FLATIRON INSTITUTE
Center for Computational Mathematics

Fast algorithms for in situ single protein matching in Cryo tomography

Manas Rachh[1], Alex Barnett[1], Leslie Greengard[1], Nikolaus Grigorieff[2]

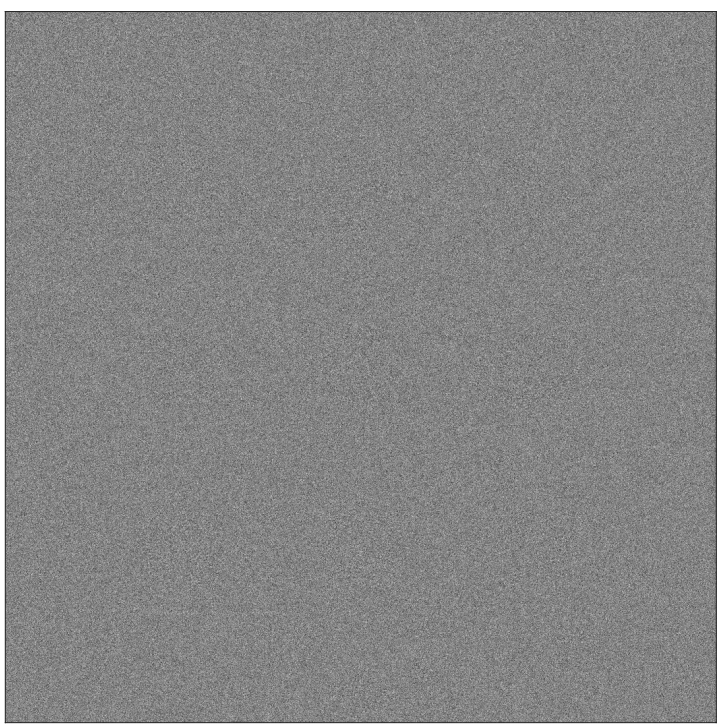
Nov 28, 2018

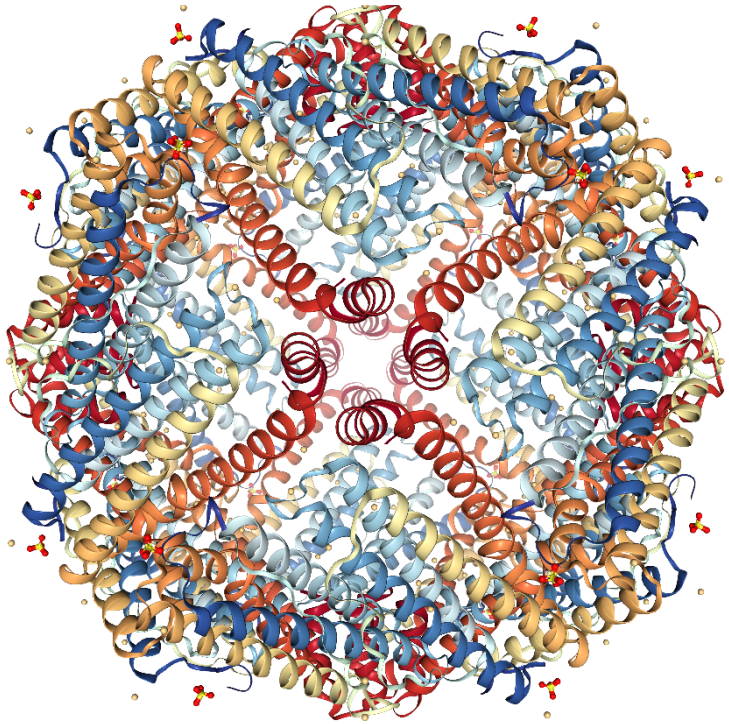
[1]Center for Computational Mathematics, Flatiron Institute, Simons Foundation

[2]Howard Hughes Medical Institute

Problem statement

Task: Find orientation and location of proteins in cryo-tomograms using template matching

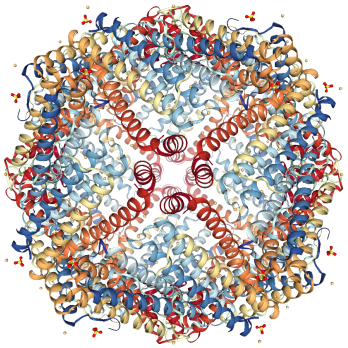




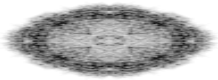
(left) A simulated tomogram with SNR = 0.025, (right) Apoferritin molecule (source: rscb pdb)

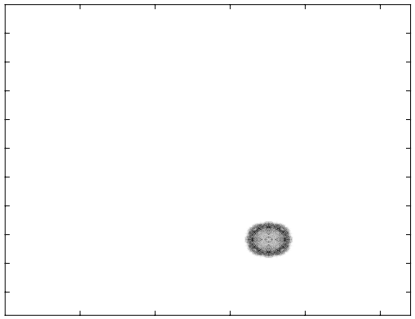
Maximum likelihood noise model - I

$R \in \mathbf{SO}(3)$



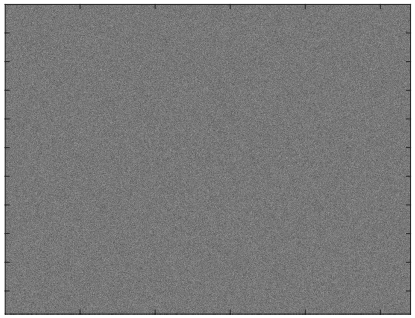
$T_R(\boldsymbol{x}) = \int V(\tilde{x}_R, \tilde{y}_R, \tilde{z}_R)\, dz$





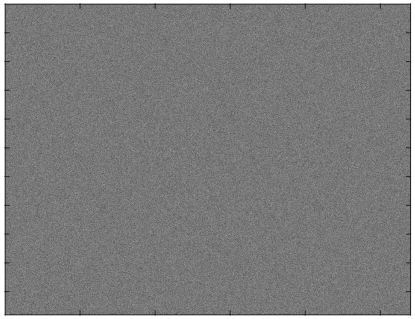
$T_R(\boldsymbol{x} + \boldsymbol{x}_0)$

$+$



$\eta(\boldsymbol{x})$

$=$



$I(\boldsymbol{x})$

Maximum likelihood noise model - II

$$I(\boldsymbol{x}) = T_R(\boldsymbol{x} + \boldsymbol{x}_0) + \eta(\boldsymbol{x}), \quad \eta \sim \mathcal{N}(0, \sigma^2), \text{ iid}$$

$$p_{\text{lik}}(I|\boldsymbol{x}_0, R) = \prod_{\boldsymbol{x}} \frac{1}{2\pi\sigma^2} \exp\left(-|T_R(\boldsymbol{x} + \boldsymbol{x}_0) - I(\boldsymbol{x})|^2/(2\sigma^2)\right)$$

$$\hat{\boldsymbol{x}}_0, \hat{R} = \text{argmin}_{\boldsymbol{x}_0, R}(-\log p_{\text{lik}}) = |I|^2 + |T_R|^2 - 2\langle I, T_R(\cdot + \boldsymbol{x}_0)\rangle \ ,$$

where

$$\langle I, T_R(\cdot + \boldsymbol{x}_0\rangle = \int I(\boldsymbol{x}) T_R(\boldsymbol{x} + \boldsymbol{x}_0)$$

Assume: $|T_R|^2 \approx \text{const}$

$$\hat{\boldsymbol{x}}_0, \hat{R} = \text{argmin}_{\boldsymbol{x}_0, R}(-\log p_{\text{lik}}) = \text{argmax}_{\boldsymbol{x}_0, R} \langle I, T_R(\cdot + \boldsymbol{x}_0)\rangle$$

Maximum likelihood estimator is the maximizer of the cross-correlation

Computational cost

- N - number of pixels in image $I \sim 1080^2$ for numerical experiments
- M - number of pixels in template $T \sim 360^2$ for apoferritin at $0.965A$ resolution
- N_r - number of rotations in $\mathrm{SO}(3) \sim 2 \times 10^6$ for $\Delta\theta = 1.7^\circ$
- Fourier slice theorem for template generation $O(N_r \cdot M)$

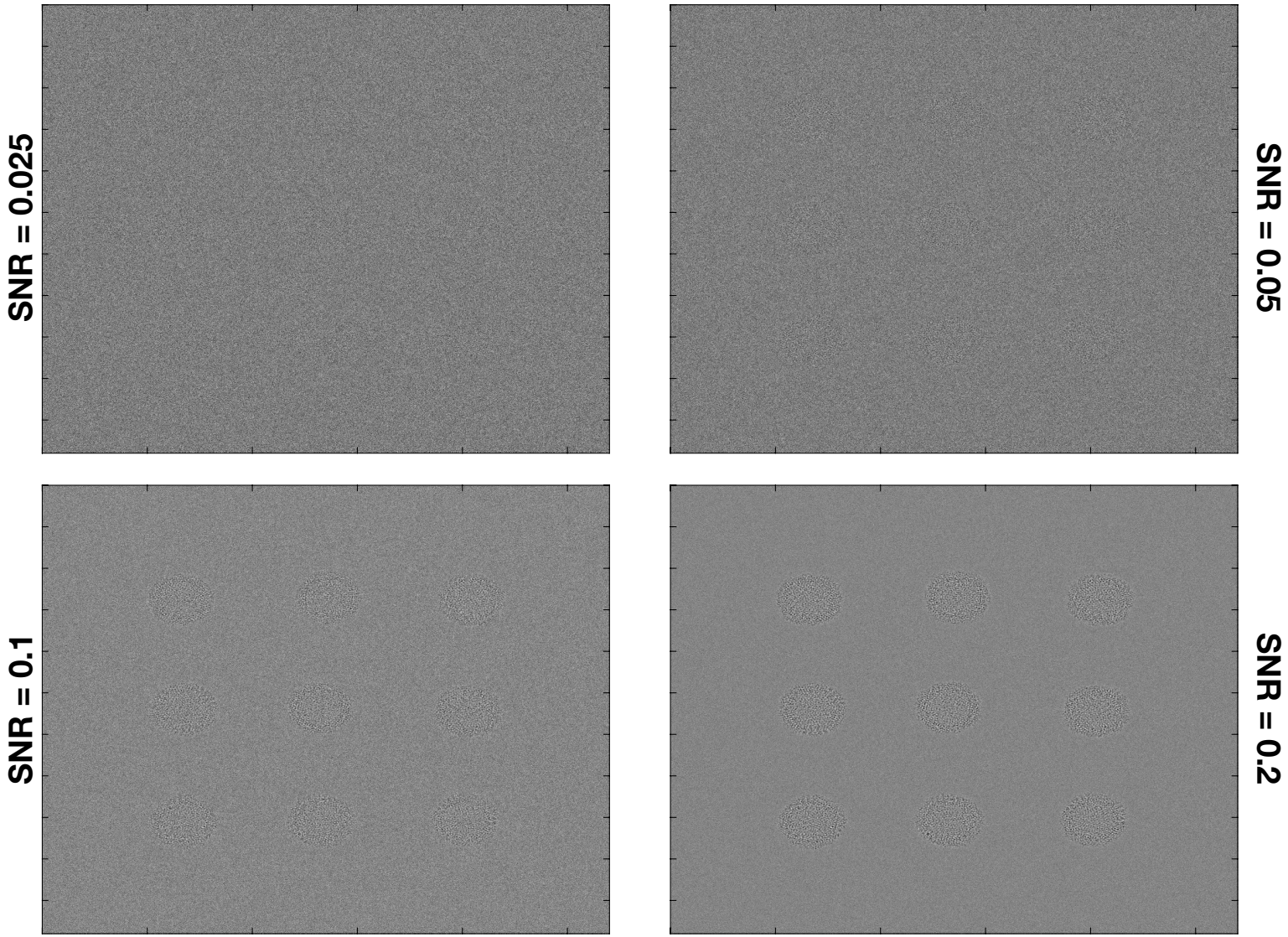
$$T_R(\boldsymbol{x}) = \int V(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, \tilde{\boldsymbol{z}})\, dz$$

- FFTs for computing cross correlations $O(N_r \cdot N)$

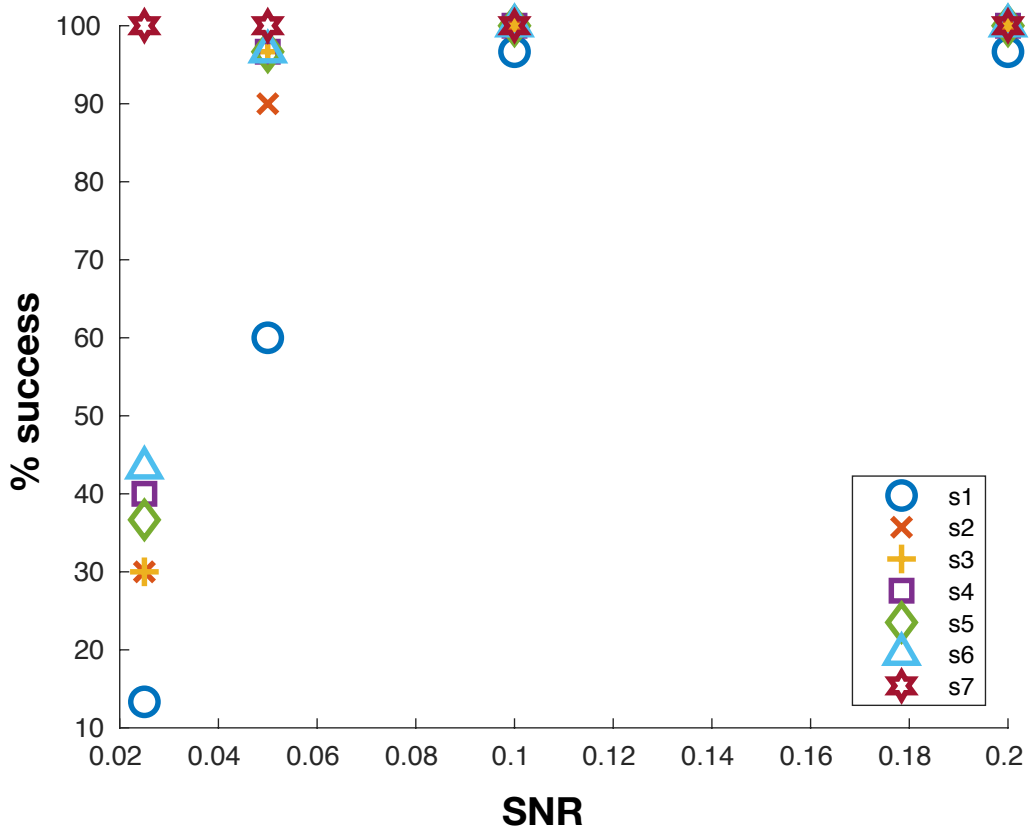
$$\langle I, T_R(\cdot + \boldsymbol{x}_0) \rangle \quad \text{for all } \boldsymbol{x}_0$$

- 100 CPU hours per template per image
- Proposed acceleration: **Frequency marching**
  Low resolution search followed by local searches at higher resolution

Typical images

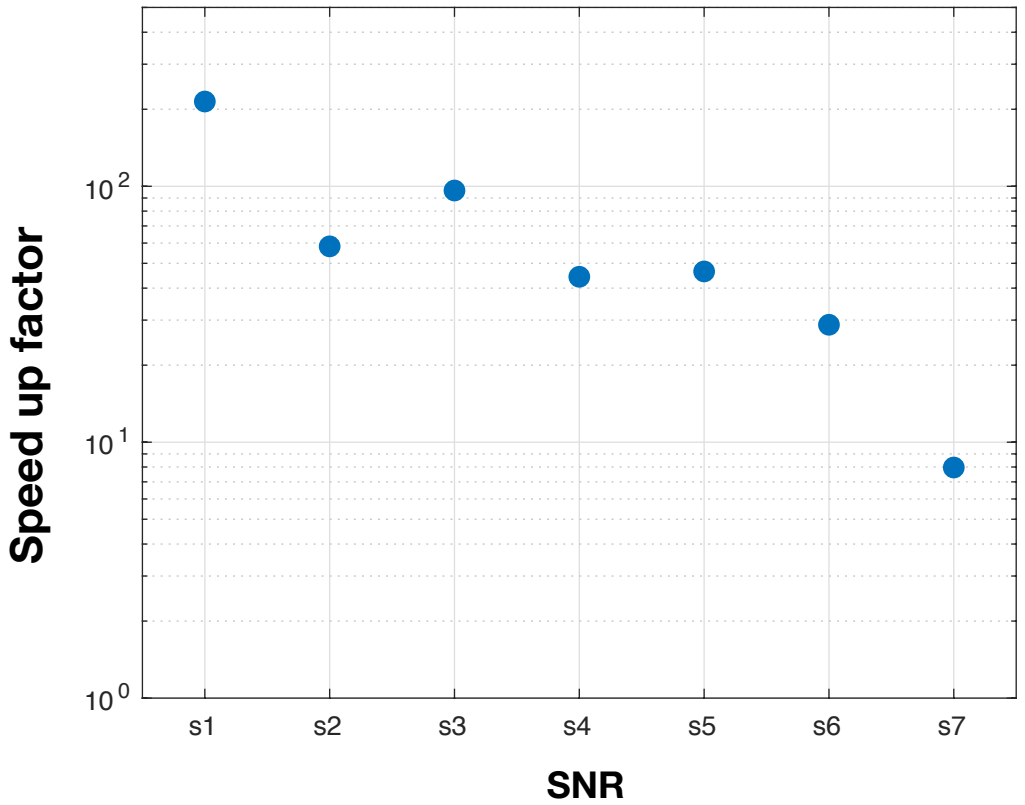


Preliminary results - Accuracy



- $s1 : \Delta\theta = 6.8^{\circ}, \Delta x = 1.93A$
- $s2 : \Delta\theta = 6.8^{\circ}, \Delta x = 0.965A$
- $s3$: initial coarse search using $s1$, then local searches around top 250 candidates
- $s4$: initial coarse search using $s2$, then local searches around top 250 candidates
- $s5$: initial coarse search using $s1$, then local searches around top 1000 candidates
- $s6$: initial coarse search using $s2$, then local searches around top 1000 candidates
- $s7 : \Delta\theta = 3.4^{\circ}, \Delta x = 0.965A$

Preliminary results - Speed



- $s1 : \Delta\theta = 6.8^\circ, \Delta x = 1.93A$
- $s2 : \Delta\theta = 6.8^\circ, \Delta x = 0.965A$
- $s3$: initial coarse search using $s1$, then local searches around top 250 candidates
- $s4$: initial coarse search using $s2$, then local searches around top 250 candidates
- $s5$: initial coarse search using $s1$, then local searches around top 1000 candidates
- $s6$: initial coarse search using $s2$, then local searches around top 1000 candidates
- $s7 : \Delta\theta = 3.4^\circ, \Delta x = 0.965A$

Ongoing work

- Faster template generation using NUFFT
- Optimization of parameters for local searches
- Exploiting symmetries of proteins
- Better maximum likelihood estimators
- Performance in crowded protein environments
- Using tilt series information available to improve SNR
- Automated and robust pipeline for simultaneous detection of collection of proteins

Ongoing work

- Faster template generation using NUFFT
- Optimization of parameters for local searches
- Exploiting symmetries of proteins
- Better maximum likelihood estimators
- Performance in crowded protein environments
- Using tilt series information available to improve SNR
- Automated and robust pipeline for simultaneous detection of collection of proteins

- Questions?

Thank you