Atomic Models from Cryo-EM Data

SLAC Cryo-EM School
January 2020

Paul Adams
Lawrence Berkeley Laboratory and Department of Bioengineering UC Berkeley
Impressive Cryo-EM Achievements

Namba Lab, Osaka

Phenix
Map Resolution

- Biggest growth is in the 3-4Å range
- Substantial number of maps in 4-5Å range

* Not all maps have an associated model
Low Resolution

PDB ID: 2gkg
Resolution: 1.00Å

PDB ID: 3k7a
Resolution: 3.80Å

- Many challenges:
  - How to interpret “featureless” maps (pattern matching, chemical constraints)
  - How to optimize models with sparse data (prior information)
Crystallographic vs. Cryo-EM Maps

Beta galactosidase at 2.2 Å

Tom Terwilliger, Los Alamos National Lab
Crystallographic vs. Cryo-EM Maps

Beta galactosidase at 2.2 Å

X-ray (PDB 3i3b)  Cryo-EM (PDB 5a1a)

Tom Terwilliger, Los Alamos National Lab
Crystallographic vs. Cryo-EM Maps

- The maps are very similar
More Accurate Low Resolution Information in Cryo-EM Maps

A

B

Original

Tom Terwilliger, Los Alamos National Lab

Phenix
Structural Biology Workflows

**X-ray/neutron crystallography**

- How good are the experimental data?
  - Data quality assessment
  - Experimental phasing
  - Molecular replacement
- Improve maps
  - Density modification
- Get a model which fits the data
  - Automatic model building
- Improve the model and determine its quality
  - Refinement/validation
  - Ligand/custom restraints
- Deposition

**Cryo-EM**

- Data quality assessment
- Map optimization
- Automatic model building
- Fitting
- Refinement/validation

Dorothee Liebschner, Lawrence Berkeley Lab

Phenix
New Tools for Cryo-EM in Phenix

- Symmetry from a map
- Rigid model docking
- Automated model building
- Automated map sharpening
- Density modification
- Real space refinement
- Map segmentation
- Model and map validation
Tutorials

• Model placement and building
  • Symmetry determination
  • Rigid body model fitting
  • Map sharpening & density modification
  • Map segmentation
  • Automated model building
  • [Focused map/model combination]

• Atomic model optimization and validation
  • Structure refinement
  • Validation
Tutorial Format

- Use graphical user interface
Tutorial Format

• Use tutorial datasets distributed with Phenix
• Should run on most laptops (2GB RAM, multiple CPUs better)
Challenges

- Automated model building
  - What is the magnification of the map? (can be 5% uncertainty)
  - What is the optimal sharpening of the map?
  - What is the region containing the molecule?
  - Low and variable resolution across maps

- Structure optimization
  - Variable resolution across maps
  - Large molecules
  - Poor initial models

- Validation
  - How to validate a model against moderate resolution maps
Automated Model Docking

Tom Terwilliger
Los Alamos National Laboratory

Pavel Afonine, Oleg Sobolev
Lawrence Berkeley National Laboratory
Automated Model Docking

- Systematic cross correlation search of rotations and translations
- Performed in reciprocal space using FFT (very fast)
- Rigid body optimization of position
Automated Model Sharpening, Density Modification, Segmentation and Model Building

Tom Terwilliger
Los Alamos National Laboratory

Pavel Afonine, Oleg Sobolev
Lawrence Berkeley National Laboratory
Automated Model Building Procedure

1. Determine optimal sharpening of the map
2. Cut out asymmetric unit of the map
3. Trace chain and build model
4. Idealize secondary structure and refine
5. Assemble and refine (protein/RNA/DNA)
6. Apply molecular symmetry and re-refine

Cryo-EM map from the yeast mitochondrial ribosome (chain I of large subunit, 3.2Å, Amunts et al., 2014)


Autobuilt model (pink)
Deposited model (green)
Automated Map Sharpening

Create series of maps with variable overall B-values

Set contour level enclosing 20% of molecular volume

Calculate surface area of contours

Count number of distinct regions enclosed by contours

Choose map with maximum of adjusted surface area

adjusted area = surface area – weight * number of regions

phenix.auto_sharpen
Automated Map Sharpening

Deposited Map

High-conductance Ca(2+)-activated K(+) channel (emd_8414 and PDB entry 5tji; Hite et al., 2017)

Autosharpened Map

B_{iso} = 260\text{Å}^2

B_{iso} = 20\text{Å}^2
Automated Map Sharpening

\[ B_{\text{iso}} = 290\text{Å}^2 \]

\[ B_{\text{iso}} = -60\text{Å}^2 \]

Deposited Map  Autosharpened Map

Cystic fibrosis transmembrane conductance regulator (emd_8461 and PDB entry 5uar; Zhang and Chen, 2016)

Phenix
Automated Map Sharpening

Terwilliger et al. Automated map sharpening by maximization of detail and connectivity. *Acta Cryst* 2018, **D74**:545-559
Map Improvement - Density Modification
Map Improvement

- Maps contain errors
- The maps can be improved by the application of real space constraints
- The Fourier coefficients are modified to produce a map most consistent with what we know about macromolecular structures:
  - Solvent density distribution (Solvent flattening)
  - Atomicity and positivity (Sayre’s equation)
  - Macromolecular density distributions (histogram matching)
  - Similarity between molecules (symmetry averaging)
Statistical Phase Improvement

- **Principle:** phase probability information from probability of the map and from experiment:
  - \( P(\varphi) = P_{\text{map probability}}(\varphi) P_{\text{experiment}}(\varphi) \)
- Phases that lead to a believable map are more probable than those that do not
- A believable map is a map that has…
  - A relatively flat solvent region
  - Symmetry (if appropriate)
  - A distribution of densities like those of model proteins
- **Method:**
  - calculate how map probability varies with the map \( \rho \)
  - deduce how map probability varies with phase \( \varphi \)
  - change map to maximize probability
  - combine with original map
Map Probability Phasing

A function that is (relatively) flat far from the origin

Function calculated from estimates of all structure factors but one (k)

Test each possible phase of structure factor k. $P(\varphi)$ is high for phase that leads to flat region

- Test all possible phases $\varphi$ for structure factor k (for each phase, calculate new map including k)
- Probability of phase $\varphi$ estimated from agreement of map with expectations
- Phase probability of reflection k from map is independent of starting phase probability because reflection k is omitted from the map
Overview of the Cryo-EM procedure

Original half-maps ➔ Averaged original half-maps

Maximum likelihood modification of Fourier coefficients

Map-phasing half-maps ➔ Averaged map-phasing half-maps

Weighted combination using error model

Final density modified full map
Density modification of 3Å apoferritin map

Improvement of cryo-EM maps by density modification. bioRxiv 845032
Improved Maps

- Apoferritin 3Å map (*Data from Kaiming Zhang, Greg Pintilie, Shanshan Li, Wah Chiu*)

![Original Map](image1.png) ![Density Modified Map](image2.png)
Improved Maps

Impact on 50 EMDataBank Maps

- Resolution (Å) at which map-model FSC = 1/2
- Average map-model FSC

Reported resolution (Å) vs. Resolution (Å) at which map-model FSC = 1/2

Reported resolution (Å) vs. Average map-model FSC
Automated Segmentation

- Use the symmetry of the map
- Identify contiguous regions representing asymmetric unit of the map
- Choose symmetry-copies that make compact molecule

Determine optimal sharpening of the map
Cut out asymmetric unit of the map
Trace chain and build model
Idealize secondary structure and refine
Assemble and refine (protein/RNA/DNA)
Apply molecular symmetry and re-refine


emd_6224 (anthrax toxin protective antigen pore at 2.9 Å; Jiang et al. 2015)
Chain Tracing

- Determine optimal sharpening of the map
- Cut out asymmetric unit of the map
- Trace chain and build model
- Idealize secondary structure and refine
- Assemble and refine (protein/RNA/DNA)
- Apply molecular symmetry and re-refine

- Variable map thresholding
- Trace protein main chain
- Identify direction of main chain by fit to density
Idealization and Refinement

- Refine and rebuild model (simulated annealing, rebuilding and combination of best parts of each model)
- Replace segments with idealized structure
- Identify hydrogen-bonding (β-sheets, α-helices) and use them as restraints in real-space refinement

Chain I, yeast mitochondrial ribosome large subunit, 3.2 Å, 3j6b
Assembly and Polymer Recognition

- Try building protein/RNA/DNA (whatever may be there)
- Choose segment type by map correlation

Determine optimal sharpening of the map
Cut out asymmetric unit of the map
Trace chain and build model
Idealize secondary structure and refine
Assemble and refine (protein/RNA/DNA)
Apply molecular symmetry and re-refine

70S ribosome at 2.9 Å
The Final Model

Determine optimal sharpening of the map
Cut out asymmetric unit of the map
Trace chain and build model
Idealize secondary structure and refine
Assemble and refine (protein/RNA/DNA)
Apply molecular symmetry and re-refine

phenix.map_to_model

30S Ribosome (1j5e, 2.9 Å)
Automated Building - Sharpening

Original

Automatically Sharpened

Phenix
Automated Building - Combining Multiple Models

Three Independently Built Models

Composite Model
Building at Low Resolution

Gamma-secretase at 4.5 Å  
(autobuilt model; emd_2677)

Gamma-secretase structure at 3.4 Å  
(autobuilt model; emd_3061)
Building at Medium/High Resolution

Proteasome at 2.8 Å  
(autobuilt model; emd_6287)

Beta-galactosidase at 2.2 Å  
(autobuilt model; emd_2984)
Autobuilding Performance
Model Building Version 2

Trace chain the way a person does

Find secondary structure

Find clear regions of density

Adjust contour level until a region just connects to another

Iterate to build up a connected chain
Model Building Version 2
Finding $C_\alpha$ and $C_\beta$ positions

Trace chain path through high density

Find $C_\beta$ positions from side-chain density

Choose $C_\alpha$ positions 3.8 Å apart and next to $C_\beta$ positions

Construct all-atom model with Pulchra* and refine

Sequence Assignment

- Determine probability of side chain at each $C_\alpha$
- Align sequence to maximize total probability for the chain
Improved Connectivity

3j9e (EMD 6240)
3.3 Å

Average chain length = 84
Improved Performance
What’s The Molecule?

- Use the highest side chain probabilities to determine a sequence (from the map)
- Search the sequence database to identify the molecule

With Xiaorun Li, Chi-min Ho & Hong Zhou, UCLA
Conclusions

• Cryo-EM maps can be improved by the modification of both phase and amplitude terms

• Automated model building is possible, but can be improved
  • Include information from secondary structure prediction, evolution etc.
  • Combine structure-modeling tools (Rosetta) with Phenix model-building

• Many challenges remain:
  • Reliably accounting for uncertainty in magnification
  • Local variation in resolution leads to uncertainties in interpretation
## 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