Phenix Tools for Cryo-EM: Refinement and Validation

Pavel Afonine

Phenix software developer

LBNL, Berkeley, California, USA

July 11, 2019
Cryo-EM Workshop, Stanford
Cryo-EM tools in Phenix

Complete set of tools for cryo-EM structure solution: from initial reconstruction to final validated model.
Structure refinement

Initial (poor) model

Refinement

Improved (refined) model
Refinement tools in Phenix

Crystallography

- Initial model
- Experimental data
- A priori knowledge

- Score
- Modify model parameters
- Improved model

phenix.refine
Available since 2005

Cryo-EM

- Initial model
- Experimental data
- A priori knowledge

- Score
- Modify model parameters
- Improved model

phenix.real_space_refine
Available since 2013
Refinement tools in Phenix

Projects

Show group: All groups

<table>
<thead>
<tr>
<th>ID</th>
<th>Last modified</th>
<th># of jobs</th>
<th>R-free</th>
</tr>
</thead>
<tbody>
<tr>
<td>ringer</td>
<td>Sep 07 2016 05:37</td>
<td>2</td>
<td>--</td>
</tr>
<tr>
<td>tmp2</td>
<td>Sep 07 2016 05:23</td>
<td>1</td>
<td>--</td>
</tr>
<tr>
<td>5gnn</td>
<td>Sep 07 2016 08:42</td>
<td>1</td>
<td>--</td>
</tr>
<tr>
<td>debug1</td>
<td>Sep 05 2016 10:51</td>
<td>2</td>
<td>0.0086</td>
</tr>
<tr>
<td>tmp4</td>
<td>Aug 18 2016 07:23</td>
<td>2</td>
<td>--</td>
</tr>
<tr>
<td>testing</td>
<td>Aug 11 2016 01:54</td>
<td>1</td>
<td>--</td>
</tr>
<tr>
<td>mich</td>
<td>Jul 29 2016 12:47</td>
<td>1</td>
<td>--</td>
</tr>
<tr>
<td>almu</td>
<td>Jul 28 2016 10:58</td>
<td>1</td>
<td>--</td>
</tr>
<tr>
<td>rchen</td>
<td>Jul 22 2016 11:10</td>
<td>1</td>
<td>--</td>
</tr>
<tr>
<td>milya</td>
<td>Jul 15 2016 12:36</td>
<td>2</td>
<td>0.1570</td>
</tr>
</tbody>
</table>

Current directory: /Users/pafonine/Desktop/work/tmp

PHENIX version dev-svn-000 Project: 5gnn

Data analysis

Experimental phasing
Molecular replacement
Model building
Refinement

phenix.refine
Automated X-ray and/or neutron refinement

Real-space refinement
Automated real-space refinement

Neutron refinement [alpha]
Alternate phenix.refine interface customized for neutron data

DEN refinement [alpha]
Deformable elastic network refinement using slow-resolution and molecular replacement steps
Automated model refinement: *phenix.real_space_refine*

- Direct refinement against the map
- No Fourier space involved

Real-space refinement in *PHENIX* for cryo-EM and crystallography

Pavel V. Afonine, Billy K. Poon, Randy J. Read, Oleg V. Sobolev, Thomas C. Terwilliger, Alexandre Urzhumtsev and Paul D. Adams
Automated model refinement: `phenix.real_space_refine`

- Best model-map fit. Any map: X-ray, neutron, EM. Any resolution
- Refined models: no poor validation metrics
- Fast (minutes – a few hours, not days or many hours)
  - Make use of multiple CPUs: as many as available
- Large convergence radius
- Easy to use: map and model in, refined model out
- Accessible: no special hardware requirements
Real-space refinement

• PDB: 5VKU
  3720 chains | 1,872,060 residues | 14,917,620 atoms

• Calculate one set of $F_{\text{calc}}$ – never finished on my laptop (run out of memory)

• Calculate real-space refinement target – several seconds

$$T = -\sum_{\text{atoms}} \rho(x_{\text{atom}}, y_{\text{atom}}, z_{\text{atom}})$$
Automated model refinement: `phenix.real_space_refine`

- Inputs
- Rigid body
- Rotamer fitting
- Model idealization
- Simulated Annealing
- Morphing
- Refine NCS operators
- Weight calculation
- Minimization
- Refinement macro-cycle
- Refined model
- Trajectory
- Log file
Automated model refinement: \textit{phenix.real_space_refine}

- Inputs
- Rigid body
  - Rotamer fitting
  - Simulated Annealing
- Model idealization
  - Refine NCS operators
  - Weight calculation
- Morphing
- Minimization
  - Trajectory
  - Log file
- Refined model
Morphing

Start model before refinement

After `phenix.real_space_refine`
Model regularization
Model regularization

• **Goal**
  - Eliminate all geometry outliers
  - Move atoms as little as possible from start position
    • Idealized model within convergence of refinement

• **Why?**
  - Refinement may not be able to refine a model with lots of bad geometries
  - Low-res data cannot validate geometry outliers
Model regularization

Before and after idealization

RMSD between two models
less than 1.5Å
Model regularization

Before...

...after model idealization
Restraints

- Lower the resolution, less detailed the map
- Need extra information to keep correct geometry during refinement

\[ T = T_{\text{DATA}} + wT_{\text{RERAINTS}} \]

\[ T_{\text{RERAINTS}} = T_{\text{BOND}} + T_{\text{ANGLE}} + T_{\text{DIHEDRAL}} + T_{\text{PLANARITY}} + T_{\text{NONBONDED}} + T_{\text{CHIRALITY}} \]

\[ T_{\text{BOND}} = \sum_{\text{all bonded pairs}} w(d_{\text{ideal}} - d_{\text{model}})^2 \]
Restraints

- Low resolution map is not sufficient to maintain secondary structure.
Restraints

• Example: refinement of a perfect α-helix into low-res map
  
  • **Using standard restraints on covalent geometry is insufficient**
  
  • Model geometry deteriorates as result of refinement
Restraints

Covalent geometry

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

Sidechain distributions

Mainchain distributions

Related structures

Internal symmetry

$T_{\text{RESTRANTS}} = T_{\text{BOND}} + T_{\text{ANGLE}} + \ldots + T_{\text{NCS}} + T_{\text{RAMACHANDRAN}} + T_{\text{REFERENCE}} + \ldots$
Validation

Model

Data

Cryo-EM or Diffraction

Model to data fit
Validation tools: Crystallography vs Cryo-EM

Exactly same

Model

Different

Data

Cryo-EM or Diffraction

Model to data fit

Similar
Validation

- Helps to save time later
- Helps to produce better models
- Helps to set correct expectations
- Minimize fraud or true mistakes

New tools for the analysis and validation of cryo-EM maps and atomic models

Pavel V. Afonine, Bruno P. Klaholz, Nigel W. Moriarty, Billy K. Poon, Oleg V. Sobolev, Thomas C. Terwilliger, Paul D. Adams, and Alexandre Urzhumtsev
Validation

The following experimental techniques were used to determine the structure:

- X-RAY DIFFRACTION

The reported resolution of this entry is 2.40 Å.

Percentile scores (ranging between 0-100) for global validation metrics of the entry are shown in the following graphic. The table shows the number of entries on which the scores are based.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Percentile Ranks</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>R_free</td>
<td></td>
<td>0.394</td>
</tr>
<tr>
<td>Clashscore</td>
<td></td>
<td>60</td>
</tr>
<tr>
<td>Ramachandran outliers</td>
<td></td>
<td>8.7%</td>
</tr>
<tr>
<td>Sidechain outliers</td>
<td></td>
<td>13.6%</td>
</tr>
<tr>
<td>RSRZ outliers</td>
<td></td>
<td>1.1%</td>
</tr>
</tbody>
</table>

The table below summarises the geometric issues observed across the polymeric chains and their fit to the electron density. The red, orange, yellow and green segments on the lower bar indicate the fraction of residues that contain outliers for >3, 2, 1 and 0 types of geometric quality criteria. A grey segment represents the fraction of residues that are not modelled. The numeric value for each fraction is indicated below the corresponding segment, with a dot representing fractions <5%.

The upper red bar (where present) indicates the fraction of residues that have poor fit to the electron density. The numeric value is given above the bar.
Validation

Validation for crystallography (X-ray, neutron) and cryo-EM
Validation

![Validation Graph](image)

- **Box**
  - Lengths (Å): 50.92, 68.34, 83.08
  - Angles (°): 90.00, 90.00, 90.00
  - Supplied Resolution (Å): 3.6
  - Resolution Estimates (Å):
    - dFSC (half maps): 0.143
    - d99 (full/half1/half2): 3.7/3.3/3.5
    - dmodel: 3.7
    - dFSC model (0/0.143/0.5): 3.4/3.5/3.5
    - Max min/max (mean): 0.42/0.81/0.03

- **Model vs. Data**
  - CC (mask): 0.83
  - CC (box): 0.55
  - CC (peaks): 0.34
  - CC (volume): 0.83
  - Mean CC for ligands: 0.86
New tools for the analysis and validation of cryo-EM maps and atomic models

Pavel V. Afonine,\textsuperscript{a,b,*} Bruno P. Klaholz,\textsuperscript{c} Nigel W. Moriarty,\textsuperscript{a} Billy K. Poon,\textsuperscript{a} Oleg V. Sobolev,\textsuperscript{a} Thomas C. Terwilliger,\textsuperscript{d,e} Paul D. Adams\textsuperscript{a,f} and Alexandre Urzhumtsev\textsuperscript{c,g}
Ramachandran plot facts

• A protein structure should conform to prior expectations

• Most (98%+) residues should have a mainchain conformation consistent with the Ramachandran distribution

• A small percentage (0.2%) of residue may show Ramachandran outliers (they are not necessarily errors!)

• Outliers can be seen in strained regions of the structure (e.g. in the active site)

• Any outliers need to be confirmed by detailed analysis
For this, a Conformation-Dependent Library (CDL) has been developed and implemented in Phenix for protein refinement. The CDL relates the expected covariant bond geometry to local backbone Ramachandran conformation. Because the expected bond geometry values in the CDL differ from those in the single-value library (especially for the N-Cα-Cβ angle), MolProbity validation now uses the CDL values for structures refined with the CDL, as detected from the REMARK 3 information of a submitted file. Similarly, for RNA, geometry targets are dependent on ribose pucker.

Cis or twisted non-trans peptides

The peptide bond that joins adjacent amino-acid residues in a protein has partial double-bond character and therefore assumes a trans, or more rarely a cis, configuration. The cis configuration is significantly more common preceding a proline and results in a unique Ramachandran distribution for cis-proline. To maintain this special relationship, we associate peptide bonds with their following residue. About 5% of prolines are cis, while only about 0.03% of all non-proline residues are genuinely cis. Recently, we were alerted to a surprising and improbable increase in the number of cis non-proline peptide bonds being modeled, as shown in the plot (updated) of Figure 9(A). These are due to model-building without consideration of prior probabilities, but also in part due to the lack of validation that flagged cis-nonPro peptides, in MolProbity or other systems. We have therefore implemented a new validation and visual markup for non-trans peptides.

Matching the PDB definition, we define a cis peptide as one with an ψ angle between $\pm 30^\circ$ and $\pm 130^\circ$, and a trans peptide as one with an ψ angle $>115^\circ$ or $<215^\circ$. We add an additional definition of “twisted peptides” for ψ angles that are more than $30^\circ$ from planar. Justifiable twisted peptides are even rarer than non-proline cis, and twisted peptides should virtually always be considered modeling errors.

MolProbity reports on non-trans peptides by providing counts of cis prolines, cis non-prolines, and twisted peptides. Counts of cis non-prolines or twisted peptides that constitute a suspiciously high percentage of the structure are flagged with yellow or red in the summary statistics chart. In the multi-criterion chart that reports on each residue individually, each non-trans residue is marked with its category (cis Pro, cis nonPro, twisted Pro, twisted nonPro) and the measured value of its omega peptide dihedral. In the multi-criterion kinemage, each non-trans peptide is marked with a surface that fills in the trapezoidal shape between the backbone trace.
Ramachandran plot

Comprehensive validation (cryo-EM) (Project: real-space-refine-8iijv)

Outlier list

<table>
<thead>
<tr>
<th>Chain</th>
<th>Residue</th>
<th>Residue type</th>
<th>Score</th>
<th>Phi</th>
<th>Psi</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>GLY</td>
<td>gly</td>
<td>0.01</td>
<td>-59.0</td>
<td>-53.5</td>
</tr>
<tr>
<td>A</td>
<td>VAL</td>
<td>Ala</td>
<td>0.05</td>
<td>117.1</td>
<td>16.9</td>
</tr>
<tr>
<td>A</td>
<td>PRO</td>
<td>trans-Pro</td>
<td>0.09</td>
<td>-137.8</td>
<td>132.5</td>
</tr>
<tr>
<td>A</td>
<td>PRO</td>
<td>trans-Pro</td>
<td>0.09</td>
<td>-138.9</td>
<td>1.5</td>
</tr>
<tr>
<td>A</td>
<td>LYS</td>
<td>General</td>
<td>0.03</td>
<td>-57.5</td>
<td>98.4</td>
</tr>
</tbody>
</table>

Ramachandran graphs:

- Ramachandran plot for all non-Pro/Gly residue
- Ramachandran plot for Glycine
- Ramachandran plot for cis-Proline
- Ramachandran plot for trans-Proline
- Ramachandran plot for pre-Proline residues
- Ramachandran plot for Ile or Val
Ramachandran plot examples

**Good**

**Poor**

**Suspicious**
Ramachandran plot

PDB code 3NOQ, 1 Å

Outliers:
(A, ILE, 152), (B, ILE, 154)

Valid Ramachandran plot outliers: justified by the data (density map)
Example: Ramachandran plot outliers

3zx9
Ramachandran plot for all non-Pro/Gly residues

Clashscore: 245
Rama outliers: 23%
Rotamer outliers: 17%
Year: 2011
Resolution: 17Å

5a9z
Ramachandran plot for all non-Pro/Gly residues

Clashscore: 197
Rama outliers: 25%
Rotamer outliers: 28%
Year: 2015
Resolution: 4.7Å
Ramachandran plot

3JA8

6EYC: re-refined (Tristan Croll)
Ramachandran plot

PDB code: 5a9z

Original

Refined with Ramachandran plot restraints
Ramachandran plot Z-score

Objectively judging the quality of a protein structure from a Ramachandran plot

Rob W.W.Hooft, Chris Sander and Gerrit Vriend

- Ramachandran Z-score is good at identifying odd-looking Ramachandran plots!
- Used in PDBREDO and WhatCheck. Implemented in Phenix (Oleg Sobolev)
- Criteria:
  - $Z < -3$: Rubbish
  - $-3 < Z < -2$: Suspicious
  - $Z > -2$: Good
Ramachandran plot Z-score: examples

6DZV

Z-score = -4.55

1US0 (0.66Å)

Z-score = 0.1

• Z<-3: Poor
• -3 < Z < -2: Suspicious
• Z > -2: Good
Ramachandran plot Z-score: examples

3JA8
Z-score = -3.5

6EYC re-refined by Tristan Croll
Z-score = -2.27

• $Z < -3$: Poor
• $-3 < Z < -2$: Suspicious
• $Z > -2$: Good
Example: side-chain rotamer outliers

4btg

Clashscore: 329
Rama outliers: 9%
Rotamer outliers: 46%
Year: 2013
Resolution: 4.4Å
Validation: model-to-map fit

3j9e (emd_6240) | 3.3Å | CC= 0.85 | Year: 2015
Validation: model-to-map fit

3a5x (emd_1641) | 4.0Å | CC <0
Model-map correlation coefficient (CC)

• Definition

  • With or w/o subtracting mean

    \[
    CC(\rho_1, \rho_2) = \left( \sum_{n} (\rho_1(n))^2 \right)^{-1/2} \left( \sum_{n} (\rho_2(n))^2 \right)^{-1/2} \left( \sum_{n} \rho_1(n) \rho_2(n) \right)
    \]

    \[
    CC(\rho_1, \rho_2) = \left( \sum_{n} (\rho_1(n) - \langle \rho_1 \rangle)^2 \right)^{-1/2} \left( \sum_{n} (\rho_2(n) - \langle \rho_2 \rangle)^2 \right)^{-1/2} \left( \sum_{n} (\rho_1(n) - \langle \rho_1 \rangle)(\rho_2(n) - \langle \rho_2 \rangle) \right)
    \]

• How model map is calculated

  • Approximation (e.g. N-gaussian)

  • Form-factors (electron, X-ray, neutron)

  • Fourier map

    • Box or sphere of Fourier map coefficients

• Region in the map used to calculate CC

  • Whole box

  • Mask around atoms

    • Atom radius
Model map

- Gaussian IAM (Independent Atom Model)
  - Universally used in crystallography (X-ray, Neutron, Electron)

- Isotropic:
  \[ \rho_{\text{atom}}(\mathbf{r}, \mathbf{r}_0, B, q) = q \sum_{k=1}^{5} a_k \left( \frac{4\pi}{b_k + B} \right)^{3/2} \exp \left( -\frac{4\pi^2 |\mathbf{r} - \mathbf{r}_0|^2}{b_k + B} \right) \]

- Anisotropic:
  \[ \rho_{\text{atom}}(\mathbf{r}, \mathbf{U}, q) = q \sum_{j=1}^{5} \frac{q a_j (4\pi)^{3/2}}{8\pi^2 \mathbf{U}_{\text{cart}} + b_j \mathbf{I}} \exp \left( -4\pi^2 (\mathbf{r} - \mathbf{r}_0)^T \mathbf{A}^T \left[ 8\pi^2 \mathbf{U}_{\text{cart}} + b_j \mathbf{I} \right]^{-1} \mathbf{A} (\mathbf{r} - \mathbf{r}_0) \right) \]

- Whole model:
  \[ \rho_{\text{MODEL}}(\mathbf{r}) = \sum_{i=1}^{N_{\text{atoms}}} \rho_{\text{atoms}}(\mathbf{r}) \]

- To account for finite resolution:
  - FT model map
  - Remove terms up to specified resolution
  - FT back to real space to get Fourier image = “Model map”
Examples: 3J5Q, resolution: 3.8 Å

<table>
<thead>
<tr>
<th>METRIC</th>
<th>Original</th>
</tr>
</thead>
<tbody>
<tr>
<td>Map CC</td>
<td>0.650</td>
</tr>
<tr>
<td>RMSD (bonds/angles)</td>
<td>0.01/1.34</td>
</tr>
<tr>
<td>Clashscore</td>
<td>100.9</td>
</tr>
<tr>
<td>Rama. outl., %</td>
<td>0.52</td>
</tr>
<tr>
<td>Rotamer outl., %</td>
<td>27.99</td>
</tr>
<tr>
<td>C-beta deviations</td>
<td>0</td>
</tr>
</tbody>
</table>
Examples: 3J5Q, 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>
**Examples: 3J6P, resolution: 8.2 Å**

<table>
<thead>
<tr>
<th>METRIC</th>
<th>Original</th>
</tr>
</thead>
<tbody>
<tr>
<td>Map CC</td>
<td>0.596</td>
</tr>
<tr>
<td>RMSD (bonds/angles)</td>
<td>0.03/2.34</td>
</tr>
<tr>
<td>Clashscore</td>
<td>92.37</td>
</tr>
<tr>
<td>Rama. outl., %</td>
<td>2.03</td>
</tr>
<tr>
<td>Rotamer outl., %</td>
<td>26.21</td>
</tr>
<tr>
<td>C-beta deviations</td>
<td>2</td>
</tr>
</tbody>
</table>
**Examples:** 3J6P, resolution: 8.2 Å

**Residues/atoms:** 949/7,501

**Refinement:** 15 min

<table>
<thead>
<tr>
<th>METRIC</th>
<th>Original</th>
<th>Phenix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Map CC</td>
<td>0.596</td>
<td>0.743</td>
</tr>
<tr>
<td>RMSD (bonds/angles)</td>
<td>0.03/2.34</td>
<td>0.00/1.11</td>
</tr>
<tr>
<td>Clashscore</td>
<td>92.37</td>
<td>34.73</td>
</tr>
<tr>
<td>Rama. outl., %</td>
<td>2.03</td>
<td>0.54</td>
</tr>
<tr>
<td>Rotamer outl., %</td>
<td>26.21</td>
<td>0</td>
</tr>
<tr>
<td>C-beta deviations</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>
Examples: 3ZEE, resolution: 6.1 Å

<table>
<thead>
<tr>
<th>METRIC</th>
<th>Original</th>
</tr>
</thead>
<tbody>
<tr>
<td>Map CC</td>
<td>0.709</td>
</tr>
<tr>
<td>RMSD (bonds/angles)</td>
<td>0.04/4.05</td>
</tr>
<tr>
<td>Clashscore</td>
<td>18.34</td>
</tr>
<tr>
<td>Rama. outl., %</td>
<td>3.66</td>
</tr>
<tr>
<td>Rotamer outl., %</td>
<td>24.64</td>
</tr>
<tr>
<td>C-beta deviations</td>
<td>637</td>
</tr>
</tbody>
</table>
Examples: 3ZEE, resolution: 6.1 Å

Residues/atoms: 4,116/32,830

Refinement: 45 min

<table>
<thead>
<tr>
<th>METRIC</th>
<th>Original</th>
<th>Phenix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Map CC</td>
<td>0.709</td>
<td>0.647</td>
</tr>
<tr>
<td>RMSD (bonds/angles)</td>
<td>0.04/4.05</td>
<td>0.01/1.23</td>
</tr>
<tr>
<td>Clashscore</td>
<td>18.34</td>
<td>18.59</td>
</tr>
<tr>
<td>Rama. outl., %</td>
<td>3.66</td>
<td>0</td>
</tr>
<tr>
<td>Rotamer outl., %</td>
<td>24.64</td>
<td>0</td>
</tr>
<tr>
<td>C-beta deviations</td>
<td>637</td>
<td>0</td>
</tr>
</tbody>
</table>
Phenix Documentation - version 1.9-1692

Phenix programs and their functions
The Phenix graphical interface
Dictionary of crystallographic and other terms
FAQs: Frequently asked questions
How to install, setup and run Phenix
Complete Phenix reference documentation
Bibliography
Index

Crystallographic Structure Solution with Phenix

Phenix Documentation for X-ray Crystallography
Checking data quality | Experimental phasing | Molecular replacement | Model building | Structure refinement
Structure validation | Ligand fitting | Making geometry restraints | Structure deposition

Phenix Documentation for Neutron Crystallography
Structure refinement | Structure validation | Making geometry restraints | Structure deposition

Phenix Documentation for Electron Microscopy (EM)
Structure refinement | Convert map to structure factors | Extract box with map and model
User support

• Feedback, questions, help

phenixbb@phenix-online.org
bugs@phenix-online.org
help@phenix-online.org

• Reporting a bug or asking for help:

  • **We can’t help you if you don’t help us to understand your problem**

  • **Do:**
    1) Make sure you can reproduce the problem using latest Phenix version
    2) Command and parameters used (series of GUI clicks that lead to problem)
    3) Input and output files
    4) Clearly explain the problem/question

PHENIX mailing list:  www.phenix-online.org