Accurate Representation of CryoEM Structures: A User Point-of-View

Corey Hryc
Previously of BCM
July 12, 2019
S2C2 CryoEM Modeling Workshop
High-Resolution Cryo-EM Maps and Models: A Crystallographer’s Perspective

“An analysis of models fitted to the cryo-EM maps indicated the presence of significant problems in almost all of them, including incorrect geometry, clashes between atoms, and discrepancies between the map density and the fitted models. ... Stricter cryo-EM structure deposition standards and their better enforcement are needed.”

- Wlodawer, 2017 Structure

Can’t stop at the map level! ... We need to work on the model!
Outline of Discussion

• User Point-of-View

• Initial model assessment and preprocess data

• Model refinement
  • My model system (P22 bacteriophage)
  • Example data T20S Proteasome

• Model Assessment / Validation
Most important aspect of model building… resolution and visible features!

(~4.4Å) Chen, 2011. *PNAS*

Non-validated C-alpha trace

(3.3Å) Hryc & Chen, 2017. *PNAS*

Can Now Build *de novo* all-atom model
Initial Model Generation Protocol

- For \textit{de novo}
  - Initial backbone trace (COOT, Pathwalker and Gorgon)
  - Fitting routines to perform initial model fitting or improvement of \textit{de novo} model
    - Chimera, Phenix, COOT, Flex-EM

- Initial model assessment
- Preprocess data
- Model Refinement
- Validation
Pathwalker Process

1. Preprocess density map
2. Seed pseudoatoms
3. Trace path
4. Find and fix SSEs
5. Compute final path and optimize geometry
6. Fix bad geometry and re-seed pseudoatoms in map
7. Full atom atom model with real space refinement

Developed by Matthew L. Baker at BCM
Pathwalker for Protein Fold
Validation

Developed by MR and ML Baker
SSE Prediction for Protein Fold Validation
Useful Pre-processing Tools

- **Pathwalker** for generating initial model
- **Remo / SABBAC** - C-alpha model to all-atom model
- **Coot** for sequence adjustments / mutations
- **Phenix.pdbtools** to operate on atom metadata
- **Chimera** - volume operations and visualization of map/model
- **EMAN2** to alter map files (like filtering)
- Text editor to change / PDB header
Useful Pre-processing Routines

Preprocess your MODEL data correctly:

- COOT for sequence adjustments
- Remove unit cell parameters (to insure map and model agree)
- space group number: 1
- map origin: (0, 0, 0)
- Strip PDB header information (remove SSE bias, let Phenix compute it)
Ready for Cryo-EM Model Optimization!

- Final models with **good fit to density** and physically **reasonable geometry** (Ramachandran distribution, rotamers, clashes)

- End result should be an **improvement of the model**

- **Restraints** should be optimized by the program and not always the user

- Refinement of model against **any density map** (Cryo-EM or X-ray)

- Large radius of **convergence** (can be used with lower resolution data)

- **Fast**: no more than one second per residue
Real Space Refinement Procedure

Inputs

Rigid body

Rotamer fitting

Model idealization

Simulated Annealing

Morphing

Refine NCS operators

Weight calculation

Minimization

Refined model

Trajectory

Log file

Rotamer gallery

Ramachandran plot

Bond distances / Fit to density

Clashes

My Model Optimization Workflow

- **Geometry minimization with NO map** to clean up unusual bond lengths and angles (great for de novo models!) (phenix.geometry_minimization)

- **RSR (Phenix) with strong SSE restraints** for one subunit / small complex

- Coot adjustments & **build up complex**

- **RSR (Phenix) with default parameters**

- Coot adjustments & **build up complex**

- Iterate and Validate
de novo model optimization for P22 phage subunit

- Real-space model optimization for cryo-EM (quick and simple)
- Fast gradient-driven model refinement
- Improve fit to density map while maintaining proper stereochemistry
- Global and local level

* Translation for visualization purposes
P22 density map and models
Resulting optimized model within ASU
Building up the complex
P22 Modeling Process

• Map generation / segmentation
• Model generation
• Individual subunit model optimization
• Expanded model optimization
• Analysis
• Model based map validation
• Model optimization and validation tutorials will be available!
Hands-on Tutorial

• **Goal**: optimize a published crystal structure (PDBid: 1YAR) (Forster, 2005) with the 2.8Å T20S proteasome density map (EMD-6287), (Campbell, 2015).

• No model was released with the EM density map and thus fitting the model to the map could possibly reveal new insights

• Real-space refinement method of the model will hopefully produce a **better** model (both stereochemistry and fit-to-density).

Preprocess -> Refine
Running Phenix.real_space_refine

Default run…

phenix.real_space_refine <input_model.pdb> <input_map.map>
resolution=X

Useful options…

run=
minimization_global*
rigid_body
simulated_annealing
local_grid_search
morphing
adp*

ncs_constraints=True / False

| ramachandran_restraints=True / False
Build up the complex...

- Half the complex was **quickly optimized**
- Stereochemistry and fit-to-density should have improved.
- Manual editing with COOT
- Next, we optimize interfaces.
- Build up the complex in Chimera.
- **Rerun** Phenix.real_space_refine
Following model optimization…
Map / Model Analysis

Honesty is appreciated for publications!
### MolProbity Statistics

<table>
<thead>
<tr>
<th>All-Atom Contacts</th>
<th>Clashscore, All Atoms</th>
<th>10.26</th>
<th>97th percentile * (N=37, 3Å-9999Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Clashescore is the number of steric overlaps (&gt;0.4Å) per 1000 atoms.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poor rotamers</td>
<td>0</td>
<td>0.00%</td>
<td>Goal: &lt;1%</td>
</tr>
<tr>
<td>Ramachandran outliers</td>
<td>46</td>
<td>1.54%</td>
<td>Goal: &lt;0.05%</td>
</tr>
<tr>
<td>Ramachandran favored</td>
<td>2754</td>
<td>91.92%</td>
<td>Goal: &gt;98%</td>
</tr>
<tr>
<td>Protein Geometry</td>
<td>MolProbity score^</td>
<td>2.02</td>
<td>100th percentile* (N=638, 3.40Å ± 0.25Å)</td>
</tr>
<tr>
<td>Cβ deviations &gt;0.25Å</td>
<td>0</td>
<td>0.00%</td>
<td>Goal: 0</td>
</tr>
<tr>
<td>Bad backbone bonds:</td>
<td>0/23401</td>
<td>0.00%</td>
<td>Goal: 0%</td>
</tr>
<tr>
<td>Bad backbone angles:</td>
<td>4/31815</td>
<td>0.01%</td>
<td>Goal: &lt;0.1%</td>
</tr>
</tbody>
</table>


---

### Molprobity Results

Judged against **all** structures (3Å-9999Å)
Map Frame Analysis

Top, Long-helix / Bottom, Three-fold

All  1-6  2-12  7-12  13-18  13-23

* Max Density Value = 1
Map Frame Analysis

* Max Density Value = 1
Gallery of Amino Acids in P22 Map

Negative Charged Side Chains
- ASP356
- GLU72

Positive Charged Side Chains
- ARG406
- HIS370
- LYS32

Polar Uncharged Side Chains
- ASN74
- GLN324

Other
- THR416
- SER167
- PRO418

Hydrophobic Side Chains
- ALA348
- ILE407
- LEU89
- VAL352
- GLY410

Other
- TRP410
- TYR186
- PHE86
- MET79
- CYS405
Average Map Values Per Atom

Aromatic residues for validation

Positive Charged Amino Acids
- Arginine (154)
- Lysine (140)

Negative Charged Amino Acids
- Aspartic Acid (224)
- Glutamic Acid (140)

Polar Uncharged Amino Acids
- Asparagine (175)
- Glutamine (126)

Special Uncharged Amino Acids
- Cysteine (7)
- Tryptophan (42)

Hydrophobic Uncharged Amino Acids
- Leucine (210)
- Isoleucine (175)

Other Polar Uncharged Side Chains

Hydrophobic Side Chains

Negative Charged Side Chains

Positive Charged Side Chains
Average Map Density-Value for Side Chains Using Frames 1-6

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Map Density Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALA</td>
<td>0.0</td>
</tr>
<tr>
<td>ARG</td>
<td>0.2</td>
</tr>
<tr>
<td>ASN</td>
<td>0.4</td>
</tr>
<tr>
<td>ASP</td>
<td>0.6</td>
</tr>
<tr>
<td>CYS</td>
<td>0.8</td>
</tr>
<tr>
<td>GLN</td>
<td>1.0</td>
</tr>
<tr>
<td>GLU</td>
<td>1.2</td>
</tr>
<tr>
<td>HIS</td>
<td>1.4</td>
</tr>
<tr>
<td>ILE</td>
<td>1.6</td>
</tr>
<tr>
<td>LEU</td>
<td>1.8</td>
</tr>
<tr>
<td>LYS</td>
<td>2.0</td>
</tr>
<tr>
<td>MET</td>
<td>2.2</td>
</tr>
<tr>
<td>PHE</td>
<td>2.4</td>
</tr>
<tr>
<td>PRO</td>
<td>2.6</td>
</tr>
<tr>
<td>SER</td>
<td>2.8</td>
</tr>
<tr>
<td>THR</td>
<td>3.0</td>
</tr>
<tr>
<td>TRP</td>
<td>3.2</td>
</tr>
<tr>
<td>TYR</td>
<td>3.4</td>
</tr>
<tr>
<td>VAL</td>
<td>3.6</td>
</tr>
</tbody>
</table>

P<0.01
Computing Map / Model FSC Curves

- Model -> map

- **Preprocess for FSC**
  - Resample the data (same box size and Å/pix) - Done in Chimera
  - Soft mask on raw map - Done in EMAN2

- Use `e2proc3d.py` from EMAN2 for the computing the FSCs
Generating a Map from the Model

Original Data
Refined / Validated
Molecular Model

All atoms are equally weighted

Apply Weights to Model
Use New, Properly Weighted Calculated Map To Compute FSC and CC

Collaboration with Dr. Michael Schmid & Dr. Dong-Hua Chen (Stanford)
Electron Scattering Table

- **Neutral** $e^-$ scattering table is available in real-space refinement and map generation from a model (reciprocal space) at low energies
  - Need charged scattering coefficients for cctbx library
  - Need higher energy scattering table
- **Simplified approach?**
  - Start by generating B-factors (Atomic Displacement Parameters - ADP)
Experimental map vs calculated map

- 3 Parameters to adjust
  - Coordinates of molecular model themselves
  - Resolution of calculated map
  - Atomic displacement parameter (ADP) - uncertainty
Map / Model

FSC Curves

- Use ADP values (B-factors) and allow for weak
- Initially set ADP values to 50
- Compute weighted calculated maps through structure factors and compare with experimental maps
Weighted Calculated Map vs. PDB2MRC

Local Correlation

- Experimental Map and its Derived Model
- Calculated Map from Model with Equal Weighted Atoms
- Calculated Map from Properly Weighted Model

Cross Correlation per Amino Acid

Amine Acid Position
Density Observations

(A) Density Surrounding Positive Amino Acids

(B) Density Surrounding Negative Amino Acids

Supporting Evidence

(Yoneykura, 2015 PNAS) Electron Crystallography
Model / Map validation through two independent maps and models

Half Data - Set 1 ~11,000 particle images

Half Data - Set 2 ~11,000 particle images
Map / Model validation with Even / Odd data sets

Map 1 (Even)

Map 2 (Odd)

Fourier Shell Correlation

Spatial Frequency (1/Å)

Even Map vs. Even Model

Odd Map vs. Even Model

3.4 Å
Assess regions of map and model uncertainty (B-factors)

**Good**
- Converged model
- Strong density

**Poor**
- Model Variation
- Weak density
And Finally... EMRinger

Evaluation of P22

Hands-on Tutorial

- Model optimization and validation tutorials will be available!
Hands-on Tutorial

Part 2 of the tutorial - Model Validation

Check Stereochemistry First!

- Check the stereochemistry improvement with Molprobity
  - `phenix.molprobity` `lyar_fit_real_space_refine.pdb`

- Do manual corrections with COOT
  - `coot` `molprobity_coot.py` `lyar_fit_real_space_refine.pdb`
Hands-on Tutorial
Compare the Map vs the Model

• Start with simple correlation

  • `phenix.map_model_cc` lyar_fit_real_space_refine.pdb
eemd_6287.map resolution=2.8

• Create a map from the model and compare to original data set with FSC with EMAN2

  • `e2pdb2mrc.py` lyar_fit_rsr_complex.pdb
  rsr_2p8A_simulated_map.mrc --apix=0.982 --res=2.8

  • `e2proc3d.py` emd_6287_masked.mrc rsr_simulated_map-vs-
  emd_6287.fsc --calcfsc=rsr_2p8A_simulated_map_rs_masked.mrc
  --apix=0.982 --res=2.8
Hands-on Tutorial

Fourier Shell Correlation

Spatial Frequency (1/Å)

0.35625 (2.81), 0.50038
Hands-on Tutorial

Compare the Half Map vs the Model

• Start with Phenix Real-Space Refinement

  • `phenix.real_space_refine` `lyar_fit_real_space_refine.pdb`  
    EVEN_map.map resolution=4.2  
    run=minimization_global+adp+simulated_annealing

• Create a map from the model (with the methods before) and then compare with the other half data set

  • `e2proc3d.py` ODD_Map.mrc rsr_EVEN_Model-vs-ODD_Map.fsc  
    --calcfsc=rsr_4p2A_EVEN_model_map.mrc  
    --apix=0.982  
    --res=4.2
Hands-on Tutorial

Compare the Map vs the Model

- Run EMRinger to assess the improvement of side-chain fit before and after

```
phenix.emringer lyar_fit_real_space_refine.pdb
emd_6287.map
```

=====Final Statistics for Model/Map Pair=====  =====Final Statistics for Model/Map Pair=====
Optimal Threshold:  0.037  Optimal Threshold:  0.028
Rotamer-Ratio:  0.665  Rotamer-Ratio:  0.799
Max Zscore:  8.295  Max Zscore:  18.907
Model Length:  1715  Model Length:  1715
EMRinger Score:  2.002979  EMRinger Score:  4.565486
Conclusion

- Model optimization greatly improved model quality of near-resolution cryo-EM
- Validation measures provide uncertainty levels and confidence in published data
- Map and model can being to be thought of as one with proper weighting
Acknowledgments

• Wah Chiu (Stanford)
• Michael F. Schmid (Stanford)
• Dong-Hua Chen (Stanford)
• Matthew L. Baker (BCM)
• Paul Adams (LBNL, PHENIX)
• Pavel Afonine (LBNL, PHENIX)
• Tom Terwilliger (Los Alamos, PHENIX)  Formerly of...
Questions / References


Formerly of...