuart McNicholas

Jake Grimmett
Toby Darling
CCP4 core team
MRC-LMB & CCP4 colleagues