The RCSB Protein Data Bank: new resources for research and education

Peter W. Rose1,*, Chunxiao Bi1, Wolfgang F. Bluhm1, Cole H. Christie1, Dimitris Dimitropoulos1, Shuchismita Dutta2, Rachel K. Green2, David S. Goodsell3, Andreas Prlić1, Martha Quesada2, Gregory B. Quinn1, Alexander G. Ramos1, John D. Westbrook2, Jasmine Young2, Christine Zardecki2, Helen M. Berman2 and Philip E. Bourne1,4,*

1San Diego Supercomputer Center, University of California San Diego, La Jolla, CA 92093-0743, 2Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, 174 Frelinghuysen Road, Piscataway, NJ 08854-8076, 3Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037 and 4Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, 9500 Gilman Drive, Mailcode 0743, La Jolla, CA 92093-0743, USA

Received September 17, 2012; Revised October 19, 2012; Accepted October 30, 2012

ABSTRACT
The Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) develops tools and resources that provide a structural view of biology for research and education. The RCSB PDB web site (http://www.rcsb.org) uses the curated 3D macromolecular data contained in the PDB archive to offer unique methods to access, report and visualize data. Recent activities have focused on improving methods for simple and complex searches of PDB data, creating specialized access to chemical component data and providing domain-based structural alignments. New educational resources are offered at the PDB-101 educational view of the main web site such as Author Profiles that display a researcher’s PDB entries in a timeline. To promote different kinds of access to the RCSB PDB, Web Services have been expanded, and an RCSB PDB Mobile application for the iPhone/iPad has been released. These improvements enable new opportunities for analyzing and understanding structure data.

INTRODUCTION
The RCSB Protein Data Bank (RCSB PDB) (1) provides access to the data in the PDB, the single archive of experimentally determined structures of nucleic acids, proteins and complex assemblies (2). The public archive currently contains >84000 entries, derived data files and related data dictionaries. With >570000 files, the PDB requires >130 GB of storage space. Data are updated weekly and loaded into the relational database that supports the web site.

The PDB is maintained by the members of the Worldwide PDB (wwPDB): RCSB PDB (USA) (1,3), PDB in Europe (PDBe, http://pdbe.org) (4), PDB Japan (PDBj, http://pdbj.org) (5) and BioMagResBank (http://bmrb.wisc.edu) (6). These member organizations host deposition, processing and distribution centers for PDB data. Data are deposited to the PDB, curated and annotated following wwPDB standards, and then made available on an FTP server. Each wwPDB partner offers unique ‘views’ of PDB data through the different query, analysis and visualization tools provided on their respective web sites.

The RCSB PDB web site currently hosts ~240000 unique visitors per month (based on the number of unique IP addresses), an increase from the 180000 visitors last reported in 2011 (3). Web site users represent a variety of interests, including students (ranging from elementary school to graduate school), academic and industrial researchers, bench scientists and programmers and web developers. To better serve these interests, the RCSB PDB home page and individual ‘Structure Summary’ pages can be customized by users by moving relevant data widgets (7) to different locations on the page, and hiding or minimizing areas of less interest. For education-focused browsing, a separate PDB-101 section offers related

*To whom correspondence should be addressed. Tel: +1 858 822 5497; Fax: +1 858 822 0873; Email: pwrose@ucsd.edu
Correspondence may also be addressed to Philip E. Bourne. Tel: +1 858 534 8301; Fax +1 858 822 0873; Email: pbourne@ucsd.edu

© The Author(s) 2012. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/3.0/), which permits non-commercial reuse, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com.
materials such as the ‘Molecule of the Month’ columns that
tell the functional story of selected macromolecules.

PDB data can be searched in many different ways. The
top menu bar can be used to perform simple searches,
including author name, molecule name, sequence or
ligand ID. ‘Advanced Search’ can be used to build
queries with multiple constraints, such as ‘find all protein
homodimers bound to DNA’. The ‘Browse Database’
option allows exploration of the PDB archive using differ-
ten hierarchical trees. Browsers are available to search
for related terms and structures based on many different
classifications, such as Biological Process, Cellular
Component, Molecular Function (8), Enzyme Com-
misson number (http://www.chem.qmul.ac.uk/iubmb),
Transporter Classification System (9), and structure
classifications SCOP (10) and CATH (11). Data distribu-
tion summaries, shown as pie charts and lists of hyperlinks,
are available for standard features of PDB entries (reso-
lution, release date, experimental method, polymer type,
organism and taxonomy). These drill-down distributions
provide another way to browse and select data from the
whole archive or any search results.

Query results can be refined, used to explore individual
structures and exported to generate interactive and tabular
reports. Tabular report features include online data
sorting, column customization, filtering and output to
other report formats. These reports also contain data
from, and links to, external resources.

User feedback is an important influence on the evolution
of the RCSB PDB resource. Recently added features, some
developed based on this feedback, are described here.

NEW WEB SITE FEATURES

Simple searches

The most common uses of the web site are simple text
searches. To further improve the text search, we have
added an autocomplete feature to guide the user to more
specific results. After typing a few letters in the top bar, a
suggestion box organizes specific result sets in different
categories. Each suggestion, which includes the number
of results, links to the set of matching structures. Some
of the suggestions use external data resources, such as the
NCBI organism taxonomy tree (8,12). These possible
matches can be especially helpful for finding structures
when using common or vague search terms, as is shown
in Figure 1 for the term ‘virus’.

The top bar search is context-specific and intelligently
detects the type of user input. Entering a sequence text
string in the search box returns possible Basic Local
Alignment Search Tool (13) search options. Chemical
formulas and SMILES strings (14) are also recognized,
e.g. the SMILES string for adenosine ‘Nc1ncnc2ncnc12’
yields choices of substructure, exact structure or structure
similarity searches. If the suggestions are not what the user
is looking for, it is still possible to perform a standard
text search of the PDB entry (in mmCIF format) by
pressing enter or clicking on the search icon.

Top bar simple searches can also be limited to specific
categories by selecting the ‘Author’, ‘Macromolecule’,
‘Sequence’ or ‘Ligand’ icon. The ‘Author’ icon restricts
searches to the names of depositors or primary citation
authors. The ‘Macromolecule’ icon returns structures
based on polymer names from the PDB and associated
entries in cross-referenced sequence databases like
UniProtKB (15). For example, typing ‘caspase’ provides
suggestions for different types of caspases. By selecting
‘caspase-1’ and examining the PDB entries returned, it
becomes obvious that the actual search is for PDB
structures with cross-references to various UniProtKB
entries for caspase-1 from different organisms. The
‘Sequence’ icon reveals a link to additional options for
selecting the method and the parameters for a sequence
search. Similarly, the ‘Ligand’ icon links to further
options, including a chemical structure editor to draw a
structure, and a form to search for ligands by name,
identifier, formula and molecular weight.

New advanced search features

Advanced Search expands on the search functionality of
the top bar searches by using additional and more specific
data categories. Advanced Search has the capability of
combining multiple searches of specific types of data in a
logical AND or OR. The result is a list of structures that
comply with ALL or ANY search criteria, respectively.

New Advanced Search options are available to search
by: ‘All/Experimental Type/Molecule Type’ to quickly
access all PDB entries or a subset based on experimental
and macromolecular type, structure determination/
phasing method (e.g. molecular replacement, MAD or
SAD), ‘Link Records’ to find structures containing inter-
residue connectivity (LINK records in PDB entries) that
cannot be inferred from the primary structure, structures
determined by electron microscopy for which experimental
data files are available in the PDB or at the Electron
Microscopy DataBank (16) and Pfam ID (17).

All Advanced Search query results can be further
refined, filtered to remove similar sequences or used to
generate reports.

Structure alignments

Sequence and structure alignments are standard methods
for analyzing the evolutionary and functional relationship
between proteins (18–23). The Protein Comparison Tool
offers a number of sequence and structure alignment
algorithms for a detailed analysis of pairwise relationships
(24). Additional algorithms are available via submission of
alignments to some of the leading external web servers
(25–28). The Protein Comparison Tool has also been
used to provide the pre-calculated alignments, updated
weekly, of a representative subset (based on sequence
identity) of the PDB (24). The first version of this tool
was based on alignments of whole protein chains. This
has recently been refined to provide alignments on a
domain basis.

The calculation based on domains extends our sequence
clustering approach. To remove redundancy, we start with a
40% sequence identity clustering procedure based on
complete polypeptide chains, and select a representative
chain from each sequence cluster (3). If the representative
chain contains multiple domains, each is included. SCOP 1.75 domain assignments are used when available; otherwise, assignments are computed using ProteinDomain Parser (PDP) (29). Pairwise alignments of the domains are performed with the jFatCat version (24) of FatCat (22).

For each PDB entry, the ‘3D Similarity’ tab provides a visual summary of the protein chains. Figure 2 highlights how the residues listed in the sequence (SEQRES) and in the atom records (ATOM) map onto the relevant parts of the UniProtKB sequence, along with annotations from DSSP (32), SCOP, PDP (29) and Pfam (33).

The results of the pre-calculated database searches are shown in a table that displays the most important calculated alignment scores (Figure 2). For multi-domain proteins, it is possible to switch between the results for different domains by selecting a domain from the pull-down menu above the table, or by clicking on a domain in the sequence image.

The results table can be sorted and filtered, and links to the 3D structure alignment in Jmol (http://www.jmol.org) (34) (Figure 2) and to information about similar domains.

Ligand reporting and visualization

Information about the chemistry and structure of all small molecule components found the PDB is contained in the Chemical Component Dictionary maintained by the wwPDB at wwpdb.org (35). As described earlier, specialized ligand queries can be made using the top bar search or Advanced Search. Special support is also offered...
for the analysis of ligands associated with PDB entries. The RCSB PDB web site builds on the functionality developed for the small molecule resource Ligand Expo (http://ligand-expo.rcsb.org) (36) by providing special support for the analysis of ligands associated with PDB entries.

Any ligands included with a PDB entry are listed in the ‘Ligand Chemical Component’ widget of the entry’s ‘Structure Summary’ page. This area displays the name and formula of each ligand, links to the summary page for the ligand and provides access to 3D visualization of the ligand in the context of that particular PDB entry using the Ligand Explorer viewer (37). For non-trivial ligands, a PoseView (38) interaction diagram shows which atoms or areas of the ligand and the polymer interact with each other, as well as the type of interaction.

‘Ligand Summary’ pages are organized into widgets highlighting different types of hyperlinked information, similar to Structure Summary pages for individual PDB entries. These widgets provide an overview of the ligand, with links to PDB entries where the component appears as a non-polymer or as a non-standard component of a polymer, links to ligand summary pages for similar ligands and stereoisomers, 2D and 3D visualization and links to many external resources. Ligand Summary pages also display information about molecules that have been annotated as having sub-components. For example, the summary page for ligand 0GM lists the sub-components with identifiers BNA, GLU, STA, LEU and TRJ that are connected with peptide-like or other bonds.

Ligand Summary Reports can be generated for query result sets and downloaded in a text file or a spreadsheet. These reports include information about the selected ligands, such as formula, molecular weight, name, SMILES string, which PDB entries are related to the ligand and how they are related. Each ligand included in the report can be expanded to show a sub-table of all related PDB entries that contain the ligand, the entries that contain the ligand as a free ligand and entries that contain the ligand as part of a polymer.

Visualization of molecular surfaces

Protein Workshop (37) is one of several 3D molecular viewers offered from the RCSB PDB web site. It offers quick default styles and views, with additional appearance options. Chains and atoms can be selected by either clicking on the structure or molecules displayed as a tree.

Protein Workshop now supports molecular surfaces to aid in the display of quaternary structure, protein–protein interactions and binding sites. Surfaces are created for all macromolecule chains in a PDB entry using the Euclidean distance transform algorithm from Xu and Zhang (39). For biological assemblies, surfaces are generated using the symmetry operation of the space group, which allows the display of even the largest assemblies in the PDB [i.e. the PBCV-1 virus capsid with 5040 chains, PDB ID 1M4X (40)] on a standard laptop computer. Surfaces can be color coded by chain, entity (unique macromolecules) and hydrophobicity. Color-blind friendly color schemes were adopted from ColorBrewer, a tool for selecting color schemes for maps (41). In addition, options to export high-resolution images with custom sizes for publications and posters are available for the three RCSB PDB viewers: Protein Workshop, Simple Viewer and Ligand Explorer.
WEB SERVICES

Web Services are used by software tools that efficiently and remotely interact with PDB data on the fly, eliminating the need for local data storage. The RCSB PDB hosts RESTful search and fetch services that return XML files in response to URL requests. Search services return PDB ID lists for queries based on Advanced Search queries. Fetch services return data (such as entity descriptions, ligand information and external annotations) for a given list of IDs. In addition to the services reported previously (3), new services are described in Table 1. For example, access to sequences released ahead of the structure is now frequently used by structure prediction servers for blind predictions (such as http://www.cameo3d.org/). More than 100 data fields can be exported in a generic way using the tabular report service. For example the URL

http://www.rcsb.org/pdb/rest/customReport?pdbids=3IP0,1M15,2XBP,3IQU,2IIM&customReportColumns=structureId,structureTitle,resolution,rFree&service=wsfile&format=csv

specifies a Web Service request for a list of PDB IDs with four data fields in the comma-separated value file format.

RCSB PDB MOBILE

A simplified interface to the RCSB PDB is available as an app for the iPhone/iPod and the iPad (Figure 3). The app offers special features, including a simplified search for macromolecule name, author name and PDB ID. Query results, displayed in a single page listing, can be filtered by author name, title and organism. A macromolecule image and the PubMed abstract (when available) for individual entries are displayed when the user selects an entry from a returned query results list.

RCSB PDB Mobile also provides a listing of the most recently released PDB entries, and can be used to explore the archive of ‘Molecule of the Month’ articles and RCSB PDB news. Users can connect to their MyPDB account, a service that allows users to store queries and structure annotations.

RCSB PDB Mobile includes an integrated molecular viewer, NDKMol, developed by collaborator Dr. Takanori Nakane, Kyoto University. The viewer presents an interactive molecular rendering using downloaded PDB format files. The user is able to modify the appearance of the rendering by changing display settings such as display style (Ribbon, C-alpha trace, strand or B-factor tube), ligand/HET atom style (sphere, stick or line), nucleotide base style (line or polygon), color scheme (spectrum, by chain by secondary structure, polar/non-polar or B-factor), symmetry mates (biological assembly or crystal packing) and several other options.

A version of the app for the Android platform is in development.

PDB-101: EDUCATIONAL FEATURES

The volume and complexity of PDB data can pose a challenge for users, particularly beginning students.

---

Table 1. Recently introduced RESTful Web Services

<table>
<thead>
<tr>
<th>Web service</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-released sequences</td>
<td>Access sequences in FASTA format for entries that have been deposited to the PDB, but are on hold until publication or a specified release date.</td>
</tr>
<tr>
<td>Custom reports</td>
<td>Create tables of sequence, structure, function, ligand information, experimental details and structure annotations in comma-separated value file, XML or MSExcel format.</td>
</tr>
<tr>
<td>Pfam annotations</td>
<td>Retrieve Pfam domain annotations, calculated by running Pfam’s Hidden Markov Models (42).</td>
</tr>
<tr>
<td>Domain-based structural alignments</td>
<td>Retrieve structural neighbors and alignment scores.</td>
</tr>
</tbody>
</table>

A full list of web services and examples are available at: http://www.rcsb.org/pdb/software/rest.do.
To support non-experts interested in exploring biomolecular structure, RCSB PDB educational resources and features (44,45) have been packaged together to form the ‘PDB-101’ web site that is accessible from the main web site via the PDB-101 logo. PDB-101 currently supports five main features: the archive of ‘Molecule of the Month’ columns, which describe biomolecular structure and function for general audiences; Educational Resources, including posters and animations; the ‘Understanding PDB Data’ resource for learning about data files and structure determination methods; the Structural View of Biology browser and Author Profiles.

Structural view of biology

The Structural View of Biology, shown at the PDB-101 landing page, was designed to encourage self-guided exploration of the PDB by non-experts. It is separated into six functional categories, such as ‘Enzymes’ and ‘Protein Synthesis’, and allows users to browse based on the biological properties typically used in biology and chemistry education. The topics can be browsed down to individual ‘Molecule of the Month’ features, which include annotated Jmol views and links to simplified summary pages highlighting specific example entries. This provides novice users with a subset of the PDB archive selected for its utility in education.

Author profiles

A unique historical and educational tool enabled by the database, ‘Author Profile’ displays a vertical timeline of the structures associated with either an individual author or a structural genomics center (Figure 4). A text search form is available to find different profiles. The structures shown are selected based on author name (deposition or primary citation author), and ordered by deposition date. Unique structures, denoted by a blue background and shown with a large image, indicate the first structure of a polymer or polymer complex deposited by the researcher. Subsequent structures that contain the same set of UniProtKB cross-reference identifiers (15) are displayed with a smaller image.

SUMMARY

We continue to build and improve RCSB PDB resources to enable a structural view of biology. New search options include search suggestions and Advanced Search options that guide the user to more specific search results. The Author Profile tool offers a new way to explore structures solved by individual authors and structural genomics centers. Structural alignments are now available for representative domains, rather than just protein chains. Ligand searching, reporting, and visualization has been improved. The addition of surfaces to the 3D viewers enables the analysis of ligand binding sites, protein–protein interactions and quaternary structure. Web Services have been expanded to include pre-release sequences and a generic mechanism to retrieve PDB data through tabular report services. To cater to the rapidly growing number of mobile users, we have deployed RCSB PDB Mobile for the iPhone and iPad, and an Android version is under development. A new educational section, PDB-101, hosts the educational content and provides a hierarchy to browse ‘Molecule of the Month’ articles. New web site releases are announced on the ‘What’s New’ widget on the home page, and in weekly news announcements.

ACKNOWLEDGEMENTS

The authors thank BioSolveIT GmbH (http://www.biosolveit.de) for access to PoseView, and ChemAxon (http://www.chemaxon.com) for providing Marvin Sketch, JChem Base and Standardizer for the chemical structure search. Dong Xu and Yang Zhang provided source code for the Euclidean distance transform algorithm for calculating molecular surfaces. Takenori Nakane developed an Objective-C version of the NDKViewer for the RCSB PDB Mobile. Access to binding affinity data was provided by Michael Gilson (BindingDB), Heather Carlson (BindingMOAD) and Renxiao Wang (PDBbind-CN). In addition, we also thank all users who provided feedback, and RCSB PDB staff, past and present, for suggestions, critical review and testing of new features. The RCSB PDB is managed by two members of the RCSB: Rutgers and UCSD, and is a member of the wwPDB.

FUNDING

National Science Foundation [NSF DBI 0829586]; National Institute of General Medical Sciences...
REFERENCES


