Received: 2 November 2024 | Revised: 19 January 2025 | Accepted: 23 January 2025

DOI: 10.1002/pro.70052

TOOLS FOR PROTEIN SCIENCE

THE PROTEIN SOCIETY WILEY

Protein Data Bank Japan: Improved tools for sequence-oriented analysis of protein structures

Gert-Jan Bekker[1] | **Chioko Nagao**[1] | **Matsuyuki Shirota**[2,3,4] | **Tsukasa Nakamura**[5,6] | **Toshiaki Katayama**[1,7] | **Daisuke Kihara**[1,5,6,8] | **Kengo Kinoshita**[2,3,4] | **Genji Kurisu**[1,9]

[1]Institute for Protein Research, Osaka University, Suita, Japan

[2]Tohoku Medical Megabank Organization, Tohoku University, Sendai, Japan

[3]Advanced Research Center for Innovations in Next-Generation Medicine, Tohoku University, Sendai, Japan

[4]Graduate School of Information Sciences, Tohoku University, Sendai, Japan

[5]Department of Biological Sciences, Purdue University, West Lafayette, Indiana, USA

[6]Structural Biology Research Center, Institute of Material Structure Science, High Energy Accelerator Research Organization, Tsukuba, Japan

[7]Database Center for Life Science, Joint Support-Center for Data Science Research, Research Organization of Information and Systems, Kashiwa, Japan

[8]Department of Computer Science, Purdue University, West Lafayette, Indiana, USA

[9]Protein Research Foundation, Minoh, Japan

Correspondence
Genji Kurisu and Gert-Jan Bekker, Institute for Protein Research, Osaka University, 3-2, Yamadaoka, Suita, Osaka 565-0871, Japan.
Email: gkurisu@protein.osaka-u.ac.jp; gertjan.bekker@protein.osaka-u.ac.jp

Funding information
National Bioscience Database Center, Grant/Award Number: JPMJND2205; National Institutes of Health, Grant/Award Number: R01GM133840; Japan Society for the Promotion of Science, Grant/Award Numbers: JP20H03229, JP21K17847; Japan Agency for Medical Research and Development, Grant/Award Numbers: JP23ama121019, JP24ama121001

Review Editor: Nir Ben-Tal

Abstract

Protein Data Bank Japan (PDBj) is the Asian hub of three-dimensional macromolecular structure data, and a founding member of the worldwide Protein Data Bank. We have accepted, processed, and distributed experimentally determined biological macromolecular structures for over two decades. Although we collaborate with RCSB PDB and BMRB in the United States, PDBe and EMDB in Europe and recently PDBc in China for our data-in activities, we have developed our own unique services and tools for searching, exploring, visualizing and analyzing protein structures. We have recently introduced a new UniProt-integrated portal to provide users with a quick overview of their target protein and shows a recommended structure with integrated data from various internal and external resources. The portal page helps users identify known genomic variations of their protein of interest and provide insights into how these modifications might impact the structure, stability and dynamics of the protein. Furthermore, the portal page also helps users to select the optimal structure to use for further analysis. We have also introduced another service to explore proteins using experimental and computational approaches, which enables experimental structural biologists to increase their insight to help them to more efficiently design their experimental studies. With these new additions, we have enhanced our service portfolio to benefit both experimental and computational structural biologists in their search to interpret protein structures, their dynamics and function.

KEYWORDS
BMRB, EMDB, PDB, protein structure, sequence analyses, UniProt

1 | INTRODUCTION

The three-dimensional structural data of biological macromolecules are collaboratively maintained by the worldwide Protein Data Bank (wwPDB) partnership. Protein Data Bank Japan (PDBj, https://pdbj.org) has accepted and processed the 3D structure data of biological macromolecules from Asia and distributed the globally collected data since 2000 (Kurisu et al. 2022). In total, roughly 23% of all PDB entries had been processed by PDBj by the end of 2023. Since our founding, PDBj has developed various original services, which are listed in Table 1. Here, we will describe updates to our original services and the introduction of several new services to assist both experimentalists and structural data users alike.

2 | OVERVIEW OF ARCHIVES MAINTAINED BY PDBj

PDBj maintains three wwPDB core archives (Protein Data Bank: PDB, Electron Microscopy Data Bank: EMDB, and Biological Magnetic Resonance Data Bank: BMRB) under the wwPDB partnership in collaboration with other wwPDB members (Burley et al. 2019), while we also maintain uniquely developed archives. The PDB data we co-maintain together with RCSB PDB in the United States and PDBe in Europe, while together with the EMDB team at EMBL-EBI in Europe, we co-maintain the EMDB archive for experimental 3DEM maps. For NMR data, PDBj collaborates with the BMRB team in the USA to maintain the BMRB archive as part of our BMRBj activities (Hoch et al. 2023). Deposition to these wwPDB core archives is handled

TABLE 1 PDBj services and tools with corresponding URLs.

Service	URL
Search PDB (PDBj Mine)	https://pdbj.org/search/pdb-filter
Chemie search	https://pdbj.org/search/chemie-filter
Search BMRB	https://bmrbj.pdbj.org/
Sequence-Navigator	https://pdbj.org/seqnavi
EM Navigator	https://pdbj.org/emnavi/
Omokage search	https://pdbj.org/omokage
wwPDB/RDF	https://rdf.wwpdb.org/
jV: Graphic Viewer	https://pdbj.org/jv/
Molmil: WebGL Molecular Viewer	https://pdbj.org/molmil2/
Yorodumi	https://pdbj.org/emnavi/
NMRToolBox	https://bmrbj.pdbj.org/en/nmr_tool_box.html
gmfit	https://pdbj.org/gmfit/
CRNPRED	https://pdbj.org/crnpred/
HOMCOS	https://homcos.pdbj.org/
eF-site	https://pdbj.org/eF-site/
eF-seek	https://pdbj.org/eF-seek/
eF-surf	https://pdbj.org/eF-surf/
ProMode Elastic	https://pdbj.org/promode-elastic
Molecule of the Month	https://numon.pdbj.org/mom/
Games	https://numon.pdbj.org/games/
Papermodels	https://numon.pdbj.org/papermodel/
OneDep(Deposition to PDB, EMDB or BMRB)	https://deposit-pdbj.wwpdb.org/deposition
Format Conversion	https://mmcif.pdbj.org/converter/
PDBx/mmCIF editor	https://pdbj.org/cif-editor/
EMPIAR-PDBj	https://empiar.pdbj.org
BSM-Arc	https://bsma.pdbj.org
XRDa	https://xrda.pdbj.org
UniProt portal	https://pdbj.org/uniprot/
Sequence Navigator Pro	https://pdbj.org/seqnavipro
Dynamics DB	https://bsma.pdbj.org/dynamicsdb/
PDBj GitLab Portal	https://gitlab.com/pdbjapan

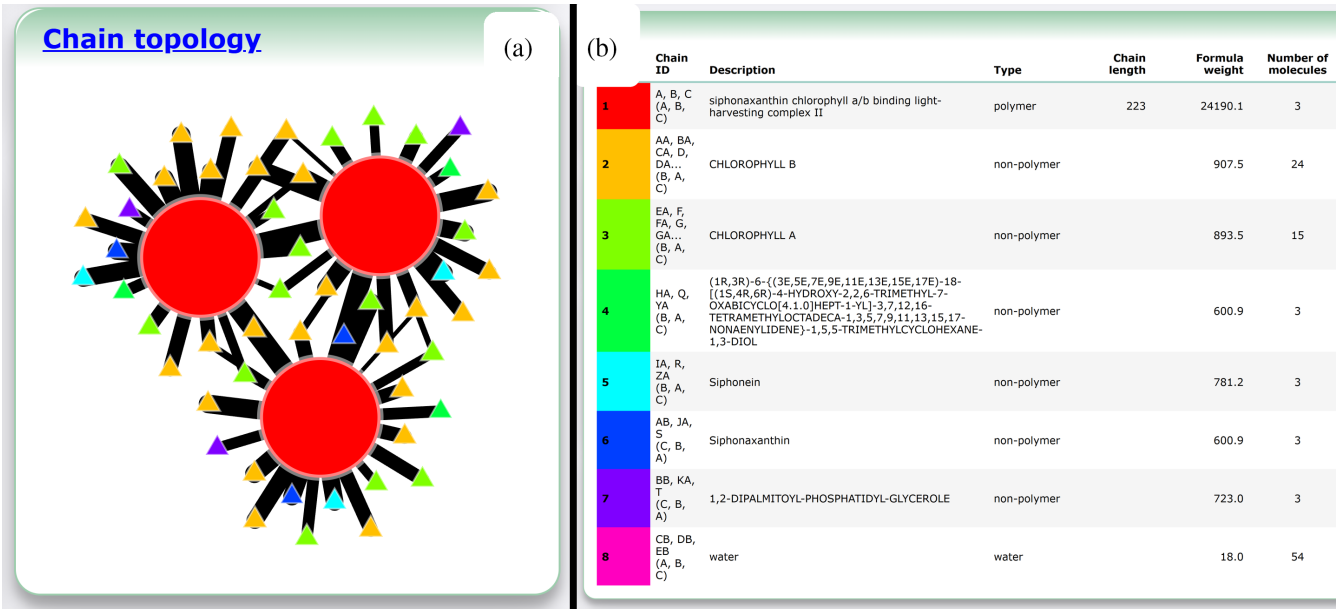
via the OneDep system, which are shared among all wwPDB partners. Here, PDBj manages the Asian depositions of experimental data submitted to the archives, corresponding to approximately 27% of the worldwide depositions over the past 5 years. Since 2018, we have also maintained a mirror of EMPIAR (the Electron Microscopy Public Image Archive: https://www.ebi.ac.uk/empiar/), in collaboration with the team at EMBL-EBI (Bekker et al. 2022). Our mirror site focuses on the PDB-related EMPIAR entries and our mirror site (EMPIAR-PDBj) is slightly different from the master archive at EMBL-EBI, with several entries not yet available from our portal site, while some of the others have compression applied to decrease their file size. In addition, to assist with depositions from Asia, we also broker the deposition of entries to EMPIAR, where we also accept HDD submissions via postal mail, which we then upload to EMBL-EBI. We have also developed two novel archives; BSM-Arc (the Biological Structural Model Archive) for computational data and XRDa (the Xtal Raw Data Archive) for experimental diffraction images. BSM-Arc (https://bsma.pdbj.org) (Bekker et al. 2020), which accepts structure models and raw data obtained via computational methods such as molecular dynamics (MD), homology modeling, or deep-learning based methods, currently consists of 51 public entries (6.6 TB). For XRDa (https://xrda.pdbj.org), which accepts raw experimental diffraction images obtained from X-ray, electron, or neutron diffraction, 174 entries have been published (9.8 TB). Thereby, PDBj collects both raw experimental data via BMRBj, EMPIAR-PDBj, and XRDa, experimental structural models via OneDep, as well as computational structural models and raw data via BSM-Arc, and is thus the only wwPDB partner that collects raw data for all experimental types and from computational sources.

PDBj also maintains several secondary databases. These secondary databases use data from the primary PDB archive and use computational methods to derive additional insights, which are stored in their respective archives. The Promode Elastic service provides information with respect to predicted dynamics of a PDB entry (or specific chains), calculated via Normal Mode Analysis, and is updated on a weekly basis (Wako et al. 2004). The eF-site service is a database containing the calculated electrostatic potentials mapped onto the molecular surfaces of functional sites (Kinoshita and Nakamura 2004). Finally, the dynamics DB is a service that provides stabilities and dynamics of proteins calculated via MD simulations.

3 | OVERVIEW OF PDBj TOOLS AND SERVICES

In addition to the primary and secondary archives that we maintain, we have also developed several tools and services (Bekker et al. 2022; Kinjo et al. 2017; Kinjo et al. 2018). Table 1 provides an overview of the available services and their links. To explore the PDB archive, we have developed the PDBj Mine service. Central to the PDBj Mine service is the Mine 2 relational database (RDB), which contains the meta-data of all PDB entries, as well as the metadata for the chemical component dictionary (chem_comp), PRD/BIRD (Biologically Interesting Molecule Reference Dictionary), validation reports (VRPT), and the EMDB data. In addition, various metadata and other calculations are also included, such as file information, release statistics, obsolete entry information, and intermolecular contact pairs. Since all this data is available within a single RDB, complex queries can be crafted to perform very precise searches, and/or extract data for a large number of entries with a single query. PostgresSQL dump files of the RDB are available from our data archive, while we also provide software to automatically maintain an up-to-date local copy of the RDB. We provide multiple interfaces for the PDBj Mine service. First is the quick search interface, which is a basic keyword search tool, but additional filtering for common queries can be applied. Second is the RDB search interface to directly search all aspects of the RDB, which can either be used via a graphical interface or via a simple SQL interface. Although these direct RDB search interfaces are much more powerful, knowledge of both the data structure (as described in the mmCIF dictionary), as well as SQL query syntax is required. Here, the graphical interface only requires users to have knowledge of the mmCIF data structure, although those unfamiliar with the mmCIF data structure can use the included help interface to search for data categories/tables instead. Therefore, by using the graphical interface, filtering using any data that is part of the PDB can be performed without requiring any SQL knowledge, but for more complex queries beyond filtering, such as alternate data extraction or for very complex filtering, the text-based SQL-query based approach is still more powerful. These search interfaces are also queryable via our REST services as described on our help page, https://pdbj.org/help/rest-interface.

To explore individual PDB entries, we provide the Mine web interface, which describes and explains various aspects of released PDB structures. Our molecular viewer, Molmil (Bekker et al. 2016), is used to provide various visualizations of a structure, including the asymmetric unit, biological unit, and electron density maps, if available. Recently, we introduced a 2D interactive representation of the structure topology based on the intermolecular contact data saved in the RDB and the characteristics of the chain (Figure 1). This representation shows a simplified version of the interactions between the various molecules in the PDB entry, to help understand the organization of molecules within the entry. Here, the chains are colored by their entity, for easier mapping between the structural details table (Figure 1b) and the topological mapping (Figure 1a).



	Chain ID	Description	Type	Chain length	Formula weight	Number of molecules
1	A, B, C (A, B, C)	siphonaxanthin chlorophyll a/b binding light-harvesting complex II	polymer	223	24190.1	3
2	AA, BA, CA, D, DA... (B, A, C)	CHLOROPHYLL B	non-polymer		907.5	24
3	EA, F, FA, G, GA... (B, A, C)	CHLOROPHYLL A	non-polymer		893.5	15
4	HA, Q, YA (B, A, C)	(1R,3R)-6-{(3E,5E,7E,9E,11E,13E,15E,17E)-18-[(1S,4R,6R)-4-HYDROXY-2,2,6-TRIMETHYL-7-OXABICYCLO[4.1.0]HEPT-1-YL]-3,7,12,16-TETRAMETHYLOCTADECA-1,3,5,7,9,11,13,15,17-NONAENYLIDENE}-1,5,5-TRIMETHYLCYCLOHEXANE-1,3-DIOL	non-polymer		600.9	3
5	IA, R, ZA (B, A, C)	Siphonein	non-polymer		781.2	3
6	AB, JA, S (C, B, A)	Siphonaxanthin	non-polymer		600.9	3
7	BB, KA, T (C, B, A)	1,2-DIPALMITOYL-PHOSPHATIDYL-GLYCEROLE	non-polymer		723.0	3
8	CB, DB, EB (A, B, C)	water	water		18.0	54

FIGURE 1 Enhancements to PDBj Mine. (a) New chain topology representation. Shown are the molecules within the entry (here, PDB ID 7WLM), with each entity uniquely colored, sized based on their mass and the visualization style depending on the molecule type. Here, proteins are represented as circles, RNA as squares, DNA as pentagons, glucans as hexagons, ions as stars and other compounds as triangles. (b) Part of the entity panel on Mine Structural details page. Here, the entities are colored to match the colors of the molecules in the chain topology representation.

Clicking on the proteins or chemical compounds in the topology viewer opens the new UniProt portal entry page for the protein or (see also below) or the corresponding Chemie page. For EM entries, we also provide links to the DAQ-Score Database, which provides a quality assessment of EM-derived structures (Nakamura et al. 2023; Terashi et al. 2022). We also expanded the interface describing additional experimental details of crystallography-derived structures (Figure S1, Supporting Information), by employing meta-data that was newly added to the mmCIF data. Recently, structures derived via integrated/hybrid methods were added to the PDB archive as the PDB-IHM, which are also visualized via our Mine interface in a similar manner to the regular PDB entries. With Chemie, we provide a search interface to the chemical compound dictionary data part of the PDB. A search interface is provided to search and filter the compound library like PDBj Mine for PDB entries, and individual entry pages are also made available, providing information about the chemical structure, visualization using Molmil and links to PDB entries that contain the chemical compound, with finally a similar interface also provided for PRD/BIRD (Biologically Interesting Molecule Reference Dictionary) entries.

We have also developed several services to explore and analyze 3DEM structures. In 2007, we started our EM Navigator service (Kinjo et al. 2012), a website to explore 3DEM data in the EMDB and PDB. The EM Navigator service produced short movies that stored representations of the 3DEM data from different orientations to help users to visualize and understand the models. The Omokage service was developed as a shape similarity search service for 3D structures of macromolecules that compares the overall shape between registered structures or a user-submitted one (Suzuki et al. 2016). The gmfit service also works on EM data and can be used to quickly fit 3D objects (either structures or density maps) using Gaussian mixture models (Kawabata 2008).

We have also developed several services to perform sequence-based analyses. To enable sequence homology searches within the PDB, we provide the Sequence Navigator service, which enables searching the PDB for homologous structures given a query sequence. Similarly for existing PDB entries, our Sequence Neighbor service enables searching for homologous structures and visualizing their superposed structures using our molecular viewer Molmil. The CRNPRED service can be used to predict characteristics of a protein such as secondary structure, contact numbers and residue-wise contact orders from the amino acid sequence (Kinjo and Nishikawa 2006). While CRNPRED uses the amino acid sequence to predict structural properties, our HOMCOS service can be used to model the quaternary structure of proteins based on homology modeling (Kawabata 2016). In addition, it can also be used to search for potential binding compounds given an amino acid sequence, or a set of binding proteins given a compound.

Molmil is a WebGL-based molecular viewer that we have been developing since 2013 and is used by PDBj for various services (Bekker et al. 2016; Bekker et al. 2022; Kinjo et al. 2017; Kinjo et al. 2018). Molmil can also be

used as a standalone viewer to load user-provided structures and MD trajectory files, without requiring any installation. We also provide an installable version called molmil-app that is also available to enable shell-based loading and headless processing (our images of PDB structures shown in the Mine PDB search and PDB entry pages are generated in this manner). During the development of Molmil, we also developed a new format called PDBx/mmJSON, which uses the same definitions (and dictionary) as PDBx/mmCIF, but encodes the data in a JSON format, which can be read by any modern programming language, without requiring a custom mmCIF parser. Furthermore, compressed mmJSON files were found to be on average 33% smaller than their compressed mmCIF counterparts (Bekker et al. 2016), leading to reduced storage, transfer and loading times. We also provide a REST service that generates files in mmJSON format for selected categories for an entry.

To help depositors grow accustomed to the mmCIF format, we have created an mmCIF editor (Bekker et al. 2019a; Bekker et al. 2022), and due to its generalized implementation, the CIF editor is also used by our archives BSM-Arc and XRDa to register and modify metadata during deposition. Like Molmil, it runs inside a web browser and does not require any installation, ensuring that users will always use the most recent version, without having to wait to install an update before every use. The editor supports two modes; a UI based mode and a manual mode that allows users to directly edit the raw mmCIF data. The mmCIF editor can load local files from the hard drive, while the modified data can finally be saved back to an mmCIF file or to an mmJSON file. To enable users familiar with the mmCIF format to manually edit the content in a similar way as they might have done with the old flat file PDB format, the editor also supports a RAW editing mode. After switching to the RAW editing mode, users can freely edit the mmCIF data manually, after which the editor will re-assimilate the modified content while validating it against the mmCIF dictionary.

4 | DYNAMICS DB

We have also created a new archive of protein stability derived from an analysis of MD simulations. Contact matrix analysis of high temperature MD simulations were previously shown to show a good correlation to experimental Tm values (Bekker et al. 2019b), as well as for binding simulations (Bekker and Kamiya 2022). Here, this analysis was performed on a subset of the PDB (9562 entries) and the data was stored in the newly created archive. We developed an interface to visualize the per-residue stability, which is shown in Figure 2. Here, in the top section, a graph that shows the stability (represented by the *R*-value) of each of the

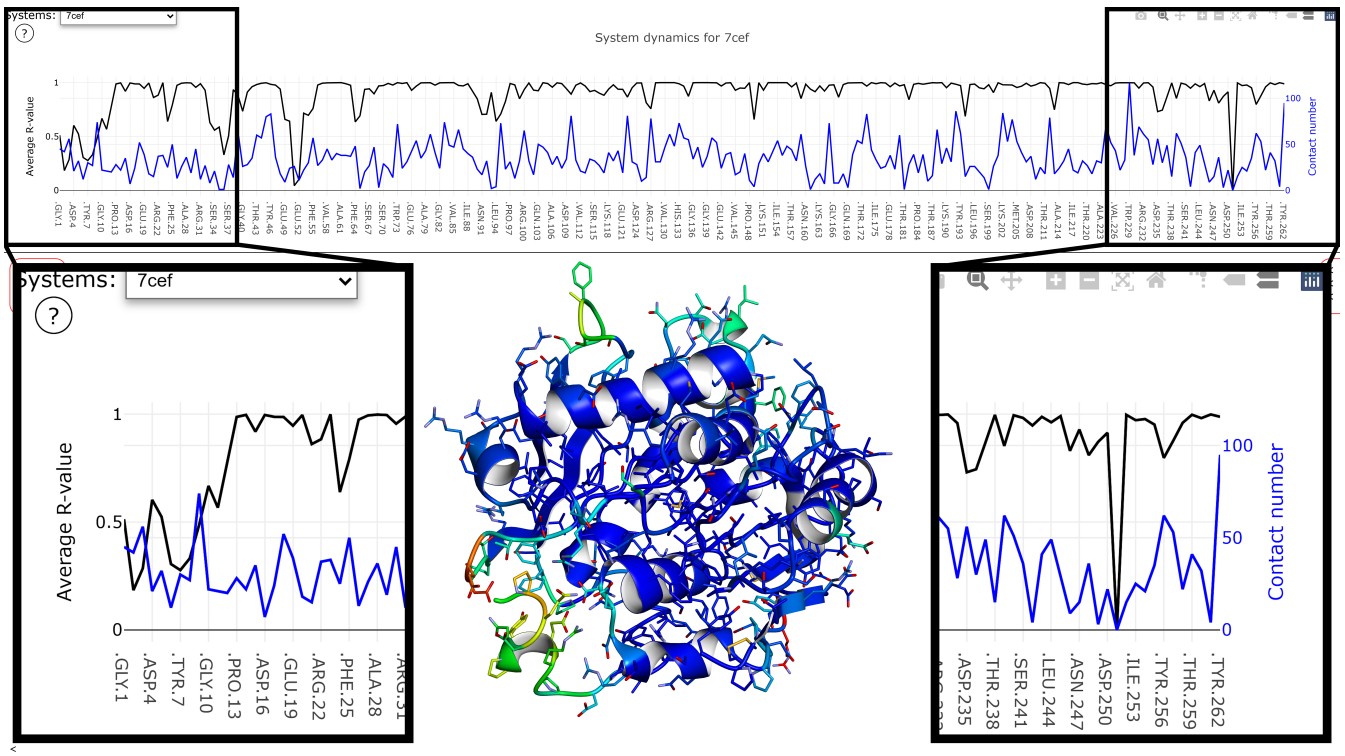


FIGURE 2 Example of dynamics DB entry. In the top section, the stability of the structure along the sequence is shown in terms of the *R*-value (black), and the number of contacts made by each residue in the representative structure (blue). In the bottom section, the representative structure is shown and visualized by our molecular viewer Molmil, with the cartoon and carbons colored based on the stability score (*R*-value). In addition, zoomed representations of the top section's left and right side are shown as insets besides the structure figure.

residues is shown (black) along with the contact number for each residue (blue). The *R*-value is a measure of the stability of the contacts measured during the MD simulations with respect to a representative structure, which was also calculated from the MD simulations and corresponds to the structure whose contact matrix is the closest to the average contact matrix. In the bottom section, Molmil is used to visualize the stability of the structure. The representative structure is shown, which is colored (cartoon and carbon atoms) based on the stability, where blue corresponds to stable residues (*R*-value = 1.0) and red to unstable residues (*R*-value = 0.0). From the PDBj Mine page, there are links available to the dynamics DB page for the entries that have been analyzed. In addition, the full list of PDB entries that have been analyzed is available from https://pdbj.org/urls/dynamicsdb.

5 | SEQUENCE EXPLORATION FOR DATA-OUT

PDBj organizes multiple luncheon seminars yearly throughout Asia, primarily in Japan. During the past 2 years, we have used questionnaires at these luncheon seminars to poll the scientific community regarding their usage of our database and requirements. Overwhelmingly, integration between our structural data and external sequence databases were among the top uses/requests (Figure 3). For example, bioinformaticians indicated that they combine multiple databases, and they wish to be able to easily combine protein structural data with genome-, variation-, pathway-, or chemical compound-data. Furthermore, users often had difficulty selecting a structure when many other structures of the same protein exist. Therefore, we set out to create a new service to tackle these issues. At its core, we wanted to create a protein portal page, which would enable users to explore various aspects of the target protein and help them decide which PDB structure would be the most useful in their endeavors.

The largest sequence database of protein sequences is the UniProt Knowledgebase (KB), which comprises of the UniProtKB/Swiss-Prot component for a reviewed and manually annotated protein dataset and the UniProtKB/Trembl component for an unreviewed and largely computationally annotated protein dataset (The UniProt Consortium et al. 2023). Combined, they represent a humongous dataset of proteins, consisting of almost 250 million proteins. Furthermore, the database that was the most frequently used/requested in our survey was UniProt. Therefore, we constructed our new portal page using the UniProt IDs to link them with the structures in the PDB archive.

In addition to the UniProt protein data, we have also integrated two genomic databases; Japanese Multi-Omics Reference Panel (jMorp) and Medical Genomics Japan Variant Database (MGeND). The jMorp database

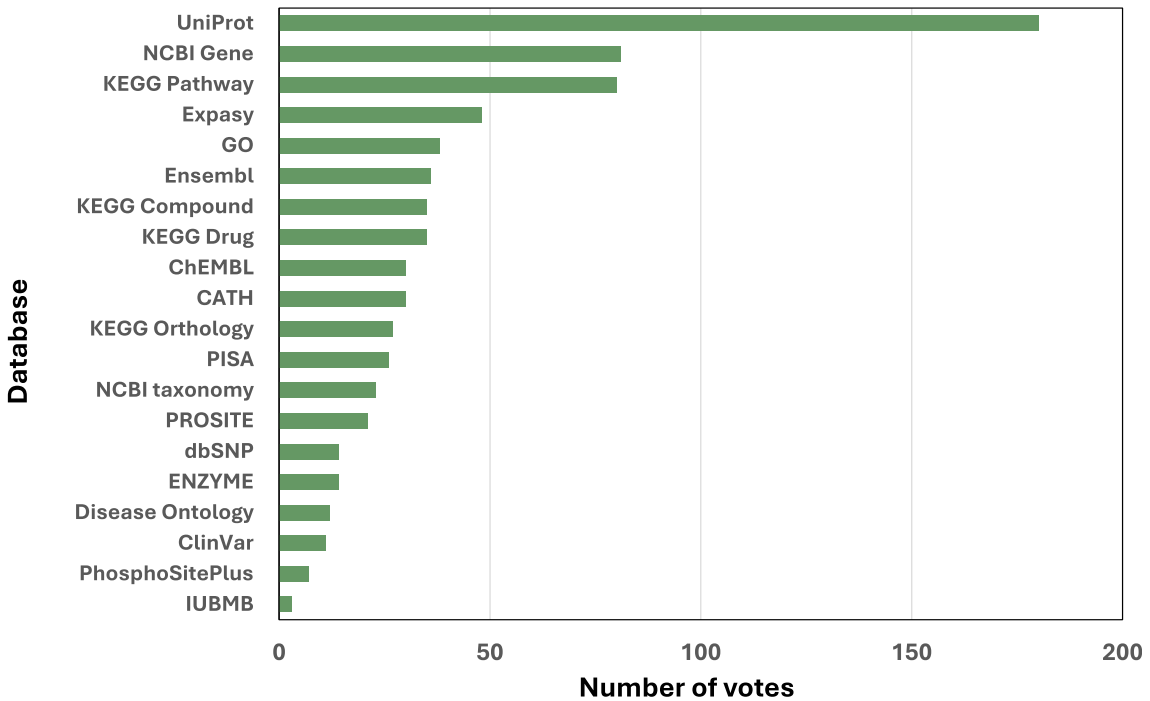


FIGURE 3 Databases the users have used and would like to use with PDB data. Combined results from 264 luncheon seminar participants at the annual meetings of the Protein Science Society of Japan, Informatics In Biology, Medicine and Pharmacology, the Crystallographic Society of Japan, and the Biophysical Society of Japan (including multiple responses).

is a secondary database produced from the analysis of data registered in the Tohoku Medical Megabank (TMM) (Tadaka et al. 2024) and provides a statistical overview of the genetic diversity of the TMM Cohort. MGeND (Kamada et al. 2019) is a curated database of genetic mutations that are involved in clinical observations. We have mapped the genomic data from jMorp and MGeND onto the corresponding UniProt entries, which subsequently allows them to be mapped to the corresponding PDB structures and we thereby integrated both resources into our new service. Currently, 3,700,745 jMorp entries are stored, in addition to 52,835 MGeND entries. These map to 64,709 and 18,335 PDB entries for jMorp and MGeND, respectively, with 65,054 PDB entries having either jMorp or MGeND entries.

The new service integrates both sequence data from UniProt, as well as structural data from the PDB. Figure 4 shows an example of the portal page for entry P20813. Here, in the right corner, the UniProt sequence is shown using our sequence viewer applet (Bekker et al. 2022; Kinjo et al. 2017; Kinjo et al. 2018). For each entry, a template PDB chain is shown from the PDB in the left interface panel in Figure 4, visualized using our molecular viewer, Molmil. However, if no structure from the PDB is available, a computational structure obtained from AlphaFold DB (AFDB, https://alphafold.ebi.ac.uk/) is shown instead (if available). In case a PDB structure was used, the secondary structure assignments from the PDB structure are used and shown in the sequence viewer panel. Otherwise, secondary structure assignments calculated by Molmil from the AFDB structure are shown. Also shown along the sequence are known ligand binding and glycosylation sites recorded in the UniProt entry, as well as the locations of the genomic variation recorded in the UniProt entry or obtained from jMorp and MGeND. This enables users to quickly see whether there is any known genomic variation at their site of interest, as well as whether there are any known structural features. Clicking on any of the residues within the sequence viewer will cause Molmil to jump to and show that residue, if available in the structure, to see where a mutation occurs in the structure and evaluate the impacts it might have on the structure and dynamics. For PDB templates, if there is a compatible entry available from one of our secondary structure archives Promode-Elastic,

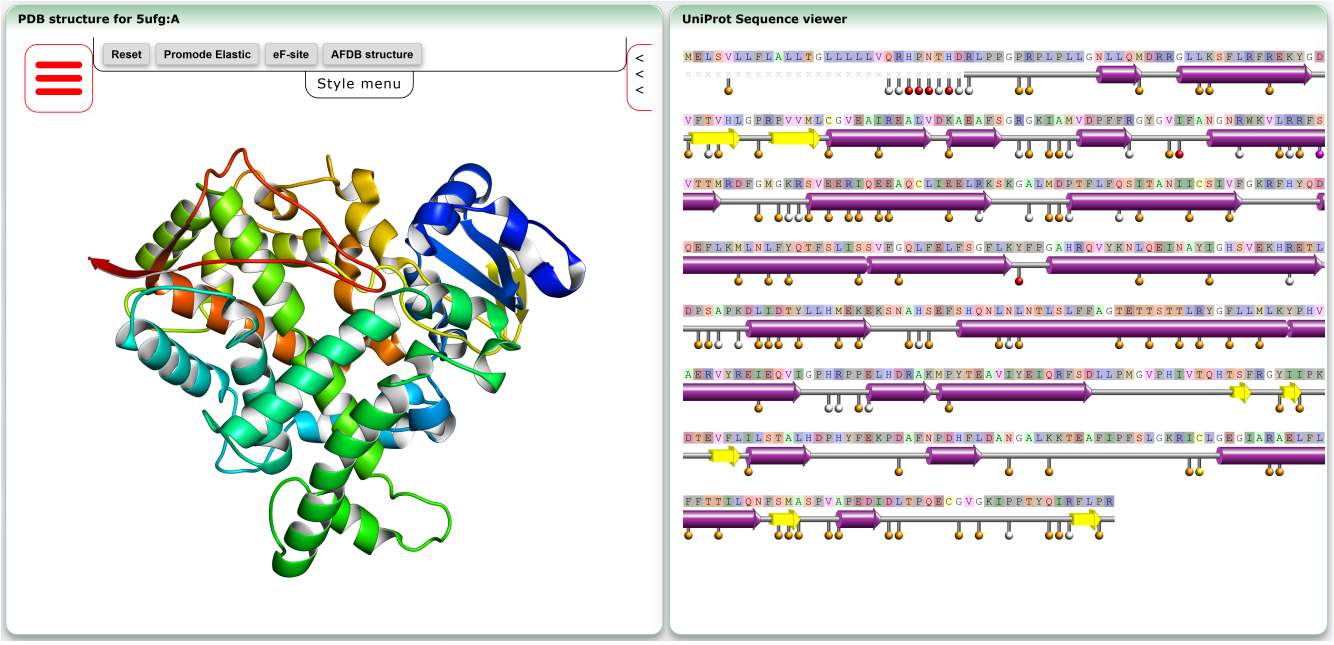


FIGURE 4 Structure and sequence visualization of the new UniProt portal entry page. As an example, the entry page corresponding to P20813 is shown (https://pdbj.org/uniprot/P20813). (a) Structure and sequence visualization. In the left interface panel (PDB structure for 5ufg: A), the representative structure is shown, colored in a blue-red gradient along the N- to C-terminus. For an AFDB structure, the structure is instead colored by its confidence score, called pLDDT. For PDB templates, if additional data is available from our secondary archives Promode-Elastic, eF-site, and Dynamics DB, buttons to toggle the visualization of these resources are shown. In addition, if DAQ scores are available, or an AFDB structure is available, buttons to toggle visualization for these resources is also shown. In the right interface panel (UniProt Sequence viewer), the sequence of the UniProt entry is shown, with the secondary structure taken from the PDB structure or AFDB structure (depending on which is shown in the structure panel). Residues that are not part of the PDB entry are indicated by gray crosses, while residues that are part of the PDB entry (i.e., expressed), but were not observed, are shown as red crosses. In addition, various sites from the UniProt entry or from external resources are indicated along the sequence. Here, orange circles correspond to mutation sites, mutations in the PDB structure correspond to red circles, sites involved in covalent bonds correspond to green circles, interaction sites correspond to blue circles, glycosylation sites correspond to cyan circles, binding sites correspond to yellow circles, post-translational modification sites correspond to magenta circles and mixed sites correspond to white circles.

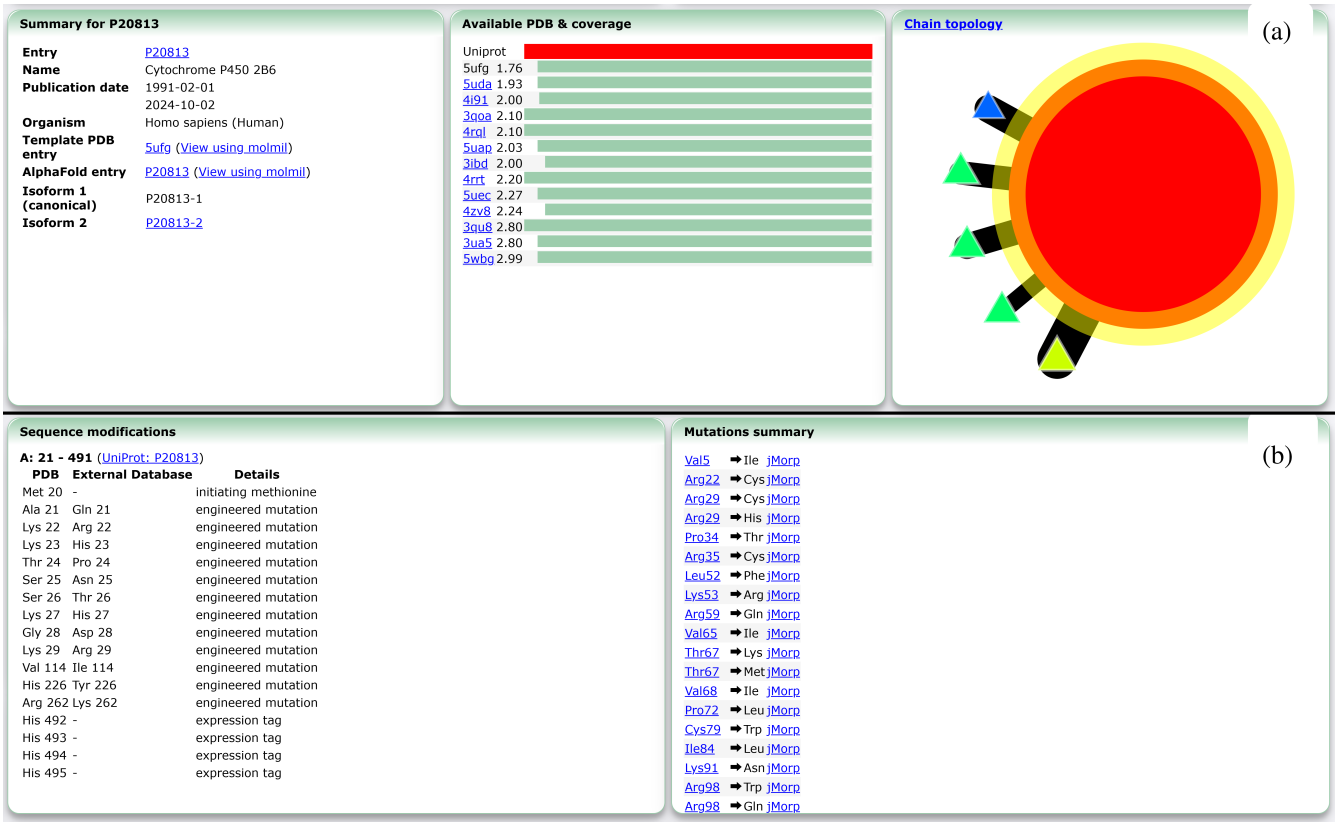


FIGURE 5 Detailed information panels of the new UniProt portal entry page. (a) Entry summary, PDB coverage, and chain topology representation. The summary panel on the left side shows basic information regarding the entry, as well as links to either the PDBj Mine page or the AFDB page of the template structure. If any isoforms are present, links to the UniProt portal entries of these isoforms are also provided. The Available PDB and coverage panel (center) lists all available PDB structures, ordered by their resolution and sequence coverage relative to the UniProt sequence. The Chain topology panel (right) shows a simplified rendering of the template PDB structure, with the chains corresponding to the UniProt entry indicated by a yellow halo. (b) Sequence modifications and mutations panels. If any modifications in the template PDB structure exist with respect to the UniProt sequence, the Sequence modifications panel is shown (left). If there are any mutations recorded in either jMorp or MGeND, these are shown in the Mutations summary panel (right). If in either case no such sites are present, the corresponding panel is not shown.

Dynamics DB, or eF-site. In addition, the AlphaFold structure (if available) can also be co-visualized, as well as coloring the structure based on the DAQ-Score, if available, to provide additional context and information.

In addition, several panels that provide detailed information are shown (Figure 5a). The UniProt summary panel shows a summary of the entry, including links to the AFDB structure page, if available, and it lists any links for isoform entry pages, if available. The Available PDB & Coverage panel lists any PDB entries that are available from the PDB for this UniProt entry with their resolution, as well as provides a visual indicator to how much of the UniProt sequence is covered by each of the PDB entries. Using this information, that is, the resolution of the PDB entries and their sequence coverage, a representative PDB structure is chosen as the template structure, although other structures, including the AFDB structure, can also be used as a template structure. For each PDB structure, a score based on the number of residues that overlap between the UniProt and PDB entry, divided by the resolution of the PDB entry is used to sort the entries list, where the top-ranking one is considered as the representative template PDB structure. The simplified intermolecular representation of the topology, which was described above for the Mine PDB entry pages, is also shown, with the chains corresponding to the UniProt entry indicated by a yellow halo. Finally, the PDB structure summary panel and the structure validation panel are shown for the selected PDB template structure (if available, Figure 5b).

For this new service, we have developed several data files. Like our other databases, we also distribute the data files for this new service. As the JSON files used by UniProt KB are not publicly distributed, we have reverse-engineered the format and generated JSON files for all entries, to both serve our own files locally, as well as to maintain compatibility with the original UniProt KB files. We have made a tarfile available of the individually compressed JSON files via our data archive (https://data.pdbj.org/uniprot/uniprot-json-latest.tar, also accessible via our rsync service). In addition, we have also placed the JSON files containing the

mapped mutations obtained from jMorp and MGeND on our archive (https://data.pdbj.org/uniprot/mutations-json/). The portal pages themselves can be linked to via the following URL: https://pdbj.org/uniprot/$UPID/$PDBID. Here, $UPID corresponds to the UniProt ID of the entry and $PDBID to the selected template PDB ID. If $PDBID is omitted, a representative PDB entry is selected based on resolution and coverage. Alternatively, $PDBID can be set to "afdb," to link to the AFDB structure template page. Finally, residues in Molmil can be pre-selected by appending the "select" parameter to the URL. Here, ranges are supported ("from-to") as well as multiple sets (e.g., "30–35, 40–45"; e.g., https://pdbj.org/uniprot/P20813?select=30-35,40-45). Finally, the mutation data has also been loaded into our Mine2 RDB, with the data stored in the misc.mutations table (https://pdbj.org/rdb/search?query=select+*+from+misc.mutations).

6 | SEQUENCE EXPLORATION FOR DATA-IN

Another new tool we recently introduced is the Sequence Navigator Pro service (Bekker et al. 2024). We have provided the Sequence Navigator service for many years to perform homology-based PDB entry searches. However, it is limited in terms of analyzing the detailed characteristics of a query sequence and the homologues discovered. Although the Sequence Navigator Pro service still performs a homology search against the PDB, it also analyzes the sequence in several other ways and packages up the results in a more usable manner. First, it also performs a homology search against the SwissProt KB sequence archive and the AFDB computational structure archive. The service then lists the matches in terms of their sequence coverage like the coverage panel in our new UniProt-based service described above (Figure S1a). Furthermore, a panel that describes the experimental details is shown, for a quick overview of the characteristics of the matching structures, as well as any present complexed ligands (Figure S1b). We also perform some predictions based on the sequence, where we predict the secondary structure using s4pred (Moffat and Jones 2021), predict disordered regions using flDPnn (Hu et al. 2021) and predict the hydropathy based on the sequence (Figure S1c). Finally, we provide an easy way to then perform keyword search against the linked literature (PubMed) of the discovered homologous PDB and SwissProt entries (Figure S1d,e). The tool can be used to provide insights into already existing structures in the PDB or AFDB, but can also be used by experimentalists to provide insights into what experimental conditions are the most likely to result in good experimental data.

7 | CONCLUSION

PDBj has developed and updated several original tools to help users to find/access/interoperate/reuse the PDB/BMRB/EMDB entries. In addition, PDBj has developed several novel archives for experimental or computationally derived data. This data has all been integrated into the PDBj website, which proves a vast amount of data. To sift through this amount of data we have created several low and mid-level tools. However, since the PDB consists of many structures, it can sometimes be challenging to find the most suitable structure for a given protein. Our new sequence-oriented services will help users identify the most suitable structure for their protein, be it an experimental structure or a computationally derived one. In addition, integration with external genomic resources provides insight into the genomic and potentially structural variability of the proteins. Finally, integration with our secondary archives provides additional information with respect to the properties of the proteins.

AUTHOR CONTRIBUTIONS

Gert-Jan Bekker: Software; writing – original draft; writing – review and editing; methodology; conceptualization; data curation; visualization; investigation; resources. **Chioko Nagao:** Writing – review and editing; conceptualization; methodology; data curation; investigation; formal analysis. **Matsuyuki Shirota:** Writing – review and editing; investigation. **Tsukasa Nakamura:** Writing – review and editing; investigation. **Toshiaki Katayama:** Writing – review and editing; investigation; conceptualization. **Daisuke Kihara:** Writing – review and editing; investigation; supervision. **Kengo Kinoshita:** Writing – review and editing; investigation; supervision. **Genji Kurisu:** Supervision; project administration; conceptualization; writing – review and editing; methodology; data curation; funding acquisition; investigation; resources.

ACKNOWLEDGMENTS

This work was supported by grants from the Database Integration Coordination Program (JPMJND2205) from the department of National Bioscience Database Center (NBDC)-JST (Japan Science and Technology Agency), and partially supported by Platform Project for Supporting Drug Discovery and Life Science Research (Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS)) from AMED under Grant Numbers JP24ama121001 and JP23ama121019. In addition, it was supported by JSPS KAKENHI grants JP20H03229 and JP21K17847, as well as by the National Institutes of Health (R01GM133840).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

ORCID

Gert-Jan Bekker https://orcid.org/0000-0001-8385-5693
Chioko Nagao https://orcid.org/0000-0002-7721-0642
Matsuyuki Shirota https://orcid.org/0000-0002-7776-9964
Tsukasa Nakamura https://orcid.org/0000-0002-6312-3070
Toshiaki Katayama https://orcid.org/0000-0003-2391-0384
Daisuke Kihara https://orcid.org/0000-0003-4091-6614
Kengo Kinoshita https://orcid.org/0000-0003-3453-2171
Genji Kurisu https://orcid.org/0000-0002-5354-0807

REFERENCES

Bekker G-J, Kamiya N. Advancing the field of computational drug design using multicanonical molecular dynamics-based dynamic docking. Biophys Rev. 2022;14:1349–58.

Bekker G-J, Kawabata T, Kurisu G. The Biological Structure Model Archive (BSM-Arc): an archive for in silico models and simulations. Biophys Rev. 2020;12:371–5.

Bekker G-J, Kudou T, Ikegawa Y, Yamashita R, Kurisu G. PDBx/mmCIF format mandatory for Protein Data Bank deposition. Nihon Kessho Gakkaishi. 2019a;61:159–60.

Bekker G-J, Ma B, Kamiya N. Thermal stability of single-domain antibodies estimated by molecular dynamics simulations. Protein Sci. 2019b;28:429–38.

Bekker G-J, Nagao C, Shirota M, Nakamura T, Katayama T, Kihara D, et al. Protein Data Bank Japan: Computational resources for analysis of protein structures. 2024; submitted.

Bekker G-J, Nakamura H, Kinjo AR. Molmil: a molecular viewer for the PDB and beyond. J Chem. 2016;8:42.

Bekker G-J, Yokochi M, Suzuki H, Ikegawa Y, Iwata T, Kudou T, et al. Protein Data Bank Japan: celebrating our 20th anniversary during a global pandemic as the Asian hub of three dimensional macromolecular structural data. Protein Sci. 2022;31:173–86.

Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, Costanzo LD, et al. Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res. 2019;47:D520–8.

Hoch JC, Baskaran K, Burr H, Chin J, Eghbalnia HR, Fujiwara T, et al. Biological Magnetic Resonance Data Bank. Nucleic Acids Res. 2023;51:D368–76.

Hu G, Katuwawala A, Wang K, Wu Z, Ghadermarzi S, Gao J, et al. flDPnn: accurate intrinsic disorder prediction with putative propensities of disorder functions. Nat Commun. 2021;12:4438.

Kamada M, Nakatsui M, Kojima R, Nohara S, Uchino E, Tanishima S, et al. MGeND: an integrated database for Japanese clinical and genomic information. Hum Genome Var. 2019;6:53.

Kawabata T. Multiple subunit fitting into a low-resolution density map of a macromolecular complex using a Gaussian mixture model. Biophys J. 2008;95:4643–58.

Kawabata T. HOMCOS: an updated server to search and model complex 3D structures. J Struct Funct Genomics. 2016;17:83–99.

Kinjo AR, Bekker G-J, Suzuki H, Tsuchiya Y, Kawabata T, Ikegawa Y, et al. Protein Data Bank Japan (PDBj): updated user interfaces, resource description framework, analysis tools for large structures. Nucleic Acids Res. 2017;45:D282–8.

Kinjo AR, Bekker G-J, Wako H, Endo S, Tsuchiya Y, Sato H, et al. New tools and functions in data-out activities at Protein Data Bank Japan (PDBj). Protein Sci. 2018;27:95–102.

Kinjo AR, Nishikawa K. CRNPRED: highly accurate prediction of one-dimensional protein structures by large-scale critical random networks. BMC Bioinformatics. 2006;7:401.

Kinjo AR, Suzuki H, Yamashita R, Ikegawa Y, Kudou T, Igarashi R, et al. Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format. Nucleic Acids Res. 2012;40:D453–60.

Kinoshita K, Nakamura H. eF-site and PDBjViewer: database and viewer for protein functional sites. Bioinformatics. 2004;20:1329–30.

Kurisu G, Bekker G-J, Nakagawa A. History of Protein Data Bank Japan: standing at the beginning of the age of structural genomics. Biophys Rev. 2022;14:1233–8.

Moffat L, Jones DT. Increasing the accuracy of single sequence prediction methods using a deep semi-supervised learning framework. Bioinformatics. 2021;37:3744–51.

Nakamura T, Wang X, Terashi G, Kihara D. DAQ-Score Database: assessment of map–model compatibility for protein structure models from cryo-EM maps. Nat Methods. 2023;20:775–6.

Suzuki H, Kawabata T, Nakamura H. Omokage search: shape similarity search service for biomolecular structures in both the PDB and EMDB. Bioinformatics. 2016;32:619–20.

Tadaka S, Kawashima J, Hishinuma E, Saito S, Okamura Y, Otsuki A, et al. jMorp: Japanese multi-omics reference panel update report 2023. Nucleic Acids Res. 2024;52:D622–32.

Terashi G, Wang X, Maddhuri Venkata Subramaniya SR, Tesmer JJG, Kihara D. Residue-wise local quality estimation for protein models from cryo-EM maps. Nat Methods. 2022;19:1116–25.

The UniProt Consortium, Bateman A, Martin M-J, Orchard S, Magrane M, Ahmad S, et al. UniProt: the universal protein knowledgebase in 2023. Nucleic Acids Res. 2023;51:D523–31.

Wako H, Kato M, Endo S. ProMode: a database of normal mode analyses on protein molecules with a full-atom model. Bioinformatics. 2004;20:2035–43.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Bekker G-J, Nagao C, Shirota M, Nakamura T, Katayama T, Kihara D, et al. Protein Data Bank Japan: Improved tools for sequence-oriented analysis of protein structures. Protein Science. 2025;34(3):e70052. https://doi.org/10.1002/pro.70052