A Consensus Method to Prepare Crystal Structures in the Protein Data Bank for Downstream Applications

A Vincelli[1], F Khatib[2]
[1]University of Massachusetts, Dartmouth, MA, [2]University of Massachusetts Dartmouth, Dartmouth, MA
ajvincelli@gmail.com

The vast majority of biomedical and drug discovery projects begin by accessing the Protein Data Bank (PDB). However, the qualities of the protein structures in the PDB (dating back to 1976) vary greatly, which may impact project duration and cost. We used six industry-standard quality metrics to create an aggregate "Quality Score" from 0% to 100%, and then scored all 120,365 protein-containing X-ray structures in the PDB. Though most structures scored well (67% of structures had a Quality Score >90%), we hypothesized that low Quality Scores may be improved to reduce the overall time and cost of projects utilizing these input proteins. We assessed the ability of Rosetta's Relax preparation algorithm to improve the Quality Scores of 500 unique structures, representing a diverse cross-section of the PDB, by Relaxing and validating each structure and then calculating its Quality Score and RMSD after Relaxation. We tested 21 variations of the Relax protocol on this dataset (n = 493), and identified a Pareto optimal Relax variant that increased all Quality Scores (from an average of 91.8 ± 3.6% to 99.4 ± 0.8%) with sub-Angstrom RMSDs from their starting structures (average Cα RMSD = 0.16 ± 0.61 Å). Future work includes incorporating feedback into our consensus method of preparing X-ray structural models, applying this method to all crystal structures in the PDB, and releasing the results and method to the public. Our Pareto optimal Relax protocol may be broadly used to prepare X-ray structures as high-quality inputs for protein engineering projects, and the prepared database may serve as a high-quality dataset for machine learning and other big data applications.