Low Resolution Refinement

ECM32 Vienna
August 2019

Paul Adams
Lawrence Berkeley Laboratory and Department of Bioengineering UC Berkeley
Macromolecular Crystallography

PDBID: 2gkg
Resolution: 1.00Å

PDB ID: 3k7a
Resolution: 3.80Å

• Many challenges, but low resolution data is increasingly an issue:
• How to interpret “featureless” maps (pattern matching, chemical constraints)
• How to optimize models with sparse data (prior information)
The Challenge of Too Few Data

- With only low resolution data we typically have too many parameters to optimize
  - Atomic coordinates, displacement parameters
- Underdetermined optimization problems lead to overfitting of the data
- To help address overfitting we can:
  - Add prior information to reduce the number of effective parameters
  - Remove parameters
- Current refinement methods do not define a reasonable chemical result in the absence of data
Improving the Observation to Parameter Ratio

- To make refinement practical the observation to parameter ratio is increased using restraints and constraints:
  - **Restraint**
    - Model property \(\sim\) ideal value
    - Adds prior observed information (reduces the number of parameters refined)
    - Inclusion of chemical information in the objective function
  - **Constraint**
    - Model property = ideal value
    - Removes one or more parameters from the model
Methods in Phenix for Improving Models

• Using prior structural knowledge as additional restraints:
  • Secondary structure
  • Protein mainchain conformations (Ramachandran)
  • Related high resolution structures as restraints
  • Multiple copies of the same molecule as restraints (*c.f. local NCS restraints in SHELX*)

• Automated correction of models during refinement using prior knowledge of stereochemistry:
  • Fixing of rotamers
  • Flipping of side chains
Reference model restraints
(Jeff Headd)
IGTX and IOHV

**IGTX**: 3.0 Å  
**IOHV**: 2.3 Å

4-aminobutyrate-aminotransferase
IGTX and IOHV

4-aminobutyrate-aminotransferase

IOHV: 2.3 Å
IGTX: 3.0 Å
Reference Model Restraints

Combines two concepts:

• Pre-correct rotamer outliers
  • Set rotamer outliers in the model to match the torsion angles of the reference model if the reference model has an acceptable rotamer at that position and there is no significant clash or density mismatch

• Generate reference torsion restraints
  • Restrain each torsion angle in the working model to the corresponding torsion angle in the reference model
    • Chains are aligned using SSM alignment to allow for sequence differences
  • Restraints take the form of a modified harmonic ‘top-out’ potential that allows for structural differences

Reference model restraints

\[ E_{total} = \sum_{i=1}^{n} E_i \]

Simple harmonic potential: \[ E_i = \omega \Delta_i^2 \]

‘Top-out’ potential: \[ E_i = \tau (1.0 - e^{-\frac{\Delta_i^2}{l^2}}) \]

\[ \tau = \omega l^2 \]
\[ \omega = \frac{1}{\sigma^2} \]

where \( \sigma \) is the ESD, \( \Delta \) is the difference between the model dihedral and reference dihedral, and \( l \) is a ‘limit’ parameter that limits how far the model dihedral may vary from the reference dihedral before being shut off.

developed by Ralf Grosse-Kunstleve
default: limit = 15.0°
The ‘limit’ parameter

default: limit = 15.0°

< limit, restrain all dihedrals to reference

>> limit, no restraint
Why torsion angles?
IGTX/1OHV reference example

**outlier correction**

5 macrocycles of *phenix.refine* w/ reference restraints

- Leu A 34
- Glu A 41

- **tp rotamer**
- **outlier**

IGTX (3.0Å)  
1OHV (2.3Å)  
IGTX w/ 1OHV reference

**restrained refinement**

R\textsubscript{free}: 0.2379 → 0.2186

ΔR: 0.833 → 0.60

MolProbity: 64\textsuperscript{th} → 96\textsuperscript{th}
Practical Example

Cyclic GMP-dependent protein kinases (PKG’s)

cAMP bound: 2.49Å

cGMP bound: 3.20Å

APO form: 2.69Å

JJ Kim et al. (2011) Crystal structures of PKG Iβ (92-227) with cGMP and cAMP reveal the molecular details of cyclic nucleotide binding. PLoS ONE.
Cyclic GMP-dependent protein kinase

cAMP bound: 2.49Å

cGMP bound: 3.20Å

APO form: 2.69Å

<table>
<thead>
<tr>
<th>Validation Criteria</th>
<th>cAMP bound</th>
<th>cGMP bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>All-Atom Contacts</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clashscore, all atoms:</td>
<td>16.53</td>
<td>56.57</td>
</tr>
<tr>
<td>Clashscore percentile</td>
<td>81st</td>
<td>15th</td>
</tr>
<tr>
<td></td>
<td>2.38%</td>
<td>18.58%</td>
</tr>
<tr>
<td></td>
<td>5.95%</td>
<td>2.02%</td>
</tr>
<tr>
<td></td>
<td>23</td>
<td>85.48%</td>
</tr>
<tr>
<td></td>
<td>0.2102</td>
<td>0.40%</td>
</tr>
<tr>
<td></td>
<td>0.2582</td>
<td>96.00%</td>
</tr>
<tr>
<td>Protein Geometry</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poor rotamers:</td>
<td>2.61%</td>
<td>24.56%</td>
</tr>
<tr>
<td>Rama outliers:</td>
<td>0.00%</td>
<td>87th</td>
</tr>
<tr>
<td>Rama favored:</td>
<td>98.80%</td>
<td>4.00%</td>
</tr>
<tr>
<td>Cβ dev. &gt; 0.25Å:</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>MolProbity score:</td>
<td>2.04</td>
<td>12th</td>
</tr>
<tr>
<td>MP score percentile</td>
<td>95th</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.11%</td>
</tr>
<tr>
<td></td>
<td>3.29</td>
<td>0.60%</td>
</tr>
<tr>
<td>Res w/ bad bonds:</td>
<td>0.00%</td>
<td>3.66%</td>
</tr>
<tr>
<td>Res w/ bad angles:</td>
<td>0.00%</td>
<td>89.02%</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>96.61%</td>
</tr>
<tr>
<td></td>
<td>2.43</td>
<td>0.60%</td>
</tr>
<tr>
<td></td>
<td>0.2264</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>0.2205</td>
<td>0.40%</td>
</tr>
<tr>
<td></td>
<td>0.2612</td>
<td>0.20%</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R-work</td>
<td>0.1960</td>
<td>12th</td>
</tr>
<tr>
<td>R-free</td>
<td>0.2264</td>
<td>89th</td>
</tr>
<tr>
<td></td>
<td>0.79%</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>0.98%</td>
<td>0.20%</td>
</tr>
<tr>
<td></td>
<td>0.2205</td>
<td>0.2166</td>
</tr>
<tr>
<td></td>
<td>0.2612</td>
<td>0.2525</td>
</tr>
</tbody>
</table>
Sources of Prior Information

Images from PumMa website (http://www.pumma.nl)

Covalent geometry

Related structures

Internal symmetry

Secondary structure

Mainchain distributions

Sidechain distributions
Torsion space NCS restraints
(Jeff Headd)
Identify rotamer outlier

Correct to corresponding rotamer in NCS-related chain by matching χ angles

‘backrub’ search, then limited χ angle torsion search

Verify rotamer is still correct match

1b04: 2.8 Å DNA ligase
molecular replacement → refinement

3hd0: 2.70 Å
endonuclease

MR w/ Phaser

\[ R_{\text{work}} = 0.3844 \]

AutoBuild

- Rebuild in place
- NCS on for rebuilding
- NCS off for refinement
- No water picking

\[ R_{\text{work}} = 0.1895 \]
\[ R_{\text{free}} = 0.2745 \]
\[ R_{\text{gap}} = 0.085 \]

phenix.refine

- 10 macrocycles
- Optimize weights
- No NCS, Cartesian NCS, torsion NCS w/ and w/o rotamer correction
3hd0 refinement

- no NCS
- Cartesian NCS
- torsion NCS
- torsion NCS w/ rotamer correction

$R_{\text{free}} = 0.2606$

$R_{\text{work}} = 0.2040$

$R_{\text{gap}} = 0.056$
Sources of Prior Information

Covalent geometry

Images from PumMa web site (http://www.pumma.nl)

Related structures

Secondary structure

Internal symmetry

Sidechain distributions

Mainchain distributions
More Prior Information

• As the number of observations decreases we need to increase the amount of prior information we include (or the number of constraints we apply)
• At the extreme - what if we had no data?
• Other fields have been trying to address this problem:
  • Structure prediction
  • Homology modelling
  • Protein folding

From: Kryshtafovych & Fidelis, Drug Discovery Today, 2009, 14:386–393

http://www.predictioncenter.org
Physically Realistic Potentials (Rosetta)
(Nat Echols & Frank DiMaio)
Rosetta

- *ab initio* model generation and model optimization
- Requires extensive computational sampling

Black - Rosetta *ab initio* models, Red - Crystal structure after Relax protocol

Why Rosetta

• Designed to recognize near-native structures among many possible models; combines empirical and physical potentials
  • All-atom force field, incorporates solvation effects, attractive forces, hydrogen bonds, knowledge-based dihedral restraints
• Can yield chemically realistic *ab initio* models without experimental data to guide assembly
• Occasionally good enough for molecular replacement
• Shown to be useful for NMR structure determination with sparse data (CS-Rosetta), MR solution improvement (MR-Rosetta), RNA structure refinement (ERRASER)

Keedy et al. (2009) *Proteins* **77**:29-49

https://www.rosettacommons.org
Complementary Algorithms

**Phenix**
- Reciprocal space X-ray target functions (ML, MLHL, LS-twin)
- Bulk solvent correction
- B-factor refinement (including TLS)
- Map calculation
- Density modification (using RESOLVE)

**Rosetta**
- Physically realistic potentials
- Repacking to remove steric clashes and building rotameric sidechains
- Torsion-angle minimization
- Real-space target (refinement against electron density)
- Fragment-based rebuilding (optional, not currently used)

Python/C++ architecture facilitates combination
Low Resolution Protocol

Protocol run 5 times in parallel and the best model selected based on R-free
Test Cases

Solved by molecular replacement with same protein from another deposition at higher resolution

Membrane Proteins

Solved using homologous proteins

3fps (3.2Å) 3k07 (3.52Å) 2x79 (3.8Å) 1isr (4.0Å)

3pwy (3.5Å) 3idq (3.7Å) 3a8n (4.5Å)
Calcium ATPase - phenix.refine

<table>
<thead>
<tr>
<th></th>
<th>R</th>
<th>R-free</th>
<th>mp score</th>
<th>RMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>start</td>
<td>0.47</td>
<td>0.51</td>
<td></td>
<td>6.1</td>
</tr>
<tr>
<td>phenix</td>
<td>0.43</td>
<td>0.48</td>
<td>2.66</td>
<td>6.2</td>
</tr>
</tbody>
</table>
Calcium ATPase - DEN

<table>
<thead>
<tr>
<th></th>
<th>R</th>
<th>R-free</th>
<th>mp score</th>
<th>RMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>start</td>
<td>0.47</td>
<td>0.51</td>
<td>3.21</td>
<td>6.1</td>
</tr>
<tr>
<td>DEN</td>
<td>0.38</td>
<td>0.44</td>
<td>3.79</td>
<td>6.1</td>
</tr>
</tbody>
</table>
Calcium ATPase - Phenix-Rosetta

<table>
<thead>
<tr>
<th></th>
<th>R</th>
<th>R-free</th>
<th>mp score</th>
<th>RMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>start</td>
<td>0.47</td>
<td>0.51</td>
<td>3.21</td>
<td>6.1</td>
</tr>
<tr>
<td>Rosetta</td>
<td>0.24</td>
<td>0.28</td>
<td>1.55</td>
<td>1.7</td>
</tr>
</tbody>
</table>
• Phenix-Rosetta model is very close to the deposited structure (even at the level of side chains) with better fit to density.
Improved Models

- Phenix-Rosetta typically has improved fit to the crystallographic data and models are closer to the known structure
- Phenix-Rosetta always has improved model quality, as judged by Molprobity
- Generally similar to DEN results but with much improved geometry, and generally faster

DiMiao et al., 2013, Nature Methods 10:1102-1104
Cryo-EM Atomic Model Optimization

Pavel Afonine, Oleg Sobolev, Nat Echols, Jeff Headd, Nigel Moriarty
Lawrence Berkeley National Laboratory

Tom Terwilliger
Los Alamos National Laboratory
Challenges

Resolution 4.5 Å
840 chains, 187,320 residues 1,443,960 atoms

Wide Resolution Range

Resolution: 11.6 Å

User data, resolution: 3.8 Å

Phenix
Poor Initial Fit
Direct Refinement Against the Map

**Table of Contents**
- PHENIX News
- Crystallographic meetings
- Expert Advice
- Fitting tips #6 - Potential misfitting by a switch of sidechain vs mainchain
- FAQ
- Short Communications
- CaBLAM: Identification and scoring of disguised secondary structure at low resolution
- Structural Classification of Allergen IgE Epitopes by Hierarchical Clustering
- New Tool: phenix.real_space_refine
- Assessment
  - cciba tools for transparent parallel job execution in Python. III. Remote access
  - phenix.ensemble_refinement: a test study of apo and holo BAC1

**Editor**
Nigel W. Moriarty, NW.Moriarty@lbl.gov

**PHENIX News**

**New programs**

FEM: Feature Enhanced Maps (Pavel V. Afonine)

Interpretation of a crystallographic map is a means of obtaining an atomic representation of a crystal structure or the map itself may serve as the crystal model. There are number of factors that affect quality of crystallographic maps that in turn affect difficulty (or even feasibility) of their interpretation and quality of resulting model of crystal structure, and include:
- finite resolution of measured reflections;
- incompleteness of data (missing reflections within the resolution range of the measured data);
- experimental errors in measured reflections;
- errors in atomic model parameters.

These factors a) result in artificial peaks in the map that may be confused with the signal and therefore erroneously interpreted in terms of atomic model, b) introduce noise that may obscure the signal and c) may distort the signal in various ways.

Another fundamentally different contributor to the difficulty of map interpretation is that not all the signal has the same strength. For example, a strong signal arising from a heavy atom derivative may easily obscure a very weak signal (that may be at or below the noise level) arising from a partially occupied very mobile ligand or residue side chain alternative conformation or even hydrogen atoms.
Real Space Refinement

• Has a long history in both X-ray crystallography and cryo-EM
  • Early crystallographic refinement programs (Diamond)
  • Alternative to reciprocal space refinement, then applied to EM maps (TNT, RSRef)
  • Regularly used in model building (O, Coot)
• New structure fitting approaches make use of real space refinement
  • Molecular dynamics flexible fitting (MDFF)
  • Deformable elastic network fitting (DireX)
  • Rosetta model building and model refinement
Refinement

- An optimization algorithm is used to minimize a target function by changing the parameters of the model
- Parameters:
  - coordinates, atomic displacements, occupancies
- Optimization algorithm:
  - minimization, simulated annealing
- Target function (Objective function):
  - Function based on electron density (real-space refinement)
  - Function based on structure factors (reciprocal-space refinement)

\[ E = E_{\text{chem}} + \lambda \sum_{hkl} \frac{1}{\sigma^2} (|F_o| - |F_c|)^2 \]
Goal for Cryo-EM Model Refinement

• Stable refinement against any density map (Cryo-EM or X-ray)

• End result should be an improvement in the model

• Large radius of convergence

• Final models with good fit to density and physically reasonable geometry (Ramachandran distribution, rotamers, packing)

• Fast: no more than one second per residue

\[ E = E_{\text{chemistry}} + \omega \sum (\rho_o - \rho_c)^2 \]
Real Space Refinement Procedure

- phenix.real_space_refine


Systematic Searching of Rotamers

• In a protein structure 99% of the side chains obey known rotameric conformations
• Often errors are fixed manually but can now be fixed automatically following structure validation
• A systematic search through rotamer space is combined with a fit-to-density score

Start

Remove misplaced waters

Finish

Fast: 0.01 – 1 second per residue

Pavel Afonine, Jeff Headd, Nat Echols

Optimization In Real Space

- Refinement against a map using minimization or other optimization method
- Minimization can get caught in local minima
- Simulated annealing is a method used to escape minima
Morphing

- Identify local translation to apply to one C$_\alpha$ atom and nearby atoms
- Smooth the local translations in window of 10 residues
- Apply the smoothed translation to all atoms in the residue

Tom Terwilliger, Los Alamos National Laboratory


Terwilliger et al., Acta Cryst. 2013, D69:2244-2250
Real Space Refinement Improves Fit to Data

- Models are moved to better fit the Cryo-EM map
Typical Results at Higher Resolution

Resolution: 3.36 Å
Residues/atoms: 10,716/82,404
Refinement: 173 min

<table>
<thead>
<tr>
<th>METRIC</th>
<th>Original</th>
<th>Phenix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Map CC</td>
<td>0.645</td>
<td>0.783</td>
</tr>
<tr>
<td>RMSD (bonds/angles)</td>
<td>0.02/2.05</td>
<td>0.01/1.21</td>
</tr>
<tr>
<td>Clashscore</td>
<td>117.1</td>
<td>18.79</td>
</tr>
<tr>
<td>Rama. outl., %</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>Rotamer outl., %</td>
<td>35.51</td>
<td>0</td>
</tr>
<tr>
<td>C-beta deviations</td>
<td>24</td>
<td>0</td>
</tr>
</tbody>
</table>

Resolution: 3.8 Å
Residues/atoms: 2,324/17,424
Refinement: 20 min

<table>
<thead>
<tr>
<th>METRIC</th>
<th>Original</th>
<th>Phenix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Map CC</td>
<td>0.650</td>
<td>0.714</td>
</tr>
<tr>
<td>RMSD (bonds/angles)</td>
<td>0.01/1.34</td>
<td>0.01/1.31</td>
</tr>
<tr>
<td>Clashscore</td>
<td>100.9</td>
<td>32.84</td>
</tr>
<tr>
<td>Rama. outl., %</td>
<td>0.52</td>
<td>0</td>
</tr>
<tr>
<td>Rotamer outl., %</td>
<td>27.99</td>
<td>0</td>
</tr>
<tr>
<td>C-beta deviations</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Lower Resolution Requires Additional Information

High Resolution

Side chains

Secondary Structure

Molecule

Low Resolution
Model Restraints

- Symmetry constraints
- Multiple symmetry groups
- Optimization of NCS operators (w.r.t density)
- Automatic expansion of monomer from MTRX records

Reference model torsion angle restraints

Secondary structure restraints

Base pairing restraints

Parallelity restraints
Improved Models from Real Space Refinement
Difference Maps

• Local scaling of map and model density, real space subtraction

• Reveal features missing from the model

phenix.real_space_diff_map model.pdb map.ccp4 resolution=3.5
Conclusions

• The application of prior or complementary information can improve refinement at low resolution for X-ray and Cryo-EM structures
  • Real space refinement is particularly powerful

• Methods from structure prediction provide additional information to improve models
  • Powerful combination of Rosetta and phenix.refine

• It is now feasible to generate good quality models even with low resolution data
  • Challenges still remain in arriving at initial models in the absence of related structures

• Many challenges remain:
  • Reliably accounting for uncertainty in magnification
  • Local variation in resolution leads to uncertainties in interpretation
  • Efficiently accounting for atomic displacements in models
# Acknowledgements

**Berkeley Laboratory**  
Pavel Afonine, Youval Dar, Nat Echols, Jeff Headd, Richard Gildea, Ralf Grosse-Kunstleve, Dorothee Liebschner, Nigel Moriarty, Nader Morshed, Billy Poon, Ian Rees, Nicholas Sauter, Oleg Sobolev, Peter Zwart

**Los Alamos Laboratory/New Mexico Consortium**  
Tom Terwilliger, Li-Wei Hung

**Baylor College of Medicine**  
Matt Baker

**Cambridge University**  
Randy Read, Airlie McCoy, Gabor Bunckozi, Tristan Croll, Rob Oeffner, Kaushik Hatti, Massimo Sammito, Duncan Stockwell, Laurent Storoni

**Duke University**  
Jane Richardson & David Richardson, Ian Davis, Vincent Chen, Jeff Headd, Chris Williams, Bryan Arendall, Bradley Hintze, Laura Murray

**UC San Francisco**  
Ben Barad, Yifan Cheng, Jaime Fraser

**University of Washington**  
Frank DiMaio, Ray Wang, David Baker

**Oak Ridge National Laboratory**  
Marat Mustyakimov, Paul Langan

**Other Collaborators**  
Corey Hryc, Zhao Wang, Wah Chiu, Pawel Janowski, David Case, Dale Tronrud, Donnie Berholz, Andy Karplus, Alexandre Urzhumtsev & Vladimir Lunin, Garib Murshudov & Alexi Vagin, Paul Emsley, Bernhard Lohkamp, Kevin Cowtan, David Abrahams, PHENIX Testers & Users

**Funding**  
- NIH/NIGMS: P01GM063210, P50GM062412, P01GM064692, R01GM071939  
- PHENIX Industrial Consortium  
- Lawrence Berkeley Laboratory