Automating Cryo-Electron Microscopy

On the fly processing/Sorting 3D heterogeneity

Donovan Webb - University of Bath - dw711@bath.ac.uk
Yuriy Chaban - eBIC, Diamond Light Source

Cryo-Electron Microscopy

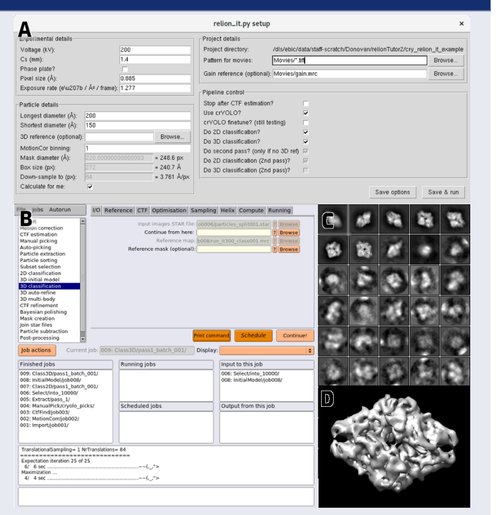
Cryo-Electron Microscopy[1] (Cryo-EM) is a popular structural biology method used to study biomolecules such as proteins and viruses. Structures found can be solved to near atomic resolution (recently even to atomic resolution[2]) and so are invaluable in understanding biomolecules mechanistic properties.

- Cryo-EM can study particles in near natural settings (in water or even in cells).
- Multiple conformations of the molecule can be observed.

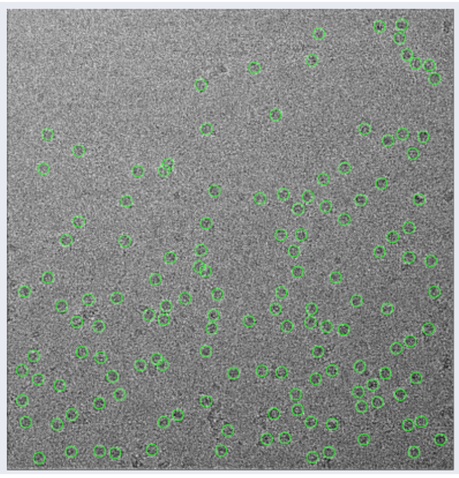
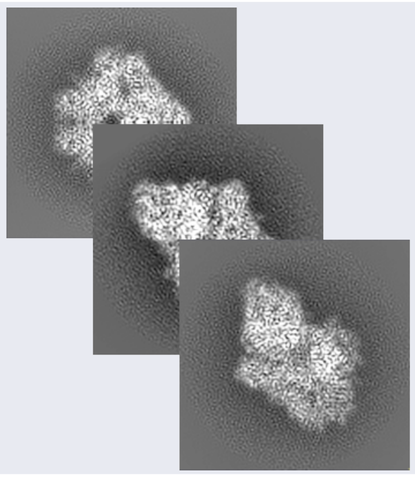
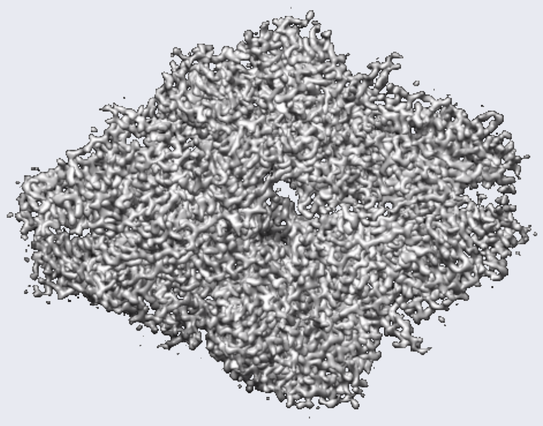
The Method

- Aqueous particle solution is plunged into liquid ethane to embed in vitrified ice.
- A transmission electron microscope, TEM, is used to collect images, known as micrographs.
- Projections of target particle are located and associated with a projection angle.
- A low dose of electrons is used as biological macromolecules are easily damaged. This leads to high noise. Consequently, >100,000 of projections are required to recover high resolution 3D structure.

Particle Picking



Data collection pipeline of β - galactosidase.
a) GUI presented to user to start pipeline.
b) Relion GUI of automatic pipeline.
c) 2D Classes found.
d) 3D reconstruction found with pipeline.

Left to right:
Typical micrograph,
Micrograph with locations of particles,
2D class averages of extracted particles,
3D reconstruction.

Motivation

- Heterogeneity in particle data sets lead to poor 3D reconstructions.
- Current methods for cleaning data sets are iterative and manual.
- An unsupervised method saves time and lowers human bias skewing models.

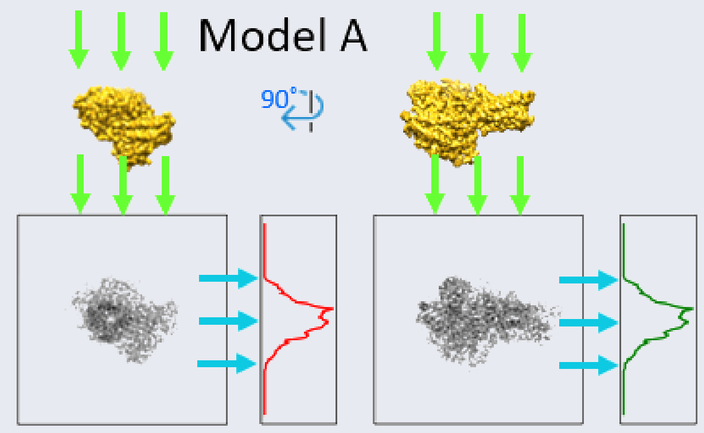
CLIC Method

Sinograms/2D Radon transform: This method relies on Radon transform **'single lines'** to relate projections to one another. A sinogram is made by finding line integrals through the image at a range of angles. An example sinogram can be seen below. The 2D Radon transform is often written with the line expressed as $\rho = x\cos\theta + y\sin\theta$, where ρ is the smallest distance to the coordinate system's origin and θ is the angle through which the line integral is calculated. The transform for a set of parameters (ρ, θ) is then defined as,

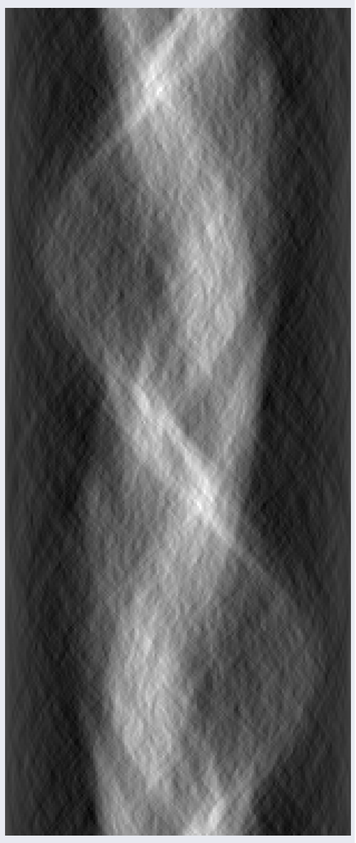
$$R(\rho,\theta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\delta(\rho - x\cos\theta - y\sin\theta)dxdy$$

where $\delta()$ is the Dirac delta function.

The importance of the Radon transform is that **sinograms produced from two 2D projections of the same 3D model will always share at least one common line**. This concept is shown below.

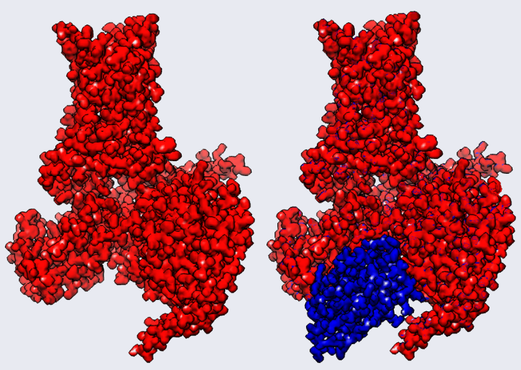


Common line between two projections from the same model.



Clustering

Projections with identical common lines are from homogeneous groups. But, in real experiments high noise causes no common line to be identical. **How to cluster projections with most similar common lines?**
We propose a new method, **C**ommon **L**ine **I**mplied **C**lustering (CLIC), consisting of statistical analysis of 1D Radon Transform profiles, followed by dimensional reduction of the data points cloud, with subsequent unsupervised hierarchical clustering of particles to recover homogeneous groups from raw data.

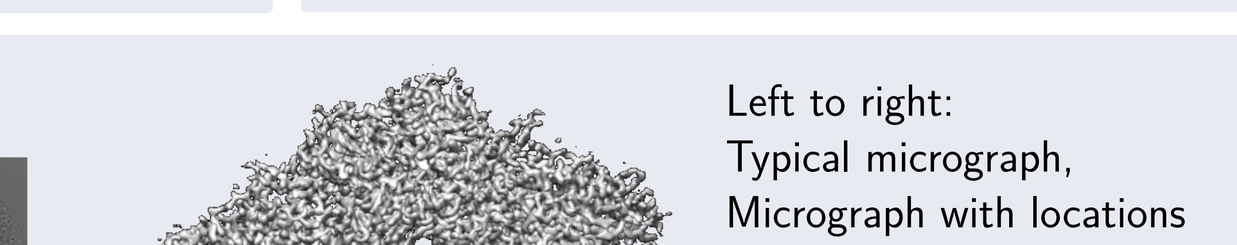


Two models used for synthetic data set. Model can be found at PDB 6gdg.[3]

CLIC Results

Dimensional reduction method

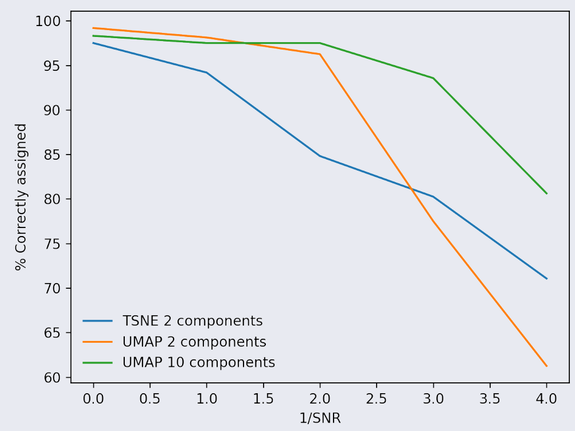
Dimensional reduction is feature extraction step - best differentiating features are found. We compared multiple linear and non-linear methods to find an appropriate technique for sinogram data. Shown below is a sample of 15 sinograms originating from 160 simulated projections from two models. Two models are of the same complex (PDB 6gdg) with one medium sized group missing, the models can be seen on bottom of the left panel.

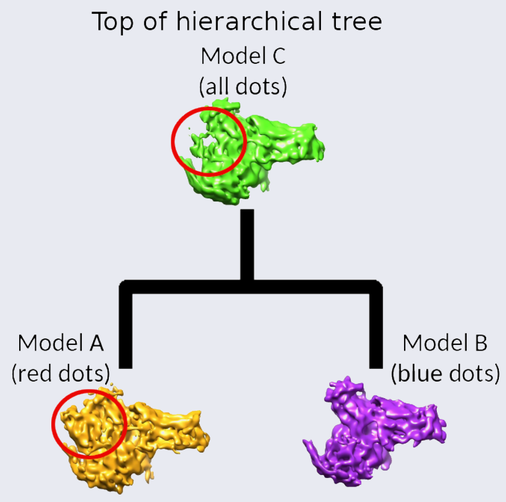


75 dimensional data reduced to 2 components. Top coloured depending on original sinogram. Bottom coloured depending on original model.
Left to right: PCA, UMAP, LLE, ISOMAP, TSNE.

Note: The points trace 'strings' in dimensionally reduced space - promising as adjacent points from the same sinogram should be similar. UMAP[4] and TSNE[5] have good separation of models in two dimensional space. As UMAP is also computationally efficient it was chosen for further testing.

Noise: Graph shows accuracy of three methods affected by increasing noise - UMAP with 10 components was found to work best at higher noise levels. This breakdown at higher noise levels, although expected, may mean experimental data will need prior denoising to separate into acceptable groups. $\Rightarrow$





$\Leftarrow$ **Final models:**
3D reconstructions were produced from the two largest clusters to validate results. On left it can be seen that the models recovered match the two original models used to create the data - the method has succeeded in grouping projections into homogeneous clusters.

Conclusion

Unsupervised clustering techniques have a potential to produce homogeneous groups by inspecting sinogram data produced from raw particles. This could lead to a more efficient approach to cryo-EM data analysis.