Microsc. Microanal. 21 (Suppl 3), 2015
doi:10.1017/S1431927615006790 Paper No. 0600

Large-Scale 3D Heterogeneity Analysis of CryoEM Data Using Likelihood-Based Classification in Frealign

Dario Oliveira dos Passos[1] and Dmitry Lyumkis[1]

[1.] The Salk Institute for Biological Studies, Laboratory of Genetics, La Jolla, CA 92037, USA

Single particle cryo-electron microscopy (cryoEM) is an important component of a structural biologist's toolkit, as improvements in instrumentation, software, automation, and specimen preparation are making this technique increasingly powerful for the analysis of large (>100 kDa) macromolecules and macromolecular complexes [1]. A typical single-particle dataset consists of many individual noisy images (often thousands or tens of thousands, and referred to as "particles"), each of which is a characteristic view of a unique macromolecular assembly. One of the biggest advantages of the methodology is the ability to analyze heterogeneous macromolecular assemblies, i.e. those that exhibit either conformational mobility within distinct parts or compositional heterogeneity exhibited by loosely and sub-stoichiometrically associated components. By employing classification techniques, it is possible to place each individual particle into one of several, potentially many, groups, according to homogeneity. Although different approaches have been proposed in the past to address specimen heterogeneity through classification (reviewed in [2], also see [3]), this problem nevertheless remains challenging, and is one of the major areas of methods development [1].

Here we describe a maximum likelihood-based method for the analysis of macromolecular heterogeneity in single particle cryoEM, which is implemented within the Frealign package [4]. Particle alignment parameters are determined by maximizing a joint likelihood that can include hierarchical priors, while classification is performed by expectation maximization of a marginal likelihood. We show that this algorithm has multiple advantages over the methods previously described, specifically:

1. The implementation within the Frealign package for refinement ensures that over-fitting is minimized through the application of a weighted target function and a resolution-limited approach to raw particle refinement and classification.
2. The refinement of particle orientations can be readily separated from the refinement of classification parameters. We show that this leads to an improvement in the accuracy of classification and additionally speeds up the classification.
3. Focused classification can be readily performed by applying a mask onto a region of interest in the dataset. This enables one to specifically analyze mobility and/or compositional heterogeneity within a defined region of interest, while ignoring heterogeneity in the rest of the density.

We provide a step-by-step protocol for implementing the algorithm with cryoEM data. We also provide experimental results with synthetic data – wherein the correct classification parameters are known – showing that separating the refinement of particle orientations from the refinement of classification parameters directly benefits and improves the accuracy of classification. We then apply the algorithm to the analysis of the Large Ribosomal Subunit-Associated Quality Control Complex (RQC), a macromolecular assembly consisting of the 60S large ribosomal subunit and multiple loosely bound components. This dataset exhibits a profound amount of heterogeneity, both compositional and conformational in nature. First, we show that it is possible to recover <3Å information for the ribosomal core (Figure 1A-C). Subsequently, we show that, using a global classification approach, it is possible to

recover several distinct conformational and compositional states of the large mobile components, including a sub-nanometer resolution reconstruction of the non-ribosomal proteins that are loosely associated to the periphery. Finally, we zoom in on the remaining areas of heterogeneity using a focused classification approach and show that it is possible to deconvolute the mobility of a 180 kDa protein bound to the large 60S ribosomal subunit, which was not possible using a global classification approach. Moreover, we show that small regions of mobility, corresponding only to a single RNA helix, can also be deconvoluted such that the RNA pitch becomes clearly evident (Figure 1D-G). Taken together, this shows that it is possible to recover structural information lost to mobility even for small regions.

References:

[1] X.-C. Bai, G. McMullan, S. H. W. Scheres, How cryo-EM is revolutionizing structural biology. *Trends Biochem. Sci.* **40**, (2015), p. 49–57.
[2] C. M. Spahn, P. A. Penczek, Exploring conformational modes of macromolecular assemblies by multiparticle cryo-EM. *Curr. Opin. Struct. Biol.* **19** (2009), p. 623–631.
[3] S. H. Scheres, A Bayesian View on Cryo-EM Structure Determination. *J. Mol. Biol.* **415**, (2012), p. 406–418.
[4] D. Lyumkis, A. F. Brilot, D. L. Theobald, N. Grigorieff, Likelihood-based classification of cryo-EM images using FREALIGN. *J. Struct. Biol.* **183**, (2013), p. 377–388.

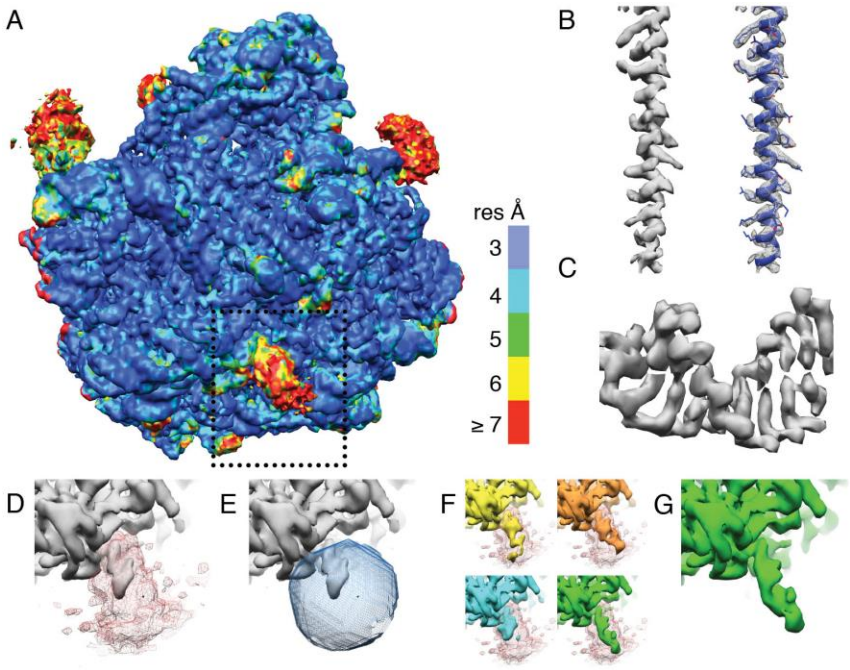

Figure 1. Analysis of Experimental RQC data using Frealign. (A) Reconstruction of the complete RQC dataset, with the 60S ribosomal subunit colored by local resolution. The dotted regions refers to panels D-G. (B,C) Density shots of ribosomal components at 3 Å resolution, including (B) an alpha-helical region with clearly resolved side-chains, and (C) an RNA helix with clearly resolved base pairs. (D) Reconstruction of the data without classification within a mobile ribosomal RNA segment. The same reconstruction, but at a lower threshold, is shown in red to indicate the extent of mobility within the region. (E) A small spherical mask used for classification is placed around the region corresponding to the mobile segment. (F) Four resulting classes from focused classification, showing different levels of mobility. (G) The best 3-D class with clear helical density for the mobile RNA segment.