Integrated resource for reproducibility in macromolecular crystallography

UNIVERSITY OF VIRGINIA

PI: MINOR, WLADEK

Grant Number: 1 U01 HG008424-01

We propose the development of a collection of data wrangling tools to store, parse, manipulate, validate, curate, analyze, and disseminate macromolecular diffraction images together with all associated relevant metadata. The proposed system will have several benefits, by (1) creating a means to improve existing structures as technology for processing diffraction image advances, (2) detecting errors (and potentially, fraud) in existing structures to ensure structure quality and reproducibility, (3) preventing the loss of data collected by structural genomics and other programs that have closed or will close, (4) providing data for analysis of diffuse diffraction effects, and (5) buildng a ""training set"" for new diffraction analysis algorithms and hardware. Biologists, bioinformaticians, and software and hardware developers will all be beneficiaries of these tools. The proposed research is designed for semantic rather than syntactic analysis of diffraction images, and has several specific goals. First, we will develop tools for automatically extracting and curating diffraction images and associated metadata, as well as producing descriptions of all data needed for reprocessing when methods for structure determination improve. Second, we will create a web-based system for organizing, searching, analyzing, and data mining of appropriate subsets of diffraction images and associated metadata in machine-readable formats. This will include a comprehensive API for programmatic access, the ability to link multiple instances into a distributed federation, and state-of-the-art compression and transfer technologies. Third, we will develop tools to automatically validate, preprocess, and score diffraction images, and to detect potential issues and errors. These tools will make use of new and existing programs for image and data analysis, contain heuristics to identify possible errors, and provide statistics to correlate errors with specific metadata. Fourth, we will create a mechanism to discover diffraction data that have not yielded X-ray structures with currently available methods. Fifth, we will set up a pilot resource incorporating all the developed tools, and collect a test data set for the development of new tools for validation and error detection. We will work closely with multiple collaborators. Most important is the RCSB Protein Data Bank (PDB), who will help us ensure the accuracy and completeness of the diffraction metadata. Other partners will include the diffuse X-ray scattering community, detector vendors, synchrotron beamline managers, members of the IUCr Diffraction Data Deposition Working Group (DDDWG) and the crystallographic community in general. Together with the RCSB PDB, we will organize workshop(s) with these communities in order to (a) improve metadata extraction and (b) better define subsets of diffraction images. By addressing the currently common, irreversible and unnecessary loss of raw diffraction data during the data reduction process, our project helps ensure that the discipline of macromolecular crystallography is capable of continuous self-improvement. PUBLIC HEALTH RELEVANCE PUBLIC HEALTH RELEVANCE: X-ray crystallography is a tool with unprecedented power to investigate life at sub-microscopic levels, by revealing the 3-D structure of atoms and bonds in biological molecules, helping us identify and learn how to treat the

molecular causes of infection and disease. The ""raw"" data of X-ray crystallography experiments are binary files called ""diffraction images,"" and due to their size (several GB per data set), these data are often quickly reduced to a simplified numeric format and discarded. This throws away a great deal of useful information, and if done incorrectly, can introduce errors in structural results that cannot be corrected. To address this problem, we propose the development of a suite of software tools to effectively ""wrangle"" diffraction images, which will lead to detection of errors and (potentially) fraud, better processing algorithms, means to recover data from incomplete structural biology projects, and ultimately, better accuracy and reproducibility of 3-D structure information for the biomedical community.