J Comput Aided Mol Des (2008) 22:707–710
DOI 10.1007/s10822-008-9234-3

WARR'S PIECE

Future structural genomics initiatives: an interview with Helen Berman, director of the Protein Data Bank

Wendy A. Warr

Received: 30 July 2008 / Accepted: 8 August 2008 / Published online: 10 September 2008

Helen M. Berman (HMB, pictured) is a Board of Governors Professor of Chemistry and Chemical Biology at Rutgers, The State University of New Jersey. Her research area is structural biology and bioinformatics, with a special focus on protein–nucleic acid interactions. She is the founder of the Nucleic Acid Database, a repository of information about the structures of nucleic acid-containing molecules; and is the co-founder and Director of the Protein Data Bank, the international repository of the structures of biological macromolecules. She is a Fellow of the American Association for the Advancement of Science and of the Biophysical Society, from which she received the Distinguished Service Award in 2000. A past president of the American Crystallographic Association, she is a recipient of the Buerger Award (2006). Dr. Berman received her A. B. in 1964 from Barnard College and a Ph.D. in 1967 from the University of Pittsburgh.

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

Interview

WAW: When did you take on the oversight of PDB from Brookhaven? What were your objectives then? To what extent have you achieved them?

HMB: The contract with the Research Collaboratory for Structural Genomics (RCSB) [1] was signed in October 1998, but there was an overlap of several months after that. I had been involved with PDB since 1971 as a co-founder. Brookhaven hosted the database from 1971 to 1998. During the 1990s there was a large increase in structures and there were more users. Then a funding review and an open competition took place and we won the contract. I have also been involved for years in projects for formalizing data formats and gathering nucleic acid structures, and I had a deep knowledge of the field. Our objectives were to make PDB a much richer resource, to produce a searchable archive, to let users make use of properties in different ways, to raise awareness, and to introduce much more rigor in the representation of structures and encode them in proper databases. We also wanted to work in a more formal and effective way with the other international resources at the European Bioinformatics Institute (EBI) and Osaka University in Japan to make PDB a more effective international resource.

WAW: People tell me that the quality of the PDB used to be poor, but it has improved nowadays. I am told

that there are no structure factors so the legacy data cannot be fully evaluated.

HMB: The structures have improved since we made it our priority within the worldwide PDB (wwPDB) organization [2] to formalize every aspect of data validation and data representation. We have remediated the data to get more uniform formatting and nomenclature in the data; all of the wwPDB annotators work together. The definitions of data items have been formalized. This is an ongoing effort but one release of remediated data has already been done and we are about to do another release. The underlying data have to be clean.

Experimental data for older structures are available for some, but not all, structures. We recently made it mandatory to submit structure factors and NMR constraints as part of the deposition process. The presence of structure factors allows the wwPDB staff and users to evaluate the quality of the protein models.

WAW: Do you have sufficient funds?

HMB: The RCSB PDB is funded by grants from the United States funding agencies, but the funding mechanisms for PDB worldwide are not what they should be for this kind of resource. There are many grants for the PDB Europe (PDBe) effort [3] at the EBI. PDB Japan (PDBj) at Osaka University [4] also has multiple grants. All wwPDB centers have discussed the issues with the funding agencies. We need collaboration mechanisms to make the funding more coherent, but I am not optimistic that this will happen. Instability with respect to funding at any one of the wwPDB centers would mean trouble for PDB and *many* users would suffer.

WAW: Everyone knows about the importance of PDB, but tell me more about the Protein Structure Initiative (PSI) [5] and why it matters.

HMB: The initiative started 8 years ago to study structures at a genomic level and develop technologies for high throughput structure determination. It was a post-genomic project, and was very courageous. The first few years were pilots, funded like research projects.

WAW: The recent review panel [6] concluded that PSI's technology development had been highly successful, did it not?

HMB: Yes. In the second phase [PSI-2] four centers were identified as production centers [7] and a few others as specialized technology centers, and these centers cooperate in a networked way to get structures out to the public. There is novelty in what they do. First of all they succeeded in implementing high throughput pipelines in ways that had never been seen before, and the structures are rapidly made available. And the quality of the structures is very high. All protocols are made public prior to the appearance of the structure in the PDB archive, even before the structure is solved [7]. This is novel data sharing.

WAW: I have just been at a cheminformatics meeting in Europe and was surprised to find that no-one had heard of PSI. Someone did say "Is that the structural genomics exercise?" Has the message not been spread widely enough?

HMB: The PSI centers were all so busy getting everything done that there was not a big emphasis on outreach. It takes years before people begin to use a data resource if there is no outreach. All the different centers had different web sites with different unusual names. We consider this a matter of great importance and are working very hard to increase the visibility, transparency and profile of PSI efforts.

WAW: Who conceived the idea of the PSI Structural Genomics Knowledgebase [8, 9] that you were appointed to run in summer 2007?

HMB: The key components of the Knowledgebase (KB) include experimental data tracking, a materials repository, homology modeling, annotation, technology development, metrics and outreach. The Knowledgebase was intended to be a "one-stop shop" with not just the structures, but with much *more* information. All the models from one structure can be leveraged. There is a portal where you can type in a sequence and get all the possible models from a variety of modeling resources. You get information about the technology developed and used by each center. The construction of the Knowledgebase started last August. The National Institutes of Health (NIH) had wanted to build it for a long time, and first asked me to direct it in June 2007. When we began work in August we set up a series of portals. The model portal was set up in Switzerland and the technology portal at Lawrence Berkeley Laboratory. The Core KB portal is at Rutgers. In October I began to discuss having a Gateway with the Nature Publishing Group. It will be launched in September 2008. A prototype of the Knowledgebase was available in February. This is all brand new. We have achieved all this in under a year.

WAW: A recent article in *Science* [10] about the PSI review panel [6] hardly mentions the Knowledgebase. Why is that?

HMB: It takes time. The *Science* article was written just when the Knowledgebase was being launched.

WAW: So you really need publicity for the Knowledgebase. Terry Stouch sent me an enthusiastic message about PSI.

HMB: Yes. Terry was telling me all the things he had to do when he wants to know something about a certain protein and I told him that he could send his sample to a technology center in Buffalo to determine crystallization conditions. And there are many capabilities built into the KB to assist in experimental design. For example, you can retrieve protocols for protein purification for target sequences. Having all the protocols out there is extraordinary. Torsten Schwede of the Swiss Institute of Bioinformatics has some of the best modeling resources behind his portal.

WAW: Why are certain structural biologists [10] not equally enthused?

HMB: It takes time. We have only been doing this for a matter of months. I am excited about the Gateway. Nature Publishing Group is providing editorial help and will write several articles a month about structural genomics-related projects. The KB site also hosts a column by David S. Goodsell about PSI structures similar to his *Molecule of the Month* column at the RCSB PDB. The Functional Sleuth section of the site presents PSI structures that are lacking functional annotation to encourage discussion about their possible functions. The KB will be a resource for all scientists studying living systems and disease.

WAW: When the PSI-2 assessment report [6] came out last December, was it too early for the Knowledgebase to be taken into account?

HMB: The PSI-2 assessment was held in September 2007 and the report came out in December. The report actually says that the Knowledgebase should have started earlier and I agree: it should have started 3 years ago. If it had, we would have had traction by now.

WAW: When will the decision be made about funding for "PSI-3"?

HMB: There is a meeting October at the NIH to discuss future structural genomics initiatives. I am looking forward to being there.

WAW: It was a major goal of PSI to obtain structures of representatives of as many protein families as possible [10, 11]. Some biologists want you to concentrate on proteins of known biological relevance. Do you think that a compromise can be or should be reached?

HMB: The family work *is* relevant to biology. Further, each center devotes 15% of its effort to a particular biomedical theme. There is also a big push in the field of metagenomics. There is a lot of discussion about all of this and a lot of misinformation. This is why I am so keen to get the Knowledgebase out there. The Gateway will get people to see that we have some gorgeous structures and novel technologies. The membrane proteins in particular are really exciting.

WAW: The Gateway launch will not really be early enough for the October meeting, though. Are people actually using the Knowledgebase?

HMB: We are measuring activity, and people *are* using it. It will take time to get further traction. We also run TargetDB [8], a database at Rutgers that tracks the status of each target. That has lots of activity. People are always looking to see what is going on there.

WAW: How many people are working on the Knowledgebase?

HMB: At Rutgers we have one software architect, one Web programmer, one database programmer and a systems support person, and we have just hired one person to do outreach. There is the technology group at Berkeley, and people in Switzerland to do the models. I manage those groups by regular meetings. This is a distributed way of working. I tend to work in an organic way: I do one thing and get it working and then I do the next thing and get that working. Then I put the things together. I also think we need ongoing community input. Workshops, such as the recent one on biological annotation of novel proteins [12], provided key input for the annotation module.

WAW: For how long do you have funding?

HMB: We go on year by year. I am assuming that we will be successful. The next funding cycle starts in July 2009. There are no guarantees in this life but I hope that we have funds until the end of PSI-2 in 2010. Our steering committee feels strongly that this should be an enduring exercise.

WAW: What functionality needs to be added?

HMB: We are currently working on standardized annotations for all structures. These can be used to determine the functions of those structures whose functions are not known. We want to use established annotations that are reliable, for example, Pfam [13], CATH [14] and SCOP [15], so that you can see all that you need for prediction of function.

WAW: How much would "PSI-3" cost, were it approved?

HMB: I have no idea, but I do believe that the PSI's high level of data sharing would be key to its success.

WAW: Is the Knowledgebase critical to the success of PSI?

HMB: Yes, it is critical, because it will show people what is going on, but the Knowledgebase goes well beyond PSI because it will be a paradigm for how structures are represented.

WAW: This initiative is clearly of great importance to you personally.

HMB: This is not about *me*. It is about creating a new community-based resource. Since I was in my twenties I have been passionately committed to open data sharing. There is no point in doing science otherwise. Science is publicly funded; the data must not get lost. It is about how science ought to be done.

WAW: You have achieved much more than most of us will in our careers but do you have other ambitions. What other plans do you have for the future?

HMB: I would like this project to work. I would like to see the Knowledgebase extended to become a model for how you represent data. I want to see much more coherent funding for data infrastructure. All these data resources exist in constant fear of disintegration and no stable funding. The funding agencies (NIH, the National Science Foundation and the Department of Energy in the United States, the Biotechnology and Biological Sciences Research Council in the United Kingdom, the Wellcome Trust etc.) are national; but science is global, and funding needs to be stable. Making this happen is my primary goal before I retire and hand over the reins. I do not give up easily.

References

1. RCSB PDB http://www.pdb.org. Accessed July 2008
2. Worldwide PDB http://www.wwpdb.org. Accessed July 2008
3. Europe PDB http://www.ebi.ac.uk/msd/. Accessed July 2008
4. Japan PDB http://www.pdbj.org/. Accessed July 2008
5. PSI http://www.nigms.nih.gov/Initiatives/PSI. Accessed July 2008
6. Report of the protein structure initiative assessment panel http://www.nigms.nih.gov/About/Council/PSIAssessment.htm. Accessed July 2008
7. Burley SK, Joachimiak A, Montelione GT, Wilson IA (2008) Structure 16:5. doi:10.1016/j.str.2007.12.002
8. Berman HM (2008) Structure 16:16. doi:10.1016/j.str.2007.12.003
9. Knowledgebase PSI http://kb.psi-structuralgenomics.org/. Accessed July 2008
10. Service RF (2008) Science 319:1610. doi:10.1126/science.319.5870.1610
11. Hendrickson WA (2007) Structure 15:1528. doi:10.1016/j.str.2007.11.006
12. Workshop on biological annotation of novel proteins. http://annotation-workshop.psi-structuralgenomics.org/. Accessed July 2008
13. Pfam http://pfam.sanger.ac.uk/. Accessed July 2008
14. CATH http://www.cathdb.info/. Accessed July 2008
15. SCOP http://scop.mrc-lmb.cam.ac.uk/scop/. Accessed July 2008