Announcing launch of Protein Data Bank China (PDBc) as an Associate Member of the Worldwide Protein Data Bank (wwPDB) Partnership

Wenqing Xu[a*], Sameer Velankar[b*], Ardan Patwardhan[c], Jeffrey C. Hoch[d], Stephen K. Burley[e,f*], and Genji Kurisu[g*]

[a]Protein Data Bank China, ShanghaiTech University and National Facility for Protein Science in Shanghai, Shanghai, People's Republic of China.

[b]Protein Data Bank in Europe, European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom.

[c]Electron Microscopy Data Bank, European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom.

[d]Biological Magnetic Resonance Data Bank, UConn Health, Farmington, CT 06030, United States.

[e]Research Collaboratory for Structural Bioinformatics Protein Data Bank, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, United States.

[f]Research Collaboratory for Structural Biology Protein Data Bank, San Diego Supercomputer Center, University of California San Diego, La Jolla, CA 92093, United States.

[g]Protein Data Bank Japan, Institute for Protein Research, Osaka University, Osaka 565-0871, Japan.

*__Correspondence e-mail__: xuwq2@shanghaitech.edu.cn, sameer@ebi.ac.uk, stephen.burley@rcsb.org, gkurisu@protein.osaka-u.ac.jp

Keywords: macromolecular crystallography; nuclear magnetic resonance; three-dimensional cryo-electron microscopy; Protein Data Bank; Biological Magnetic Resonance Bank; Electron Microscopy Data Bank; worldwide Protein Data Bank

The Protein Data Bank (PDB) (Protein Data Bank, 1971) is the single global archive of atomic-level, three-dimensional (3D) structures of biological macromolecule experimentally determined by macromolecular crystallography (MX), nuclear magnetic resonance (NMR) spectroscopy, or three-dimensional cryo-electron microscopy (3DEM). The PDB is growing continuously, with new structure depositions coming in from Asia increasing rapidly. In 2022, the Worldwide Protein Data Bank (wwPDB, wwpdb.org) partners welcomed Protein Data Bank China (PDBc based in the National Facility for Protein Science in Shanghai which is associated with Shanghai Advanced Research Institute of Chinese Academy of Sciences, and the SIAIS and iHuman Institutes of ShanghaiTech University; PDBc.org.cn) to the organization as an Associate Member. This letter describes the history of the wwPDB, recently-established mechanisms for adding new wwPDB data centers, and the processes developed to bring PDBc into the partnership.

The PDB archive is maintained as a global public good by the wwPDB (Berman *et al.,* 2003), an international non-governmental organization that was founded in 2003 by three regionally-funded data centers located in the United States (Research Collaboratory for Structural Bioinformatics Protein Data Bank, RCSB PDB; RCSB.org), the United Kingdom (Protein Data Bank in Europe, PDBe; PDBe.org), and Japan (Protein Data Bank Japan, PDBj; PDBj.org). Two method-focused wwPDB data centers, Biological Magnetic Resonance Bank (BMRB, BMRB.io) and Electron

Microscopy Data Bank (EMDB, www.ebi.ac.uk/emdb/) joined the wwPDB in 2006 and 2021, respectively (wwPDB consortium, 2019). wwPDB partners adhere to the FAIR principles of Findability, Accessibility, Interoperability, and Reusability (Wilkinson *et al.*, 2016), and ensure that all archival data can be accessed at no charge and with no limitations on usage under the most permissive Creative Commons CC0 1.0 Universal License (creativecommons.org/publicdomain/zero/1.0/). At present, wwPDB members jointly manage three Core Archives, overseen by wwPDB-designated Archive Keepers, including PDB (Archive Keeper: RCSB PDB), EMDB (Archive Keeper: EMDB), and BMRB (Archive Keeper: BMRB).

wwPDB operations are guided by an international advisory committee (wwPDB AC, www.wwpdb.org/about/advisory), consisting of an independent chair, representatives appointed by each wwPDB data center, and experts drawn from the MX, 3DEM, and NMR communities. The wwPDB AC meets annually to review the health and well-being of the three Core Archives. Annual meeting presentations and reports are publicly available online.

When the wwPDB was launched 20 years ago, significant growth in the PDB archive was eagerly anticipated. At that time, each wwPDB data center accepted 3D biostructure depositions through independently operated data deposition/validation/biocuration (data-in) systems. To accommodate the increased demand for data archiving, wwPDB members initiated the development of a single global software system for data-in and started planning for the extension of the franchise to new PDB data centers for management of depositions coming from emerging economies. Both steps were strongly endorsed by the wwPDB AC.

In 2015, the wwPDB launched OneDep (Young *et al.*, 2017) as a single global system for complete deposition, rigorous validation, and expert biocuration of 3D biostructures (Young *et al.*, 2018). The OneDep software system also supports archive-wide remediation of existing PDB structures and secure transfer of information among wwPDB data centers. Each wwPDB data center manages its own instance of OneDep. Initially, incoming structures were assigned to wwPDB data centers on a regional basis as follows: RCSB PDB – America and Oceania; PDBe – Europe and Africa; and PDBj – Asia and the Middle East. During 2017, RCSB PDB, PDBe, and PDBj processed 6,206, 4,044, and 2,799 new depositions, respectively (Total: 13,049; MX: 11,915; 3DEM: 674; NMR: 460). In 2022, RCSB PDB, PDBe, and PDBj processed 7,020, 4,811, and 4,214 new depositions, respectively (Total: 16,344; MX:10,650; 3DEM: 5,407; NMR: 287). Annual depositions of 3DEM experimental maps to EMDB were as follows: 1,448 in 2017 (including 774 map only, no atomic coordinates), and 8,519 in 2022 (including 3,112 map only, no atomic coordinates). Annual NMR experimental data depositions for macromolecular structures to BMRB were 460 in 2017, and 287 in 2022. Growth in the number of PDB depositions originating from Asia is primarily the result of increased investment in research and development within the People's Republic of China (PRC). In 2022, structural biologists working in the PRC contributed 3,118 new PDB structures (~65% of all depositions from Asia). Reflecting the impact of the resolution revolution, 3DEM depositions to PDB coming from the PRC in 2022 numbered 1,538 (*versus* 1,539 MX and 41 NMR depositions coming from the PRC in 2022).

wwPDB operations are governed by an international agreement (www.wwpdb.org/about/agreement) that was most recently renewed at the beginning of 2021. The new "Charter of the wwPDB" defined all five existing members as Core Members of the wwPDB, and established mechanisms for admitting new Associate Members and advancing them to Core Member status. These procedural changes were reviewed and endorsed by the wwPDB AC. Effective February 1st, 2022, PDBc was admitted to the wwPDB partnership as its first Associate Member. During the remainder of 2022, to support the launch of PDBc, formal training including remote online orientation and policy training was provided by RCSB PDB, PDBe, and

EMDB. This was, followed by on-site hands-on biocuration training in Osaka provided by PDBj. RCSB PDB, PDBe, and EMDB helped to mitigate the impact on PDBj by processing overflow depositions during PDBc on-site training.

Following two months of PDBc on-site training at PDBj, two PDBc biocurators returned to Shanghai and began processing PDB depositions originating from within the PRC using a new PDBc OneDep instance (currently managed by PDBj). As of the end of 2022, PDBc biocurators had processed 297 new PDB depositions, representing ~10% of 3D biostructure depositions from the PRC for the year (~7% of all Asian depositions). During 2023, it is expected that PDBc will process most, possibly all, depositions made by structural biologists working in the PRC to all three wwPDB Core Archives.

The wwPDB is excited to formally announce the launch of PDBc coinciding with the 20th anniversary of continuous wwPDB operations serving tens of thousands of structural biologists around the world and many millions of PDB data consumers based in nearly every sovereign country recognized by the United Nations.

Funding Statement

PDBc is supported by the Shanghai Advanced Research Institute (SARI), Chinese Academy of Sciences and the ShanghaiTech University. PDBe and EMDB are supported by the European Molecular Biology Laboratory-European Bioinformatics Institute and by Wellcome Trust (104948/Z/14/Z, 221327/Z/20/Z). BMRB is supported by the US National Institutes of Health (R01GM109046, P.I.: J.C. Hoch). RCSB PDB is jointly funded by the National Science Foundation (DBI-1832184, PI: S.K. Burley), the US Department of Energy (DE-SC0019749, PI: S.K. Burley), and the National Cancer Institute, the National Institute of Allergy and Infectious Diseases, and the National Institute of General Medical Sciences of the National Institutes of Health (R01GM133198, PI: S.K. Burley). PDBj is supported by grants from the Database Integration Coordination Program from the department of NBDC program, Japan Science and Technology Agency (JPMJND2205, PI: G. Kurisu), and by the Platform Project for Supporting Drug Discovery and Life Science Research (Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS)) from AMED (23ama121001, P.I. G. Kurisu).

References

Berman, H.M., Henrick, K. & Nakamura, H. (2003). *Nat. Struct. Biol*. **10**, 980

Gore, S., Sanz Garcia, 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

Protein Data Bank (1971). *Nature New Biol*. **233**, 223.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J., Groth, P., Goble, C., Grethe, J. S., Heringa, J., t Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S. A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. & Mons, B. (2016). *Sci Data*, 3, 1–9.

wwPDB consortium (2019). *Nucleic Acids Res.* **47**, D520-D528.

Young, J. Y., Westbrook, J. D., Feng, Z., Sala, R., Peisach, E., Oldfield, T. J., Sen, S., Gutmanas, A., Armstrong, D. R., Berrisford, J. M., Chen, L., Chen, M., Di Costanzo, L., Dimitropoulos, D., Gao, G., Ghosh, S., Gore, S., Guranovic, V., Hendrickx, P. M. S., Hudson, B. P., Igarashi, R., Ikegawa, Y., Kobayashi, N., Lawson, C. L., Liang, Y., Mading, S., Mak, L., Mir, M. S., Mukhopadhyay, A., Patwardhan, A., Persikova, I., Rinaldi, L., Sanz-Garcia, E., Sekharan, M. R., Shao, C., Swaminathan, G. J., Tan, L., Ulrich, E. L., van Ginkel, G., Yamashita, R., Yang, H., Zhuravleva, M. A., Quesada, M., Kleywegt, G. J., Berman, H. M., Markley, J. L., Nakamura, H., Velankar, S. & Burley, S. K. (2017). *Structure*, **25**, 536–545.

Young, J.Y., Westbrook, J.D., Feng, Z., Peisach, E., Persikova, I., Sala, R., Sen, S., Berrisford, J. M., Swaminathan, G. J., Oldfield, T. J., Gutmanas, A., Igarashi, R., Armstrong, D.R., Baskaran, K., Chen, L., Chen, M., Clark, A.R., Costanzo, L.D., Dimitropoulos, D., Gao, G., Ghosh, S., Gore, S., Guranovic, V., Hendrickx, P.M.S., Hudson, B.P., Ikegawa, Y., Kengaku, Y., Lawson, C.L., Liang, Y., Mak, L., Mukhopadhyay, A., Narayanan, B., Nishiyama, K., Patwardhan, A., Sahni, G., Sanz-García, E., Sato, J., Sekharan, M.R., Shao, C., Smart, O.S., Tan, L., v. Ginkel, G., Yang, H., Zhuravleva, M.A., Markley, J.L., Nakamura, H., Kurisu, G., Kleywegt, G.J., Velankar, S., Berman, H.M., Burley, S.K. (2018). *Database,* **2018**, bay002