A machine learning method for resolving heterogeneity in CryoEM single particle reconstruction

Muyuan Chen
2018-03
Single particle reconstruction
Heterogeneity in Single particle reconstruction

Low resolution in heterogeneous regions
Classical strategy: multi-model refinement

Works, but is suboptimal...
Determination of conformational landscape
A even simpler simulated dataset

- 3000 particles from 11 classes
- Low resolution SNR: ~1
- 2 - 3.5um defocus
Start from a single model refinement
Gaussian representation of density map

\[ \text{map}(x) = \sum_{j=0}^{n} A_j e^{-\frac{(x - \mathbf{p}_j)^2}{\sigma_j^2}} \]

- Reduce complexity
- Easier to model continuous motion
Particle-projection comparison

Raw particle

Target resolution

Corresponding projection

Particle-projection FRC
Gradient calculation

Goal:
To improve particle-projection matching (FRC)
Goal:
To improve particle-projection matching (FRC)

For each Gaussian, which direction should it move to improve the particle-projection FRC on each particle?
Gradient per particle per Gaussian

2D gradient vectors on projection plane
Gather gradient vectors of all particles in 3D

3D gradient vectors projected on x-y plane
Statistics on gradient vectors

Average amplitude of gradient vector of each Gaussian indicates local flexibility
Statistics on gradient vectors

Connect gradient vectors from the same particle
Statistics on gradient vectors

Gradient vectors from each image are correlated
PCA extracts eigen-motion vectors of the system

Most particles span on this motion trajectory

This is one vector of length (#Gaussian x 3), which shows a global motion of all Gaussians
Eigen-motion trajectory
Map each particle on the motion trajectory

\[ \text{map}_i(x) = \sum_{j=0}^{n} A_j e^{\frac{x-(p_j+c_i v_j)}{\sigma_j^2}} \]

- \( \text{map}_i(x) \): Density map corresponding to the \( i \)th particle
- \( A_j, p_j, \sigma_j \): Amplitude, position, sigma of the \( j \)th Gaussian
- \( v_j \): Eigen-motion vector of the \( j \)th Gaussian
- \( C_i \): Conformation position of the \( i \)th particle on the motion trajectory

For each particle, optimize \( C_i \) with gradient descent
Reconstruct 3D map with particles of similar conformations

planar motion in this simulated dataset (invisible at z direction) exaggerates top view error…
Relationship with neural network

For each particle:

- Euler angle (alt, az, phi)
- Rotation matrix (3x3 matrix)
  - Multiply
  - Position of Gaussians (3xn matrix)
- Transformed Gaussian positions (3xn matrix)
- Projection
- Projection image
- FRC
- Raw particle

Gradient
Relationship with neural network

One layer neural network

Input (x) \rightarrow Multiply \rightarrow W x + b \rightarrow Output (y) \rightarrow Activation function

Weights (W) \rightarrow Loss function \rightarrow Target output (y')
Relationship with neural network

For each particle:

- Implement under deep learning framework
- Efficient GPU utilization
- Symbolic gradient calculation

- Euler angle
- Rotation matrix (3x3 matrix)
- Multiply
- Position of Gaussians
- Transformed Gaussian positions (3xn matrix)
- Activation function
- Projection
- Projection image
- Multiply
- Weights
- FRC
- Target output
- Raw particle
Performance on real examples
Ribosome (EMPIAR-10107)

Compositional heterogeneity (50Å, 64 Gaussians): gradient with respect to amplitude of each Gaussian

\[ \text{Grad}_j = \frac{\partial FRC}{\partial A_j} \]
Ribosome (EMPIAR-10107)

Conformational heterogeneity (40Å, 128 Gaussians)

Color: eigen-motion amplitudes

Gaussian model
Global motion

Gaussian model

Reconstructed maps from classified particles
Focus on local regions
Focus on local regions
Focus on local regions
GroEL

- From Roh 2017, PNAS
- Filtered to 7Å
- 1344 Gaussians
- Motion of helices
- Symmetry breaking
- Correlation of conformation between subunits
Advantages

VS multi-model refinement:
• Deterministic
• Handles continuous motion

VS image based manifold mapping:
• Lower requirement for dataset
• Simple, expandable framework
• Solves global and local motion

Limitations

• Requires determined orientations from single model refinement
• Linear and short motion trajectory
• Low resolution due to limited GPU memory
Advantages

VS multi-model refinement:

- Deterministic
- Handles continuous motion

VS image based manifold mapping:

- Lower requirement for dataset
- Simple, expandable framework
- Solves global and local motion

Limitations and future directions

- Requires determined orientations from single model refinement
  - Iteratively optimize conformation and orientation
- Linear and short motion trajectory
  - Replace PCA with stacked autoencoder for trajectory calculation
- Low resolution due to limited GPU memory
  - Better hardware and software platforms
Availability

- EMAN2/examples
  - build.ali_lst.py
    build list file with alignment information from existing refinement
  - gmm_heterog.py or gmm_heterog_tensorflow.py
    main program (theano or tensorflow implementation)
- Tutorial coming soon…
Acknowledgement

• PIs:
  • Steven Ludtke, Baylor College of Medicine

• NIH grants:
  • R01GM080139

Special thanks to Hui Ye for looking after my cat…
Thank you
Both are linear in Gaussian representation.
Relationship with neural network

For each particle:
- Implement under deep learning framework
- Efficient GPU utilization
- Symbolic gradient calculation