New Challenges for Processing Heterogeneity

Nikolaus Grigorieff
Heterogeneity and Biology

- Translocation, Brilot et al. 2013
  - Kinesin power stroke, Sindelar & Downing 2010
  - Spliceosome, Wahl et al. 2009
  - Glutamate receptor, Dürr et al. 2014

Spliceosome, Wahl et al. 2009

GroEL/GroES ATP cycle, Clare et al. 2012

Kinesin power stroke, Sindelar & Downing 2010
Types of Heterogeneity

- **Compositional**
- **Conformational**
  - discrete
  - continuous
- **General**
Classification Goal

Group images based on their similarity.
A Hypothetical Experiment

Early experiments in transportation

Larson, The Far Side
Wishful Thinking

HeLa cells  Blender  EM grid  3D structures

What are the challenges?

commons.wikimedia.org  amazon.com  emresolutions.com  Wilhelm et al. 2014
Challenge: Size of Dataset

- Assume 1000 different molecular species with $M_w > 100$ kDa
- Assume linear histogram with maximum concentration difference of 100-fold
- Require minimum of 30,000 particles per species

Required dataset: $1000 \times 100/2 \times 30,000 = 1.5$ billion particles
Challenge: Processing Time

- Assume 1.5 billion particles
- Assume $n \log n$ dependence on particle number (fast sorting), 8h/7h for 2D/3D classification of 130,000 particles
  - 2D classification: 19 years
  - 3D classification: 17 years
Challenge: Small Classes

- Assume that smallest population is 100x smaller than largest population
- Larger classes tend to ‘attract’ particles from smaller classes (Yang et al. 2012, ISAC)
  - Detectability will depend on size & shape of molecule/complex
  - Particles may be discarded in 2D classification that might be assignable in 3D
Challenge: Convergence

- Incomplete separation of classes

70S ribosome + EF-G

Brilot et al. 2013
Challenge: Detection

40S ribosomal subunit bound to CSFV-IRES, DHX29 and eIF3

- Computationally expensive
- Very sensitive to particle misalignments
- Noisy/low resolution

26317 particles (one class out of 630k particles)
40k bootstrap volumes
Challenge: Reproducibility

TRPV1 channel

Dataset: 88915 particles (300 kV, K2)

Relion Refinement & classification
35645 particles (40%)

Frealign Refinement & classification
38326 particles (44%)

Overlap: 23230 particles (~60%)

Liao et al. 2013
Challenge: Interpretation

• Current techniques classify pixels, not features
• Classes may still be mixtures
• States may be missing
• Results are irreproducible

➢ Structural interpretation may be difficult
Challenge: Continuous States

Clathrin cage bound to auxilin and Hsc70

\[ Q = \begin{pmatrix} a & 0 & 0 \\ 0 & a & 0 \\ 0 & 0 & c \end{pmatrix} \]

Fotin et al. 2004, Xing et al. 2010

<table>
<thead>
<tr>
<th>Model</th>
<th>FSC at 22 Å (σ = 0.016)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Delta a = -\Delta c ) const. surface</td>
<td>0.157</td>
</tr>
<tr>
<td>( \Delta a = -\sqrt{\Delta c} ) const. volume</td>
<td>0.145</td>
</tr>
<tr>
<td>No deformation</td>
<td>0.107</td>
</tr>
<tr>
<td>( \Delta a = 5\Delta c )</td>
<td>0.108</td>
</tr>
</tbody>
</table>
Normal Modes

70S ribosome + EF-G

Normal mode corresponding to ratcheting

Reconstruction from bins with *

50S

30S

EF-G region

Jin et al. 2014
Alignment With Masks

80S ribosome + Sec61

60S ribosome + Sec61

Voorhees et al. 2014
Masking And Filtering

$V_0$ motor of a eukaryotic V-ATPase

Mazhab-Jafari et al 2016
Slo2.2, a Na\(^+\)-dependent K\(^+\) channel
Challenge: Junk Classes

- Frealign refinement & classification
- EMAN2 initial map
- K-means classification

356,211 particles
F20, K2

43% 49% 50 Å
32% 25% 80,000 particles
25% 26% 3.8 Å resolution

VSV polymerase
240 kDa

➢ Junk may not affect all classes equally

Liang et al. 2015
Challenge: Preferred Views

a) 0° tilt
b) 10° tilt
c) 20° tilt
d) 30° tilt
e) 40° tilt
f) 50° tilt

Tan et al. 2017
Prokaryotic CIC Cl:\textsuperscript{-} channel

Dutzler et al. 2002/2003
Challenge: Ab-Initio 3D

D2 460 kDa
Start | Cycle 9 | Cycle 17 | Cycle 40 0.7 h

C1 240 kDa
Start | Cycle 9 | Cycle 27 | Cycle 40 4.2 h

O 440 kDa
Start | Cycle 9 | Cycle 25 | Cycle 40 0.3 h
Computational Resources
Computational Imaging System for Transmission Electron Microscopy

Tim Grant

Alexis Rohou
### cisTEM GUI

<table>
<thead>
<tr>
<th>Processing step</th>
<th>Details</th>
<th>Time (hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Movie processing</td>
<td>1539 movies, 38 frames, super-resolution</td>
<td>1.3</td>
</tr>
<tr>
<td>CTF determination</td>
<td>using frame averages</td>
<td>0.01</td>
</tr>
<tr>
<td>Particle picking</td>
<td>181,574 particles</td>
<td>0.1</td>
</tr>
<tr>
<td>2D classification</td>
<td>50 classes, 17 selected with 138,975 particles</td>
<td>0.9</td>
</tr>
<tr>
<td>Ab initio 3D reconstruction</td>
<td>40 iterations</td>
<td>0.7</td>
</tr>
<tr>
<td>Auto refinement</td>
<td>8 iterations, final resolution 2.2 Å</td>
<td>1.1</td>
</tr>
<tr>
<td>Manual refinement</td>
<td>1 iteration, final resolution 2.1 Å</td>
<td>0.3</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>4.4</strong></td>
</tr>
</tbody>
</table>

**44 CPU cores, no GPU**
Flexible Architecture

- Workstation
  - GUI
  - cisTE M
  - Job controller
  - Slave jobs

- Cluster Head
  - cisTE M
  - Job controller
  - Cluster Nodes
  - Slave jobs
Challenge: Processing Time

- Assume 1.5 billion particles
- Assume $n \log n$ dependence on particle number, 0.9h for 2D classification of 180,000 particles on 44 CPU cores

- 2D classification: 5h on 5000 CPU cores
Finding Molecules in a Heterogeneous Mess
3D Template Matching

Templates match visible features

tomogram

Frangakis et al. 2002
Herpes virus entering a synaptosome

Maurer et al. 2008
High Resolution Fingerprints

Close-up high-resolution EM image

NMDA receptor

AMPA receptor
Finding Molecules

Apo ferritin

Cryo-EM image

Close to focus

Correlation map

Rickgauer et al. 2017
Finding Asymmetric Units

60 asymmetric units: 13 VP6 + 2 VP2

Rickgauer et al. 2017

0.3 μm underfocus

Correlation map

720 kDa

+ defocus search

75% of expected positions found
Finding RNA Polymerase

DLP

Icosahedron

Experimental density

15,265 vertices averaged

RNA polymerase (VP1, 115 kDa)

Template

5-fold

Rickgauer et al. 2017
• Current molecular weight limit:
  – ~300 kDa when orientations are not constrained
  – ~100 kDa with constraints (e.g. membrane)
• If images are perfect: limit lowered to 30 kDa.
• Positional accuracy:
  – 1 Å horizontally
  – ~20 Å vertically
Summary and Questions

• How do we detect heterogeneity?
  – Search for weak/blurred density, calculate variance maps.

• How do we make sure it does not lead us to the incorrect result?
  – Carful biochemistry, repeat analysis with different starting conditions, check that the results make structural/biological sense.

• How to distinguish conformational vs. compositional variability?
  – Biochemistry, classification, modeling, possibly 3D MSA of bootstrap volumes.

• What are the prospects for getting to atomic resolution for a small and heterogeneous particle?
  – Guess: 50 kDa particle with 10-20 kDa heterogeneity should be possible.

• Are there some samples that will never be amenable to high resolution reconstruction?
  – Very likely, for example if a particle contains large unstructured domains.

Bottom line
Better biochemistry, bigger datasets, bigger computers, better algorithms
Acknowledgements

Template matching
- Peter Rickgauer
- Winfried Denk

cisTEM
- Tim Grant
- Alexis Rohou

Janelia cryo-EM
- Zhiheng Yu
- Chuan Hong
- Rick Huang