LA-UR-12-23983

Title: E-Science and Protein Crystallography

Author(s): Miller, Laniece E.
Powell, James E. Jr.

Intended for: Report

Los Alamos
NATIONAL LABORATORY
EST. 1943

Disclaimer:
Los Alamos National Laboratory, an affirmative action/equal opportunity employer,is operated by the Los Alamos National Security, LLC for the National NuclearSecurity Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Departmentof Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

E-Science and Protein Crystallography at LANL

L. E. Miller, J. E. Powell

Dr. Zoë Fisher is the instrument scientist for the Protein Crystallography Station (PCS) at the Los Alamos Neutron Science Center's (LANSC) Lujan Neutron Scattering Center. She helps schedule researchers who intend to use the instrument to collect data, and provides in depth support for their activities. Users submit proposals for beam/instrument time via LANSCE proposal review system. In 2012, there were about 20 proposals submitted for this instrument. The instrument scientists review the proposals online. Accepted proposals are scheduled via an aggregate calendar which takes into account staff and resource availability, and the scientist is notified via email when their proposal is accepted and their requested time is scheduled. The entire PCS data acquisition and processing workflow is streamlined through various locally developed and commercial software packages. One 24 hour period produces one 200 Mb file, giving a total of maybe 2-5 Gb of data for the entire run. This data is then transferred to a hard disk in Dr. Fisher's office where she views the data with the customer and compresses the data to a text format which she sends them. This compression translates the data from an electron density to structural coordinates, which are the products submitted to a protein structure database.

As noted above, the raw experimental data is stored onsite at LANSCE on workstations maintained by the instrument scientist. It is extraordinarily rare for anyone to request this data, although the remote possibility of an audit by a funding organization motivates its limited preservation. The raw data is not rigorously backed up, but only stored on a single hard drive. Interestingly, only about 50% of the experimental data actually ends up deposited and described in peer reviewed publications; the data that is not published tends to either not be viable structures or is calibration data.

Dr. Fisher does protein crystallography research using both neutron and x-ray scattering techniques. Many of the major funders as well as the major journals dealing with protein crystallography require deposition of the structural data in the Protein Data Bank (PDB). Files formatted for the PDB are automatically generated when the data is compressed. The header files in the PDB included experimental conditions of the experiment as well as experimental methods. Depending on the completeness and how 'hot' of a topic, it may not be needed to contact the original experimenter about using the data. Having said that, not all of the data is accurate and does requires some back and forth with the creators of the data. The RCSB PDB staff at Rutgers University goes through all submissions and works with the submitters to verify that the data meets their minimum standards of completeness and robustness.

The Protein Data Bank (PDB) was initially created by Walter Hamilton at Brookhaven National Laboratory in 1971 after discussions about the value of scientists having access to structural biology data. Originally a partnership between Brookhaven and the Cambridge Crystallographic Data Center, the idea was conceived as a global initiative, which is certainly has become with partner sites in the US, Europe, and Japan. The PDB now contains structures determined from many different experimental techniques (Berman et al. 2012). Deposited structures are assigned a unique ID, and the structures are embargoed until the publication that references and describes them is published. The PDB staff often monitors these publications and takes the initiative to release protein structures when papers describing them are published.

Dr. Fisher records setup and experimental details in word documents and inserts printed copies into paper lab notebooks. These details appear in the final published papers and the header files for structures in the PDB. Analysis of data collected at the PCS is performed with a combination of locally developed tools and commercial products which are capable of outputting data suitable for importing into the PDB. While the original output data from the LANL instrument is stored indefinitely on a hard disk, the analysis results in a text file that, as described above, which represents the structure of the protein, which can be modeled and explored via tools that scientists in this domain have access to and are familiar with.

The entire process is well understood and well-supported by software used by researchers in this field. The incorporation of the PDB into research->analysis->publication is embraced by the international community of researchers in this field. There are mirror depository sites for the PDB in several countries. Curation of the submitted protein structures is rigorous, although Dr. Fisher noted that some structures are rushed to publication with what she termed “bogus filler”, which is possible since protein structures are 50-70% water. This is one of the things other scientists who look at deposited data will look for.

A case study on HIV protease was conducted on the data in the PDB. There was no further data collection beyond the existing published data. The main conclusion regarding the data was the need for some level of quality control before using the data, but that useful conclusions could be made strictly using PDB data (Venkatakrishnan et al. 2012).

Figure 1: A flowchart indicating the general workflow of data from the PCS to curation in the PDB.

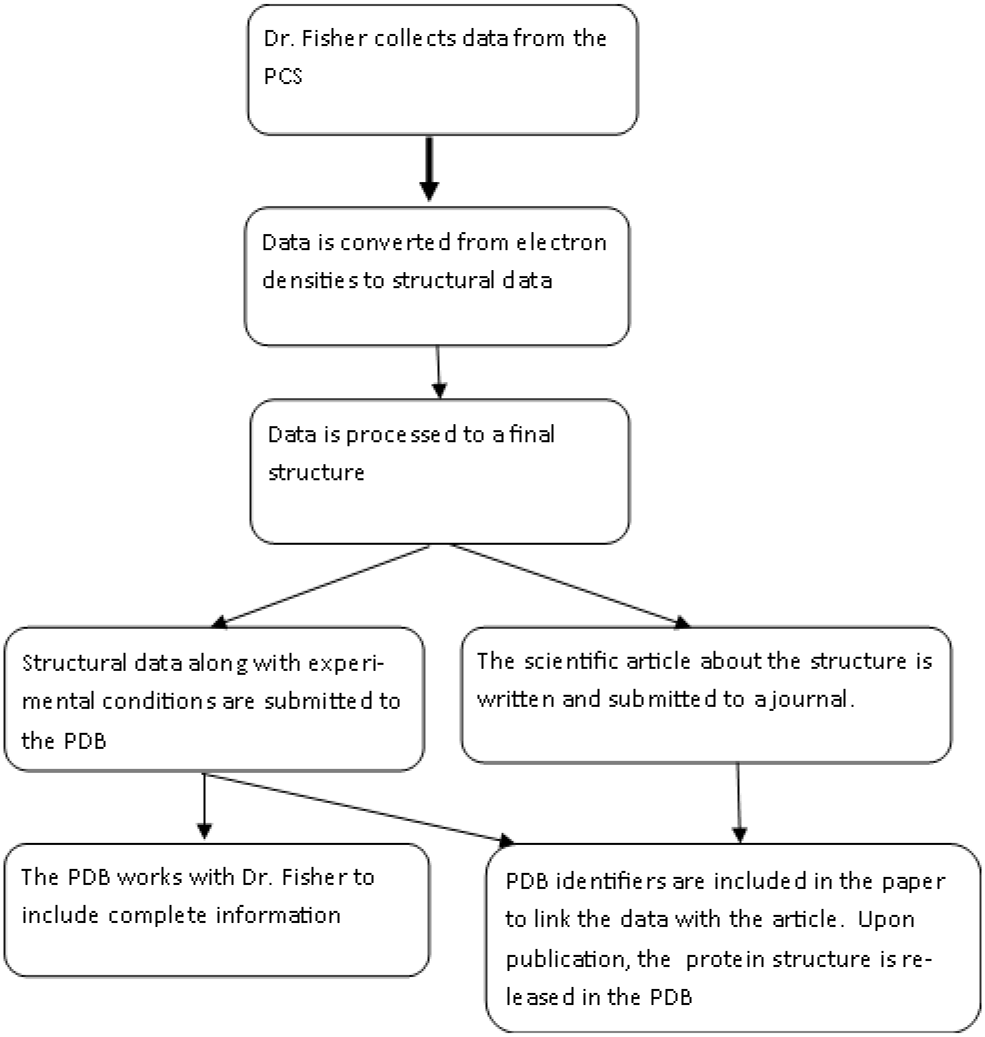
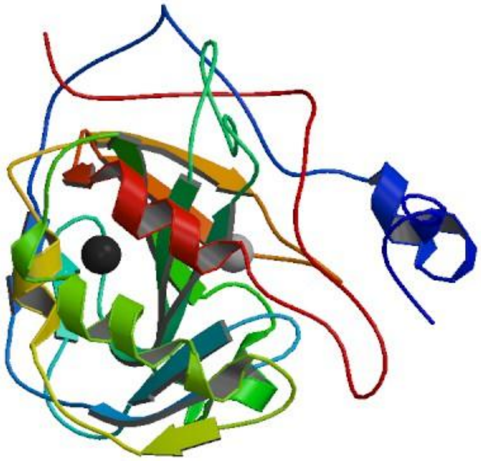

Dr. Fisher collects data from the PCS
Data is converted from electron densities to structural data
Data is processed to a final structure
Structural data along with experimental conditions are submitted to the PDB
The scientific article about the structure is written and submitted to a journal.
The PDB works with Dr. Fisher to include complete information
PDB identifiers are included in the paper to link the data with the article. Upon publication, the protein structure is released in the PDB

Figure 2: An excerpt of a PDB header, featuring journal "metadata"

```
HEADER    HYDROLASE (SERINE PROTEINASE)          13-APR-88   1SGT      1SGT   3
COMPND    TRYPSIN (/SGT$) (E.C.3.4.21.4)                               1SGT   4
SOURCE    (STREPTOMYCES $GRISEUS, STRAIN K1)                           1SGT   5
AUTHOR    R.J.READ,M.N.G.JAMES                                         1SGT   6
REVDAT   1   16-JUL-88 1SGT    0                                       1SGT   7
JRNL        AUTH   R.J.READ,M.N.G.JAMES                                1SGT   8
JRNL        TITL   REFINED CRYSTAL STRUCTURE OF STREPTOMYCES $GRISEUS  1SGT   9
JRNL        TITL 2 TRYPSIN AT 1.7 ANGSTROMS RESOLUTION                 1SGT  10
JRNL        REF    J.MOL.BIOL.                 V. 200   523 1988       1SGT  11
JRNL        REFN   ASTM JMOBAK  UK ISSN 0022-2836               070    1SGT  12
```

Figure 3: A link to an entry for a protein analyzed by the researcher we interviewed, along with a visual model of this protein

http://www.pdb.org/pdb/explore/explore.do?structureId=3v3f



We would like to acknowledge Dr. Zoë Fisher for taking the time out to show up their process and her workflow. Most of this write up is based off our conversation with her.

Berman, H. M., G. J. Kleywegt, H. Nakamura & J. L. Markley (2012) The Protein Data Bank at 40: Reflecting on the Past to Prepare for the Future. *Structure,* 20**,** 391-396.

Venkatakrishnan, B., M. L. Palii, M. Agbandje-McKenna & R. McKenna (2012) Mining the Protein Data Bank to Differentiate Error from Structural Variation in Clustered Static Structures: An Examination of HIV Protease. *Viruses-Basel,* 4**,** 348-362.