Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Deposition of structural data redux

Alexander Wlodawer

Macromolecular Crystallography Laboratory, National Cancer Institute at Frederick, Frederick, MD 21702, USA

Correspondence e-mail: wlodawer@ncifcrf.gov

Received 10 January 2007
Accepted 11 January 2007

© 2007 International Union of Crystallography
Printed in Denmark – all rights reserved

It has been almost exactly ten years since the publication of a 'revolutionary' suggestion that macromolecular coordinates resulting from crystallographic and NMR studies should be deposited in the Protein Data Bank and released immediately upon publication of the relevant papers (Wlodawer, 1997). That suggestion was directed primarily to the journals, but its targets were also authors, funding agencies, and the International Union of Crystallography (IUCr). After some further discussion (Wlodawer *et al.*, 1998), the previous policy that allowed long hold periods for both the coordinates and structure factors (IUCr Commission on Biological Macromolecules, 1989) was modified in a substantial way. IUCr implemented a new policy that disallows coordinate holds beyond the date of publication, while allowing six-month hold on the structure factors (IUCr Commission on Biological Macromolecules, 2000). That policy is mandatory for the IUCr journals and provides a template for policies of almost all other journals that publish macromolecular structures. The two major organizations that fund structural biology in the United States, namely NIH and HHMI, also implemented policies that demanded deposition and release upon publication of coordinates resulting from all research that was supported by them (the NIH policy can be found at http://grants2.nih.gov/grants/guide/notice-files/not99-010.html). Although some of my very distinguished colleagues that should remain nameless have warned me that the requirement of deposition of coordinates 'will destroy structural biology as we know it', these dire predictions did not materialize and prompt deposition of the coordinates is no longer a matter of controversy, but rather completely routine.

However, the question of deposition of supporting data, such as crystallographic structure factors or NMR restraints, has not been addressed directly in the NIH document, other than by reference to the IUCr policy on that matter. In addition, it is not entirely clear what enforcement mechanisms should be put in place, and who should police the depositions. It is quite clear that the journals should be taking a leading role in the enforcement, since they control acceptance of the papers describing macromolecular structures, and thus carry a big stick. Their record of enforcement is, unfortunately, rather patchy. Another important player is of course the Protein Data Bank (Berman *et al.*, 2000), but, as the name indicates, a bank has to take all deposits, and also cannot force its customers to make them. The funding agencies are unlikely to pursue routinely the violators of their policy, since they cannot be expected to track individual depositions of structural data.

Is the deposition of structure factors important? From my own experience I must state unequivocally that the availability of such data is absolutely crucial, sometimes even for the authors. When the first set of protein coordinates that were refined at truly atomic resolution of 1 Å (Wlodawer *et al.*, 1984) was deposited in the PDB (accession code 5pti), we neglected to include the structure factors. Since the structure of BPTI was used in the late 1980s to develop a number of important experimental and theoretical approaches to structural biology, we received a number of requests for these primary data. Unfortunately, to our horror we soon realised that none of the authors kept a copy, so these unique data are now permanently lost. I wish we had them deposited in the PDB.

However, the availability of structure factors is even more important if there could be any doubts about the correctness of the reported structures. A recent fiasco related to the inverted structure of the proteins MsbA and EmrE (Miller, 2006) provides a very good example. A number of coordinate data sets deposited in the PDB (1jsq, 1pf4, 1z2r, 1s7b, and 2f2m) were not accompanied by structure factors, so other scientists could not perform any verification of the models that turned out to be seriously wrong. It is quite possible that data processing errors that led to these erroneous structures could have been discovered much sooner than five years after the first one of them was published (Chang & Roth, 2001). About 20% of coordinate sets deposited in the PDB since the beginning of 2000 are either not accompanied by structure factors at all, or the latter are still on hold. I do not consider that situation to be acceptable.

Is there still any question on whether deposition of structure factors might give an unfair advantage to scientists from major research centers who might jump into the fray and prevent the authors of the original work from claiming their right rewards? I do not think so. First, whether enforced or not, the policy of depositing the structure factors (with a possible six-month hold) is already on the books. Second, when a paper is already published, it is unlikely that the availability of structure factors could really hurt the competitiveness of the authors, unless there was a problem that could be detected by others. But that is exactly why these data should be generally available!

I would thus like to make a number of suggestions directed to authors, journals, funding bodies, the IUCr and the PDB. First, I feel that the PDB should not accept deposition of coordinates not accompanied by structure factors. If they do, then they are abetting violation of policies that are already binding, since there seems to be a general agreement that structure factors must be made available, with the discussion concerning only the timing of their release. Second, the journals should be much more vigilant in enforcing the rules regarding deposition of structural data. The authors should no longer be allowed to provide PDB accession codes only after a paper has been accepted, but should include them when a manuscript is submitted. Of course, the coordinates and structure factors would not yet have to be released at that stage, but the current approach often leads to delays of several weeks between publication of a paper (these days, often a pre-publication on the web) and the actual availability of the coordinates (to say nothing about structure factors that might not be accessible for much longer). As far as I understand, it is possible to request that the PDB does not publish the title of the submission before actual release of data (many records in the current directory of unreleased entries have the title N/A), so such deposition would not even have to alert competition to the existence of 'hot' structures.

However, I strongly urge the IUCr to reconsider and revise its current policy that allows six-month hold of the structure factors, and instead to treat them exactly in the same way as the coordinates. I am convinced that the current policy regarding the hold has outlived its usefulness, and I would like to initiate discussion about its possible amendment. The Union has an important role to play here, since, as mentioned above, its recommendations are generally followed by the funding agencies and the journals. However, I would also urge the latter to introduce such changes immediately, even without waiting for a policy change on the part of the IUCr. The journals, in particular, should become serious about at least enforcing the policies that already exist.

The PDB should become much more active in assuring the scientific community that the coordinates and structure factors in their repository are accurate, properly annotated and fully cross-referenced to their respective publications. I could write pages about that subject, but let me give just a few examples. At this time the PDB seems not to flag structures that should have clearly raised some serious questions. For example, the structure of the eIF4A helicase (PDB code 2g9n) has a number of peptide torsion angles deviating from 0 or 180° by as much as 50°, and at least one peptide is *cis* in one molecule in the asymmetric unit, and *trans* in the other. The structure of *Vibrio cholerae* L-asparaginase (PDB code 2hyk) has at least a PDB-provided caveat indicating a chirality error, although other obvious errors should also have been flagged during the process of validation. In addition, the PDB does not seem to require from the depositors that they provide all the relevant experimental details, and does not check if the numerology is significant. An example of the former problem is provided by the recent deposition of the coordinates of multidrug transporter SAV1866 (PDB code 2hyd), in which such important information as the software used in data processing and structure solution is missing. This is particularly ironic since that structure was the basis of discovery that the structures of MsbA and EmrE were wrong, apparently becasue of a software glitch (Miller, 2006). And what is the meaning of the r.m.s. deviation of bond lengths of 0.004631 Å (by my count, about 1/10th of the size of an electron)? Such incredible precision (and exactly the same number!) is claimed by two related coordinate sets, 1gmu and 1gmw. I do not think that prevention of any of the problems outlined above would put any undue burden on the overworked PDB. Numerology and the lack of data can be automatically detected and/or corrected or flagged by validation software that is run in any case. Even more routine insertion of a 'CAVEAT' record into the coordinates can be automated, as long as there would be an agreement of what makes a coordinate set suspicious. Let us remember that many of the current depositors might be proficient in running crystallographic programs, but have little or no understanding of crystallography as such. Even more important is the fact that major consumers of the coordinates are not structural biologists, but experts in other fields. Thus, many (or maybe most) users of the coordinates might not have a clue of what a torsion angle ω of 90° means, especially in a structure refined at 3 Å resolution, to give a ridiculous example that should be very clear to all macromolecular crystallographers.

Finally, I would like to raise a related point, namely access of reviewers of crystallographic and NMR papers to the primary data. I am aware of suggestions that both coordinates and structure factors should be included in material provided to the reviewers, but I do not think that this is feasible, since it would raise a number of questions regarding confidentiality of data before publication. However, if the suggestion that provision of PDB codes becomes mandatory in submitted papers becomes accepted, then it might be sufficient to require that the PDB validation report be attached to each paper as supplementary material, only for the use of reviewers. Such reports would not violate confidentiality of the data, but would still provide the reviewers with more information necessary to evaluate the quality of structures. It would not be that difficult for the PDB to generate a 'report for reviewers' as an extra file in the validation process. Such a file should contain the complete header (so that the reviewer might be able to see all the NULL records!), as well as the outliers. Since these data are already generated in an automated way, this suggestion would not put any new burden on the PDB. However, I am certain that giving the reviewers better tools would enhance the credibility of the field of structural biology.

To summarize, much has changed during the last ten years and the policies that were adequate then seem no longer to benefit the scientific community in the same way. Thus, time has probably come to take another look at them, and modify them accordingly. The IUCr

is the right organization to initiate such a change, but the journals and funding agencies might wish to act even sooner. I hope that my proposal, controversial as it might appear to some (as was the previous one ten years ago), could be a basis of starting a thorough discussion of this important matter.

Disclaimer. Although the suggestions contained here have been extensively discussed with several of my colleagues, all the blame for their publication rests solely with the author. In particular, the opinions expressed here do not reflect in any way the policies of the National Cancer Institute and/or the National Institutes of Health.

References

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.
Chang, G. & Roth, C. B. (2001). *Science* **293**, 1793–1800.
IUCr Commission on Biological Macromolecules (2000). *Acta Cryst.* D**56**, 2.
IUCr Commission on Biological Macromolecules (1989). *Acta Cryst.* A**45**, 658.
Miller, G. (2006). *Science*, **314**, 1856–1857.
Wlodawer, A. (1997). *Nature Struct. Biol.* **4**, 173–174.
Wlodawer, A., Davies, D., Petsko, G., Rossmann, M., Olson, A. & Sussman, J. L. (1998). *Science*, **279**, 306–307.
Wlodawer, A., Walter, J., Huber, R. & Sjolin, L. (1984). *J. Mol. Biol.* **180**, 301–329.