Algorithmic Advances in Single Particle Cryo-EM Data Processing

Ali Punjani[1,2], Haowei Zhang[2], John Rubinstein[3], Marcus Brubaker[4], David Fleet[1]

[1.] University of Toronto, Department of Computer Science, Toronto, Canada.
[2.] Structura Biotechnology Inc., Toronto, Canada.
[3.] Molecular Structure and Function Program, The Hospital for Sick Children Research Institute, Toronto, Canada. Department of Biochemistry, The University of Toronto, Toronto, Ontario, Canada. Department of Medical Biophysics, The University of Toronto, Toronto, Ontario, Canada.
[4.] Department of Electrical Engineering and Computer Science, York University, Toronto, Ontario, Canada

Single particle cryo-EM (cryo-electron microscopy) allows high-resolution imaging of macromolecular complexes in close-to-native state, at near-atomic resolutions. Cryo-EM has undergone several technological breakthroughs in microscopy, electron detectors, and image processing that have enabled its use recently in solving high-resolution structures of difficult proteins and complexes. As cryo-EM makes rapid progress, several key challenges remain in the quest for higher resolutions on challenging targets, as well as in the widespread and routine use of cryo-EM. Here, we introduce new methods in the cryo-EM data processing pipeline, along with implementations within the cryoSPARC software system [1], to address some of these challenges.

The image processing pipeline in single particle cryo-EM is required to solve the 3D electron density of a target molecule, in potentially many conformational states, from noisy 2D images collected using cryo-TEM. Each collected image is a movie of dose-fractionated frames that require motion estimation and correction. The corrected images (micrographs) are then used to estimate the microscope CTF during the exposure, as well as to find and pick out single particles. The single particles are extracted from the micrographs, and then are sorted and filtered using 2D classification methods. The resulting filtered particle stacks are used to perform *ab initio* 3D structure determination of potentially multiple discrete states or targets. These coarse structures are then further classified and refined in 3D to yield interpretable molecular density maps and achieve state of the art resolutions.

Here we introduce new robust methods for multiple stages in the processing pipeline, along with fast implementations of these new algorithms in the cryoSPARC software system.

For motion correction, we develop an improved variant of the *alignparts_lmbfgs* [2] algorithm that can estimate and correct for the motion of each individual particle in a micrograph. This new algorithm has the advantages of using a non-parametric motion trajectory model that can capture trajectories not modelled well as linear [3] or polynomial [4] motion, and a method to automatically tune parameters governing the smoothness of estimated trajectories. Together these improvements yield a robust, parameter-free algorithm with improved refinement resolutions on benchmark datasets, while requiring less than 10 seconds of processing time on a single GPU, per raw exposure movie.

For refining higher resolution structures, we introduce a new algorithm called non-uniform refinement that iteratively accounts for regions of a structure that have disordered or flexible density causing local loss of resolution. Accounting for these regions and dynamically estimating their locations can

significantly improve resolution in other regions as well as overall map quality by impacting the alignment of particles and reducing the tendency for refinement algorithms to over-fit disordered regions.

References:

[1] A Punjani *et al,* Nature Methods **14** (2017), p. 290–296
[2] JL Rubinstein and MA Brubaker J. Struct Biol **192** (2015), p. 188-95
[3] SHW Scheres, eLife **3** (2014), e03665
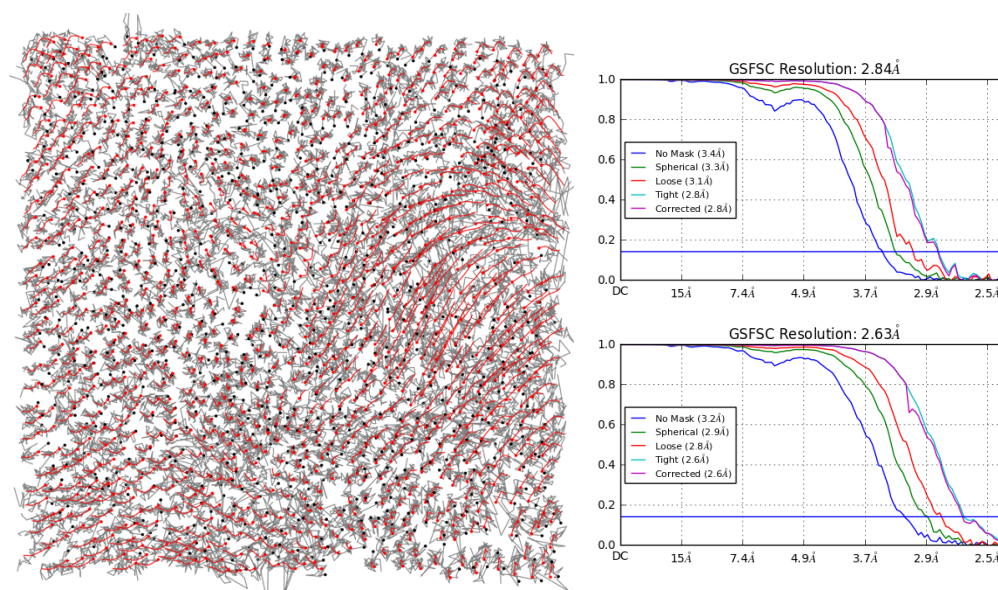[4] SQ Zheng, Nature Methods **14** (2017), p. 331–332

Figure 1. Left: Example of per-particle estimated motion trajectories on a typical single particle cryo-EM movie, showing both the raw estimated trajectories (gray) and the automatically optimally smoothed trajectories (red), both shown at 40x scale. There is clearly non-rigid motion in different parts of the ice, and multiple changes in direction over the duration of the exposure. Right: Fourier Shell Correlation (FSC) plots for (top) a refinement done with linear per-particle motion trajectories (i.e. particle polishing) and (bottom) our new per-particle motion trajectories, showing a clear improvement from 2.8A to 2.6A resolution. All results are on a sample of T20S Proteasome data from EMPIAR-10025.