

# Computational methods for single-particle cryo-electron microscopy

Daniel Hogan and Hugo Kitano

CS371 presentation

15 February 2017

# What is Cryo-EM?



Gaining traction in recent years due to better cameras

# What is Cryo-EM?



Gaining traction in recent years due to better cameras
Crystallization avoided!

- can change conformation

# What is Cryo-EM?



Gaining traction in recent years due to better cameras
Crystallization avoided!

- can change conformation
- difficult for larger molecules

# What is Cryo-EM?



Gaining traction in recent years due to better cameras
Crystallization avoided!

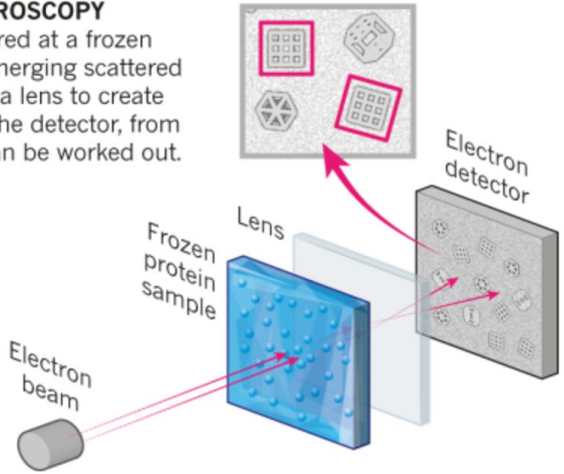
- can change conformation
- difficult for larger molecules

Lower resolution, but easier reconstruction problems

# Setup



**CRYO-ELECTRON MICROSCOPY**
A beam of electron is fired at a frozen protein solution. The emerging scattered electrons pass through a lens to create a magnified image on the detector, from which their structure can be worked out.




Nature

# Two Steps to Reconstruct a 3D structure



Refine the 2D images

# Two Steps to Reconstruct a 3D structure



Refine the 2D images

- align movie frames to account for movement

# Two Steps to Reconstruct a 3D structure



Refine the 2D images

- align movie frames to account for movement
- cluster images that look similar together to average them

# Two Steps to Reconstruct a 3D structure



Refine the 2D images

- align movie frames to account for movement
- cluster images that look similar together to average them

3D reconstructions

- Combine our 2D projections into a 3D structure

# Two Steps to Reconstruct a 3D structure



Refine the 2D images

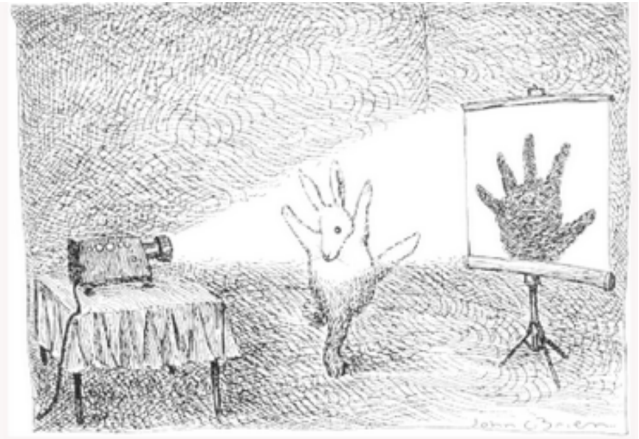
- align movie frames to account for movement
- cluster images that look similar together to average them

3D reconstructions

- Combine our 2D projections into a 3D structure
- Back-projection is difficult!

# Bunny





From Joachim Frank, *Three-dimensional electron microscopy of macromolecular assemblies: Visualization of biological molecules in their native state*, 2006

# Analysis difficulties



- noisy images

# Analysis difficulties



- noisy images
- random protein orientations

# Analysis difficulties



- noisy images
- random protein orientations
- 3D reconstruction

# Analysis difficulties



- noisy images
- random protein orientations
- 3D reconstruction
- risk of overfitting data

# Clustering



In order to create a 3D reconstruction, the 2D projections need to be clustered

# Clustering



In order to create a 3D reconstruction, the 2D projections need to be clustered

Chicken and egg problem ("ill-posed"):

# Clustering



In order to create a 3D reconstruction, the 2D projections need to be clustered
Chicken and egg problem ("ill-posed"):

- orientation information is necessary for cluster determination

# Clustering

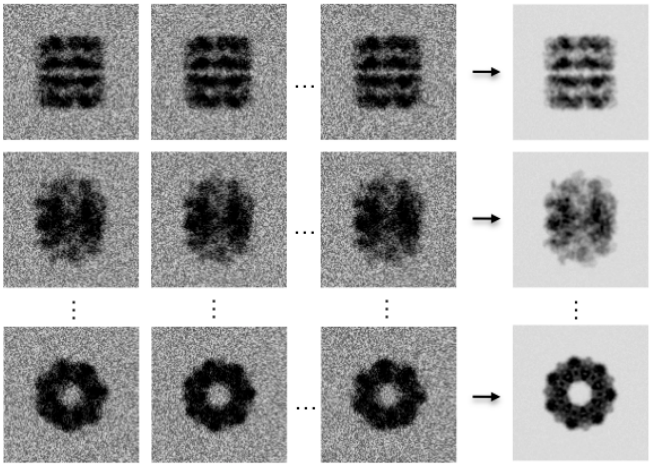


In order to create a 3D reconstruction, the 2D projections need to be clustered
Chicken and egg problem ("ill-posed"):

- orientation information is necessary for cluster determination
- cluster information makes orientation determination tractable

# Clustering

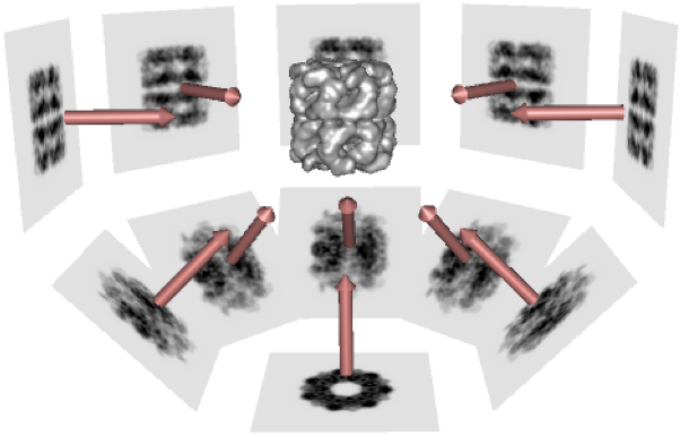




Pintilie `http://people.csail.mit.edu/gdp/cryoem.html`

# Back projection



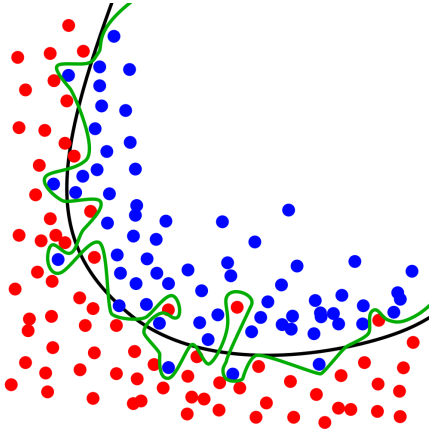


Pintilie `http://people.csail.mit.edu/gdp/cryoem.html`

# Overfitting



Random noise becomes part of the model



Wikipedia https://en.wikipedia.org/wiki/File:Overfitting.svg

# Smoothing



Smoothing is a powerful way to reduce overfitting, but it's currently done via ad hoc filtering

# Smoothing



Smoothing is a powerful way to reduce overfitting, but it's currently done via ad hoc filtering

- arbitary decisions using unstandardized heuristics, causes overfitting as well

# Smoothing



Smoothing is a powerful way to reduce overfitting, but it's currently done via ad hoc filtering

- arbitary decisions using unstandardized heuristics, causes overfitting as well
- separate steps of particle alignment, class averaging, filtering, and 3D reconstruction

# MAP estimator



- We will try to maximize a single probability function that takes into account all of the steps

# MAP estimator



- We will try to maximize a single probability function that takes into account all of the steps
- *Maximum a priori* estimation, which uses prior information to make our prediction:

$$\hat{\theta}_{\text{MAP}} = \underset{\theta}{\text{argmax}}\ P\left(\theta|D\right)$$

$$\hat{\theta}_{\text{MAP}} = \underset{\theta}{\text{argmax}}\ P\left(D|\theta\right) P\left(\theta\right)$$

# Bayesian refinement algorithm



This is very difficult!

$$V_l^{(n+1)} = \frac{\sum_{i=1}^{N} \int_\phi \Gamma_{i\phi}^{(n)} \sum_{j=1}^{J} \mathbf{P}^{\phi}_{lj}{}^{T} \frac{CTF_{ij} X_{ij}}{\sigma_{ij}^{2(n)}} \, d\phi}{\sum_{i=1}^{N} \int_\phi \Gamma_{i\phi}^{(n)} \sum_{j=1}^{J} \mathbf{P}^{\phi}_{lj}{}^{T} \frac{CTF_{ij} X_{ij}}{\sigma_{ij}^{2(n)}} \, d\phi + \frac{1}{\tau_l^{2(n)}}}$$

$$\sigma_{ij}^{2(n+1)} = \frac{1}{2} \int_\phi \Gamma_{i\phi}^{(n)} \left| X_{ij} - CTF_{ij} \sum_{l=1}^{L} \mathbf{P}^{\phi}_{jl} V_l^{(n)} \right|^2 d\phi$$

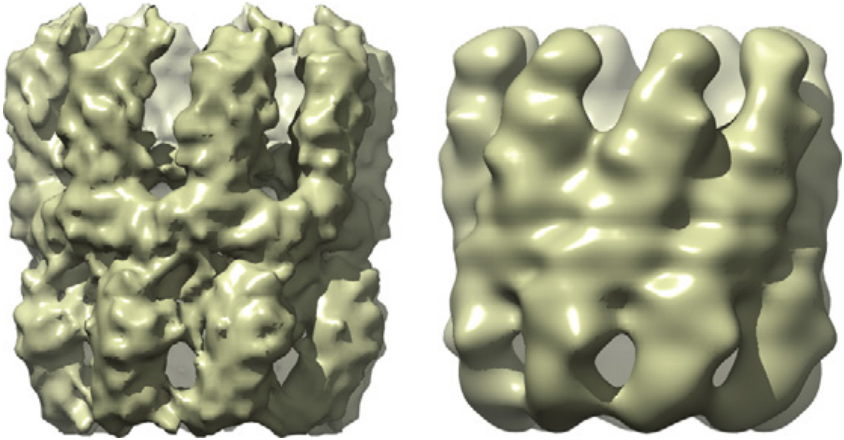
$$\tau_l^{2(n+1)} = \frac{1}{2} \left| V_l^{(n+1)} \right|^2$$

where

$$\Gamma_{i\phi}^{(n)} = \frac{P\left(X_i | \phi, \Theta^{(n)}, Y\right) P\left(\phi | \Theta^{(n)}, Y\right)}{\int_{\phi'} P\left(X_i | \phi', \Theta^{(n)}, Y\right) P\left(\phi' | \Theta^{(n)}, Y\right) \, d\phi'}$$

# Less overfitting

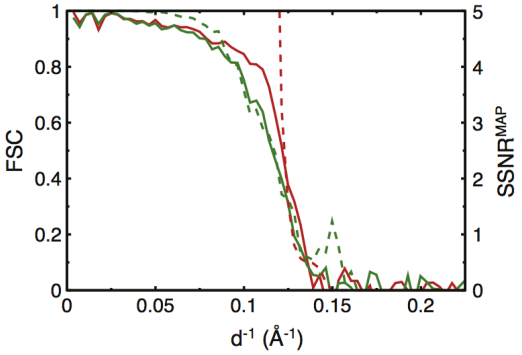
## Overfitted vs. MAP



Scheres http://dx.doi.org/10.1016/j.jmb.2011.11.010

# Greater objectivity

The new approach (red) has higher resolution and greater objectivity than the old (green)




Scheres http://dx.doi.org/10.1016/j.jmb.2011.11.010

# Future improvements



- Better microscopes and detectors will lead to less noise

# Future improvements



- Better microscopes and detectors will lead to less noise
- More information about the relative orientations (especially for symmetric molecules)

# Future improvements



- Better microscopes and detectors will lead to less noise
- More information about the relative orientations (especially for symmetric molecules)
- Regularization and the use of prior information (used here!)

# Ribosomes



- Responsible for the synthesis of protein using a mRNA template

# Ribosomes



- Responsible for the synthesis of protein using a mRNA template
- Two subunits
  - Large subunit, composed of three rRNAs and 46 proteins
  - Small subunit, composed of one rRNA and 33 proteins

# Ribosomes



- Responsible for the synthesis of protein using a mRNA template
- Two subunits
  - Large subunit, composed of three rRNAs and 46 proteins
  - Small subunit, composed of one rRNA and 33 proteins
- The subunits rotate during each step elongation

# Time series



Objective: a series of structures of the ribosome to construct a time series

# Time series



Objective: a series of structures of the ribosome to construct a time series

- Purify ribosomes

# Time series



Objective: a series of structures of the ribosome to construct a time series

- Purify ribosomes
- Cryofix and image

# Time series



Objective: a series of structures of the ribosome to construct a time series

- Purify ribosomes
- Cryofix and image
- Categorize by orientation and conformation

# Time series



Objective: a series of structures of the ribosome to construct a time series

- Purify ribosomes
- Cryofix and image
- Categorize by orientation and conformation
- Determine structures

# Time series

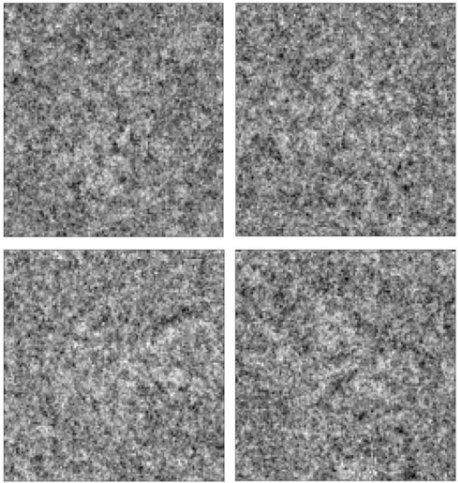


Objective: a series of structures of the ribosome to construct a time series

- Purify ribosomes
- Cryofix and image
- Categorize by orientation and conformation
- Determine structures
- Construct a time series

# Raw images





Dashti, *et al.* http://dx.doi.org/10.1073/pnas.1419276111

# Data



- ∼4,700 micrographs

# Data



- ∼4,700 micrographs
- ∼1,100,000 particles found algorithmically

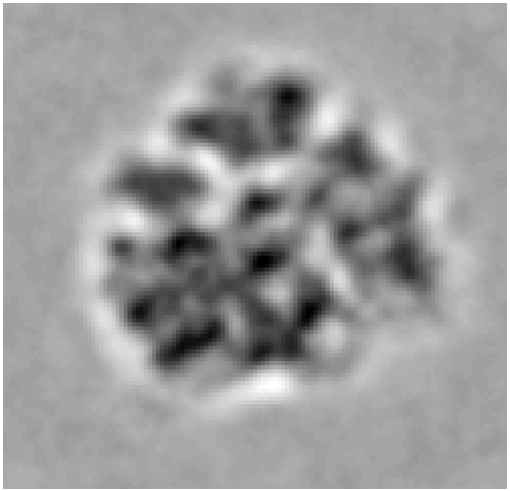
# Data



- ∼4,700 micrographs
- ∼1,100,000 particles found algorithmically
- ∼850,000 particles after manual selection

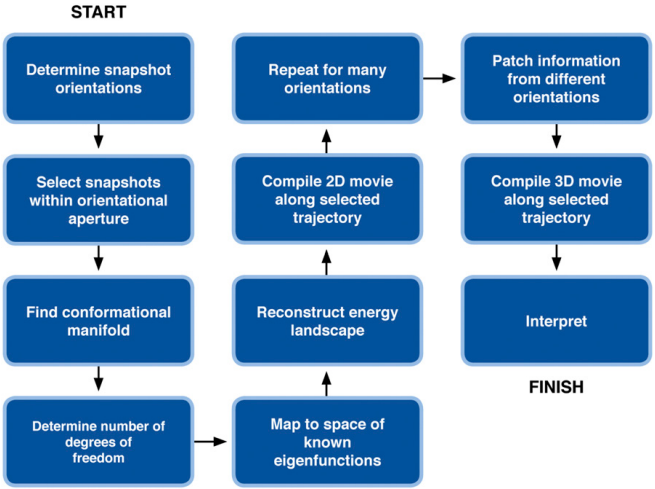
# Oriented image





Dashti, *et al.* http://dx.doi.org/10.1073/pnas.1419276111

# Analysis procedure

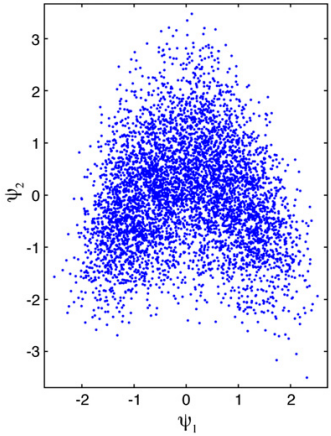



Dashti, *et al.* http://dx.doi.org/10.1073/pnas.1419276111

# Conformational manifold



Determined by a non-linear analog of PCA




Dashti, *et al.* http://dx.doi.org/10.1073/pnas.1419276111

# Analysis details



- 50 distinct conformations were identified

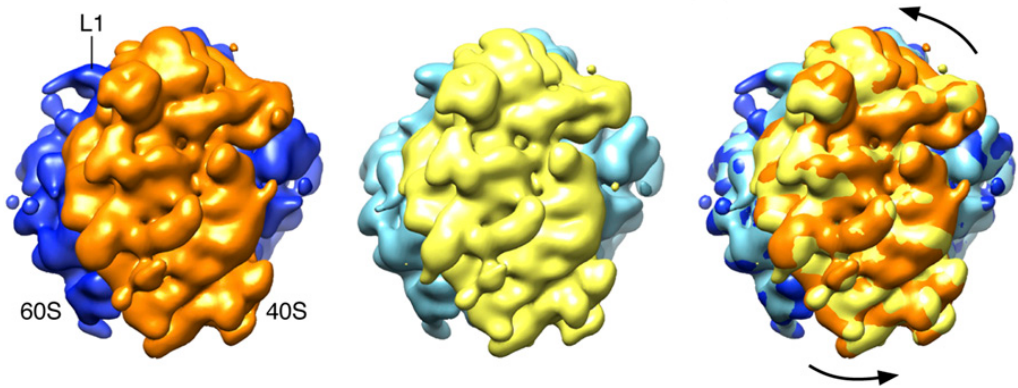
# Analysis details



- 50 distinct conformations were identified
- Ordering was inferred from similarity

# Structures

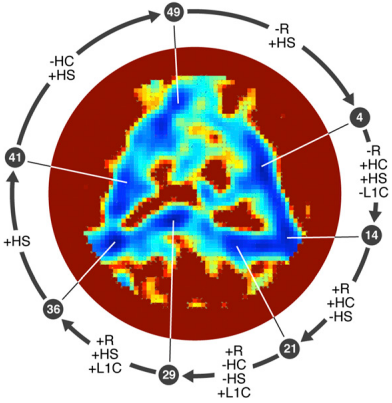


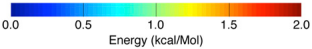



Dashti, *et al.* http://dx.doi.org/10.1073/pnas.1419276111

# Ribosome trajectory

## Free energy inferred by relative populations




+R: increase intersubunit rotation
+HC: increase 40S head closure
+HS: increase 40S head swivel
+L1C: increase L1-stalk closing

Dashti, *et al.* http://dx.doi.org/10.1073/pnas.1419276111

# Concerns



- Lack of detail on the preparation of ribosomes

# Concerns



- Lack of detail on the preparation of ribosomes
  - The imaged ribosomes were “not engaged in translation”
  - But ribosomal subunits do not bind together in the absence of mRNA

# Concerns



- Lack of detail on the preparation of ribosomes
  - The imaged ribosomes were “not engaged in translation”
  - But ribosomal subunits do not bind together in the absence of mRNA
- The ribosomes were manually selected from the micrographs, introducing a potential source of bias

# Concerns



- Lack of detail on the preparation of ribosomes
  - The imaged ribosomes were “not engaged in translation”
  - But ribosomal subunits do not bind together in the absence of mRNA
- The ribosomes were manually selected from the micrographs, introducing a potential source of bias
- Selecting images based on orientation before conformation