Utilizing Scipion-ED for 3DED data processing

Laura C. Pacoste[1], Viktor E.G. Bengtsson[1], José Miguel de la Rosa-Trevín[2], Gerhard Hofer[1], Hongyi Xu[1], Xiaodong Zou[1]

[1]Stockholm Univeristy, Stockholm, Sweden;
[2]Department of Biochemistry and Biophysics, Science for Life Laboratory, Stockholm University;

laura.pacoste@mmk.su.se

Three-dimensional electron diffraction (3DED) techniques for structure determination has gained traction over the past few years (Gemmi *et al.*, 2019). Rapid development such as increasing acquisition speed and automated data collection leads to large amounts of data that needs to be processed. At the same time, gained interest and implementation of 3DED as a standard practice has increased the demand for straightforward processing tools that can be used by scientist at the novice level for the specific data processing methods. To face these challenges, an extension of Scipion (de la Rosa-Trevín *et al.*, 2016) for processing of 3DED data using DIALS has been developed under the name Scipion-ed (Bengtsson *et al.*, 2021). In this work, the usefulness of Scipion-ed for processing a large number of 3DED datasets has been demonstrated. A total of 52 datasets were collected on as-grown tetragonal lysozyme ($P\,4_3\,2_1\,2$) crystals through the continuous rotation electron diffraction method (cRED), also known as microcrystal electron diffraction (MicroED). Parallel workflows were generated in Scipion-ed for each dataset. The quality of each dataset was examined after scaling. Since the average completeness amongst all the datasets were 24 %, multiple datasets had to be merged to increase the completeness for the structure solution and refinement. Three different strategies were applied to find the appropriate datasets to merge. **Strategy 1** included scaling and merging of datasets with the most favourable overall merging statistics with regard to three formulated criteria (CC(1/2) > 0.8*, I/SigI > 2 and R_meas < 0.60). 14 of the processed datasets fulfilled all criteria and were scaled and merged accordingly. **Strategy 2** focused on maximizing the completeness of the final merged reflection file, without consideration of the reflection statistics of the higher resolution data. On top of the 14 previously selected datasets, additional datasets were added consecutively. Datasets that did not contribute to increased completeness were removed. **Strategy 3** included merging all datasets regardless of the contribution of each individual dataset to the completeness. The different merging strategies were evaluated with respect to the ability to resolve non-modelled features in the electrostatic potential map. This was done by refining the data against a modified model lacking the Trp28 residue. A previously solved X-ray diffraction model of the tetragonal lysozyme structure (PDB: 193L, Vaney *et al.*, 1996) was used as search model. After the refinement, the Trp28 residue added and real-space refined against the un-modelled electrostatic potential region representing the location of the residue. The final model (including Trp28) was validated against the electrostatic potential map that was refined against the modified model (without Trp28). Strategy 1 resulted in the highest correlation coefficient (CC) for the Trp28 residue (CC_{trp}= 0.974), along with the lowest R-value (R-work/R-free = 0.210/0.307) for the final structure model. Strategy 1 had the lowest completeness (74.5%) but the highest overall CC(1/2) (0.993*) and with a CC(1/2) > 0.330 down to 2.7 Å, compared to strategy 2 and 3 where the corresponding limits were 3.0 and 4.3 Å. At the same time, strategy 2 and 3 resulted in a higher overall completeness (85.8% and 87.2% respectively). Strategy 2 gave a slightly better CC of the Trp28 residue (CC_{trp} = 0.942) compared to merging all the datasets (Stratergy 3, CC_{trp} = 0.933). However, Strategy 3 resulted in a lower R-value (R-work/R-free = 0.224/0.303) compared to Strategy 2 (R-work/R-free = 0.232/0.340) indicating that the overall fit of the model was better. All the strategies resulted in very similar final structure models with regards to modelling of the Trp28 residue, indicating that the difference in the CC_{trp} is due to the differences in the electrostatic potential map. The results suggests that it is favourable to be more selective when merging datasets with considerations to the reflection statistics at higher resolution, despite limiting the completeness. The effect of different merging strategies should be investigated further to find the appropriate balance between completeness and resolution cut off but is beyond the scope of this study. We have demonstrated the usefulness of the Scipion-ed interface when investigating different strategies in parallel, as well as processing and merging large amounts of datasets, which is the standard procedure for collecting MicroED data of highly beam sensitive materials.

Keywords: 3DED;scipion-ed