Structural database resources for biological macromolecules

Luciano A. Abriata

Corresponding author: Luciano A. Abriata, Laboratory for Biomolecular Modeling, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, and Swiss Institute of Bioinformatics, Switzerland AAB014 Station 19 - 1015 Lausanne, Switzerland. E-mail: luciano.abriata@epfl.ch

Abstract
This Briefing reviews the widely used, currently active, up-to-date databases derived from the worldwide Protein Data Bank (PDB) to facilitate browsing, finding and exploring its entries. These databases contain visualization and analysis tools tailored to specific kinds of molecules and interactions, often including also complex metrics precomputed by experts or external programs, and connections to sequence and functional annotation databases. Importantly, updates of most of these databases involves steps of curation and error checks based on specific expertise about the subject molecules or interactions, and removal of sequence redundancy, both leading to better data sets for mining studies compared with the full list of raw PDB entries. The article presents the databases in groups such as those aimed to facilitate browsing through PDB entries, their molecules and their general information, those built to link protein structure with sequence and dynamics, those specific for transmembrane proteins, nucleic acids, interactions of biomacromolecules with each other and with small molecules or metal ions, and those concerning specific structural features or specific protein families. A few webservers directly connected to active databases, and a few databases that have been discontinued but would be important to have back, are also briefly commented on. Along the Briefing, sample cases where these databases have been used to aid structural studies or advance our knowledge about biological macromolecules are referenced. A few specific examples are also given where using these databases is easier and more informative than using raw PDB data.

Key words: protein; nucleic acids; ligand binding; interactions; dynamics; PDB mining

Introduction
The worldwide Protein Data Bank [1] (referred here simply as ‘PDB’) is a partnership of servers for the collation, maintenance and distribution of macromolecular structure data (Figure 1A), which stand as the primary data resource in structural biology, containing all structures of biological macromolecules determined by NMR, X-ray or neutron diffraction and cryo-electron microscopy. PDB entries include structures of isolated proteins, nucleic acids, their complexes with each other as well as with lipids, cofactors, substrate mimics, regulators, inhibitors, etc., adding up to >117 000 entries by April 2016 (for a recent discussion of extensive statistics, see the review by Berman et al. [2]). Naturally, each PDB entry brings important insights into the structural and functional biochemistry related to the original subject that motivated the study. But on top of that, the data bank as a whole is a reservoir of broad rich information about biomolecular structure, dynamics and conformational variability, interactions, hydration, etc., and somehow also reflects the state of the art of structure determination methods and programs.

The worldwide PDB is a large and complex database, and each of its entries contains large amounts of data besides the already rich information of its atomic coordinates. Thus, despite the PDB servers having powerful querying interfaces, it turns out that for several goals it is often easier, faster and even more informative to resort to the specialized databases derived from the PDB. These PDB-derived databases (Figure 1B) are made of PDB entries prefiltered by the types of molecule(s) they contain,
the level of sequence redundancy and parameters about the quality of the structures; many are cross-annotated with other types of data and classifications, and most output graphical information and even interactive 3D visualizations of the relevant molecules besides the alphanumerical results. Importantly, most PDB-derived databases have the added value that they are built, maintained and updated by experts in a specific field of structural biology, therefore they perform analyses and calculations on the coordinates that would be cumbersome for a nonexpert user to carry out.

This Briefing describes the most used databases derived from the PDB, which are currently active, curated and updated, giving database-specific remarks and general comments. These databases (Figure 1) simplify the process of finding and analyzing specific sequences, molecules or their interactions, and facilitate browsing and mining PDB entries related by a property, interaction or molecule of interest. Specific examples about the utility of these databases are referenced to the literature, and some specific test cases are presented (Figures 2–5).

A few important Web servers related to databases and a few discontinued databases of special importance are also mentioned.

**PDB data centers and PDB-derived databases that facilitate browsing through PDB entries at a global level**

PDB entries are available chiefly from three data centers: the Research Collaboratory for Structural Bioinformatics PDB [3], PDB Europe [4] and PDB Japan [5], whose contents are centralized by the worldwide PDB but each having its own series of original ways to search, browse and display data, as well as different associated analysis tools and connections to external databases. Both RCSB PDB and PDB Europe have extensive search and download facilities, simplified programmatic access,
online 3D visualization capabilities and internal tools for performing sequence, motif and structure analysis, structural alignments, etc. Both have also pre-built pictures of each PDB entry, those of PDB Europe being more varied in their colors, orientations and rendering of different units and ligands, and of better quality. Importantly, both RCSB and PDBe display the most likely biological assembly as determined by the PISA server, which is helpful to better validate and even find overlooked interactions [6, 7]. On its side, the Japanese PDB site [5] contains two unique resources: eF-site [8], a database of precomputed molecular electrostatic potential surfaces for PDB entries useful to quickly visualize electrostatic surfaces without having to run lengthy Poisson-Boltzmann solvers; and the associated eF-seek server [9], which attempts to predict functional sites by searching for molecular surfaces of similar shapes and electrostatic signatures. Last, each of the sites mentioned in this paragraph contains its own educational resources about protein structures for experts and nonexperts.

Aiming to facilitate browsing and structure visualization online through a picture-rich interface, PDBsum [10] lists brief descriptions of all PDB entries including precomputed views of most molecular components, easy-to-browse displays of structure resolution and R-factors, protein sequences with annotated PFAM and CATH domains, wire diagrams and Ramachandran plots for protein components, varied information about ligands and metal sites, and summaries of interactions between molecular components. PDBsum further offers precomputed predictions about potential pores and tunnels, quick links to search for PDB entries with similar sequences, and graphics-aided display of sequence variants annotated with predicted changes in interactions and solvent accessibility. PDBsum contains also direct links to the main PDB entries at RCSB PDB and PDB Europe, to the literature where the entries have been cited, to other databases that summarize information about PDB entries, to databases and servers of quality check reports, to databases of annotations about secondary and quaternary structures, motifs, domains, functions, sequence alignments, ontology terms, possible orientations in membranes and to community-annotated resources, among others. Part of the rich connectivity among the databases covered up to this point is schematized through an example in Figure 2A.

With a graphical way to browse the PDB, PDB-Explorer provides an online interactive map built from a high-dimensional fingerprint of atom pairs that reflects protein shapes, mapped to two dimensions through principal components analysis. The PDB-Explorer interface loads in less than a minute and then allows searching for structures similar to those of a given query provided in atomic coordinates with smooth and essentially instantaneous feedback [11].

Last in this section, PDB_SELECT [12] compiles minimal lists of representative X-ray structures at a sequence identity cutoff of 30%, of the highest available quality (measured as a combination of resolution and R-factor). This database is particularly useful for mining studies as it reduces the number of PDB entries to analyze, minimizing redundancy as well as noise and errors on the mined values. Similarly, PDB-REPRDB is a database of representative protein chains [13].

## Detecting errors and rebuilding PDB structures

PDBREPORT [14] is a database that describes structural problems in PDB entries. It summarizes anomalies and errors in...
structures of the PDB computed by WHAT_CHECK, in the form of
text and graphics reporting on differences between positions or
angles of multiple copies of a molecule, presence of ligands of
unknown topologies, outliers in Ramachandran plots, unex-
pected and missing atoms, chain breaks, suspicious B-factors
and occupancies, unusual geometries (bond lengths, angles, tor-
sions, planarity of aromatic molecules and puckering of proline
residues and carbohydrates, unusual backbone conformations),
unusual packing including unsatisfied hydrogen bonds, poten-
tial problems with solvent molecules and ions, possible histi-
dine/asparagine/glutamine flips and more.

PDB_REDO [15] is a database of automatically re-refined PDB
entries. This is important because many PDB entries are old and
suffer from problems that modern software, methods and
knowledge about biomolecular structure can fix. This server is
integrated directly into the Coot and Yasara programs for pro-
tein crystallography, facilitating the comparison of original and
optimized structures and electron density maps. As an example
of its importance beyond the curation of specific errors in PDB
structures, high-throughput analyses based on PDB_REDO led to
a large compilation of peptide planes predicted to be flipped
and peptide bonds predicted to be swapped between trans and
 cis conformations in the PDB [16].

All these databases are advised for preliminary checks of
PDB structures before launching calculations that rely heavily
on accurate coordinates and structure completion, for example,
when setting up molecular dynamics simulations. In case the
user cannot find an entry of interest, the WHYNOT database
(http://www.cmbi.ru.nl/WHY_NOT2) attempts to explain why a
given PDB entry is not available in the PDB-derived databases
from the Vriend group, PDBREPORT, B-factor Data Bank (BDB),
PDB_REDO, PDBFINDER and PDB_SELECT [17].

PDB-derived databases that connect protein sequence, structure and dynamics

PDBFINDER [18] is a particularly useful database that provides
precomputed secondary structures directly in text format
aligned to protein sequences, for each entry of the PDB, alto-
gether in a single text file. A new version, PDBFINDER II, in-
cludes also information about chain breaks, quality indicators,
B-factors and much more, also in a single compact text file.
These databases allow fast searches through sequences and
residue-specific information, largely simplifying searches and
comparisons of such kind of data through simple Linux scripts
as exemplified in Figure 3. Related to the problem of mapping
sequences, secondary structures and PDB structures, ChSeq is a
database focused on chameleon sequences, i.e. sequences that
adopt radically different conformations across PDB entries [19].

There are also PDB-derived data sets focused on protein con-
formational variability, among others PCDB [20], CoDNaS [21]
and PDBFlex [22]. These data sets take advantage of the fact that
several proteins have been crystallized in different conforma-
tions arising from varying point group crystals, pH, precipi-
tants, bound ligands, mutations, etc. The idea behind these
databases is that the conformational variability observed in dif-
ferent structures of a given protein somehow reflects the under-
lying conformational states that the protein can adopt [23], with
the caveat that some might not truly exist in the intrinsic dy-
namic landscape of the free protein in standard conditions but
rather be forced by the special conditions in which the structure
was solved (although this is somewhat considered in CoDNaS
[21]). In practice, it turns out that conformational variability as
observed across PDB structures has multiple applications, for
example, to explore the role of conformational flexibility in pro-
tein function, regulation and even evolution [24–26], to project
molecular dynamics trajectories [27] and to better discriminate
the effects of mutations on protein structure and even stability
[28].

One of the latest servers dedicated to protein variability in
the context of flexibility, PDBFlex [22], provides several unique
visualization capabilities that clearly highlight dynamics from
structures. Especially enticing is the display of an animation
that summarizes the collective protein dynamics, which are
often the most functionally relevant ones and which are hard to
observe for example in molecular dynamics simulations be-
cause they occur in slow timescales. Moreover, PDBFlex disen-
tangles local from global variability, both of which can be
inspected in plots interactively connected to the PDB structures
of the corresponding structures. Another unique feature is that
the user can download the sets of aligned protein structures in
each cluster.

B-factors are additional atom-specific outputs from the pro-
cess of structure refinement from X-ray diffraction data, often
interpreted in terms of internal atomic motions to extrapolate
information about protein dynamics. But B-factors are affected
by variables other than true dynamics, hence caution must be
taken for their interpretation. The B-factor data bank (BDB)
gathers all PDB entries with consistent B-factors that reflect
true dynamics [29].

Last in this section, the now discontinued database ProDDO
[30] was particularly interesting because it was built by text
mining the PDB files for keywords related to protein dynamics,
such as ‘disorder’, ‘gap’ (referring to unresolved residues), ‘un-
folded’, etc. A rebuild of this lexicon-based database could be-
come important in the context of automated annotations of
protein properties.

Protein flexibility and disorder from
X-ray and NMR structures

Some small protein segments and peptides of high flexibility
are resolved in NMR and X-ray structures when bound to pro-
teins, if this binding restricts motions. In this regard, an inter-
esting database, ComSin [31], compares structures of protein
complexes with structures of the isolated proteins focusing on
order–disorder transitions, with the outcome that the unbound
proteins tend to, but not always, have more disordered residues.
A related important database, of less structural content but
with more annotations, is MobiDB [32].

Many proteins and peptides that do not crystallize and/or
characterized by extensive flexibility are still represented in the
PDB, mostly solved by solution-state NMR. The Biological
Magnetic Resonance Data Bank [33] (BMRB, or biomagresbank)
is the hub that collects raw NMR observables for biomolecules,
not limited to restraints for protein structure calculation.
Entries with NMR structural restraints are interconnected to the
corresponding PDB entries; indeed submission of new NMR
structures to the PDB is entangled with submission of NMR data
to the BMRB [34]. Also, the PACSY programs provide a relational
database system to browse information from the PDB, BMRB
and SCOP databases through SQL queries [35].

In the highly dynamic extreme of the flexibility spectrum,
intrinsically disordered proteins and peptides simply lack struc-
tures that can be captured neither by crystallography nor by
NMR, and are therefore poorly represented in the PDB. Still,
their structural heterogeneity can be treated through combina-
tions of computational methods coupled to experimental
observations from solution-state techniques like NMR, fluoresc-
ence, Electron Paramagnetic Resonance and Small Angle X-ray
Scattering. Although not derived from the PDB, the Protein
Ensemble Database [36] gathers these structural ensembles in
the same format as normal PDB files, together with the experi-
mental restraints and algorithms used to generate the ensem-
bles. As such, this server is a valuable resource for research
involving disordered peptides and proteins too flexible to be
found in the PDB.

**Databases of membrane proteins**

Membrane proteins represent around 20–30% of the protein-
coding genes in an organism [37, 38], and many are vital be-
cause they are at the heart of molecular sensing mechanisms,
cell-to-cell communication processes and even the essential
process of respiration itself. Membrane proteins are harder to
work with in the laboratory, so they are much less represented
than soluble proteins in the PDB. But on the other hand, they
are the main human protein targets of current drugs [39].
Because of this, making the most out of existent structures by
mining the PDB, and using this information for simulations,
modeling and driving experiments, is of utmost importance.
There are indeed several PDB-derived databases that focus on
membrane proteins.

PDBTM [40] was the first comprehensive database of trans-
membrane proteins from the PDB, with currently >2600 entries.
It consists basically of a list of PDB files that can be browsed one
by one, downloaded entirely or downloaded by groups (alpha or
beta proteins, redundant or nonredundant data sets). From the
same lab, the TOPDB database cross-references sequences and
PDB structures of membrane proteins with information about
membrane protein topologies obtained through experiments and
bioinformatic predictions [41].

Similar to PDBTM, mpstruc (http://blanco.biomol.uci.edu/
mpstruc/) also maintains an up-to-date list of membrane pro-
teins of known 3D structure, but provides finer classifications
into monotropic proteins, transmembrane β-barrels and trans-
membrane alpha helical proteins, each further classified ac-
cording to their structures, functions and families.

One of the most popular such databases is OPM, the
Orientation of Proteins in Membranes database [42]. OPM pro-
pides PDB coordinates of integral membrane proteins, some
peripheral proteins and membrane-active peptides, pre-
oriented relative to the membrane normal for membranes of
variable thickness. Orientations are optimized by minimizing
protein transfer energies from water to membrane as computed
with an implicit solvent model [43]; and usefully, a related Web
server allows orientation of user-uploaded structures. OPM pro-
pides reasonable orientations in most cases, but if necessary
they may be refined manually and/or through coarse-grained
MD simulations. It is directly connected to the CHARMM-GUI
server [44], together simplifying largely the setup of atomistic
and coarse-grained MD simulations for membrane proteins
from the PDB (Figure 2B).

The new database MemProtMD [45] contains proteins pre-
equilibrated into explicit membranes using coarse-grained self-
assembly MD simulations. This approach is in principle
expected to be more accurate than the implicit solvent model of
the OPM server and should therefore deal better with complex
cases like that of peripheral (i.e. nonintegral) membrane pro-
teins. From these coarse-grained models it is straightforward to
inspect the lipid environment of a membrane protein and to
setup more complicated simulations, even of atomistic level.
The process of building the MemProtMD database illuminated
new structural aspects about how proteins stay embedded in
membranes, how membranes respond with deformations,
amino acid–lipid interactions and lipid-binding protein sites,
different distributions of amino acids along the membrane nor-
amal, and more; even a protocol for the identification of novel
membrane proteins from new structures emerged from that
work [45].

Connecting structure and sequence spaces for transmem-
brane proteins, the TMalphaDB and TMbetaDB Web servers
[46] allow searches of amino acid sequences (including wild-
cards) in PDB files of alpha and beta membrane proteins. After
searching, the servers allow interactive display of the matched
sequences and their structures in the found PDB entries, as
well as extraction of unique sequences and calculation of
backbone and sidechain torsional angles for the matched pep-
tide. Interesting effects of some residues on transmembrane
secondary structures were unveiled by analyses with these
servers [46].

**Databases specialized on nucleic acids**

Nucleic acids, especially RNAs, have also been the focus of
specialized databases. The first database of nucleic acid structures,
NDB [47], has been around since the early 1990s and is today the
main reference in the field. It currently contains >8000 experi-
mental structures of DNA and RNA molecules (including DNA–
RNA pairs) extracted and curated from the PDB. The structures
are annotated with information specific to nucleic acids, hardly
accessible from the original PDB entry. On incorporation of PDB
entries to NDB, the structures are analyzed through computa-
tion of structural geometries (hydrogen bonding and base-
pairing patterns, extraction of regular motifs, etc.), and classi-
fied and made searchable at the sequence, secondary structure
and structure levels. Users can carry out further specific ana-
lyses and online visualizations in 2D and 3D using several tools
available at the NDB server, with links to other resources (two
examples in Figure 4). Moreover, new tools were introduced in
2014 to analyze RNA sequences, align RNA and DNA molecules
and calculate and visualize RNA structural geometries and
base-pairing patterns, which are more complex and varied than
those for DNA. The NDB server also contains statistics about
ideal geometries for bases and sugars as well as for different
hydrogen bonding pairing patterns, including a new catalog of
base pairing in RNA, an educational section, and software for
further offline analyses.

Other relevant databases specific for RNA combine data
from several sources, including the NDB and occasionally di-
rectly the PDB sites, to provide more comprehensive annota-
tions, which are in many cases combined with sequence-
based predictions [48–52]. Most of these servers allow searches
from RNA sequences and secondary structures; some offer the
possibility of searching and analyzing RNA coordinates from
an uploaded PDB file or filter results based on geometries,
neither of which is supported by the NDB at the moment. A
particularly informative database about RNA structure is the
RNA 3D hub [51], which is in fact the source of information
for part of the precomputed structural analyses of RNA mol-
ecules at NDB.
The most interesting aspect of iPFAM is that a user can query through high-throughput analysis of interactions in the PDB. Related, although not a structural database, the iPFAM database integrates structures of all complexes involving protein, RNA and nucleic acids, a classification that the server performs. Searches can be done by PDB ID, SCOP family of proteins. It can be downloaded entirely, browsed or searched by PFAM identifier for all the interactions established by proteins of that family with proteins of all PFAM families as seen at the structural level.

Besides interactions between macromolecules, PDB entries also often contain small molecules bound to proteins, such as additives for crystallization, substrates, substrate analogs, regulators, lipids for proteins purified from membranes, and drugs or drug candidates added to proteins before crystallization or soaked into crystals. Given that the analysis of intermolecular interactions is most useful in the context of functional information about how the interaction modulates activity, databases about protein-bound ligands typically cross data from multiple sources to that of the PDB. The sc-PDB [58] is a database of ‘druggable’ binding sites from the PDB. Its construction and updates filter out solvent molecules, detergent, ions and other common additives used for protein crystallization, thus enriching relevant binding sites. It is annotated by crossing information from Uniprot and GO, includes several classification criteria and chemoinformatic descriptors of the ligands and binding sites. The output is rich in geometric and chemical data about the interactions between ligands and proteins, and including enticing yet simple 2D diagrams of the binding site besides the usual 3D visualization (example in Figure 5, top part).

Still related to targeting protein function with small molecules, the Pocketome [59] is an automatically updated database of small-molecule binding sites that includes in most cases more than one PDB entry per site, allowing the user to inspect the structural variability of each binding site (example following from scPDB in Figure 5, bottom part). From the same authors, PeptiSite [60] is a database of peptide-binding sites that also groups entries into ensembles allowing inspection of structural variability in the docked peptides and binding sites. Last, PDTD [61] is a database of protein targets with known structures, built from the PDB based on actual functional data extracted from the literature and from several databases of therapeutic compounds. It is dominated by enzymes as targets, but includes also receptors, transport proteins and many others; and it homogeneously covers targets related to varied diseases. PDTD

Structural databases of intermolecular interactions

Interactions between biomolecules are the cornerstone of cellular structure, signal transduction and biochemistry. Moreover, small molecules can strongly regulate protein function, interaction and localization through binding, as exploited in drug design.

When interactions are strong enough, there is a fair chance of solving structures of the bound biomolecules by Cryo-EM, X-ray diffraction or NMR spectroscopy. When interactions are not too strong, NMR can still provide ensembles of possible structures with variable confidence. For weak binding, or for cases of strong binding where Cryo-EM, X-ray or NMR failed, only computational modeling techniques based on sparse data can be used to achieve a structural model of the complex [53] (pure molecular dynamics simulations in the usual 3D visualization (example in Figure 5, top part).

Continuing with nucleic acids from the previous section, NPIDB [55] focuses on interactions between nucleic acids and proteins. It can be downloaded entirely, browsed or searched by PDB entry, PFAM or SCOP classification of the protein partners, or by GO terms, among others. The work leading to the development of NPIDB yielded unprecedented classification of protein-nucleic acid interactions defined by the contacts established between different structural elements of the intervening proteins and nucleic acids, a classification that the server performs.

Protein–protein interactions were the scope of the original version of the DOMMINO database [56], whose latest version integrates structures of all complexes involving protein, RNA and DNA molecules. Searches can be done by PDB ID, SCOP family of the intervening proteins or interaction type, among others, and the output can be filtered by number of intermolecular contacts. Related, although not a structural database, the iPFAM database [57] of protein–protein and protein–ligand interactions was built through high-throughput analysis of interactions in the PDB. The most interesting aspect of iPFAM is that a user can query a

Figure 4. How PDB-derived databases for nucleic acids help highlight the structural complexity of RNA over that of DNA. For an example of protein-bound DNA (A) 100% of the base pairs are in standard Watson-Crick bonding with anti-parallel strand orientation (‘cWW’) according to NDB. This simplicity stems from the canonical B conformation adopted by this piece of DNA. But the exemplified tRNA molecule (B) has around 25% of its nucleotides in one of -in this case- 8 noncanonical pairing motifs (from NDB). This is owing to the complex 3D structure as observed in the online 3D visualization or in the simple circular representation at the RNA 3D hub. For a color version of this figure please visit the online article.
is integrated to the TarFisDock [62] Web server, with which a user can easily dock a small molecule to all PDTD entries.

There are also databases specialized on interactions involving specific kinds of proteins. For example, KLIFS [63] is specialized on kinases, currently containing around 250 different proteins and almost 2000 unique ligands taken from over 3000 PDB entries, including also sequence and structure alignments and annotations of the pockets targeted by the ligands, the conformations of key structural elements of kinases like the DFG loop and more. Another example is the Antigen–Antibody Interaction Database [64], which collects molecular interactions between antigens and antibodies at atomic/residue levels classified by interaction type, and whose output includes information about the antibody regions involved in binding and an online visualization tool.

Metal sites in proteins and nucleic acids

Around 30–40% of the proteomes consists of metal-binding proteins [65]. This includes enzymes whose activities require one or more metal ions, proteins that require metals to achieve their native structures and as more recently identified, proteins of the metal homeostasis systems, which control intracellular metal levels and that deliver metal ions to functional and structural sites. Now discontinued, Scripp's Metalloprotein Database and Browser was the first PDB-derived database of protein metal sites [66]. But other resourceful databases are available, widely used for point applications and also for mining and thus better understanding protein metal sites [67–70], a knowledge that in turn helps to better refine the metal sites of newly solved metalloprotein structures [68, 71, 72].
Metal Interactions in Protein Structures (MIPS [73]) allows to easily find, download and visualize all PDB entries that contain a given metal ion (also monoatomic anions), filtering them by the types of molecules interacting with the ion and by structure quality and degree of redundancy.

MetalDB [74] is a database specialized on metal sites, developed by one of the world leading groups in structural metallobiology. Its search is limited to PDB IDs only, but has more complete visualization capabilities and provides some information unavailable from MIPS, including automatic calculation of coordination numbers, coordination geometries and of protein and nonprotein metal ligands, plus CATH, SCOP and Pfam annotations. A related server from the same group, MetalIS(3) [75], allows to search metal sites structurally similar to the metal site of a given structure (from the PDB or user-uploaded) throughout the whole PDB.

Metal ions in Nucleic Acids [76] is designed to search for metal ions bound to nucleic acids. The search interface is complex, but this allows detailed queries specifying which bases to allow in the inner and outer coordination spheres, restraints on their distances to the metal ion, restraints on the relative positions of different bases bound to the metal ion and more. The output includes a list of atoms in the inner and outer coordination spheres, their distances to the metal ion, the bases they belong to and online visualizations.

Last, CheckMyMetal [77] is not strictly a database but rather a Web server of fast response such that it works as a database, powerful to validate and detect inconsistencies in metal sites of the PDB or in user-uploaded structures.

Databases of specific structural features of proteins

Posttranslational modifications are key players in the regulation of protein function, intracellular localization and turnover. Despite being minor compared with protein sizes, they provoke radical alterations in binding specificities and often entire loop refolding. While most databases and Web servers about posttranslational modifications are dedicated to their compilation and prediction, PTM-SD is probably the only one that focuses globally on the structures of protein amino acid modifications as retrieved from the PDB [78]. It crosses information from the PDB, UniProt, PTMCuration [79] (which curates modifications on-the-fly on Swiss-Prot updates) and dbPTM [80] (a complete repository of posttranslational modifications but lacking structural information), facilitating sequence-structure-function analyses.

Knots in the backbone traces of proteins are relatively poorly understood, but important especially regarding the field of protein folding [81]. The pKnot database/server allows browsing through PDB entries that contain knots in their backbone traces, searching sequences in the database of PDB entries with knots (including homology models when no perfect sequence match is found), and also searching for knots in user-uploaded structures. The main output includes online visualization of the knot in the context of the full protein, and classification of its knot(s) into one of so-far four types identified by the developers.

Tandem repeats in proteins are also relatively poorly understood; in particular, the degeneracy of the basic repeat units makes tandem repeats especially hard to find, define with precision and mine. RepeatsDB is a structural database of tandem repeats in proteins, built through automatic detection followed by manual curation by a group of experts in repeat proteins [82]. By building this database the authors could provide a draft structure-based classification of repeats including coiled coils, β-solenoids and β-propellers as some of the most represented examples, and proposed a large number of potential new repeats.

Protein loops are another important element under intense study, as they are often dynamic and are usually directly related to function. Moreover, they differ from secondary structures in that they do not adopt regular conformations, and therefore, they are not easy to model even though they are accurately predicted from sequences. The ArchDB [83] database provides a simple interface to perform complex PDB searches of loops that connect specific secondary structure elements (i.e. two helices, a helix and a β-strand, etc.) with specific geometries and numbers of spacing residues. The detailed output includes secondary structures, surface accessibility, geometric parameters describing the loop and online 3D visualization, and is helpful for loop engineering. The work that led to ArchDB further derived a novel structural classification of loops based on Ramachandran and sequence patterns. A more dynamical classification of loops based on the predicted timescales of their fluctuations can be obtained through a Web server developed from extensive MD simulations [84].

Other relevant databases of biological macromolecules

The Glycan Fragment Database (GFDB [85]) focuses on protein-bound oligosaccharides, a special kind of posttranslational modification especially ubiquitous in the extracellular domains of membrane proteins where it is often involved in molecular recognition. GFDB allows users to search for specific glycan sequences in a set precompiled from the PDB, to retrieve structural information. This database was the starting point for recent parameterization of carbohydrates in the CHARMM force field [86], allowing the treatment of glycoproteins in molecular dynamics simulations.

In recent years, the large number of PDB entries for some proteins has allowed protein-specific databases to appear, like the ones for kinases or antibody-antigen pairs reviewed above. Another example is UbSRD [87], which compiles structures of ubiquitin, ubiquitin-like folds and their interaction partners, browsable in different ways. Similar in spirit but broader in terms of molecules, SATPdb [88] focuses on peptides of therapeutic interest (antimicrobial, hemolytic, anticancer, etc.) with structures available in the PDB, extended with high-level molecular modeling when structures are unavailable but can be reliably estimated by prediction methods. SATPdb can be browsed by different criteria such as therapeutic function, secondary structures and other properties. Interesting statistics emerged from building and updates of SATPdb; for example, showing that most peptides in the database have more than one therapeutic activity.

Remarks on other relevant Web services

Finally, a few comments on Web services that do not contain structures of biological macromolecules but are extremely relevant to computational and experimental research in biochemistry and structural biology. One is EMDDataBank [89], the main repository for primary electron microscopy data, important as cryo-EM structures rapidly populate the PDB providing unprecedented structural data for large macromolecular assemblies. The EMDB and PDB sites are indeed highly interconnected, just like the BMRB [34] for NMR structures and the Electron Density Database [90] for X-ray structures. Another novel server that is...
not a database but is related to database building and could revolutionize data mining studies is PatternQuery, a web scripting-based application to extract structure patterns from PDB files [91].

Last, whereas this Briefing has covered structural database resources for biological macromolecules, it is important to highlight the also extremely useful databases containing chemical and functional information, ontologies and structures of millions of small molecules. Some of these databases focus on molecules with biological activities, others focus on metabolites, and others simply attempt to cover the full chemical space or subsets of drug-like compounds. Some of the best known servers for small molecules are PubChem, ChEMBL, ChEBI, ZINC, the Human Metabolome Database and The Chemical Space Project among the most comprehensive academic options [92–97].

Conclusion

It is hoped that this Briefing would recapitulate most of the currently up-to-date (as of April 2016) databases derived from the PDB, which researchers should have in mind and profit from to better and more easily analyze their structures and mine PDB data. An important remark is that most of these databases were built by specialists in the relevant fields and molecular types. Also important, these databases provide immediate response and online visualization capabilities (although some depend on external plug-ins so their developers should consider replacing them by options like JSmol [98]) and are highly interconnected to each other, both aspects making them accessible and interactive.

Key Points

- The worldwide PDB is the main repository for structural data of biomolecules, but its complexity often obscures browsing, finding and mining its entries efficiently, accurately and without bias from for example redundancy or structure quality.
- Specialized databases built down from the PDB through mining methodologies, curation (sometimes even manual) and connections to other data repositories, facilitate browsing and finding specific kinds of molecules and molecular features as well as connecting structures to sequence, dynamics, interactions and function.
- The specialized databases described here are built by experts in the relevant techniques, molecular types and interactions. Thus, they often contain precomputed descriptors, for example, about molecular geometries, that would be cumbersome to calculate for nonexperts.
- Most described databases feature online displays of relevant data and online structure visualization facilities optimized to show the relevant molecules and interactions.

Acknowledgement

I acknowledge EMBO for a Long-Term Postdoctoral Fellowship.

References


