Improving Unit-Cell Distance Algorithms for Clustering MX Images

Herbert J. Bernstein[1], Lawrence C. Andrews[2], Jean Jakoncic[3], Alexei Soares[3], Nicholas K. Sauter[4]

[1]Rochester Institute of Technology, c/o NSLS-II, Brookhaven National Laboratory, Upton, NY, USA
[2]Kirkland, WA, USA
[3]Brookhaven National Laboratory, Upton, NY, USA
[4]Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Unit-cell-based clustering is an established technique in serial crystallography, as performed both at xfels and at high brightness synchrotrons. In some cases the image-by-image distance between cells is the only metric used in defining clusters. In many cases, cell-based clustering is used to establish clusters of sufficient size to achieve reasonable completeness in each cluster, and then a transition is made to reflection-based clustering. In either case, it is critical to the success of the overall clustering process that the cell-based clustering be based on an accurate estimate of the distances among lattices and that the clustering technique used be able to recognize meaningful gaps among clusters. The most accurate and effective technique for such cell-based clustering in recent years has been the Niggli-reduction-based Andrews-Bernstein NCDist cell distance [Andrews, Bernstein 2014] used in [Zeldin *et al.* 2015], but the computational burden is high because of the complexity of the space. The Niggli Cone used in NCDist has two major components (+++ and ---), is not convex, and has 15 boundaries. Delaunay reduction produces the reduced cell in a single convex component, with only 9 boundaries, of which only 7 are well-populated, reducingthe number of cases to be considered by a ratio of 49 in 225. And then recasting to the earlier unordered Selling formalism of just the inner products simplifies the space further, leaving only 6 boundaries and reducing the original number of cases by a ratio of 36 in 225.

At the same time, transitioning from the currently popular Ward hierarchical clustering to the newer HDBSCAN hierarchical clustering provides further speed and accuracy improvements.

Work supported in part by NIH (GM117126 and P41GM111244), DOE (DE-AC02-98CH10886 and E-SC0012704), and Dectris.