Introduction to Three Dimensional Structure Determination of Macromolecules by Cryo-Electron Microscopy

Amit Singer

Princeton University, Department of Mathematics and PACM

July 23, 2014
Single Particle Reconstruction using cryo-EM

Schematic drawing of the imaging process:

The cryo-EM problem:
The Resolution Revolution

Werner Kühlbrandt

Precise knowledge of the structure of macromolecules in the cell is essential for understanding how they function. Structures of large macromolecules can now be obtained at near-atomic resolution by averaging thousands of electron microscope images recorded before radiation damage accumulates. This is what Amunts et al. have done in their research article on page 1485 of this issue (1), reporting the structure of the large subunit of the mitochondrial ribosome at 3.2 Å resolution by electron cryo-microscopy (cryo-EM). Together with other recent high-resolution cryo-EM structures (2–4) (see the figure), this achievement heralds the beginning of a new era in molecular biology, where structures at near-atomic resolution are no longer the prerogative of x-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy.

Ribosomes are ancient, massive protein-RNA complexes that translate the linear genetic code into three-dimensional proteins.

Near-atomic resolution with cryo-EM. (A) The large subunit of the yeast mitochondrial ribosome at 3.2 Å reported by Amunts et al. In the detailed view below, the base pairs of an RNA double helix and a magnesium ion (blue) are clearly resolved. (B) TRPV1 ion channel at 3.4 Å (2), with a detailed view of residues lining the ion pore on the four-fold axis of the tetrameric channel. (C) F₄₂₀-reducing [NiFe] hydrogenase at 3.36 Å (3). The detail shows an α helix in the FrhA subunit with resolved side chains. The maps are not drawn to scale.
X-ray crystallography is one of the greatest innovations of the 20th century

W. Kühlbrandt, Science 343, 1443, March 28 2014:

“Does the resolution revolution in cryo-EM mean that the era of x-ray protein crystallography is coming to an end? Definitely not. For the foreseeable future, small proteins—in cryo-EM, anything below 100 kD counts as small—and resolutions of 2Å or better will remain the domain of x-rays. But for large, fragile, or flexible structures (such as membrane protein complexes) that are difficult to prepare yet hold the key to central biomedical questions, the new technology is a major breakthrough. In the future, it may no longer be necessary to crystallize large, well-defined complexes such as ribosomes. Instead, their structures can be determined elegantly and quickly by cryo-EM. These are exciting times.”
Big “Movie” Data, Publicly Available

http://www.ebi.ac.uk/pdbe/emdb/empiar/

---

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Title</th>
<th>Authors</th>
<th>Related EMDB/PDB entries</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>10002</td>
<td>S. cerevisiae 80S ribosome direct electron detector dataset</td>
<td>Bai XC, Fernandez IS, McMullan G, Scheres SH</td>
<td>EMD-2275</td>
<td>260GB</td>
</tr>
<tr>
<td>10003</td>
<td>Molecular Architecture of the Unliganded Membrane-Bound HIV-1 Envelope Glycoprotein Trimer: 10 focal pair micrographs</td>
<td>Mao Y, Wang L, Sodroski J</td>
<td>EMD-5447</td>
<td>1.3GB</td>
</tr>
<tr>
<td>10004</td>
<td>Pre-fusion structure of trimeric HIV-1 envelope glycoprotein determined by cryo-electron microscopy</td>
<td>Bartesaghi A, Merk A, Borgnia MJ, Milne JLS, Subramaniam S</td>
<td>EMD-2484</td>
<td>119GB</td>
</tr>
<tr>
<td>10005</td>
<td>TRPV1 dataset taken on a K2 direct electron detector</td>
<td>Liao M, Cao E, Julius D, Cheng Y</td>
<td>EMD-5778</td>
<td>280GB</td>
</tr>
<tr>
<td>10006</td>
<td>2D crystal images of the potassium channel MloK1 with and without cAMP ligand (with cAMP: PDB 4chv and EMD-</td>
<td>Kowal J, Chami M, Baumgartner P, Arheit M, Chiu P-L, Rangl M, Scheuring S, Schroeder GF, Nimigean CM,</td>
<td>EMD-2526, 4chv, 4chw,</td>
<td>12GB</td>
</tr>
</tbody>
</table>
Open source toolbox, publicly available:
http://spr.math.princeton.edu/
E. coli 50S ribosomal subunit

27,000 particle images provided by Dr. Fred Sigworth, Yale Medical School

3D reconstruction by S, Lanhui Wang (now data scientist at TripAdvisor), and Jane Zhao (now postdoc at Courant Institute, NYU)
Data-Driven Discovery: The cryo-EM experience

Scientific Questions

Methods

Applied and Pure Mathematics
- Tomography
- Convex optimization
- Random matrix theory
- Representation theory
- Signal Processing
- Numerical linear algebra
- Information Theory

Computer Science
- Theoretical CS
- Unique Games, Max-Cut
- Semidefinite Programming
- Nearest neighbors search
- Randomized Algorithms
- Computer graphics, vision
- Software engineering

Scientific Discovery

Statistics and Machine Learning
- Maximum Likelihood Estimation
- High dimensional statistics
- Manifold learning
- Principal component analysis
- Expectation-Maximization

Data Acquisition, Technological Advances

Dr. Fred Sigworth (Yale)
Dr. Joachim Frank (Columbia)
Image Formation Model and Inverse Problem

Projection images $I_i(x, y) = \int_{-\infty}^{\infty} \phi(xR_i^1 + yR_i^2 + zR_i^3) \, dz + \text{"noise"}$.  

$\phi : \mathbb{R}^3 \mapsto \mathbb{R}$ is the electric potential of the molecule.  

Cryo-EM problem: Find $\phi$ and $R_1, \ldots, R_n$ given $I_1, \ldots, I_n$.  

$$R_i = \begin{bmatrix} - & R_i^1 & - \\ - & R_i^2 & - \\ - & R_i^3 & - \end{bmatrix} \in \text{SO}(3)$$
Toy Example
Typical Reconstruction Procedure: Iterative Refinement

- Alternating minimization or expectation-maximization, starting from an initial guess $\phi_0$ for the 3-D structure

$$I_i = P(R_i \cdot \phi) + \epsilon_i, \quad i = 1, \ldots, n.$$  

- $R_i \cdot \phi(r) = \phi(R_i^{-1}r)$ is the left group action
- $P$ is integration in the $z$-direction.

- Converges to a local optimum, not necessarily the global one.
- Model bias is a well-known pitfall
  “Einstein from noise” (Henderson, PNAS 2013)

- Is “reference free” orientation assignment and reconstruction possible?
Orientation Estimation: Fourier projection-slice theorem

Projection $I_i$  

Projection $I_j$  

$\hat{I}_i$  

$\hat{I}_j$  

3D Fourier space  

$R_i c_{ij} c_{ij} = (x_{ij}, y_{ij}, 0)^T$  

$R_i c_{ij} = R_j c_{ji}$
Angular Reconstitution  (Vainshtein and Goncharov 1986, Van Heel 1987)
Experiments with simulated noisy projections

- Each projection is 129x129 pixels.

$$\text{SNR} = \frac{\text{Var(Signal)}}{\text{Var(Noise)}}$$

- (a) Clean
- (b) SNR=2^0
- (c) SNR=2^{-1}
- (d) SNR=2^{-2}
- (e) SNR=2^{-3}
- (f) SNR=2^{-4}
- (g) SNR=2^{-5}
- (h) SNR=2^{-6}
- (i) SNR=2^{-7}
- (j) SNR=2^{-8}
Define common line as being correctly identified if both radial lines deviate by no more than $10^\circ$ from true directions.

<table>
<thead>
<tr>
<th>$\log_2(\text{SNR})$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.997</td>
</tr>
<tr>
<td>0</td>
<td>0.980</td>
</tr>
<tr>
<td>-1</td>
<td>0.956</td>
</tr>
<tr>
<td>-2</td>
<td>0.890</td>
</tr>
<tr>
<td>-3</td>
<td>0.764</td>
</tr>
<tr>
<td>-4</td>
<td>0.575</td>
</tr>
<tr>
<td>-5</td>
<td>0.345</td>
</tr>
<tr>
<td>-6</td>
<td>0.157</td>
</tr>
<tr>
<td>-7</td>
<td>0.064</td>
</tr>
<tr>
<td>-8</td>
<td>0.028</td>
</tr>
<tr>
<td>-9</td>
<td>0.019</td>
</tr>
</tbody>
</table>
Outline for this week

1. Orientation assignment
2. Heterogeneity
3. Class averaging and symmetry detection
Orientation Estimation

- $n = 3$: Vainshtein and Goncharov 1986, van Heel 1987
- $n > 3$: S, Shkolnisky SIAM Imaging 2011
  - Minimizing a quadratic cost: $\sum_{i,j=1}^{n} \| R_i c_{ij} - R_j c_{ji} \|^2$
  - Quadratic constraints: $R_i^T R_i = I_{3 \times 3}$
  - Semidefinite Programming (SDP) Relaxation, Spectral Relaxation
- $n > 3$: Maximum likelihood using SDP: Bandeira, Charikar, Chen, S (in preparation)
The heterogeneity problem

What if the molecule has more than one possible structure?


- Covariance matrix estimation of the 3-D structures from their 2-D projections
- Katsevich, Katsevich, S (submitted)
Class averaging for image denoising

- Rotation invariant representation (steerable PCA, bispectrum)
- Vector diffusion maps

Generalization of Laplacian Eigenmaps (Belkin, Niyogi 2003) and Diffusion Maps (Coifman, Lafon 2006)

Graph Connection Laplacian

Experimental images (70S) courtesy of Dr. Joachim Frank (Columbia)

Class averages by vector diffusion maps (averaging with 20 nearest neighbors)
Algorithmic Pipeline

- **Particle Picking**: manual, automatic or experimental image segmentation.

- **Class Averaging**: classify images with similar viewing directions, register and average to improve their signal-to-noise ratio (SNR).

- **Orientation Estimation**: common lines

- **Three-dimensional Reconstruction**: a 3D volume is generated by a tomographic inversion algorithm.

- **Iterative Refinement**

**Extensions**

- Non-trivial point-group symmetries
- Heterogeneity (structural variability)

