Grouping 3D Structure Conformations using Network Analysis on 2D Cryo-EM Projection Images

Szu-Chi Chung[1*], Hung-Yi Wu[1], Wei-Hau Chang[2] and I-Ping Tu[1]

[1] Institute of Statistical Science, Academia Sinica, Taiwan.

[2] Institute of Chemistry, Academia Sinica, Taiwan.

*** Corresponding author: steve2003121@gmail.com**

KEY WORDS: Cryogenic Electron microscopy, heterogeneity problem, common line distance, community detection.

Cryogenic Electron microscopy (cryo-EM) is one of the most promising instruments for determining the structures of macromolecular protein complexes in near-atomic resolution. Nevertheless, there are still open challenges unsolved and here we addressed the heterogeneity problem inherent in cryo-EM data set. Originally, single particle analysis assumes the projection images come from the same molecule with the same structure conformation, but in fact, some molecules have various conformation states in the solution even after purification. Tradition approaches address this problem at 3D level. Thus, they require the information of 3D orientations and a consensus 3D structure before starting analyze the 3D variability which are computationally expensive. Therefore, there is a need to develop a new algorithm that preserves the accuracy while improving speed and scalability with the growing dataset size.

Classification algorithms based on cross-correlation often serve as the first step to boost the signal-to-noise ratio by averaging the particles within the same class in the literature. However, as we have observed in Figure 1, the cross-correlation carries little information about the heterogeneity within images. Here, a two-stage approach is proposed to address the heterogeneity problem while obviating the need for both averaging and 3D information. Firstly, we construct a k-nearest-neighbor graph based on the pairwise common-line distances among a small fraction of images and run community detection algorithms to partition the images into several communities. Secondly, we assign the remaining images to their nearest community based on common line distances. A novel criterion that measures the similarity between an image and a community based on 1D projections is also developed. We test this approach on two synthetic heterogeneous data sets, the first one is the benchmark dataset in heterogeneity problem [1] and the other one is the first dataset containing multiple conformational states [2]. For both dataset we achieve an accuracy of over 99%, which makes a record among the approaches. Figure 1 shows our analysis results on the second dataset. It is clear that our graph sucessfully separates five conformations. This new avenue to classify 3D heterogeneity at 2D level would potentially solve the scalability issue due to the fast-growing data acquisition rate.

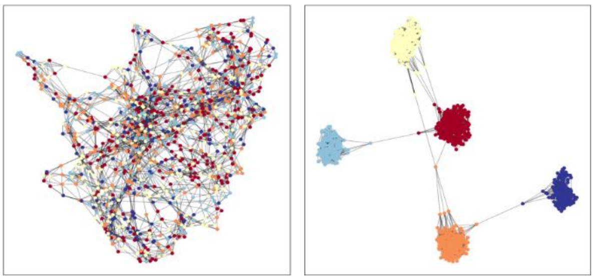

Figure 1: Left: The graph constructed based on cross-correlation. Right: The graph constructed using our approach. Different conformations are colored with different colors.

[1] Herman, Gabor T., and Miroslaw Kalinowski. "Classification of heterogeneous electron microscopic projections into homogeneous subsets." Ultramicroscopy 108.4, 327-338 (2008).

[2] Penczek, Pawel A., Marek Kimmel, and Christian MT Spahn. "Identifying conformational states of macromolecules by eigen-analysis of resampled cryo-EM images." Structure 19.11, 1582-1590 (2011).