<https://academic.oup.com/nar/article/45/D1/D271/2333880>

ABSTRACT :

The Protein Data Bank (PDB; http://www.rcsb.org/pdb/ ) is the single worldwide archive of structural data of biological macromolecules. This paper describes the goals of the PDB, the systems in place for data deposition and access, how to obtain further information, and near-term plans for the future development of the resource.

INTRODUCTION:

The Protein Data Bank (PDB) was established at Brookhaven National Laboratories (BNL) (1) in 1971 as an archive for biological macromolecular crystal structures. In the beginning the archive held seven structures, and with each year a handful more were deposited. In the 1980s the number of deposited structures began to increase dramatically. This was due to the improved technology for all aspects of the crystallographic process, the addition of structures determined by nuclear magnetic resonance (NMR) methods, and changes in the community views about data sharing. By the early 1990s the majority of journals required a PDB accession code and at least one funding agency (National Institute of General Medical Sciences) adopted the guidelines published by the International Union of Crystallography (IUCr) requiring data deposition for all structures. The mode of access to PDB data has changed over the years as a result of improved technology, notably the availability of the WWW replacing distribution solely via magnetic media. Further, the need to analyze diverse data sets required the development of modern data management systems. Initial use of the PDB had been limited to a small group of experts involved in structural research. Today depositors to the PDB have varying expertise in the techniques of X-ray crystal structure determination, NMR, cryoelectron microscopy and theoretical modeling. Users are a very diverse group of researchers in biology, chemistry and computer scientists, educators, and students at all levels. The tremendous influx of data soon to be fueled by the structural genomics initiative, and the increased recognition of the value of the data toward understanding biological function, demand new ways to collect, organize and distribute the data. In October 1998, the management of the PDB became the responsibility of the Research Collaboratory for Structural Bioinformatics (RCSB). In general terms, the vision of the RCSB is to create a resource based on the most modern technology that facilitates the use and analysis of structural data and thus creates an enabling resource for biological research. Specifically in this paper, we describe the current procedures for data deposition, data processing and data distribution of PDB data by the RCSB. In addition, we address the issues of data uniformity. We conclude with some current developments of the PDB.

DATA ACQUISITION AND PROCESSING:

A key component of creating the public archive of information is the efficient capture and curation of the data—data processing. Data processing consists of data deposition, annotation and validation. These steps are part of the fully documented and integrated data processing system shown in Figure 1. In the present system (Fig. 2), data (atomic coordinates, structure factors and NMR restraints) may be submitted via email or via the AutoDep Input Tool (ADIT; http://pdb.rutgers. edu/adit/ ) developed by the RCSB. ADIT, which is also used to process the entries, is built on top of the mmCIF dictionary which is an ontology of 1700 terms that define the macromolecular structure and the crystallographic experiment (2,3), and a data processing program called MAXIT (MAcromolecular EXchange Input Tool). This integrated system helps to ensure that the data submitted are consistent with the mmCIF dictionary which defines data types, enumerates ranges of allowable values where possible and describes allowable relationships between data values. After a structure has been deposited using ADIT, a PDB identifier is sent to the author automatically and immediately (Fig. 1, Step 1). This is the first stage in which information about the structure is loaded into the internal core database (see section on the PDB Database Resource). The entry is then annotated as described in the validation section below. This process involves using ADIT to help diagnose errors or inconsistencies in the files. The completely annotated entry as it will appear in the PDB resource, together with the validation information, is sent back to the depositor (Step 2). After reviewing the processed file, the author sends any revisions (Step 3). Depending on the nature of these revisions, Steps 2 and 3 may be repeated. Once approval is received from the author (Step 4), the entry and the tables in the internal core database are ready for distribution. The schema of this core database is a subset of the conceptual schema specified by the mmCIF dictionary. All aspects of data processing, including communications with the author, are recorded and stored in the correspondence archive. This makes it possible for the PDB staff to retrieve information about any aspect of the deposition process and to closely monitor the efficiency of PDB operations. Current status information, comprised of a list of authors, title and release category, is stored for each entry in the core database and is made accessible for query via the WWW interface (http://www.rcsb.org/pdb/status.html ). Entries before release are categorized as ‘in processing’ (PROC), ‘in depositor review’ (WAIT), ‘to be held until publication’ (HPUB) or ‘on hold until a depositor-specified date’ (HOLD).

Diagram

Description automatically generated

Validation:

Validation refers to the procedure for assessing the quality of deposited atomic models (structure validation) and for assessing how well these models fit the experimental data (experimental validation). The PDB validates structures using accepted community standards as part of ADIT’s integrated data processing system. The following checks are run and are summarized in a letter that is communicated directly to the depositor: Covalent bond distances and angles. Proteins are compared against standard values from Engh and Huber (5); nucleic acid bases are compared against standard values from Clowney et al. (6); sugar and phosphates are compared against standard values from Gelbin et al. (7). Stereochemical validation. All chiral centers of proteins and nucleic acids are checked for correct stereochemistry. Atom nomenclature. The nomenclature of all atoms is checked for compliance with IUPAC standards (8) and is adjusted if necessary. Close contacts. The distances between all atoms within the asymmetric unit of crystal structures and the unique molecule of NMR structures are calculated. For crystal structures, contacts between symmetry-related molecules are checked as well. Ligand and atom nomenclature. Residue and atom nomenclature is compared against the PDB dictionary (ftp://ftp.rcsb. org/pub/pdb/data/monomers/het\_dictionary.txt ) for all ligands as well as standard residues and bases. Unrecognized ligand groups are flagged and any discrepancies in known ligands are listed as extra or missing atoms. Sequence comparison. The sequence given in the PDB SEQRES records is compared against the sequence derived from the coordinate records. This information is displayed in a table where any differences or missing residues are marked. During structure processing, the sequence database references given by DBREF and SEQADV are checked for accuracy. If no reference is given, a BLAST (9) search is used to find the best match. Any conflict between the PDB SEQRES records and the sequence derived from the coordinate records is resolved by comparison with various sequence databases. Distant waters. The distances between all water oxygen atoms and all polar atoms (oxygen and nitrogen) of the macromolecules, ligands and solvent in the asymmetric unit are calculated. Distant solvent atoms are repositioned using crystallographic symmetry such that they fall within the solvation sphere of the macromolecule. In almost all cases, serious errors detected by these checks are corrected through annotation and correspondence with the authors. It is also possible to run these validation checks against structures before they are deposited. A validation server (http://pdb.rutgers.edu/validate/ ) has been made available for this purpose. In addition to the summary report letter, the server also provides output from PROCHECK (10), NUCheck (Rutgers University, 1998) and SFCHECK (11). A summary atlas page and molecular graphics are also produced. The PDB will continually review the checking methods used and will integrate new procedures as they are developed by the PDB and members of the scientific community.

Other data deposition centers:

The PDB is working with other groups to set up deposition centers. This enables people at other sites to more easily deposit their data via the Internet. Because it is critical that the final archive is kept uniform, the content and format of the final files as well as the methods used to check them must be the same. At present, the European Bioinformatics Institute (EBI) processes data that are submitted to them via AutoDep (http://autodep.ebi.ac.uk/ ). Once these data are processed they are sent to the RCSB in PDB format for inclusion in the central archive. Before this system was put in place it was tested to ensure consistency among entries in the PDB archive. In the future, the data will be exchanged in mmCIF format using a common exchange dictionary, which along with standardized annotation procedures will ensure a high degree of uniformity in the archival data. Structures deposited and processed at the EBI represent ~20% of all data deposited. Data deposition will also soon be available from an ADIT Web site at The Institute for Protein Research at Osaka University in Japan. At first, structures deposited at this site will be processed by the PDB staff. In time, the staff at Osaka will complete the data processing for these entries and send the files to the PDB for release.

The database architecture:

In recognition of the fact that no single architecture can fully express and efficiently make available the information content of the PDB, an integrated system of heterogeneous databases has been created that store and organize the structural data. At present there are five major components (Fig. 3): • The core relational database managed by Sybase (Sybase SQL server release 11.0, Emeryville, CA) provides the central physical storage for the primary experimental and coordinate data described in Table 1. The core PDB relational database contains all deposited information in a tabular form that can be accessed across any number of structures. The final curated data files (in PDB and mmCIF formats) and data dictionaries are the archival data and are present as ASCII files in the ftp archive. • The POM (Property Object Model)-based databases, which consist of indexed objects containing native (e.g., atomic coordinates) and derived properties (e.g., calculated secondary structure assignments and property profiles). Some properties require no derivation, for example, B factors; others must be derived, for example, exposure of each amino acid residue (13) or C contact maps. Properties requiring significant computation time, such as structure neighbors (14), are precalculated when the database is incremented to save considerable user access time. • The Biological Macromolecule Crystallization Database (BMCD; 15) is organized as a relational database within Sybase and contains three general categories of literature derived information: macromolecular, crystal and summary data. • The Netscape LDAP server is used to index the textual content of the PDB in a structured format and provides support for keyword searches. It is critical that the intricacies of the underlying physical databases be transparent to the user. In the current implementation, communication among databases has been accomplished using the Common Gateway Interface (CGI). An integrated Web interface dispatches a query to the appropriate database(s), which then execute the query. Each database returns the PDB identifiers that satisfy the query, and the CGI program integrates the results. Complex queries are performed by repeating the process and having the interface program perform the appropriate Boolean operation(s) on the collection of query results. A variety of output options are then available for use with the final list of selected structures. The CGI approach [and in the future a CORBA (Common Object Request Broker Architecture)-based approach] will permit other databases to be integrated into this system, for example extended data on different protein families. The same approach could also be applied to include NMR data found in the BMRB or data found in other community databases.

Database query:

Three distinct query interfaces are available for the query of data within PDB: Status Query (http://www.rcsb.org/pdb/status.html ), SearchLite (http://www.rcsb.org/pdb/searchlite.html ) and SearchFields (http://www.rscb.org/pdb/queryForm.cgi ). Table 3 summarizes the current query and analysis capabilities of the PDB. Figure 4 illustrates how the various query options are organized. SearchLite, which provides a single form field for keyword searches, was introduced in February 1999. All textual information within the PDB files as well as dates and some experimental data are accessible via simple or structured queries. SearchFields, accessible since May 1999, is a customizable query form that allows searching over many different data items including compound, citation authors, sequence (via a FASTA search; 16) and release or deposition dates. Two user interfaces provide extensive information for result sets from SearchLite or SearchFields queries. The ‘Query Result Browser’ interface allows for access to some general information, more detailed information in tabular format, and the possibility to download whole sets of data files for result sets consisting of multiple PDB entries. The ‘Structure Explorer’ interface provides information about individual structures as well as cross-links to many external resources for macromolecular structure data (Table 4). Both interfaces are accessible to other data resources through the simple CGI application programmer interface (API) described at http://www. rcsb.org/pdb/linking.html

The website usage has climbed dramatically since the system was first introduced in February 1999 (Table 5). As of November 1, 1999, the main PDB site receives, on average, greater than one hit per second and greater than one query per minute.

CURRENT DEVELOPMENTS:

In the coming months, the PDB plans to continue to improve and develop all aspects of data processing. Deposition will be made easier, and annotation will be more automated. In addition, software for data deposition and validation will be made available for in-laboratory use. The PDB will also continue to develop ways of exchanging information between databases. The PDB is leading the Object Management Group Life Sciences Initiative’s efforts to define a CORBA interface definition for the representation of macromolecular structure data. This is a standard developed under a strict procedure to ensure maximum input by members of various academic and industrial research communities. At this stage, proposals for the interface definition, including a working prototype that uses the standard, are being accepted. For further details refer to http://www.omg.org/cgi-bin/doc?lifesci/ 99-08-15 . The finalized standard interface will facilitate the query and exchange of structural information not just at the level of complete structures, but at finer levels of detail. The standard being proposed by the PDB will conform closely to the mmCIF standard. It is recognized that other forms of data representation are desirable, for example using eXtensible Markup Language (XML). The PDB will continue to work with mmCIF as the underlying standard from which CORBA and XML representations can be generated as dictated by the needs of the community. The PDB will also develop the means and methods of communications with the broad PDB user community via the Web. To date we have developed prototype protein documentaries (19) that explore this new medium in describing structure– function relationships in proteins. It is also possible to develop educational materials that will run using a recent Web browser (20). Finally it is recognized that structures exist both in the public and private domains. To this end we are planning on providing a subset of database tools for local use. Users will be able to load both public and proprietary data and use the same search and exploratory tools used at PDB resources. The PDB does not exist in isolation, rather each structure represents a point in a spectrum of information that runs from the recognition of an open reading frame to a fully understood role of the single or multiple biological functions of that molecule. The available information that exists on this spectrum changes over time. Recognizing this, the PDB has developed a scheme for the dynamic update of a variety of links on each structure to whatever else can be automatically located on the Internet. This information is itself stored in a database and can be queried. This feature will appear in the coming months to supplement the existing list of static links to a small number of the more well known related Internet resources.

<https://nvlpubs.nist.gov/nistpubs/jres/101/3/j3abol.pdf>

The Protein Data Bank: Current Status and Future Challenges:

The Protein Data Bank (PDB) is an archive of experimentally determined three-dimensional structures of proteins, nucleic acids, and other biological macromolecules [1, 2]. PDB has a 25 year history of service to a global community of researchers, educators, and students in a variety of scientific disciplines [3]. The common interest shared by this community is a need to access information that can relate the biological functions of macromolecules to their three-dimensional structure. PDB is now being replaced by the 3DB, Three-Dimensional Database of Biomolecular Structures, which will continue to operate from Brookhaven National Laboratory.

The challenge facing the new 3DB is to keep abreast of the increasing flow of data, to maintain the archives as error-free as possible, and to organize and present this information in ways that facilitate data retrieval, knowledge exploration, and hypothesis testing without interrupting current services. The PDB introduced substantial enhancements to both data management and archive access in the past two years, and is well on the way to converting to a more powerful system that combines the advantages of object oriented and relational database systems. 3DB will transform PDB from a data bank serving solely as a data repository into a highly sophisticated knowledge-based system for archiving and accessing structural information. The process will be evolutionary, insulating users from drastic changes and providing both a high degree of compatibility with existing software and a consistent user interface for casual browsers. Development is under way for 3DB to operate as a direct-deposition archive. Mechanisms are provided for depositors to submit data with minimal staff intervention. Data archived in 3DB is managed using the Relational Database Management System (RDBMS) from SYBASE1 [4]. The new database (3DBase) is being developed with a view towards being a member of a federation of biological databases. Collaborative international centers are also being established to assist in data deposition, archiving, and distribution activities.

Resource Status—1995:

Rapid developments in preparation of crystals of macromolecules and in experimental techniques for structure analysis have led to a revolution in structural biology. These factors have contributed significantly to an enormous increase in the number of laboratories performing structural studies of macromolecules to atomic resolution. Advances include: 1) recombinant DNA techniques that permit almost any protein or nucleic acid to be produced in large amounts; 2) faster and better x-ray detectors; 3) real-time interactive computer graphics systems, together with automated methods for structure determination and refinement; 4) synchrotron radiation, allowing the use of extremely tiny crystals, Multiple Wavelength Anomalous Dispersion (MAD) phasing, and time-resolved studies via Laue techniques; 5) NMR methods permitting structure determination of macromolecules solution; and 6) electron microscopy (EM) techniques, for obtaining high-resolution structures of two-dimensional crystals. These dramatic advances produced an abrupt transition from the linear growth of 15–25 new structures deposited per year in the PDB before 1987 to a rapid exponential growth reaching the current rate of approximately 25 deposits per week (Fig. 1). This rapid increase overwhelmed PDB staff resources and data processing procedures and, by mid-1993, a backlog of some 800 coordinate entries had accumulated. By January 1994 this backlog was eliminated by increased automation of processing and the addition of new staff. In all, more than 3000 of the nearly 4000 current PDB coordinate entries (approximately 75 %) have been processed since 1991. Table 1 is a summary of the contents of PDB. Present staff now keep abreast of the deposition rate with a timeline of three months from receipt to final archiving, which includes the time that the entry is with the depositor for checking. This timeline is comparable to the publication schedules of the fastest scientific journals.

Image of depositions of PDB every year

In the same period, the proliferation and increasing power of computers, the introduction of relatively inexpensive interactive graphics, and the growth of computer networks greatly increased the demand for access to PDB data (Fig. 2). The requirements of molecular biologists, drug designers, and others in academia and industry were often fundamentally different from those of crystallographers and computational chemists, who had been the major users of the PDB since the 1970s.

PDB entries are accessible by FTP, the World Wide Web (WWW), and on CD-ROM. PC users of the CDROM are provided with the browser, PDB-SHELL [5], built using the FoxPro RDBMS [6]. In addition to its browsing mechanisms, PDB-SHELL provides direct access to the public-domain molecular viewing program RasMol [7]. Recent enhancements to PDB’s WWW server (http://www.pdb.bnl.gov) have greatly improved the accessibility and utility of the archive over the Internet. This includes the release of PDBBrowse [8, 9], which is accessible through the World Wide Web. PDBBrowse incorporates a number of features that make it easy to access information found in PDB entries. Multiple search strings covering various fields corresponding to PDB record types such as compound, header, author, biological source, and heterogen data, are supported. These searches support Boolean ‘‘and’’, ‘‘or’’, and ‘‘not’’ operators. Entries selected can be retrieved automatically, and the molecular structures can be displayed using RasMol or other viewers. Entries include links to information resources such as SWISSPROT [10], BMRB [11], the Enzyme Commission Database [12], and the Entrez Reference Database [13]. Internet access to the archives has become the primary mode of retrieving entries from the PDB. However, we continue to receive a considerable number of orders for the CD-ROM product. We anticipate that this will continue to be true for a variety of reasons. For example, network performance still remains poor in a number of locations, and these disks, released quarterly, provide local access to the contents of the archives. Some of these network access difficulties may be easily overcome by installing a copy of the PDB FTP and WWW servers using mirroring software. With this software all files in the PDB are stored locally and changes are automatically reflected on a daily basis.

The 3DBase—A Relational Database Management System for 3DB (seems outdated)

3DBase is constructed with the SYBASE RDBMS, the Object-Protocol Model (OPM), and the OPM data management tools [14] developed by Dr. Victor Markowitz’s group at Lawrence Berkeley National Laboratory. SYBASE provides a powerful and robust environment for data management; the OPM tools allow rapid development of SYBASE databases; and OPM’s object-oriented view provides a scientifically intuitive representation of data. Along with a graphical schema editor, Markowitz’s group distributes a number of other development tools; foremost is a schema translator that generates SQL statements for building tables, indices, and constraint rules and triggers. This development effort attempts to address the needs of the diverse user community served by the PDB. The schema supports queries related to crystallographic as well as molecular biology questions. The database is being designed with the idea that in the near future it will be federated with other biological databases. Our expectation is that through federation, complex queries may be submitted to our database for which answers that originate from several databases may be easily returned. Interoperability is addressed through the use of schema sharing with other OPM-based databases and support for a variety of data interchange formats in query results. In addition to providing users with a powerful environment capable of complex ad hoc queries, 3DBase will also facilitate management of the growing archives, which are expected to contain over 30 000 structural reports by the year 2000. This work is being done as a collaboration among the following groups:

The Protein Data Bank—Brookhaven National Laboratory • BioInformatics Unit—Weizmann Institute of Science • OPM Data Management Tools Project—Lawrence Berkeley Laboratories • The Genome Data Base - Johns Hopkins University

4. Data Deposition

3DB will operate as a direct-deposition archive, providing mechanisms that will allow depositors to load data with minimal staff intervention. This strategy is essential if 3DB is to meet present projections of exponential growth in depositions against a fixed staff size. This is particularly challenging due to the complexity of the data being handled, the need for a common viewpoint of the entry description, and the community requirement that these data be accessible immediately upon receipt. With direct deposition, there will be a concomitant need to increase the power of data validation procedures. These procedures must reflect current models for identifying errors and must be as complete as possible. Quality control issues assume a more central and difficult role in direct deposition strategies. Distributed data must be of the highest quality; otherwise users will lose their trust in the archived data and will have to revalidate data received from 3DB before using them, clearly an unproductive scenario.

4.2 Development of Automatic Deposition and Validation

3DB must overcome many challenges for direct deposition to work. In a recent workshop held to assess the needs of 3DB users, crystallographers and NMR spectroscopists were unanimous in their desire to have a system that did not require additional work on their part when depositing data. On the other hand, consumers (which included these same depositors) were vocal in their desire for entries to contain more information than what is currently available within the PDB. We are striving to develop a suite of deposition and validation programs that accommodates these somewhat conflicting desires while ensuring that the archives maintain the highest standard of accuracy. A schematic diagram of the automatic deposition process is depicted in Fig. 4. AutoDep, the new automatic deposition program, is designed to simplify the deposition process. It includes a convenient and interactive electronic deposition form that guides the author in providing information. It also contains tools for data verification and validation and is able to flag errors in syntax or spelling. A considerable variety of information, which must be supplied by the authors, is archived about each structure. The form requests the same information as the electronic deposition form, but helps ease the burden of filling it out by populating fields using data from existing PDB entries or other computer-generated output (e.g., X-PLOR output). These data can then be reviewed and modified. Checks against other databases are an important and evolving part of this process. For example, names of organisms are checked against the taxonomy database of the National Center for Biotechnology Information (NCBI) [22], chemical names against IUPAC nomenclature tables [23], and author names and citations against MEDLINE [24] standard residue and heterogen dictionary is being developed to be used in the data entry and checking process. We are also adopting programs developed by the Cambridge Crystallographic Data Center (CCDC) [26] to handle heterogens automatically. In addition to the deposition form that is filled out through AutoDep, authors must submit the coordinate file and other experimental data files for processing and archiving. Facilities are provided by AutoDep that help simplify this process. An FTP script is generated that takes author-specified local files and uploads these data to the PDB server. The completed form is then converted automatically into a PDB formatted file and, along with the coordinate data, is submitted to a set of validation programs for checking and further annotation. These programs are designed to check: 1) the quality, consistency, and completeness of the experimental data; 2) possible violations of physical or stereochemical constraints (e.g., two atoms in the same place, appropriate bond angles, etc.); 3) compliance with our data dictionary (syntax checks); and 4) the correspondence of the experimental data to the derived structure (in the near future). Development of the validation suite will evolve with advice from the community and encompass programs currently in use, written both within and outside the PDB. The validation software automatically generates and includes in the entry measures of data quality and consistency as well as annotations giving details of apparent inconsistencies and outliers from normal values. This output is returned to the depositor for review. Entries whose data quality and consistency meet appropriate standards may then be sent by the depositor directly for automatic entry into the database. Entries that do not pass the quality and consistency checks may be revised by the depositor to correct inadvertent errors. Alternatively, the depositor may decide to do more experimental work in order to resolve problems. Apparent inconsistencies or outliers remaining in a submitted entry must be explained by the depositor in an annotation. In the most interesting cases, unusual features are a valid and important part of the structure. However, all such entries will be reviewed for possible errors by 3DB staff, who may discuss any important issues with the depositor. The 3DB staff will then forward acceptable entries to the database. To make automatic deposition as easy as possible, we are working with developers of software commonly used by our depositors. By modifying these programs to produce compliant data files and performing validation and consistency checks before submission, it may be possible to bypass most of the tedious steps in deposition. We are already working with Dr. Axel Bru¨nger to use procedures available through X-PLOR [27] to replace part of the validation suite for structures produced by x-ray crystallography and NMR. Diagnostic output will be included automatically as annotations in the entry. A limited version of X-PLOR will be available from 3DB to all depositors for validation purposes only. Validation of coordinate data against experimental x-ray crystallographic data requires access to structure factor data, which are requested by PDB, the International Union of Crystallography (IUCr), and some journals but are not always supplied by the depositor. We are working toward building a consensus in the community that structure factor data are a necessary component of deposits of structures derived by x-ray crystallography. Statistics such as number of F’s and R-values vs. sin(theta)/lambda, will be calculated and included in the 3DB entry as annotations to the experiment. In order to make it easier for depositors to submit structure factors (as well as to exchange these data between laboratories), the PDB, in close collaboration with a number of macromolecular crystallographers, has (CitDB when it becomes available). FASTA/BLAST programs are run against the SWISSPROT and PIR databases to verify protein sequences, and variant and mutant sequences are checked against the Protein Mutant Database (PMD) [25]. Links between the PDB/3DB entry and these databases are established in the process. To handle the increasing number of entries with nonstandard residues (heterogens), a developed a standard interchange format for these data. This standard is in CIF and was chosen both for simplicity of design and for being clearly self-defining, i.e., that the file contains sufficient information to be read and understood by either a program or a person. Details of this format are available through the PDB WWW server. A consensus is still developing in the NMR community as to what types of experimental data should be deposited and what kinds of validation and consistency checks should be performed. Structural data produced by other methods may also have special features that should be archived or checked. Requirements for the types of data to be deposited and proper ways of checking the validity and consistency of the data will be developed in cooperation with the experimental community for each type of structure data archived by the 3DB.

5. Accessing Data in 3DBase—User Queries and Report Generation

Primary access to 3DBase will be via the network using general purpose graphical user interfaces like Mosaic or Netscape. Access will also be available through the use of software developed by third parties (commercial developers). As diagrammed in Fig. 5, user queries will be addressed to the Query Analyzer (3DB-QA), a program module running at the server site that will parse queries and pass them on to 3DBase. Query results will be returned through the Output Generator (3DB-OG) in the format requested by the user. Queries placed over the network will generally be in the form of URLs, which are easily generated from hypertext links, HTML-based forms, or by programs or scripts using the National Center for Supercomputing Applications libraries [28] for more sophisticated applications. As part of the query the user may specify the format of the response, as we do at the present time in the PDB WWW browser. The response frequently will be in the form of an HTML document, but it can also be a PDB- or CIF-formatted file [29, 30]. The information returned may be either a complete or partial entry, or information from linked databases or external programs. A 3DBase browser has been built using Dr. Stan Letovsky’s Genera system [31]. Users specify search criteria by filling out an HTML form. Software at BNL processes this form and generates the required SQL. System performance is improved by using stored SYBASE SQL procedures that access each predefined object. The fields available are similar to those in our PDBBrowse program. For those familiar with (or willing to learn about) the OPM protocol, access to the object layer will be provided using a high level OPM-based query language. As part of the 3DB open database policy, direct access to the underlying RDBMS will be allowed and actively supported. These queries are not parsed by the 3DB-QA module, so better response time can be expected. This provides third party developers with the opportunity to either incorporate SQL clients in their products or to learn more about the OPM protocol and thereby gain access to all of the benefits that the Object model affords, e.g., active external links, programs, etc. As depicted in Fig. 5, the output generator will return query results using a variety of data interchange formats. PDB will continue to support its current format for the foreseeable future. We plan to extend this format to allow us to represent objects being stored in 3DBase. In addition, a ‘‘raw format’’ is being provided which returns an attribute/value pair. This form is easily parsed and is more compact than PDB format.

<https://www.researchgate.net/publication/365726240_RCSB_Protein_Data_Bank_RCSBorg_delivery_of_experimentally-determined_PDB_structures_alongside_one_million_computed_structure_models_of_proteins_from_artificial_intelligencemachine_learning>

ABSTRACT:

The Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB), founding member of the Worldwide Protein Data Bank (wwPDB), is the US data center for the openaccess PDB archive. As wwPDB-designated Archive Keeper, RCSB PDB is also responsible for PDB data security. Annually, RCSB PDB serves >10 000 depositors of three-dimensional (3D) biostructures working on all permanently inhabited continents. RCSB PDB delivers data from its researchfocused RCSB.org web portal to many millions of PDB data consumers based in virtually every United Nations-recognized country, territory, etc. This Database Issue contribution describes upgrades to the research-focused RCSB.org web portal that created a one-stop-shop for open access to ∼200 000 experimentally-determined PDB structures of biological macromolecules alongside >1 000 000 incorporated Computed Structure Models (CSMs) predicted using artificial intelligence/machine learning methods. RCSB.org is a ‘living data resource.’ Every PDB structure and CSM is integrated weekly with related functional annotations from external biodata resources, providing up-to-date information for the entire corpus of 3D biostructure data freely available from RCSB.org with no usage limitations. Within RCSB.org, PDB structures and the CSMs are clearly identified as to their provenance and reliability. Both are fully searchable, and can be analyzed and visualized using the full complement of RCSB.org web portal capabilities.

INTRODUCTION

On 20 October 2022, the Protein Data Bank (PDB) marked its 51st anniversary of continuous operations (1). As one of the most intensively used open-access biodata resources worldwide, it has been accredited by CoreTrustSeal (coretrustseal.org). In addition to the 60 000 or more structural biologists who generously contribute their data to the archive, the PDB is utilized by many millions of basic and applied researchers, educators, and students working across fundamental biology, biomedicine, bioengineering, biotechnology and energy sciences (2–28). Other database resources numbering ∼450, many of which have been highlighted in Nucleic Acids Research (29,30), download, integrate and distribute PDB data (30). Collectively, they enjoy open access to nearly 200 000 consistently archived, rigorously validated and expertly biocurated experimentallydetermined three-dimensional (3D) structures of biological macromolecules (proteins, nucleic acids, carbohydrates) and their complexes with one another and small molecule ligands (e.g. enzyme co-factors, approved drugs, investigational agents). Because ‘function follows form’ in biology, 3D biostructures archived in the PDB have enabled myriad important scientific breakthroughs by basic and applied researchers (11,31–36). Open access to PDB data without limitations on usage also allowed structural bioinformatics to develop as a vibrant sub-discipline of computational biology. Inspired by the work of Anfinsen who showed that the sequence of a polypeptide chain determines its shape or fold (37), members of this emerging field strove for decades to predict 3D structures of proteins accurately. Initial successes were realized using homology or comparative protein structure modeling, which depends on use of an experimentallydetermined structure with a similar amino acid sequence (∼40% identity or greater) to use as a modeling template or scaffold (reviewed in (38)). As PDB archival holdings grew and the field advanced, template-free protein structure prediction became possible for very small globular proteins, fostered by two ongoing community-led blind challenges (i.e. Critical Assessment of Structure Prediction (CASP (39)), Continuous Automated Model EvaluatiOn (CAMEO (40))). The 2020 CASP challenge witnessed a sea change in structural bioinformatics. Google DeepMind emerged as the top performer with its Alpha Fold 2 software that uses artificial intelligence/machine learning (AI/ML) to predict 3D structures of proteins with accuracies comparable to that of low-resolution experimental methods (41). Subsequently, the Rosetta team led by David A. Baker (University of Washington/Howard Hughes Medical Institute) released RoseTTAFold (42), which also uses AI/ML methods to generate computed structure models (CSMs) of proteins with reported accuracies comparable to that of AlphaFold 2. At the time of writing, CSMs for nearly every protein sequence represented in UniProt (43) are publicly accessible from AlphaFold DB (41,44,45). Some CSMs generated by computational biologists operating independently of DeepMind (using RoseTTAFold, AlphaFold 2, etc.) are available from the open-access ModelArchive (modelarchive.org). More than one million of these public-domain CSMs are now being delivered alongside ∼200 000 PDB structures by the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB, RCSB.org (46–49)). RCSB PDB was a founding member of the Worldwide Protein Data Bank (wwPDB, wwpdb.org) partnership (50,51), which has jointly managed the Protein Data Bank archive since 2003. Core RCSB PDB operations are funded by the National Science Foundation, National Institutes of Health, and US Department of Energy. RCSB PDB is headquartered at Rutgers, The State University of New Jersey, with additional performance sites at the University of California San Diego and the University of California San Francisco. Like its wwPDB partners, RCSB PDB is committed to the FAIR (Findability, Accessibility, Interoperability and Reusability (52)) and FACT (Fairness, Accuracy, Confidentiality and Transparency (53)) Principles emblematic of responsible data stewardship in the modern era. As the US data center of the wwPDB, RCSB PDB is responsible for managing deposition, validation, and biocuration of new experimentally-determined biostructures contributed by researchers working in the Americas and Oceania. Additional wwPDB Full Members include Protein Data Bank in Europe (PDBe, PDBe.org, (54)); Protein Data Bank Japan (PDBj, PDBj.org, (55)); the Electron Microscopy Data Bank (EMDB, emdb-empiar.org, (56,57)); and the Biological Magnetic Resonance Bank (BMRB, bmrb.io, (58,59)). Protein Data Bank China (PDBc) was recently admitted to the wwPDB as an Associate Member. In its role as wwPDBdesignated PDB Archive Keeper, RCSB PDB is responsible for weekly updates of the archive and safeguarding both digital information and a physical archive of correspondence, etc. The replacement cost of the entire PDB archive is conservatively estimated at ∼US$20 billion, assuming an average cost of ∼US$100 000 for regenerating each experimental structure. In order to continue serving the needs and interests of the diverse community of PDB users, an assortment of new features and tools have been developed and integrated into the RCSB PDB research-focused RCSB.org web portal, as described previously (47–49,60). A significant software development project was undertaken to overhaul the information management services underlying RCSB.org since our last Nucleic Acids Research Database Issue publication (47). In this comprehensive redesign, we developed a one-stop-shop for studying 3D biostructures by extending RCSB.org web portal functionality to support parallel delivery of more than one million CSMs publiclyavailable from AlphaFold DB (alphafold.ebi.ac.uk) and ModelArchive (modelarchive.org) together with nearly 200 000 experimentally-determined structures stored in the growing PDB archive. These CSMs reflect great advances made in the field and are not comparable to the theoretical models that were removed from the main PDB archive in 2002. While experimentally-determined PDB structures will remain the ‘gold standard’ at RCSB.org, integrated access to these models will be of great value to those studying 3D biological macromolecules. (N.B.: Criteria for inclusion of 3D biostructures in the PDB remain unchanged. They must be based on actual experimental measurements on sample specimens of the biological macromolecule(s) comprising the structure. For full details, see wwpdb.org.) This initial release of one million CSMs reflects the number of models available at the time this software development project was initiated. It does not include the recent release at AlphaFold DB of a new set of CSMs corresponding to the whole non-redundant UniProt database (ca. 200 million entries). The breakdown of CSMs currently integrated within RCSB.org is: • From AlphaFold DB: Generated by DeepMind using AlphaFold 2 • Model organism proteomes: 326 175 protein structures from 48 different model organisms • Global health proteomes: 238 274 protein structures from various disease-causing organisms • Swiss-Prot sequences (43): 542 380 protein structures, 430 961 of which are in addition to those already in the first two sets • MANE (Matched Annotation from NCBI and EMBL-EBI) sequences (61): 17 334 protein structures, 3844 of which are in addition to those from the above three sets • From ModelArchive: 1106 models of core eukaryotic protein complexes produced by the Baker lab (62). Generated using a combination of RoseTTAFold and AlphaFold 2.

Expansion of the purview of RCSB.org was enabled by interoperation of the PDBx/mmCIF data standard, which underpins the PDB archive and RCSB.org services (see below), with the related ModelCIF data standard for CSMs (see below). RCSB.org now provides PDB data consumers with access to CSMs covering the entire human proteome, and those of many model organisms, selected pathogens, organisms relevant to bioenergy research (44), and protein complexes from select studies (62). Importantly, to maintain clear distinction between experimental structures and computational models, PDB structures and CSMs are identified as to their respective provenance and reliability. In addition to the RCSB.org web portal features described here, the newly integrated data are also available via RCSB PDB APIs: (Data, Search, and 1D-coordinates (63)), dramatically enhancing the ability of programmatic users to use CSMs in their workflows. An upcoming article will describe the new programmatic developments in more detail.

RESULTS

Motivation

As of mid-2022, the PDB housed nearly 200 000 3D biostructures, encompassing proteins from organisms representing all kingdoms of life (Figure 1). Archival holdings of eukaryotic protein structures exceeded 105 000, with more than half being human in origin. Bacterial protein structures were also numerous, totaling nearly 66 000 (∼10% of which came from E. coli). Archaeal protein structures were the least numerous (totaling ∼5500). Notwithstanding the importance of model organisms in basic and applied research in biology, PDB coverage is decidedly limited, with mouse protein structures being most numerous at ∼8000 structures. Rigorously validated and expertly biocurated PDB structures have been long-considered a ‘gold-standard’ in the biosciences and in-fact made AI/ML prediction of protein structures possible (64). Powerful tools developed by RCSB PDB for searching and analysis (including sequence, structure, structure motif, sequence motif), and visualization (including 1D-3D views of annotations, Mol\*) help drive research and education in the biosciences worldwide. The value of integrating PDB structures and CSMs within RCSB.org is as follows:

1. One-stop-shop data delivery should help ∼99% of RCSB.org web portal users, who are not structural biologists and are often frustrated that their protein(s) of interest is not represented in the PDB archive as an experimentally-determined structure. 2. Inclusion of CSMs provides all users with structural information for full-length polypeptide chains. Structural biologists will be able to use this information to identify, express, and purify compact globular domains that are more likely to be crystallizable for macromolecular crystallography (MX) or sufficiently soluble to study via solution nuclear magnetic resonance (NMR) spectroscopy. Other users will have more information with which to develop testable hypotheses and design experiments to probe the functional importance of disordered segments of polypeptide chains. 3. One-stop-shop data delivery should help structural biologists accelerate structure determination by 3D electron microscopy (3DEM) and integrative or hybrid methods. 4. All users stand to benefit from RCSB.org capabilities supporting contextual examination of CSMs through a one-stop-shop offering parallel delivery of PDB structures and CSMs

The decision to deliver CSMs alongside PDB structures is in no way intended to send the message that they are equivalent in accuracy. Analyses carried out by RCSB PDB and published in 2022 have shown that even confidently predicted CSMs are not as accurate as experimentally determined structures coming from macromolecular crystallography (at 3.5 A resolution or better ( ˚ 65)). PDB structures should be used preferentially whenever they are available. Moreover, most CSMs publicly available at the time of writing are those of monomeric proteins (even when they are known to exist within homo- or hetero-oligomers or complex assemblies in their physiological state). Similarly, these CSMs do not typically include information about bound ligands (e.g. enzyme co-factors, substrate analogs, inhibitors, investigational agents, approved drugs, nucleic acids). Parallel delivery of PDB structures and CSMs through the RCSB.org web portal ‘one-stop shop’ should allow all users to analyze, visualize and explore CSMs in the context of experimentally-determined structures of closelyrelated proteins to better understand their biological and biochemical functions.

Data integration: combining external information with both PDB structures and CSMs

The RCSB.org web portal provides added value to users going well beyond the content of the archive itself. In addition to serving 3D structural data, their supporting data files, and metadata, RCSB PDB integrates information from trusted data resources (Table 1) to provide insights and details about the chemistry, sequence, 3D structure determination method, structure, functions, and evolution of the molecule(s) being studied. These data, annotations, and classifications also provide contexts for applying this knowledge to address questions in biology, medicine, bioenergy, biotechnology, evolution and more. Integrating 3D structure data with external information ensures that the RCSB.org web portal operates as a ‘living data resource’. It is not uncommon for new biological or biochemical functions of a macromolecule to come to light, or new disease-causing mutations to be identified after 3D structure data are deposited to PDB or repositories for CSMs. New findings are integrated within RCSB.org on a weekly basis, thereby ensuring public access to current information. When multiple reliable data resources provide annotations/information about a specific biomolecular property, feature, or function, the RCSB PDB integrates all relevant data together with suitable provenance. Access to these data, enables users to identify and explore details that best meet their research interests/needs. For example, currently membrane protein annotations are integrated from four different data resources

Searching, analyzing, visualizing, and exploring PDB structures and CSMs with RCSB.org

Upon reaching the RCSB.org home page, users can query, organize, visualize, analyze, compare and explore PDB structures and CSMs side-by-side. Searching 3D structure information can encompass PDB structures and CSMs or be limited to PDB structures only. Either PDB structures or CSMs can be excluded from the search results. The two types of structure information accessible via RCSB.org are clearly distinguished from each other (Figure 2). Top bar searching and data delivery for PDB structures and CSMs. Figure 3 identifies key navigational features that provide users with access to Top Bar Search on the RCSB.org home page (top, upper panel), Advanced Search (middle panel), and Browse Annotations (lower panel). Top Bar Search (also referred to as Basic Search) appears throughout the RCSB.org portal (Figure 4). The default option (3D Structures) searches PDB structure data only; CSMs can be included when the toggle switch is activated and its color changes from gray to cyan. Entering a keyword (e.g. molecule name, database entry ID (PDB, UniProt, AlphaFold DB, ModelArchive), author name (PDB structures only)) will launch autosuggestions organized by data category. Select one of the suggestions to launch the search. Note that when using CSM identifiers (e.g. from AlphaFold DB or ModelArchive) the toggle switch to include CSMs must be activated. Sequence searches can be run by entering single-letter code sequences for protein, DNA or RNA polymers and executing the query (hitting return or clicking the magnifying glass icon). This sequence-based search uses the MMseq2 software (108) to identify similar protein or nucleic acid sequences. Another option supports free text searches, which are carried out most expeditiously when the phrase of interest is enclosed within double quotes. Otherwise, structures containing any of the text words in the query will be returned and may include false positives. Top Bar Search can also be used to search documentation and news announcements available on both RCSB.org and the RCSB PDB outreach and education web portal (PDB101.RCSB.org, (109)) by changing the search type from 3D Structures to Documentation on the left of the search box.

Additional Mol\* 3D visualization options

The most common way to explore 3D structures available from RCSB.org is to visualize them. Within RCSB.org, 3D structures may be visualized using a web-native visualization tool known as Mol\* (110,115). This tool can be accessed by clicking on the ‘Structure’ link below the thumbnail image of the structure or by clicking on the tab ‘3D View’ on the top of the page (Figure 5A). Mol\* has also been implemented within other RCSB.org tools as follows: 1. Linked to a 1D (sequence) browser that can be accessed by clicking on the 1D-3D link, below the thumbnail images of the 3D structure (Figures 5A and C). This feature allows users to map and display a variety of annotations integrated from various bioinformatics resources on the 3D structure. 2. As part of the Pairwise 3D Structure Alignment tool to display regions of match between two or more structures being compared or the whole superimposed structure(s) (Figure 7). This tool can be used to compare 3D structures that are not available from the RCSB.org (e.g. 3D structure atomic coordinates stored on a local computer, using the file upload option; CSM atomic coordinates from AlphaFold DB an external data resource, using the Web Link option).

Note: In both Mol\* implementations, clicking on the Expanded Viewport button in the vertical toggle menu in the Mol\* 3D canvas expands the Mol\* window, providing access to all options and tool functionalities. Finally, a standalone implementation of Mol\* (https:// www.rcsb.org/3d-view) is available for visualizing and analyzing 3D structures not accessible within RCSB.org. The overall layout of the tool is the same with a right-hand Controls panel. The Open File options allow upload of a locally saved file, while the Download Structure options allow specification of a structural biology resource (e.g. AlphaFold DB structures not currently available from RCSB.org). Multiple structures can be uploaded to this implementation of the tool for superposition and analysis. Standalone Mol\* also provides a convenient platform to upload and view a previously saved Session using the Sessions → Download/Open options.

FUTURE DIRECTIONS

As the PDB archive entered its 52nd year, RCSB PDB embarked on comprehensive analyses of its diverse user communities (i.e. basic and applied researchers, educators, and students spanning fundamental biology, biomedicine, bioenergy, bioengineering and biotechnology), and strategic reviews of how it (i) Delivers Data In and Data Out services efficiently to a growing user base, now numbering many millions worldwide; (ii) Works with wwPDB partners to process, rigorously validate, and expertly biocurate the growing number of increasingly complex PDB depositions received annually (projected at ∼16 500 for 2022); (iii) Manages and safeguards the growing PDB archive in its role as wwPDB-designated Archive Keeper; (iv) Enables efficient searching, analysis, visualization, and exploration of hundreds of thousands of experimentally-determined PDB structures integrated with more than one million CSMs through its RCSB.org research-focused web portal; and (v) Supports user training, education, and outreach through its PDB101.RCSB.org introductory web portal.

Additional challenges lying ahead for RCSB PDB include, but are by no means limited to the following:

A. Rapid growth in public-domain CSMs of individual polypeptide chains, already numbering >200 million at the time of writing; B. Anticipated advances in AI/ML-based prediction of structures of multi-protein complexes and those of protein-ligand complexes; C. Continued development of biomolecular structure determination methods using X-ray Free Electron Lasers, revealing the microscopic details of chemical reactions in real time; D. Growth in the number and complexity of atomic-level cryo-electron tomography structures of macromolecular machines imaged within cryogenically preserved cells and tissues; E. Integration of PDB structures and CSMs with complementary information coming from correlative light microscopy and related imaging methods across length scales ranging from atoms to small molecules to individual biomolecules to macromolecular assemblies to organelles to cells and ultimately tissues; F. Merging of the PDB-Dev (pdb-dev.wwpdb.org) prototype archiving system for integrative (or hybrid) methods structures with the PDB archive; and G. Federating other biodata resources, such as the SmallAngle Scattering Database (SASBDB, sasbdb.org) and the Proteomics Identification Database (PRIDE, ebi.ac.uk/pride), with the PDB, EMDB and BMRB core archives jointly managed by the wwPDB partnership.

Policy changes recently promulgated by the Office of Science and Technology (OSTP) in the United States (US) are also likely to affect future RCSB PDB operations. The Executive Office of President Joe Biden has called on the federal agencies with research and development expenditures to update their public access policies as soon as possible (126), and no later than 31 December 2025, to make publications and their supporting data (e.g. biomolecular structure information stored in the PDB archive) resulting from federally funded research publicly accessible without an embargo on their free and public release. This announcement is expected to accelerate progress towards full open sharing of data generated with federal research funding in the United States. It will add considerable weight to awareness campaigns undertaken by non-governmental organizations such as CoreTrustSeal (coretrustseal.org) and the Global Biodata Coalition (globalbiodata.org). The recent OSTP announcement also begs the question as to how heavilyused, open-access data resources, such as the PDB archive, should be sustainably funded at levels commensurate with the central roles they play in biological and biomedical research and education ecosystems worldwide (127,128).

<https://www.researchgate.net/publication/361873960_Exploring_protein_symmetry_at_the_RCSB_Protein_Data_Bank>

Introduction:

The PDB is a core resource central to the global biodata ecosystem serving many millions of users drawn from diverse scientific and educational communities. It provides a permanent and expertly curated data archive [21– 25] for structural biologists to disseminate their results, promotes reproducibility of the structural biology scientific literature, and makes biomolecular structure information freely available to a wide community of researchers, educators, students, and the general public without limitations on data usage. The PDB was established in 1971 at Brookhaven National Laboratory as the first open-access, digital-data resource in biology [26]. Since 2003, the PDB has been managed by the Worldwide Protein Data Bank partnership (wwPDB; wwPDB.org) [27,28]. Member organizations of the wwPDB (RCSB Protein Data Bank, RCSB PDB; Protein Data Bank in Europe, PDBe; Protein Data Bank Japan, PDBj; Electron Microscopy Data Bank, EMDB; and Biological Magnetic Resonance Bank, BMRB) together curate and annotate 3D biostructure data deposited by scientists from around the globe, and make it publicly, freely, and easily available through user-friendly web portals and host services. RCSB PDB, a founding member of the wwPDB, is responsible for US PDB operations, and serves as the wwPDB-designated PDB Archive Keeper. The RCSB PDB web portal (RCSB.org) supports millions of users worldwide [29–31]. In 2021, the website was visited each month by an average of ∼757 000 unique visitors according to Google Analytics, with ∼4.7 million unique visitors annually. A total of 257.71 TB of data were accessed. In 2021, 1.8 billion data files in various file formats, including structure files, experimental data files, chemical and molecular reference data files, and validation reports, were downloaded and/or viewed from RCSB PDB-hosted FTP and websites. Additional data were downloaded from wwPDB partners PDBe and PDBj for a total of 2.3 billion data files. This research-focused website provides tools and services that support users across scientific disciplines to access, analyze, and visualize up-to-date structural views of proteins and nucleic acids important to fundamental biology, biomedicine, and bioenergy sciences.

<https://onlinelibrary.wiley.com/doi/full/10.1002/pro.4213>

Abstract

The Research Collaboratory for Structural Bioinformatics Protein Data Bank(RCSB PDB), funded by the US National Science Foundation, National Insti-tutes of Health, and Department of Energy, has served structural biologistsand Protein Data Bank (PDB) data consumers worldwide since 1999. RCSBPDB, a founding member of the Worldwide Protein Data Bank (wwPDB)partnership, is the US data center for the global PDB archive housing biomo-lecular structure data. RCSB PDB is also responsible for the security of PDBdata, as the wwPDB-designated Archive Keeper. Annually, RCSB PDB servestens of thousands of three-dimensional (3D) macromolecular structure datadepositors (using macromolecularcrystallography, nuclear magnetic

INTRODUCTIONThe Protein Data Bank (PDB) has been serving global sci-ence for more than 50 years. It was established onOctober 20th, 1971 as the first open-access digital dataresource in biology with just seven protein structures.1Thanks to the generosity of more than 50,000 structuralbiologists working on every inhabited continent, thearchive has grown to more than 180,000 structures ofproteins and nucleic acids (DNA and RNA). Today, thePDB archive is jointly managed by the Worldwide Pro-tein Data Bank (wwPDB, wwpdb.org) partnership,2,3which was founded in 2003 by the US-funded RCSB Pro-tein Data Bank (RCSB PDB), the Protein Data Bank inEurope (PDBe), and Protein Data Bank Japan (PDBj).Current wwPDB members also include the ElectronMicroscopy Data Bank (EMDB) and the Biological Mag-netic Resonance Bank (BMRB). Millions of PDB dataconsumers worldwide working in fundamental biology,biomedicine, bioengineering, biotechnology, and energysciences enjoy no-cost access to 3D biostructure informa-tion with no limitations on data usage.The Research Collaboratory for Structural Bioinfor-matics Protein Data Bank (RCSB PDB; RCSB.org)4,5isjointly funded by the National Science Foundation, theNational Institutes of Health, and the US Department ofEnergy. Safeguarding and nurturing the PDB archive andproviding open access to PDB data are the responsibilityof four coordinated RCSB PDB“services,”encompassingdata deposition; archive management and access; dataexploration; and outreach and education. RCSB PDB, likeits wwPDB partners, is committed to the FAIR (Findability,Accessibility, Interoperability, and Reusability)6and FACT(Fairness, Accuracy, Confidentiality, and Transparency)7Principles emblematic of responsible data stewardship inthe modern era. It is no exaggeration to state that the PDBwas“walking the walk”decades before people began“talking the talk”regarding these important concepts.Service 1—Data deposition: The global wwPDBOneDep8software system manages deposition, validation,expert biocuration, and remediation of macromolecularcrystallography (MX), 3D electron microscopy (3DEM),nuclear magnetic resonance (NMR) spectroscopy, andmicro-electron diffraction (microED) structures, experi-mental data, and related metadata. Within the wwPDB,RCSB PDB supports PDB data depositors working in theAmericas and Oceania ensuring completeness and accu-racy of the ever-growing body of 3D structure data.Service 2—Archive management and access:Initsrole as the wwPDB-designated Archive Keeper, RCSBPDB safeguards the PDB archive and maintains thePDBx/mmCIF data dictionary9,10that enables organiza-tion and searching of archived data. Programmaticaccess to PDB data is available via FTP and applicationprogramming interfaces (APIs). Strict adherence to thePDBx/mmCIF data standard enables facile integrationof 3D structure information with >50 trusted externaldata resources.188BURLEYET AL.

Service 3—Data Exploration:Toolsfordatasearching,browsing, visualization, custom report generation, andanalysis are made freely available on our research-focusedRCSB.org web portal with no limitations on usage.Service 4—Outreach and Education: RCSB PDB has along history of delivering outreach and educationresources focused on structural biology and its impactacross the sciences via its introductory PDB101.RCSB.orgweb portal (reviewed in this issue11).Two other important elements of RCSB PDB opera-tions are the Customer Service Help Desk, responsible forsupporting 3D structure depositors and PDB data con-sumers around the world, and the Infrastructure Team,which works to ensure >99% 247365 service avail-ability uptime. The status of RCSB PDB servers, micro-services, and application programming interfaces (APIs)is monitored by NS1 (NS1.com) and publicly available ona real time basis at status.rcsb.org.Recent redesign of RCSB PDB data architecture andoverhaul of our research-focused RCSB.org web portalwith many added new features have been described pre-viously.5,12This invited Protein Science Tools Issue con-tribution describes the unprecedented growth of the PDBarchive during 2020, continued evolution of the PDBx/mmCIF data standard, archive-wide remediation ofcarbohydrate-containing structures, and deployment ofnew RCSB.org tools and features completed during 2021

RESULTS

2.1 PDB data metrics and trends

The first year of the COVID-19 pandemic witnessedunprecedented growth in the PDB archive. Thissection presents impressive structure deposition andrelease metrics for 2020 and describes the ongoing impactof the 3DEM resolution revolution.13PDB structures are contributed annually by tens ofthousands of structural biologist depositors worldwidethrough the wwPDB OneDep software system (deposit.wwpdb.org) for structure deposition,8rigorousvalidation,14,15and expert biocuration.16OneDep cur-rently supports 3D macromolecular structures deter-mined using MX, 3DEM, NMR, and microEDexperimental methods. Incoming structures are processedat regional wwPDB data centers allocated on the basis ofthe depositor's geographic location.During calendar year 2020, wwPDB partnersprocessed a record 15,436 experimental structure deposi-tions to the PDB archive. During this same period, RCSBPDB processed47% of the global depositions (primarilyfrom the Americas and Oceania, and GroupDep17users),PDBe processed31% of global depositions (from Europeand Africa), and PDBj processed22% of global deposi-tions (from Asia and the Middle East) (Figure 1a,b).FIGURE 1PDB data deposition and release metrics. (a) Depositor geographic locations in 2020. (b) Structure deposition processing bywwPDB regional data centers in 2020. (c) Annual rates of PDB archive growth (logarithmic scale) for 3DEM (dashed line), NMR (dottedline), MX (dashed-dotted line), and all methods (total, solid line)BURLEYET AL.189 Consequently, the number of new structures publiclyreleased into the PDB reached another record high of14,031 during 2020. Among these newly deposited struc-tures, more than are 1,000 SARS-CoV-2 (causative agentof the COVID-19 pandemic) related protein structures,reflecting enormous efforts made by the structural biol-ogy community to understand and fight the pandemic. Acomprehensive enumeration of freely available SARS-CoV-2 related resources provided by RCSB PDB can befound at rcsb.org/covid19.Figure 1c illustrates annual growth of the PDB archivebroken down by experimental method. 3DEM and relatedtechnologies continue to evolve rapidly bringing new capa-bilities to the structural biology community. 3DEM struc-tures in the archive increased by60% between 2019 and2020 (from 1,452 to 2,390). Equally impressive, 3DEMmedian structure resolution limits improved from 9 Å in2010 to 3.5 Å in 2020. microED structure depositions havealso increased with median resolution limits improving inrecent years (from 8 Å to 1.2 Å). New NMR structuresreleased annually have plateaued at400 over the pastfew years. Notwithstanding rumors of the imminentdemise of protein crystallography, the number of newlydeposited MX structures continues to grow annually. Themuch-heralded success of Google DeepMind AlphaFold 218for protein structure prediction will only increase the effi-ciency of experimental structural biologists, yielding morePDB depositions of protein–ligand complexes and largemulti-protein assemblies, neither of which can be predictedtoday at accuracy levels comparable to experiment.

PDB ARCHIVE MANAGEMENT

3.1 PDB data hierarchy

Data stored in the PDB archive are categorized accordingto the following definitions:•Entry: All data pertaining to a particular structuredeposited into the PDB constitute an archival Entry,identified with a unique PDB ID (currently four alpha-numeric characters, e.g., 1q2w).•Entity: Each chemically unique molecule constitutinga PDB Entry is defined as an Entity (including Poly-mer, Branched, or Non-polymer), and labeled with anumeric Entity ID.

Polymer Entities are composed of smaller chemicalbuilding blocks linked together by covalent bonds(e.g., proteins or polypeptides, DNA or poly-deoxyribonucleotides, RNA or polyribonucleotides),which are identified by individually numberedamino acids or nucleotides covalently linked in theorder defined by the polymer sequence.

Branched Entities are either linear or branched car-bohydrates and are composed of saccharide unitscovalently linked via one or more glycosidic bonds.

Non-polymer Entities are small chemicals (enzymecofactors, ligands, water molecules, etc.). EveryNon-polymer Entity has a unique wwPDB ChemicalComponent Dictionary (CCD)19ID (one to threecharacter alphanumeric code). The CCD providesnomenclature standards and chemical descriptionsfor all small-chemical ligands and biopolymer com-ponents represented in the PDB archive.N.B.: Every PDB Entry contains at least one PolymerEntity or one Branched Entity (either linear orbranched oligosaccharides).

Instance: There can be multiple Instances of any par-ticular Entity within a PDB structure.

Each Instance or“copy”of a Polymer Entity islabeled with a unique Chain ID (one or more alpha-numeric characters, e.g., A, AA,...).

Each Instance of a Branched Entity is similarlylabeled with a unique Chain ID.

Each Non-polymer Entity is identified with theChain ID of the spatially nearest Polymer Entity. Their Instances are distinguished with uniquenumbering.

Assembly: Polymer Entity Instances (or Chains) com-monly occur in nature as components of larger macro-molecular Assemblies, ranging in size and complexity.Each Assembly in a PDB structure is assigned a uniquenumeric Assembly ID

3.2 Recent PDBx/mmCIF data standardimprovements

The semantic foundation for PDB data architecture isdefined in the PDBx/mmCIF dictionary.10,20PDBx/mmCIF is the macromolecular extension of an earliercommunity data standard, the Crystallization Informa-tion Framework (cif.iucr.org),21developed under the aus-pices of the International Union of Crystallography fordescription of small molecule X-ray diffraction studies.The PDBx/mmCIF data standard is maintained bythe wwPDB organization in collaboration with wwPDBPDBx/mmCIF Working Group domain experts recruitedfrom the scientific community (hereafter Working Group, wwpdb.org/task/mmcif). Content dictionaries and Work-ing Group discussions are hosted on the GitHub platform(github.com/pdbxmmcifwg). The PDBx/mmCIF webresource (mmcif.wwpdb.org) supports browse and searchaccess to standard terminology. Because the WorkingGroup includes developers for many of the widely usedstructure determination software systems, this groupplays a vital role in ensuring that data produced by theseprograms comply with the PDBx/mmCIF data standard,generating complete and correct data files for PDBdeposition.The wwPDB and the Working Group collaborate ondeveloping terminologies for new and rapidly evolvingmethodologies (e.g., X-ray Free Electron Laser SerialCrystallography and 3DEM), and improving representa-tions for existing data content (e.g., carbohydrate remedi-ation). Most recently, the Working Group has focused onmodernizing content descriptions for processed X-ray dif-fraction data, including extensions describing anisotropicdiffraction limits, unmerged reflection data, and newquality metrics of anomalous diffraction data (wwpdb.org/news/news?year=2021#60638da1931d5660393084c3).Deposition and delivery of this extended content will sig-nificantly enhance our ability to assess experimental dataquality, thereby improving every PDB data consumer'sability to Find and Reuse relevant PDB Entries.

Recent RCSB PDB data architectureimprovements

In 2020, RCSB PDB launched a significant upgrade ofits data delivery service delivery architecture12at RCSB.org.5This web resource upgrade transformed a legacymonolithic data delivery application into a distributeddeployment of individual microservices, each with asingle responsibility. Within the new architecture, dataaccess services (data.rcsb.org) provide both Representa-tional State Transfer (REST) and GraphQL (graphql.org) API access to a data warehouse hosted in aMongoDB document-oriented database (mongodb.com). In the initial release, Advanced Search QueryBuilder functionality encompassed text, PDB data attri-butes, 3D structure, sequence, biopolymer sequencemotif, and chemical similarity. Every search function isimplemented as an independent service. A separate sea-rch aggregation service is responsible for launchingeach search function, and combining and deliveringtheir integrated search results to front-end services andpublic programmatic search APIs (search.rcsb.org).Various operational benefits accrue from reliance on adata architecture in which each service has a singleresponsibility, notably greater flexibility in scaling the deployment of services in response to changes in userload and significant reductions in the time required todevelop, test, and deploy new features.Since re-launch of our RCSB.org web portal, we havecontinued to develop the new data architecture and aug-ment website features. The Sequence Motif search func-tion has been extended with a new 3D Structure Motifsearch capability.22Our Chemical Search function hasalso been extended with the ability to perform exhaustivesubstructure searching (eyesopen.com) across the smallmolecules represented in the PDB archive. Both of thesenew search services are also managed by the SearchAggregator service and available through our public sea-rch APIs. An important feature of the Search Aggregatoris the capability to deliver search results at different levelsof molecular granularity. For example, in our firstdeployment, search results could be shown as eitherdeposited structure Entries, Assemblies, or distinct Poly-mer Entities. These search result types have been aug-mented to include distinct Non-polymer molecularconstituents, plus the chemical and molecular definitionsfrom the CCD.To monitor the new service architecture, we recentlydeveloped a system for processing service logs and indexingthe time course of successful and failed access requests,while respecting PDB data consumer privacy. These statis-tics are essential for monitoring the health of RCSB PDBproduction services. They also provide simple analyticinformation, such as data file downloads or service access.The new metrics provide a rigorous basis for refining andextending our new search and data access services toenable greater FAIRness. This work built atop widely usedopen-source telemetry tools (e.g., ElasticSearch, Kibana,Metricbeat, and Filebeat),23using the open-source ElasticCommon Schema Standard24for encoding log data. N.B.:Historically, when open source tools have becomeunavailable they were replaced with other open source orlow-cost/no-cost proprietary tools.

3.6 Recent advances in RCSB PDB dataintegration

To make PDB data more Findable and Interoperable,RCSB PDB integrates the content of each expertly bio-curated Entry with information from more than 50 exter-nal data resources (rcsb.org/docs/general-help/data-from-external-resources-integrated-into-rcsb-pdb). Itemsof external data content are integrated into a data schemathat defines the organization of the RCSB PDB data ware-house. External data integration represents an integralpart of the weekly new Entry release workflow, whichmanages loading of new data into the PDB archive andour data warehouse. From there, it becomes available toRCSB PDB front-end services, public data access APIs,and our text search indexing service.26In response tocommunity input, we have continued to integrate newtrusted external data resources (Table 1)

Mol\* molecular graphicsvisualization

The most common use of PDB data is to visualize or“see”the 3D shapes and interactions of biological macro-molecules in order to understand biochemical and bio-logical function. In 2020, Mol\*42was deployed as thedefault RCSB.org molecular graphics tool for visualizingmacromolecules, carbohydrates, and small-moleculeligands. This open-source molecular graphics softwaresystem was developed as a community project, co-led by RCSB PDB and the Protein Data Bank in Europe (PDBe;PDBe.org). Mol\* is a web-native graphics tool for interro-gating 3D macromolecular structure data from the PDBor computed structure models. It works entirely withinthe PDB data consumer's internet browser, obviating theneed to download, install, or maintain any external soft-ware. Importantly, Mol\* supports integration of informa-tion from other bioinformatics resources to provide newinsights (e.g., about active site amino acids, known muta-tions, locations of post-translational modifications). Inturn, these insights can help develop new hypotheses forresearch and facilitate analysis and/or interpretation ofobservations and experimental results.Mol\* is a versatile tool driven by an intuitive graphi-cal user interface (GUI). It enables users to easily visual-ize entire polypeptide or nucleic acid chains, whole biological Assemblies (some including millions of non-hydrogen atoms), or specific atoms or groups of atoms ina particular biological macromolecule. With a few mouseclicks, it can present 3D structure data in a variety ofcommonly used molecular representational styles. It canalso display molecular surfaces, and non-covalent inter-actions with bound ligands, ions, drugs, and inhibitors.The GUI enables rapid display of specific biomolecularfeatures, comparison of related structures, and launch ofarchive-wide queries for Instances containing specified3D structure motifs of amino acids or nucleotides (seebelow).We exemplify some new Mol\* features in Figure 2ausing a 3DEM structure of SARS-CoV-2 spike protein(PDB ID 6vxx).43The spike protein interacts with itshuman cell-surface receptor, angiotensin convertingenzyme 2 or ACE2, to facilitate cellular invasion/infec-tion. It is the target of antibodies and T-cells produced bythe host immune system in response to viral infection orvaccination. Various engineered anti-spike protein anti-bodies have received emergency use authorization fortreatment of mild to moderate COVID-19 infections inadults and pediatric patients (e.g., bamlanivimab andetesevimab, Eli Lilly and Co.). Disturbingly, substitutionsthat change only a few amino acids on the spike proteinyielded more transmissible SARS-CoV-2 variants that inturn spread rapidly around the world. Such variants haveput the enormous number of individuals not yet fullyvaccinated and the smaller, but not insignificant, numberof individuals who cannot be vaccinated (for medical orreligious reasons) at risk of serious illness requiring hos-pitalization or death.Locations of some of the substitutions seen in theSARS-CoV-2 spike protein of the B.1.617.2 or Delta“vari-ant of concern”designated by the US Centers for DiseaseControl and Prevention (cdc.gov/coronavirus/2019-ncov/variants/variant-info.html) are marked with red hemi-spheres in Figure 2b. The full tally of Delta variant substi-tutions includes T19R, V70F, T95I, G142D, E156-, F157-,R158G, A222V, W258L, K417N, L452R, T478K, D614G,P681R, and D950 (where“–”indicates an amino aciddeletion)(cdc.gov/coronavirus/2019-ncov/variants/variantinfo.html). One particular amino acid substitutionat position 452 of the spike protein that changes a leucineto an arginine (L452R) has been detected in more thanfive SARS-CoV-2 variants. Mol\* enables quick and easyexamination of differences in molecular interactions fornative versus variant spike protein structures (Figure 2c,d). Similar analysis of key inter- and intra-molecularinteractions of different viral variants can facilitate bothbasic research and design of new diagnostic tools andtherapeutic interventions.

<https://akjournals.com/view/journals/1886/11/4/article-p77.xml>

ABSTRACT

The Research Collaboratory for Structural Bioinformatics Protein Data Bank (RSCB PDB) provides a wide range of digital data regarding biology and biomedicine. This huge internet resource involves a wide range of important biological data, obtained from experiments around the globe by different scientists. The Worldwide Protein Data Bank (wwPDB) represents a brilliant collection of 3D structure data associated with important and vital biomolecules including nucleic acids (RNAs and DNAs) and proteins. Moreover, this database accumulates knowledge regarding function and evolution of biomacromolecules which supports different disciplines such as biotechnology. 3D structure, functional characteristics and phylogenetic properties of biomacromolecules give a deep understanding of the biomolecules’ characteristics. An important advantage of the wwPDB database is the data updating time, which is done every week. This updating process helps users to have the newest data and information for their projects. The data and information in wwPDB can be a great support to have an accurate imagination and illustrations of the biomacromolecules in biotechnology. As demonstrated by the SARS-CoV-2 pandemic, rapidly reliable and accessible biological data for microbiology, immunology, vaccinology, and drug development are critical to address many healthcare-related challenges that are facing humanity. The aim of this paper is to introduce the readers to wwPDB, and to highlight the importance of this database in biotechnology, with the expectation that the number of scientists interested in the utilization of Protein Data Bank’s resources will increase substantially in the coming years.

INTRODUCTION

The Protein Data Bank (PDB) is known as an international virtual data core, which serves as a fundamental information source in association with atomic structures, crystallography and three-dimensional (3D) structures of biomolecules, including nucleic acids and proteins (e.g., enzymes, immunoglycoproteins, adhesins) which are applicable for education and research. In this regard, biotechnology, biopharmaceutics, bioengineering, biomedicine, biology are disciplines that are directly dependent on the use of PDB [1–7]. Indeed, the data and information regarding crystallography and 3D structures of biomolecules released by PDB enable us to have an effective prognostication about the biochemical, biophysical and physicochemical properties comprising affinities and bonds of the related macromolecules and small biomolecules [2, 8–10]. Since 1971, the PDB as the first global open access recourse, which serves invaluable digital data for free. This international public good, supports vital data and information to visualize the biological structures and the related bindings between macro- and small biomolecules. Since 2013, the management of PDB is in accordance with the FAIR (the acronym depicts: Findable, Accessible, Interoperable, Reusable) guiding principles for scientific data [2, 11]. Figure 1 shows the timeline of PDB progression (https://www.rcsb.org/ pages/about-us/history) [2, 12–18]. Interestingly, the open access “treasure” of PDB archives and represents several thousands of biomolecules to global users. Atomic and molecular structures of biological molecules together with their complexes (biomolecule-specific ligand(s)) are archived in PDB. Simultaneously, the PDB archive gets bigger and bigger every year. Up to now, the PDB is recognized as a high-managed resource for effective biodata. The FAIR principles are guaranteed via the application of OneDep software system. This software system controls the input structure data receiving by PDB data ecosystem for being validated, standard and biocurated. This process makes the data representing by PDB as findable, accessible, interoperable and reusable [11, 19–21]. Since the establishment of wwPDB [21] in 2003 (Fig. 1) up to now, several biocurators have been recruited by wwPDB centers in different continents such as Asia, Europe and the Americas. A collection of basic sciences and skills comprising enzymology, biophysics, computational chemistry, biochemistry, small molecule crystallography, electron microscopy, macromolecular crystallography and nuclear magnetic resonance (NMR) spectrometry supports the structural biology as the front line aim and goal of the PDB archive [19]. Even during the severe acute respiratory syndrome–related coronavirus (SARS-CoV-2) pandemic era, more than 2000 structures associated with the causative agent of the coronavirus disease (COVID-19) were released and have become accessible for global users for free. A brief collection of PDB deposits is available on SARS-CoV-2 related structures page (https://covid-19.bioreproducibility. org/) [7]. The structural properties of different organisms e.g., COVID-19 released by PDB archives give us this opportunity to find out the spatial conformation of ligands, ligand binding sites, protein-protein interactions and amino acid substitutions regarding different viral proteins. The related data may also be represented by other centers and websites rather than PDB (https://www.rcsb.org/news?year 52020&article55e74d55d2d410731e9944f52&feature5true), including the COVID-19 Data Portal (https://www. covid19dataportal.org/) and PDBe-KB COVID-19 Data Portal (https://www.ebi.ac.uk/pdbe/covid-19) among others. Moreover, chemical, functional and energetic characteristics are effective data, which may be gained from PDB to describe the potential capabilities for each individual molecule. These properties belonging to each structure and organisms may support us to determine the potential drug targets for drug design and vaccine preparation [22]. As an important documentary evidences, 210 new molecular entities (NMEs) were discovered and developed during a period of 2010–2016 and then were approved by the US Food and Drug Administration (FDA). The primary 3D structural data and information belonging to all of these NMEs compartments, were first produced and released via PDB archive. The representation of the related structures encouraged pharma companies to finance in drug discovery and development [2, 23]. Due to this fact, the aim of this review article is to show the vital importance of RCSB PDB as a virtual information “treasure” for research in biotechnology.

METHODS (LITERATURE SEARCH)

The design of the present manuscript is a narrative review, with the aim of critically analyzing and contextualizing the present knowledge and future perspectives on PDB. To formulate the present manuscript, a literature search was performed by the authors in the PubMed/MEDLINE, SCOPUS, EMBASE, and Web of Science databases up to 1st of September, 2021. No restrictions on article type, language or year of publication were set. The authors examined the primary search results and selected papers based on their suitability to be included in this review paper. After the selection of appropriate articles, the reference lists of these papers were also screened for relevant articles. Additionally, in case of some sub-topics of the review, authors also used references from their personal collection, totaling in n 5 106 references.

PROTEIN DATA BANK (PDB)

The establishment of PDB in 1971 as an effective global open access resource for biological digital data was initiated by the introduction of only seven structures of proteins; and now at the time of writing this article PDB houses >182,600 biological macromolecule structures (https://www.rcsb.org/) pertaining to DNAs, proteins, RNAs, these biological molecules complexes with other molecules (e.g., drugs). The foundation of PDB as a unique feature was happened for the first time in the world’s science history. Nowadays, PDB is identified as a remarkable gold standard and a great investment for archiving digital data regarding 3D structures of biological molecules. Therefore, PDB currently is known as an outstanding reference for researchers, trainers and students in the fields of applied and basic sciences associated with biology and biomedicine [23, 24]. For ensuring the highly validation and well-expertized biocurated of archived 3D macromolecular structures in PDB, the International consortium of wwPDB (RCSB PDB [25], PDB in Europe (PDBe) [26], PDB Japan (PDBj) [17] and Biological Magnetic Resonance Data Bank (BMRB) [27, 28]) (Fig. 1) has launched the OneDep software system which is known as a deposition-biocuration-validation tool [29]. These evaluations are achieved through professional expertized processes e.g., 3D cryo-electron microscopy (3DEM), X-ray crystallography and NMR [29]. Indeed, OneDep covers the wwPDB consortium through its unified software tool for deposition, biocuration and validation of the represented archived data associated with macromolecular structures [28]. To promote the validation and the quality of archived structures data in the wwPDB archive, availability of raw experimental data is enforced. OneDep system controls any ambiguity issues associated with experimental data and/or atomic models. This process facilitates the following handling processes for depositors to check and accomplishing any correction regarding a PDB deposition. Further doubtful issues will be rechecked by the manuscript reviewers or via wwPDB biocurators. To reduce the duration of validation process and to convene the validation task forces (VTFs) and effective validation metrics, the wwPDB has recruited a the OneDep software tool (https://deposit.wwpdb.org) for depositors server (https:// validate.wwpdb.org/) [29] to check the experimental methodology containing electron microscopy [30], electron crystallography [31], solid-state- and solution NMR [31, 32], neutron diffraction [33], X-Ray diffraction [34, 35], fiber diffraction [24].

PROTEIN DATA BANK (PDB)

The establishment of PDB in 1971 as an effective global open access resource for biological digital data was initiated by the introduction of only seven structures of proteins; and now at the time of writing this article PDB houses >182,600 biological macromolecule structures (https://www.rcsb.org/) pertaining to DNAs, proteins, RNAs, these biological molecules complexes with other molecules (e.g., drugs). The foundation of PDB as a unique feature was happened for the first time in the world’s science history. Nowadays, PDB is identified as a remarkable gold standard and a great investment for archiving digital data regarding 3D structures of biological molecules. Therefore, PDB currently is known as an outstanding reference for researchers, trainers and students in the fields of applied and basic sciences associated with biology and biomedicine [23, 24]. For ensuring the highly validation and well-expertized biocurated of archived 3D macromolecular structures in PDB, the International consortium of wwPDB (RCSB PDB [25], PDB in Europe (PDBe) [26], PDB Japan (PDBj) [17] and Biological Magnetic Resonance Data Bank (BMRB) [27, 28]) (Fig. 1) has launched the OneDep software system which is known as a deposition-biocuration-validation tool [29]. These evaluations are achieved through professional expertized processes e.g., 3D cryo-electron microscopy (3DEM), X-ray crystallography and NMR [29]. Indeed, OneDep covers the wwPDB consortium through its unified software tool for deposition, biocuration and validation of the represented archived data associated with macromolecular structures [28]. To promote the validation and the quality of archived structures data in the wwPDB archive, availability of raw experimental data is enforced. OneDep system controls any ambiguity issues associated with experimental data and/or atomic models. This process facilitates the following handling processes for depositors to check and accomplishing any correction regarding a PDB deposition. Further doubtful issues will be rechecked by the manuscript reviewers or via wwPDB biocurators. To reduce the duration of validation process and to convene the validation task forces (VTFs) and effective validation metrics, the wwPDB has recruited a the OneDep software tool

PROTEIN DATA BANK IN EUROPE (PDBE): AN EFFECTIVE PARTNER OF WWPDB

As a partner of PDB consortium, PDBe collaborates with different resources of bioinformatics to enrich its data center. PDBe represents a collection of bioinformatic data through the project of Structure Integration with Function, Taxonomy and Sequence (SIFTS, http://pdbe.org/sifts/) [38]. The SIFTS project provides huge amounts of data pertaining to protein sequences and structures and annotations. This project bridges the core resources of PDBe and the Universal Protein Resource (UniProt) Knowledgebase (UniProtKB, http://uniprot.org) at the European Bioinformatics Institute (EMBL-EBI; http://www.ebi.ac.uk) [38, 39]. A portion of annotation resources which cover the SIFTS project data are consisted of CATH (https://www.cathdb.info) [40], Ensembl (www.ensembl.org) [41], Gene3D (http://gene3d.biochem. ucl.ac.uk/Gene3D/) [40, 42], Gene Ontology Annotation (GO/GOA) (http://www.ebi.ac.uk/GOA) [43], HomoloGene (https://www.ncbi.nlm.nih.gov/homologene) [44], Integrated relational Enzyme database (IntEnz) (http://www.ebi. ac.uk/intenz) [45], Integrative classification of Protein sequences (InterPro) (https://www.ebi.ac.uk/interpro/) [46], Protein families database (Pfam) (http://pfam.xfam.org/) [47], NCBI Taxonomy (https://www.ncbi.nlm.nih.gov/ taxonomy/) [48], PubMed (http://www.ncbi.nlm.nih.gov/ pubmed) [49] and Structural Classification of Proteins (SCOP) (http://scop.mrc-lmb.cam.ac.uk) [50]. In addition to SIFTS, FunPDBe is another project which supports Protein Data Bank in Europe-Knowledge Base (PDBe-KB) (https://pdbe-kb.org). In another word, the PDBe-KB contains all the data belongs to the projects of SIFTS and FunPDBe. The functional annotations and predictions associated with molecular structures data in the PDB archive are merged and compared through PDBe-KB [51]. Indeed, PDBe-KB supports the enhancement of annotations visibility disseminated by data resources and simultaneously decreases the splitting of annotations [51]. The structural data belonging to PDB are applied via a huge number of scientific software tools and data resources. In parallel with this feature, several numbers of these data resources promote the biological context of macromolecular structures through adding a wide range of effective annotations associated with biophysical and biochemical characteristics relating to data [51]. Due to this knowledge, biomacromolecular tunnels and pores, molecular pockets and channels [52], ligand binding sites [53–55], interactions between biomolecar complexes [56], structural and functional analyses of single nucleotide polymorphisms (SNPs) in biomolecules [57] and proteins catalytic sites [58, 59]. It is important that, several effective centers for bioinformatics e.g., InterPro [46], MobiDB (https://mobidb.org/) [60], PDBsum [61], PDBj [62], Pfam [47], RSCB PDB [63, 64], Reactome (https://reactome.org) [65], SCOP2 [50, 66] and UniProt [67] count on SIFTS as an active resource data to represent fruitful links between PDB consortium and the other biological bioinformatic digital data for serving their global users with up-to-date data and information [38]. The PDBe at the European Molecular Biology Laboratory (EMBL)-European Bioinformatics Institute (EBI) manages PDBe-KB; an activity which is covered by ELIXIR 3DBioInfo community [16, 68, 69]. Molecular recognition of inhibitors, signaling molecules and adaptors and substrates determine the strength of protein functions. Molecular dynamics and the dynamic characteristics of protein molecules are directly involved in spatial configuration and folding and unfolding activities of proteins. In this regard, a mass of software tools and systems has been designed and made [70– 74]. The annotations pertaining to structural and functional data associated with proteins represent an effective activity in the field of protein engineering (e.g., antibodies and enzymes). Due to this fact, the canonical structures were identified in spatial configurations of antibodies’ 3D structures within their hypervariable domains. Indeed, the pivotal role of biocomputational methods in determination of canonical structures in 3D structures belonging to immunoglobulin molecules led to influential progression in predictive procedures through the bioinformatic and computational tools and techniques to obtain effective and accurate structural data in antibodies and other proteins. The effective and strong employment of bioinformatic and biocomputational procedures and methodologies in protein engineering resulted in development and progression in biotechnology through the establishment of a significant number of biotechnological companies to represent influent clinical procedures, tools and methodologies for advanced research fields [68, 75, 76]. ELIXIR encompasses a wide range of platforms which is able to support different digital data centers around Europe. The PDBe and InterPro – as the core digital resources of ELIXIR – are linked to other important annotation and structure prediction resources including CATH-Gene3D [42], FUGUE [77], GenTHREADER [78], PHYRE [79], SUPERFAMILY [80] and SWISS-MODEL [81]. Moreover, since 2018 BRENDA enzyme data base (https://www. brenda-enzymes.org) is known as the ELIXIR core data resource (https://elixir-europe.org/platforms/data/core-dataresources), too [82, 83]. BRENDA as a continuous curated system releases effective and reliable data, updated categorization of enzymes and simultaneously involves new identified enzymes. BRENDA shares new and high-quality data to support the needs of global users in the fields of biotechnology, systems biology, pharmaceutics, and medicine [82]. The core data resource of BRENDA belongs to German Network for Bioinformatics Infrastructure (de.NBI (https://www.denbi.de/)) which is covered by the German Node of ELIXIR [82, 84]. The availability, 3D visualization and structural analyses of macromalecules constitute the core of structural biology and structural bioinformatics. Hence, the recruitment of MolpViewer as a part of the Molp open-source project supports the development of a common library and tools for web-based molecular visualization, graphics and analyses. This software tool covers services for the structural biology and structural bioinformatics to feed international PDB consortium [68, 73, 85].