IUCrJ
ISSN 2052-2525
BIOLOGY | MEDICINE

Sharing Big Data

Marek Grabowski and Wladek Minor*

Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA 22903, USA.
*Correspondence e-mail: wladek@iwonka.med.virginia.edu

Edited by S. S. Hasnain, University of Liverpool, England

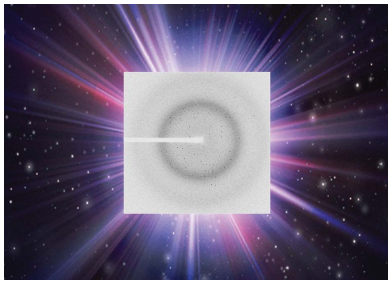
Keywords: Big Data; reproducibility; data sharing; metadata; open science.

Experimental reproducibility is the cornerstone of scientific research, upon which all progress rests. However, recent systematic surveys have revealed that a large fraction of representative sets of studies published in biomedical journals cannot be reproduced in another laboratory. This increased focus on reproducibility has likely contributed to the growing rate of retractions among scientific publications (Cokol *et al.*, 2008).

In contrast to many other areas of biomedical research, macromolecular crystallography has always been at the forefront of 'reproducible research' and 'open science', long before these approaches became widely appreciated and practiced. From the outset, crystallographers have embraced two fundamental tenets regarding crystallographic data: the preservation of relevant 'data' and making the data publicly available. Initially, the relevant data – the 'D' in the PDB (Protein Data Bank) – was limited to atomic coordinates, and was later supplemented by the 'header' containing the metadata describing the parameters of data collection (Berman *et al.*, 2000). Since the 1980s, deposition of structure coordinates into the PDB has been a requirement for the publication of a structure in scientific journals. As of 2006, structural deposits include structure factors, which permit recalculation of electron density maps. Yet the primary, 'raw' data of macromolecular crystallography, the sets of X-ray diffraction images used to derive the structure-factor files and atomic coordinates, typically have not been preserved, or if they have been preserved, have not been publicly available. In some cases, these data have been retained in 'data silos' reportedly kept by synchrotron facilities, individual crystallographers, or pharma companies. It is the experience of the authors of this commentary that only a very small fraction of requests for original diffraction images sent directly to authors of structures resulted in access to the data.

Traditionally, several factors have been regarded as very difficult challenges for creation of public repositories of diffraction data: (1) the sheer size of the data, which is 2–3 orders of magnitude greater than structure factors, (2) difficulties in organizing, acquiring, curating and managing the associated metadata, and (3) the deep-rooted tendency of researchers to keep their data private. Progress in storage technologies has almost eliminated the first challenge – the cost of hardware required to accommodate diffraction data has dropped significantly. The other two challenges still remain formidable, as discussed in the paper by Kroon-Batenburg, Helliwell, McMahon and Terwilliger in this issue of **IUCrJ** (Kroon-Batenburg *et al.*, 2017). The authors are long-term advocates of raw diffraction data preservation and are founding members of the Diffraction Data Working Group (DDWG). The paper presents an overview of several initiatives that have emerged in recent years to create public repositories of diffraction data and the diverse ways in which they approach these challenges.

The initiatives surveyed in the paper include dedicated diffraction data repositories such as the SBGrid (Meyer *et al.*, 2016) and the IRRMC (Grabowski *et al.*, 2016); institutional repositories such as the one run by the University of Manchester (Tanley, Schreurs, Kroon-Batenburg & Helliwell, 2012); general-purpose repositories for scientific data such as Zenodo and Research Gate; and synchrotron repositories such as the Synchrotron.Store (Meyer *et al.*, 2014). Between themselves, these resources now contain about 6600 publicly available diffraction datasets, corresponding to nearly 3600 diffraction experiments which have resulted in a PDB deposition. In addition, there are likely thousands more datasets sequestered in 'dark data', non-public data silos at synchrotron facilities or big pharma. Altering the traditional reluctance of researchers to share their data may best be addressed by funding-agency mandates stipulating that all data supporting publicly funded publications should be made publicly available. The two largest public repositories (Grabowski *et al.*, 2016, Meyer *et al.*, 2016) have used a variety of incentives to attract submissions from the community, amassing significant numbers of diffraction experiments (3118 and 253, respectively).



All this data would be, however, largely unusable without essential metadata, such as the identity of the sample and data collection parameters, which are necessary for optimal data reduction and structure determination. The latter is usually recorded in the headers of diffraction images – in one of the more than 200 formats defined by manufacturers of detectors and some synchrotron beamlines. The problem is that the metadata in the headers are sometimes missing, inconsistent, or just plain wrong. Herein lies the 'metadata challenge', common to many branches of science. A number of ongoing efforts are directed toward creating scientific ontologies and standardizing metadata (Ashburner *et al.*, 2000; Musen *et al.*, 2015). As the paper explains, structural biology has been in the vanguard of such efforts. The Crystallographic Information Framework (CIF) defined in the 1990s has created an ontology for structural biology (including the imgCIF dictionary for data collection) and defined a 'holistic metadata framework for crystallography'. Unfortunately, as the authors note, perhaps with a bit of understatement, 'not all real-world workflows use CIF as their actual mechanism for capturing data and metadata'.

The ultimate goal of an ideal repository is to provide a full description of diffraction experiment in the form that would allow others to easily perform data reduction and structure re-determination for every new deposit in the PDB. A re-examination of raw data and/or structure factors may bring a significant improvement in structure quality, as shown by the case of cisplatin-protein complexes discussed in the paper. In particular, a re-processing of the original diffraction data for cisplatin bound to hen egg lysozyme (Tanley, Schreurs, Kroon-Batenburg, Meredith *et al.*, 2012) accessible at the University of Manchester repository, and the subsequent re-refinement by a different group, improved the resolution of the crystal structure from 2.4 to 2.0 Å (Shabalin *et al.*, 2015) and a subsequent follow-up by the original group was able to improve resolution further to 1.7 Å (Tanley *et al.*, 2016). The 'data debate' which ensued in this case provides a nice illustration of the benefits of sharing data and trying different interpretations, even though, as the authors emphasize there is currently a lack of uniform community standards as to what is 'the best' interpretation. In this particular case, data sharing gave the same apparent improvement as could have been gained by performing new data collection on a hypothetical new powerful detector on a super-modern beamline and on a new generation synchrotron.

An often-overlooked Achilles heel of most biomedical databases and repositories is that negative results are often very well hidden or non-existent. The diffraction experiment that results in one protein–inhibitor complex is often the outcome of hundreds of diffraction data sets. The IRRMC has a mechanism that allows the deposition of data sets that did not bring expected results (for example an empty active site). For obvious reasons, such datasets not only have to be associated with a detailed description of the protein, but must also include which compounds were introduced and *via* which method. The negative information from screening results could identify 'blind alleys' and significantly speed-up drug discovery.

None of the current diffraction data repositories provides much, if any, information about the macromolecular sample. Ideally, one would like to have structural data integrated with data from expression, purification, and crystallization – as well as information about biomedical experiments. Creation of such integrated resources remains a dream that may be a major goal of BD2K type of program.

Acknowledgements

We would like to thank David R. Cooper for critical reading and discussion. The authors acknowledge support from NIH BD2K grant HG008424.

References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. & Sherlock, G. (2000). *Nat. Genet.* **25**, 25–29.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Cokol, M., Ozbay, F. & Rodriguez-Esteban, R. (2008). *EMBO Rep.* **9**, 2.

Grabowski, M., Langner, K. M., Cymborowski, M., Porebski, P. J., Sroka, P., Zheng, H., Cooper, D. R., Zimmerman, M. D., Elsliger, M.-A., Burley, S. K. & Minor, W. (2016). *Acta Cryst.* D**72**, 1181–1193.

Kroon-Batenburg, L. M. J., Helliwell, J. R., McMahon, B. & Terwilliger, T. C. (2017). *IUCrJ*, **4**, 87–99.

Meyer, G. R., Aragão, D., Mudie, N. J., Caradoc-Davies, T. T., McGowan, S., Bertling, P. J., Groenewegen, D., Quenette, S. M., Bond, C. S., Buckle, A. M. & Androulakis, S. (2014). *Acta Cryst.* D**70**, 2510–2519.

Meyer, P. A., Socias, S., Key, J., Ransey, E., Tjon, E. C., Buschiazzo, A., Lei, M., Botka, C., Withrow, J., Neau, D., Rajashankar, K., Anderson, K. S., Baxter, R. H., Blacklow, S. C., Boggon, T. J., Bonvin, A. M., Borek, D., Brett, T. J., Caflisch, A., Chang, C. I., Chazin, W. J., Corbett, K. D., Cosgrove, M. S., Crosson, S., Dhe-Paganon, S., Di Cera, E., Drennan, C. L., Eck, M. J., Eichman, B. F., Fan, Q. R., Ferré-D'Amaré, A. R., Christopher Fromme, J., Garcia, K. C., Gaudet, R., Gong, P., Harrison, S. C., Heldwein, E. E., Jia, Z., Keenan, R. J., Kruse, A. C., Kvansakul, M., McLellan, J. S., Modis, Y., Nam, Y., Otwinowski, Z., Pai, E. F., Pereira, P. J., Petosa, C., Raman, C. S., Rapoport, T. A., Roll-Mecak, A., Rosen, M. K., Rudenko, G., Schlessinger, J., Schwartz, T. U., Shamoo, Y., Sondermann, H., Tao, Y. J., Tolia, N. H., Tsodikov, O. V., Westover, K. D., Wu, H., Foster, I., Fraser, J. S., Maia, F. R., Gonen, T., Kirchhausen, T., Diederichs, K., Crosas, M. & Sliz, P. (2016). *Nat. Commun.* **7**, 10882.

Musen, M. A., Bean, C. A., Cheung, K. H., Dumontier, M., Durante, K. A., Gevaert, O., Gonzalez-Beltran, A., Khatri, P., Kleinstein, S. H., O'Connor, M. J., Pouliot, Y., Rocca-Serra, P., Sansone, S. A., Wiser, J. A. & team, C. (2015). *J. Am. Med. Inform. Assoc.* **22**, 1148–1152.

Prinz, F., Schlange, T. & Asadullah, K. (2011). *Nat. Rev. Drug Discov.* **10**, 712.

Shabalin, I., Dauter, Z., Jaskolski, M., Minor, W. & Wlodawer, A. (2015). *Acta Cryst.* D**71**, 1965–1979.

Tanley, S. W. M., Schreurs, A. M. M., Kroon-Batenburg, L. M. J. & Helliwell, J. R. (2012). *Acta Cryst.* F**68**, 1300–1306.

Tanley, S. W. M., Schreurs, A. M. M., Kroon-Batenburg, L. M. J. & Helliwell, J. R. (2016). *Acta Cryst.* F**72**, 253–254.

Tanley, S. W. M., Schreurs, A. M. M., Kroon-Batenburg, L. M. J., Meredith, J., Prendergast, R., Walsh, D., Bryant, P., Levy, C. & Helliwell, J. R. (2012). *Acta Cryst.* D**68**, 601–612.