J Comput Aided Mol Des (2011) 25:791–793
DOI 10.1007/s10822-011-9460-y

WARR'S PIECE

Protein Data Bank Japan (PDBj): an interview with Haruki Nakamura of Osaka University

Wendy A. Warr

Received: 11 July 2011 / Accepted: 13 July 2011 / Published online: 28 July 2011

Prof. Haruki Nakamura (HN) has B.Sc. and Ph.D. degrees in physics from the University of Tokyo. After 7 years as a research associate at the University of Tokyo, he worked briefly at the University of Leeds, England, then in August 1987 became Research Director in the Protein Engineering Research Institute (PERI) at Osaka University. In April 1996, he became Research Director in the Department of Bioinformatics, Biomolecular Engineering Research Institute (BERI), the successor to PERI. Since April 1999 he has been a Professor in the Laboratory of Protein Informatics, Research Center for Structural Biology, Institute for Protein Research, at Osaka University.

Interview

WAW: Tell me something about the history of PDBj [1]. When was PDBj inaugurated? How did it come to be centered on Osaka and how is it funded?

W. A. Warr (✉)
Wendy Warr & Associates, 6 Berwick Court, Holmes Chapel, Cheshire CW4 7HZ, UK
e-mail: wendy@warr.com

HN: PDBj started operating in 2000, so we have managed PDBj for more than 11 years. From the very beginning of PDB at Brookhaven, Japanese crystallographers had made a center for protein crystallography in our Institute, the Institute for Protein Research (IPR), at Osaka University [2], where they had built a service to distribute the magnetic tapes of the PDB coordinates data to every University in Japan. This was before Internet data distribution started. So, IPR has always been the site to contact about PDB in Japan.

In 2000, IPR contacted RCSB PDB [3] and prepared to start curation and processing of the deposited data directly. In 2001, IPR got a 5 year grant from the Japanese government, and formally started PDBj. In 2003, we were co-founders of the wwPDB [4], after long discussions with Helen Berman in the US and Janet Thornton and Kim Henrick in the UK. In 2006, we renewed our PDBj project for another 5 years with a grant from the Japanese government, and this April, 2011, we got another grant, for 3 years, until March 2014.

WAW: Do you have confidence that the resource is secure in the long term? Do you think that wwPDB should be funded globally?

HN: In fact, the current grant comes from the new Japanese database institute, the National Bioscience Database Center (NBDC) [5], which has just recently been founded, in April 2011. The concept of NBDC has a long history. It has taken nearly 10 years to establish an integrated life science database center in Japan (like NBCI in the US or EBI in Europe). I was on the government committee that established NBDC. There have been very many arguments about the foundation of NBDC, because many life science researchers think of databases as an "air", which always exists around them, but they do not think about who will

have to make the effort to build databases or how they are managed and by which grants. We have run many campaigns using scientific journalism and symposia at annual meetings of molecular biology societies. Finally, NBDC has been founded, and in the future we hope that a more stable and cross-ministry database center will be established. The current NBDC is supported only by MEXT, the Ministry of Education, Culture, Sports, Science and Technology in Japan. The Ministries for Economics, Agriculture, and National Health also ought to be involved, but how to do the integration is an issue.

The essential problem of database management is the difficulty of making a business model for its sustainable management. The size of the database increases year by year, suggesting its management would require more and more budget. Nobody is happy with this situation, so the globally sustained system of the wwPDB is one good answer for a free and public database. Another way may be from a new system, where every data producer, including experimental and computational researchers, medical doctors, and even hospital patients, pays a very little amount, like a consumption tax, to each government, in order to construct and maintain the databases for the future of science.

WAW: What is the size and organizational structure of the group at PDBj?

There are 17 people in IPR at the moment: seven university staff and ten PDBj staff. Another three PDBj people will join very soon, now that the PDBj project has been renewed. In addition, we have another five people outside of IPR (professors at other universities), who collaborate with us in developing database services and tools.

WAW: How does PDBj work with other PDB collaborators? What are the synergies?

HN: One of the most essential collaborations is to curate, edit and process the deposited coordinates data. Since our annotators started work in 2000, they have processed a total of 17,186 PDB depositions, which is nearly a quarter of all the depositions (70,868 in all) [6]. We also have the PDBj Biomagnetic Resonance Bank (BMRB) group, who collect and process the chemical shift data deposited to PDBj-BMRB. Between 2006 and 2010, 738 chemical shift records have been processed: 21% of the entire BMRB data (3,536 chemical shifts).

The other important synergy is data validation, in particular for data description. Since PDBj started its PDB activity (even before the formal foundation of the wwPDB) we have focused on the remediation of the PDB data description, and we have worked on establishing PDBML, which is the canonical XML version of the PDB data, in collaboration with the RCSB PDB group [7]. Since then, our database viewer (unique to PDBj) has been based on the PDBML, and so validation of the data description is critical. We have always done validation, and sent the validation report to repair the original PDB data in all of the formats. The oceanic region version of our viewer is in English. Japanese pages are also available and they can be searched by Japanese language keywords. There are also some Korean pages, and Chinese pages, with both full and simplified character sets.

Frequent discussions and exchange of information among many different levels of the staff of the wwPDB take place every day; this makes the wwPDB a really unified global organization. For example, the new data in PDB are updated weekly at the same time (midnight, GMT, every Wednesday) from all the wwPDB member sites, PDBj, PDBe, and RCSB PDB. It works well and gives an impression of a good synergy to users.

WAW: The four wwPDB partners (RCSB PDB, PDBe, PDBj, and BMRB) are working on a common software tool for deposition and annotation of X-ray, NMR and EM data and structure models [8]. What is PDBj's role in this enormous project?

HN: As I said, PDBj is strong in validation of the data description, and this is also important for the common deposition and annotation procedure. Our PDBj team has been developing new software tools to validate the deposited data description, and to show what is wrong in the descriptions, by both the mmCIF and PDBML formats. This work is ongoing: currently, processing differs at the different PDB sites.

WAW: When it comes to "data out" [8], PDBe, RCSB and PDBj all produce their own set of unique, value-added services, tools and activities, based on the local environment, strengths and interests. Tell me about PDBj's unique tools and services.

HN: We consider that the PDB data should be used very widely with other biological databases in an integrated manner. Therefore, PDB data must be available and be applied with modern informatics techniques, such as using a Semantic Web framework, as Ross King's group once suggested in 2007 [9]. In fact, the mmCIF and PDBML are, in principle, in accordance with the correct ontology with their well-defined dictionary, and we have recently started to provide data in a Resource Description Framework (RDF) format in which most entities are dereferenceable by unique Uniform Resource Identifiers (URI) [10]. Providing a unique URI to each category element of PDB entries in the RDF format implies that the entire PDB is accessible on the Web, and that the category elements are linked to one another as well as to external resources. A complete ontology of the RDF in Web Ontology Language (OWL) is

also provided, which is based on the established PDB exchange dictionary so that specialists in structural biology or structural bioinformatics can readily understand the data structures. This new option will be presented at the International Union of Crystallography (IUCr) 2011 meeting to be held in Madrid this August.

Our other unique services are searches by analog data: similar shapes or folds [11], similar molecular surfaces [12], and similar local atomic dispositions for ligand binding sites (in proteins and nucleic acids) [13, 14], so that the biochemical and biological functions can be estimated through those similarities. We have also developed EM-Navi [15], which allows users to view the electron microscopy pictures in the 3D-EM database, simultaneously with the PDB data. In addition, we have developed an encyclopedia of protein structures and functions (eProtS) [16], in both English and Japanese, which is used regularly by many university and high-school students. Even the general public is interested in the articles in eProtS.

WAW: wwPDB has convened expert validation task forces for X-ray, NMR and EM to advise on the most suitable criteria to use for validating structure entries when they are deposited [8]. The wwPDB sites are expected to provide detailed quality statistics for every crystal structure in the PDB (following the recommendations of the wwPDB X-ray Validation Task Force). What is PDBj's role in this?

HN: Our annotators follow the standard to examine the quality of the structures, with daily email exchanges, discussion with the validation task force, and visits to and from the other wwPDB members, so that the structure and metadata qualities are always shared among the wwPDB members. For historical reasons there are no Japanese members of the task force but we have good relationships with the collaborators. For example, Gert Vriend, one of the task force members, visited and stayed in our institute recently.

WAW: Scientists across the world sympathize with you, our Japanese colleagues, in view of the tremendous losses you have suffered as a result of the devastating earthquake, the tsunami, and the nuclear crisis. At Osaka do you foresee any specific repercussions of Japan's current economic problems?

HN: The tsunami wrought terrible destruction in the northern part of Japan, and killed up to twenty thousand people. Moreover, the unstable situation of the Fukushima nuclear plant is still a big worry, from which we now expect two serious problems. One is lack of electric power this summer. This is particularly serious in the Tokyo area, where many computers may not work. To address this situation, databases may be distributed using the cloud system, as NBDC now plans. Even in Osaka, some electric power cuts are expected, but the PDBj may not stop at all. The other serious problem is some decrease in the budget for the science and engineering field, including education costs for national universities. In addition, this disaster has caused a paradigm shift in development for science and technology: many Japanese people now consider that future development for science and technology should be sustainable enough, to accord with a global society and with nature.

WAW: I hope that PDBj holds firm throughout the crisis and emerges even stronger. Thank you so much for allowing me to interview you.

References

1. Protein Data Bank Japan (PDBj) http://pdbj.org/pdbj_aboutus.html. Accessed June 2011
2. Institute for Protein Research (IPR), Osaka University http://www.protein.osaka-u.ac.jp/index_e.php. Accessed June 2011
3. Research collaboratory for structural genomics (RCSB) PDB http://www.pdb.org. Accessed June 2011
4. Berman HM, Henrick K, Nakamura H (2003) Announcing the worldwide Protein Data Bank. Nat Struct Biol 10(12):980
5. National Bioscience Database Center (NBDC) http://bioscience dbc.jp/?lng=en&pg=1. Accessed June 2011
6. Statistics for PDB structures that are deposited and processed by year and site http://www.wwpdb.org/stats.html. Accessed June 2011
7. Westbrook J, Ito N, Nakamura H, Henrick K, Berman HM (2005) PDBML: the representation of archival macromolecular structure data in XML. Bioinformatics 21(7):988–992
8. Warr WA (2010) Collaboration, competition, validation and plans for the future. An interview with Gerard Kleywegt, Head of the Protein Data Bank in Europe. J Comput-Aided Mol Des 24:887–890
9. Schierz AC, Soldatova LN, King RD (2007) Overhauling the PDB. Nat Biotechnol 25:437–442
10. PDBj RDF http://pdbj.org/rdf. Accessed June 2011
11. Structure Navigator http://pdbj.org/stnavix. Accessed June 2011
12. Electrostatic surface of functional site (eF-site) http://pdbj.org/efsite; eF-seek. http://pdbj.org/efseek. Accessed June 2011
13. Kinjo AR, Nakamura H (2009) Comprehensive structural classification of ligand binding motifs in proteins. Structure 17:234–246
14. Geometric indexing with refined alignment finder (GIRAF) http://pdbj.org/giraf/. Accessed June 2011
15. Electron microscopy navigator (EM-Navi) http://pdbj.org/emnavi/?&lang=en. Accessed June 2011
16. Encyclopedia of protein structures (eProtS) http://www.pdbj.org/eprots/index_en.cgi. Accessed June 2011