J Comput Aided Mol Des (2012) 26:165–167
DOI 10.1007/s10822-012-9555-0

WARR'S PIECE

An interview with Phil Bourne, associate director of the RCSB protein data bank

W. A. Warr

Received: 7 February 2012 / Accepted: 16 February 2012 / Published online: 25 February 2012

Philip E. Bourne PhD [PEB] is a Professor in the Department of Pharmacology and Skaggs School of Pharmacy and Pharmaceutical Sciences at the University of California San Diego, Associate Director of the RCSB Protein Data Bank and an Adjunct Professor at the Sanford Burnham Institute. He received his Ph.D. in chemistry from the Flinders University of South Australia in 1980. In the early 1980s he was a postdoctoral fellow in structural biology, first at the University of Sheffield, UK and later at Columbia University, New York. During the late 1980s as first the Director of the Columbia University Cancer Center Computer Facility and later as Director of the Medical School Computer Facility he worked in computational biology and medical informatics. In the early 1990s he joined the Howard Hughes Medical Institute and worked on developing high performance hardware and software for computational structural biology. He moved to the University of California San Diego in 1995 to work on structural bioinformatics. He is also developing new methods for early stage drug discovery, scientific visualization and scholarly communication. He is the co-founder and Editor-in-Chief of the open access journal *PLoS Computational Biology*.

W. A. Warr (✉)
Wendy Warr & Associates, Holmes Chapel Crewe, Cheshire CW4 7HZ, UK
e-mail: wendy@warr.com

Interview

WAW: I have done a series of interviews [1–3] with partners in wwPDB [4] and you are the latest. Tell me where you fit in.

PEB: The RCSB PDB [5] is led by Helen Berman at Rutgers University and my group here at the University of California at San Diego is on a subcontract. You can think of us as "data in, data out". The group at Rutgers does the "data in" part: they validate and annotate incoming structures and then release the processed data to San Diego for putting on the website and integrating with other data: "data out". Both groups collaborate on extensive educational outreach efforts. Many external resources are fully integrated [6]: literature sources such as PubMed, mappings to classifications such as Pfam, SCOP and CATH, ligand databases such as BindingDB, biological details such as the Gene Ontology, experimental data, and the genome resource Entrez Gene. The website and associated services have recently been redesigned [7, 8]. We have 275,000 unique users a month and seven structures downloaded every second. The users are not bound by geography and we try and cater for users at different experience levels: hence our "PDB 101" pages and PDBmobile efforts. Features such as Molecule of the Month created by David Goodsell at Scripps appeal right down to secondary school level students using smart phones. Ease of use is of paramount importance in our basic services.

WAW: Gerard Kleywegt [2] told me that collaboration on "data in" is crucial to ensure a single, uniform global archive, while serving the data back to users by each of the three wwPDB sites independently allows for diversity.

PEB: Yes, the wwPDB sites [4, 9] collaborate: we all accord to one set of data processing standards, for example. That standard depends on mmCIF [6]. Until recently mmCIF has not been widely used in the community. Users prefer the simpler PDB format. However, as structural biology has advanced and structures have got larger and more complex the PDB format no longer works well. For example the 99,999-atom limit as a result of fixed field atom numbering means that big structures must be accommodated by more than one PDB file. There are also important data management features of mmCIF, like the fact that it is self-describing, that I will not go into further here [10–13]. My personal belief is that mmCIF (or its XML equivalent) will become analogous to HTML in the Web world; you rarely look at HTML directly, but use a tool such as a Web browser to use the format effectively. The new wwPDB submission system planned for introduction in the fourth quarter of 2012 uses the mmCIF-based standard PDBx.

WAW: And what about "competition"?

PEB: If there is any competition it is in "data out", but even then each site, RCSB PDB [5], PDBe [14] or PDBj [15] tends to specialize, catering for somewhat different audiences. Nevertheless, we share information and tools to insure the highest possible data quality, regardless of where the user goes to get the data. Recently there has been a task force on validation [16] and the wwPDB partners will implement the recommendations of that task force. The notion of competition comes in part from the fact that the three PDB sites are funded by agencies in their respective geographic regions. PDB's funding is not like UniProt's [17]. UniProt is funded by the U. S. National Institutes of Health, the U. S. National Science Foundation, the European Commission and the Swiss federal government but every PDB site is funded by different agencies, so it is goodwill by the partners that makes the wwPDB work.

WAW: One of your other major commitments is to open access publishing and open data. You also founded SciVee [18] to distribute scientific videos.

PEB: Publishing has to move beyond the PDF and the current simple use of HTML to represent journal pages to fully utilize the interactive and social power of the Internet in achieving maximum accessibility, reproducibility and comprehension of science. Open access to the literature, open software and open data are a key part of this change. Developments like Mendeley [19] and Utopia [20] which augment the PDF are a wonderful indicator of what is to come. The Force11 virtual community [21, 22] was founded recently to help this transformation of scholarly communications through advanced use of computers and the Web, and I am proud to be part of it. A recent example is what *PLoS Computational Biology* is doing with Wikipedia. Computational biology is poorly covered in Wikipedia, but *PLoS* now has a media wiki site where people can publish topic pages in the education section of the journal where they become the copy of record; these pages are then released to Wikipedia to become a living version of the same content where it will hopefully be expanded by the community as the science advances.

WAW: What about open data and software specifically?

PEB: PDB data have always been open. NSF and NIH now have data sharing policies [23, 24] which is an important step in the evolution of open science. The wwPDB are assigning DOIs to data and *PLoS* is preparing to use Dryad [25] as a repository for supplemental data submitted with papers. Dryad is an international repository of data underlying peer-reviewed articles. It allows scientists to validate published findings, explore new analysis methodologies and reuse data.

It is common in journals to see the statement "the software is available from the authors". Often it is *not* available; this is a broken model. *PLoS Computational Biology* has a software section for which reviewers must be able to run the software and readers must be able to access the software through a shared repository. A publication I am now working on suggests that we should be publishing workflows. At present you cannot easily reverse engineer an article. In other words, the scientific pillar of reproducibility is a term which is used very loosely. I hope soon to start a new *PLoS* venture which will have data pages with metadata about data sets and scientists will be rewarded for submitting data sets, especially benchmark data sets. For open data to be more common we must reward those who submit data. The PDB, I believe does a good job at showcasing such data and I am proud to be part of that effort.

WAW: That brings us back full circle to the PDB which was the main topic of this interview. Thank you so much for sharing your full and frank opinions with me.

References

1. Warr WA (2008) J Comput-Aided Mol Des 22(10):707–710
2. Warr WA (2010) J Comput-Aided Mol Des 24(11):887–890
3. Warr WA (2011) J Comput-Aided Mol Des 25(9):791–793
4. Worldwide PDB http://www.wwpdb.org. Accessed 3 Feb 2012
5. RCSB PDB http://www.pdb.org. Accessed 3 Feb 2012
6. Bourne PE et al (2011) WIREs Comput Mol Sci 1:782–789
7. Rose PW et al (2011) Nucleic Acids Res 39:D392–D401
8. Rose PW et al. (2011) NCI-nature pathway interaction database http://pid.nci.nih.gov/archive/2011/04/primer.shtml. Accessed 2 Feb 2011

9. Berman HM, Henrick K, Nakamura H (2003) Nat Struct Biol 10:980
10. PDB file formats http://www.rcsb.org/pdb/static.do?p=file_formats/index.jsp. Accessed 2 Feb 2011
11. Bourne PE, Berman HM, Watenpaugh K, Westbrook JD, Fitzgerald PMD (1997) Meth Enzymol 277:571–590
12. Fitzgerald PMD, Westbrook JD, Bourne PE, McMahon B, Watenpaugh KD, Berman HM (2005) In: Hall SR, McMahon B (eds) International tables for crystallography. Springer, Dordrecht, pp 295–443
13. Westbrook JD, Fitzgerald PMD (2009) In: Bourne PE, Gu J (eds) Structural bioinformatics, 2nd edn. Wiley, New York, NY, Chapter 10, pp 271–291
14. Europe PDB http://www.ebi.ac.uk/PDBe/. Accessed 3 Feb 2012
15. Japan PDB http://www.pdbj.org/. Accessed 3 Feb 2012
16. Read RJ et al (2011) Structure 19:1395–1412
17. UniProt http://www.uniprot.org/help/about. Accessed 3 Feb 2012
18. SciVee http://www.scivee.tv/. Accessed 3 Feb 2012
19. Mendeley http://www.mendeley.com/. Accessed 3 Feb 2012
20. Utopia http://utopia.cs.man.ac.uk/. Accessed 3 Feb 2012
21. The Future of Research Communication and e-Scholarship. Force 11 http://www.force11.org/. Accessed 3 Feb 2012
22. Force11 white paper http://www.force11.org/white_paper. Accessed 3 Feb 2012
23. NSF data sharing policy http://www.nsf.gov/bfa/dias/policy/dmp.jsp. Accessed 3 Feb 2012
24. NIH data sharing policy http://grants.nih.gov/grants/policy/data_sharing/. Accessed 3 Feb 2012
25. Dryad http://ils.unc.edu/mrc/dryad_repository/. Accessed 3 Feb 2012