A Template-free Pipeline for Recovering Structures in Cryo-electron tomography

Xiangrui Zeng

Computational Biology Department
School of Computer Science
Carnegie Mellon University

MMBioS Monthly Scientific Meeting
November 1, 2018 2:00-3:00 pm
Tilt series and reconstruction

Steven and Belnap, *Current Protocols in Protein Science*, 2005
Example slices of 3D tomogram


Grant Jensen Lab at Caltech
Subtomogram extraction

A subtomogram containing one macromolecule
Challenge 1: Orientation

Particles can adopt all different orientations in 3D space

Image from Philippe Ringler lab at C-CINA
Challenges 2: Missing wedge effect

Reconstruction distortion due to limited range of tilt angles

Palmer and Iowe, *Ultramicroscopy*, 2014
Challenges 3: Noise

Subtomograms are very noisy

Mammalian ribosome ground truth \((40^3)\)  

Mammalian ribosome experimental data
Challenges 4: Molecular crowding

Standard template-based structure recovery pipeline

**Drawbacks**

1. Rely on prior knowledge, unable to detect novel structures
2. Could be biased
3. Unsystematic for large datasets
A template-free pipeline

Overview of techniques we developed

- Subtomogram segmentation
  - Semantic segmentation
- Coarse structural separation
  - Pose normalization
  - Deep structural feature extraction
  - Convolutional Autoencoder
  - CNN classifier
    - Deeper models
    - Model compression
- Averaging and classification
  - Fast alignment + maximum likelihood averaging
- Others
  - Saliency detection
  - Generative models of pseudo molecular structures
  - Multi-task learning
CNN based subtomogram classification

Input subtomogram

Output classes

Complex 1
Complex 2
Complex n
None

Xu et al. 2017, Che et al. 2018, Guo et al. 2018
Subtomogram segmentation

Neighbor structure
Target macromolecule

True structure
Subtomogram

Input
3D Segmentation Network
output

Segmented region of interest

Liu et al. ICIP 2018
Segmentation network model
Multi-task learning model

Shared Layers

Regression Layers

Segmentation Layers

Classification Layers

Input

Output

True structure

Subtomogram

PDB ID: 1QO1

Liu et al. BMVC 2018
Autoencoder based pattern compression
Autoencoder: large scale subtomogram separation
Pose normalization

Standard PCA

$$W = \sum_i X_i X_i^T$$

PCA on weighted covariance matrix

$$W = \int_x \phi(x)^2 \ xx^T$$
Autoencoder model network

(a)

Input
32-3×3×3-1 Conv
ReLU
2×2×2-1 MaxPool
32-3×3×3-1 Conv
ReLU
2×2×2-1 MaxPool
Flatten
FC-32
ReLU
K-means clustering of encoded image patches

Unsupervised clustering

Decoder3D network

Input
FC-8×8×8×32
ReLU
2×2×2-1 UpSample
32-3×3×3-1 Conv
ReLU
2×2×2-1 UpSample
32-3×3×3-1 Conv
ReLU
1-3×3×3-1 Conv
Linear

Encoder3D network
Pattern detected

Surface patch  Surface patch  Large globule  Small globule

Interaction between cellular components
Pattern embedded

Ribosome like macromolecule
Vesicular membrane
Tomogram boundary
Carbon edge
N subtomograms are transformed (rotated and translated), clustered, and averaged into K different classes.
Existing method 1: Maximum likelihood

- Optimizes the probability of observing the data given a data model
  \[ X_i = R_{\phi_i}A_{\kappa_i} + G_i \quad \forall i = 1, \ldots, N \]

- Uses the EM algorithm
- Exhaustive scanning in 6D parametric space of rigid transformation is needed
  - In principle, computationally infeasible
Existing method 2: Fast alignment

- Optimizes a correlation score between a subtomogram X and an average A:
  \[ c(\phi^ro, \phi^{tr}) = \frac{\sum_j w_j^2 X_j \exp(2\pi i \xi_j^T \phi^{tr}) (R_{\phi^ro} A)_j}{\sqrt{\sum_j w_j^2 [R_{\phi^ro} (A \circ A)]_j}} \]

- Uses the fast rotational matching algorithm
- Searches sub-optimal rigid transformations using local maximum under constraints
- Not sufficiently robust to low SNR and missing wedge effects
FAML: Integration of two methods

Approximate the integral with sub-optimal transformations:

\[
\int f(\phi, X, A) \, d\phi \approx \sum_{\phi \in \Phi} f(\phi, X, A) \, \tilde{v}(\phi, \oplus)
\]

Where \( \tilde{v}(\phi, \oplus) := \frac{|v(\phi, \oplus)|}{\sum_{\phi' \in \Phi} |v(\phi', \oplus)|} \) is the normalized hypervolume of \( \emptyset \).

Voronoi weights are approximated using Monte-Carlo sampling.
FAML: Algorithm

Initialize model parameters \( \Theta = (A, \alpha, \sigma, \xi) \) from the distribution of \( X \)

For M iterations:

Compute a list of sub-optimal rigid transformations using FA
Compute their Voronoi weights
Update \( \alpha \), the proportion of particles belonging to different classes
Update \( \sigma \), the standard deviation of Gaussian noise
Update \( \xi \), the standard deviation of translation parameters (3D Gaussian)
Update \( A \), the underlying true structure (subtomogram average)
Results

Low SNR

Orientation preferred

TMV (Kunz et al 2015)

Ribosome (Guo et al 2018)
Results

We randomly selected 400 GroEL/GroES subtomograms from an original dataset of 780.

FAML:
- GroEL: $r = 0.87$
- GroEL-GroES: $r = 0.78$

FA:
- GroEL: $r = 0.40$
- GroEL-GroES: $r = 0.24$

ML (on original dataset):
- GroEL: $r = 0.88$
- GroEL-GroES: $r = 0.81$
Computing time

Table 1: Computing time used for FA, FAML and RELION methods. $32^3$ in parenthesis denotes the testing subtomograms are of size $32^3$.

<table>
<thead>
<tr>
<th>Method</th>
<th>Iterations to converge</th>
<th>Mean time per iteration</th>
<th>Total time</th>
</tr>
</thead>
<tbody>
<tr>
<td>FA ($32^3$)</td>
<td>11</td>
<td>7 s</td>
<td>78 s</td>
</tr>
<tr>
<td>FAML ($32^3$)</td>
<td>10</td>
<td>56 s</td>
<td>562 s</td>
</tr>
<tr>
<td>RELION ($32^3$)</td>
<td>8</td>
<td>340 s</td>
<td>2720 s</td>
</tr>
<tr>
<td>FA ($64^3$)</td>
<td>4</td>
<td>27 s</td>
<td>106 s</td>
</tr>
<tr>
<td>FAML ($64^3$)</td>
<td>3</td>
<td>150 s</td>
<td>451 s</td>
</tr>
<tr>
<td>RELION ($64^3$)</td>
<td>6</td>
<td>340 s</td>
<td>2041 s</td>
</tr>
<tr>
<td>FA ($128^3$)</td>
<td>5</td>
<td>143 s</td>
<td>717 s</td>
</tr>
<tr>
<td>FAML ($128^3$)</td>
<td>4</td>
<td>449 s</td>
<td>1794 s</td>
</tr>
<tr>
<td>RELION ($128^3$)</td>
<td>3</td>
<td>921 s</td>
<td>2764 s</td>
</tr>
</tbody>
</table>

- FAML is slower than FA but takes less iterations to converge
- FAML is 2 to 5 times faster than Relion using their default parameters
Hypothesis testing for structure identification
Pipeline results on a rat neuron tomogram
Future works
Future works

1. Saliency based particle picking method
2. Better tomogram reconstruction algorithm
3. Fast systematic structure identification method
4. Models for recovered structures
5. Spatial statistical models for macromolecule distribution
Acknowledgement

Funding support:
NIH P41 GM103712
Samuel and Emma Winters Foundation

People
Dr. Min Xu
Dr. Robert Murphy
Dr. Zachary Freyberg
Dr. Tzviya Zeev-Ben-Mordehai
References


