STRUCTURAL BIOLOGY

ISSN 2059-7983

Everyone is using biological structures, but how does one find the structure(s) one wants?

Charles S. Bond[a]* and Joel L. Sussman[b]

[a]University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia, and [b]Department of Chemical and Structural Biology, Weizmann Institute of Science, 76100 Rehovot, Israel. *Correspondence e-mail: charles.bond@uwa.edu.au

Keywords: biological structures; structural databases; Protein Data Bank; UniProt; AlphaFold DB.

Atomic structures of biological molecules have never been so available and ubiquitous, but this raises questions as to how they are made appropriately accessible for optimal use. The variety of uses of structures is huge, ranging from convenient illustrations in talks to small-molecule docking, understanding of ligand specificity and cofactor binding, expression construct design, comparative conformational analyses, prediction of complex structures, mutant design, phylogeny, and much more.

We have put ourselves in the shoes of a newcomer to biology and here consider what some of the needs of the broader scientific community might be, and how they might be addressed.

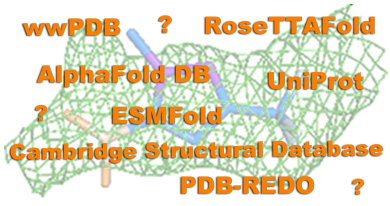
Dear Structural Biologists,

I have just begun my PhD studies in a cell biology lab and am using a lot of databases to plan my project. I have so many questions! The first ones are about the Protein Data Bank (Berman et al., 2000). What an amazing resource – my supervisor told me that the PDB was the first major open database in molecular biology and that it has led the field of data validation for decades (Gore et al., 2017; Helliwell et al., 2019).

My supervisor asked me to summarize the structural information available for my target protein, which is part of a complex found in all prokaryotes. As a non-expert in structural biology, I find it difficult to determine which of the various structures is most reliable and most relevant. I typically access the PDB entries directly from a sequence database like UniProt (The UniProt Consortium, 2021), or via a link generated by various online servers, but I see that I have to find each of the relevant papers (where they exist!) to try to understand which structure is the most reliable. Having that information readily available in the PDB could make it much more useful. To be honest, starting from UniProt, if I instead follow the links to AlphaFold DB (Varadi et al., 2024), everything is very comparable and in a common format, and even the disordered regions are included in the protein structure! What is the advantage to me of using the PDB over AlphaFold DB, or models from RoseTTAFold (Baek et al., 2021) or ESMFold (Lin et al., 2023)?

Another protein I plan to work on forms polymeric filaments (it's a member of the RecA superfamily). When I look for this protein and its orthologues in the PDB, the assigned biological assembly is often everything from monomer and dimer to dodecamer. Still, the arrangement in the crystal usually looks like a biologically relevant filament. Sometimes, they are labeled as helical, but not always. Having looked at other filament-forming proteins, it seems the same. Is there a way to search the PDB for all proteins that form 'infinite' 1D, 2D or 3D lattices that are functionally relevant?

I wondered about this because I used another amazingly useful database in my undergraduate project, the Cambridge Structural Database (Groom et al., 2016), and



found the data interrogation tools really useful for my project (Bruno et al., 2002). Is there a way that I can interrogate the PDB for geometric information without having to download the whole thing?

In a recent lecture on 'Protein Structure and Function', the professor described how proteins are not rigid but actually show flexibility and conformational change, such as in the movie of the catalytic cycle of nucleoside monophosphate kinases (Vonrhein et al., 1995). Is there any way of using the PDB to compile sets of structures of the same protein and make a 'morph' between them?

One of my fellow students is a real crystallography nerd and decided to try re-refining a structure she found in the PDB. The latest refinement tools have helped her produce a structure with much better statistics than the original one, and she even found some new small molecules in it! Is there any way that this kind of edit can be included in the PDB? Our supervisor pointed out that versioning was introduced to the PDB in 2017 to allow updated structures, but it seems – sensibly – that this is only open to the original authors. We did find PDB-REDO (Joosten et al., 2014), which is an automated solution to this problem, but sometimes biological or chemical knowledge can be crucial in curating the results of these re-refinements. Because her project is on a related enzyme, my friend found a paper (Wlodawer et al., 2024) which seemed to have the same problem: how to make re-refined and re-curated structural data available.

Maybe it is a bit impertinent to suggest, but could the wider community engage in a brainstorming workshop to consider the future of biological structure databases?

Yours,

Wilson B. Student

The difficulties outlined above put a focus on a pivotal moment in structural biology. Since the release of the now Nobel Prize-winning AI-based revolutionary tools that predict 3D structures of biological macromolecules and complexes with typically high accuracy, the appearance of AI-generated 3D structures has become ubiquitous in molecular and cell biology conference presentations, even by those with no structural biology experience. Structural biology is no longer an isolated niche field but is now firmly in the mainstream of biology, many fields of chemistry, and even physics. We suggest that the Protein Data Bank, one of the first community-wide scientific resources, which spearheaded the open deposition and sharing of scientific information concerning protein structures since 1972 must radically rethink its role in this world where experimental and predicted structures are used almost indistinguishably. Over 200 million 3D structures are currently available in AlphaFoldDB, which dwarfs the >225 000 experimentally determined structures deposited in the PDB. Many scientists, even structural biologists, turn initially to a database such as UniProt first and let it direct them to many key databases related to sequence, function, and location of the biological macromolecule *etc.*, and from there to structure databases. With so much experience in distributing such valuable 3D structural information to all corners of the world, the time is ripe for the PDB to reconsider how these data are presented. We suggest a two-step process.

(1) Open worldwide, virtual consultation and discussion as to what the current databases could do differently or additionally. This might include sustainable financial models for database maintenance.

(2) A focused face-to-face meeting including scientists from a broad spectrum of disciplines, reflecting the significance of new experimental and computational structural biology techniques, cryoEM, dynamics, engineering, evolution, intrinsically disordered proteins and RNA ensemble structures. In 1975, the Asilomar Conference on Recombinant DNA resulted in scientists taking the lead in new scientific policies, and such a science community-led conference could brainstorm and set the future direction of biological structure databases.

There is an urgent need to completely rethink how biomolecular structural databases should be organized and accessed to address the rapidly developing needs of non-structural and structural biologists.

References

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., Opperman, D. J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M. K., Dalwadi, U., Yip, C. K., Burke, J. E., Garcia, K. C., Grishin, N. V., Adams, P. D., Read, R. J. & Baker, D. (2021). *Science*, **373**, 871–876.

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.

Bruno, I. J., Cole, J. C., Edgington, P. R., Kessler, M., Macrae, C. F., McCabe, P., Pearson, J. & Taylor, R. (2002). *Acta Cryst.* B**58**, 389–397.

Gore, S., Sanz García, E., Hendrickx, P. M. S., Gutmanas, A., Westbrook, J. D., Yang, H., Feng, Z., Baskaran, K., Berrisford, J. M., Hudson, B. P., Ikegawa, Y., Kobayashi, N., Lawson, C. L., Mading, S., Mak, L., Mukhopadhyay, A., Oldfield, T. J., Patwardhan, A., Peisach, E., Sahni, G., Sekharan, M. R., Sen, S., Shao, C., Smart, O. S., Ulrich, E. L., Yamashita, R., Quesada, M., Young, J. Y., Nakamura, H., Markley, J. L., Berman, H. M., Burley, S. K., Velankar, S. & Kleywegt, G. J. (2017). *Structure*, **25**, 1916–1927.

Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. (2016). *Acta Cryst.* B**72**, 171–179.

Helliwell, J. R., Minor, W., Weiss, M. S., Garman, E. F., Read, R. J., Newman, J., van Raaij, M. J., Hajdu, J. & Baker, E. N. (2019). *J. Appl. Cryst.* **52**, 495–497.

Joosten, R. P., Long, F., Murshudov, G. N. & Perrakis, A. (2014). *IUCrJ*, **1**, 213–220.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S. & Rives, A. (2023). *Science*, **379**, 1123–1130.

The UniProt Consortium (2021). *Nucleic Acids Res.* **49**, D480–D489.

Varadi, M., Bertoni, D., Magana, P., Paramval, U., Pidruchna, I., Radhakrishnan, M., Tsenkov, M., Nair, S., Mirdita, M., Yeo, J., Kovalevskiy, O., Tunyasuvunakool, K., Laydon, A., Žídek, A., Tomlinson, H., Hariharan, D., Abrahamson, J., Green, T., Jumper, J., Birney, E., Steinegger, M., Hassabis, D. & Velankar, S. (2024). *Nucleic Acids Res.* **52**, D368–D375.

Vonrhein, C., Schlauderer, G. J. & Schulz, G. E. (1995). *Structure*, **3**, 483–490.

Wlodawer, A., Dauter, Z., Lubkowski, J., Loch, J. I., Brzezinski, D., Gilski, M. & Jaskolski, M. (2024). *Acta Cryst.* D**80**, 506–527.