Protein Data Bank

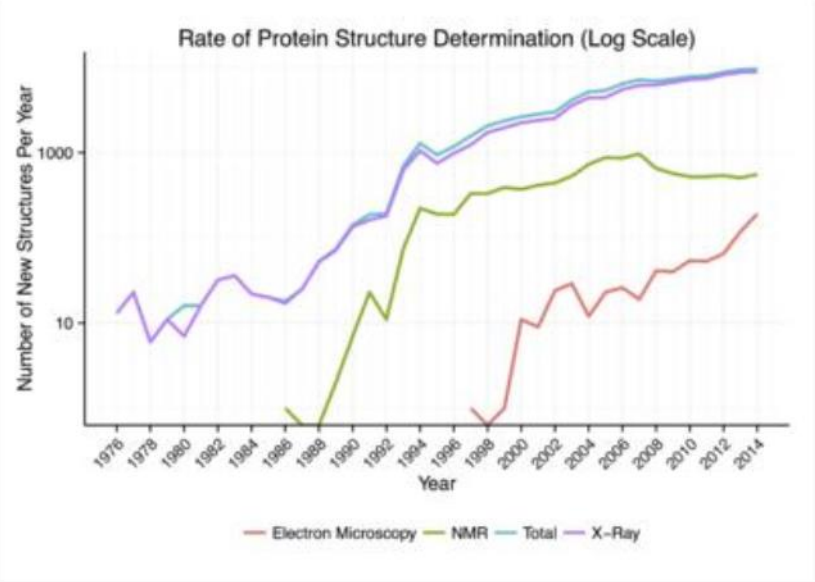
The **Protein Data Bank** (**PDB**) is a crystallographic database for the three-dimensional structural data of large biological molecules, such as proteins and nucleic acids. The data, typically obtained by X-ray crystallography,NMR spectroscopy, or, increasingly, cryo-electron microscopy, and submitted by biologists and biochemists from around the world, are freely accessible on the Internet via the websites of its member organisations (PDBe, PDBj, and RCSB). The PDB is overseen by an organization called the Worldwide Protein Data Bank, wwPDB.

The PDB is a key resource in areas of structural biology, such as structural genomics. Most major scientific journals, and some funding agencies, now require scientists to submit their structure data to the PDB. Many other databases use protein structures deposited in the PDB. For example, SCOP and CATH classify protein structures, while PDBsum provides a graphic overview of PDB entries using information from other sources, such as Gene ontology

Two forces converged to initiate the PDB: 1) a small but growing collection of sets of protein structure data determined by X-ray diffraction; and 2) the newly available (1968) molecular graphics display, the Brookhaven RAster Display (BRAD), to visualize these protein structures in 3-D. In 1969, with the sponsorship of Walter Hamilton at the Brookhaven National Laboratory, Edgar Meyer (Texas A&M University) began to write software to store atomic coordinate files in a common format to make them available for geometric and graphical evaluation. By 1971, one of Meyer's programs, SEARCH, enabled researchers to remotely access information from the database to study protein structures offline. SEARCH was instrumental in enabling networking, thus marking the functional beginning of the PDB.

Upon Hamilton's death in 1973, Tom Koeztle took over direction of the PDB for the subsequent 20 years. In January 1994, Joel Sussman of Israel's Weizmann Institute of Science was appointed head of the PDB. In October 1998, the PDB was transferred to the Research Collaboratory for Structural Bioinformatics (RCSB); the transfer was completed in June 1999. The new director was Helen M. Berman of Rutgers University (one of the member institutions of the RCSB). In 2003, with the formation of the wwPDB, the PDB became an international organization. The founding members are PDBe (Europe), RCSB (USA), and

PDBj (Japan). The BMRB joined in 2006. Each of the four members of wwPDB can act as deposition, data processing and distribution centers for PDB data. The data processing refers to the fact that wwPDB staff review and annotate each submitted entry. The data are then automatically checked for plausibility (the source code for this validation software has been made available to the public at no charge)



Rate of Protein Structure Determination by Method and Year

The PDB database is updated weekly (UTC+0 Wednesday). Likewise, the PDB holdings list is also updated weekly. As of 27 December 2015, the breakdown of current holdings is as follows:

Experimental Method	Proteins	Nucleic Acids	Protein/Nucleic Acid complexes	Other	Total
X-ray diffraction	95636	1694	4817	4	102151
NMR	9840	1135	231	8	11214
Electron microscopy	666	29	227	0	922
Hybrid	83	3	2	1	89
Other	170	4	6	13	193
Total:	106293	2865	5283	26	**114569**

91,748 structures in the PDB have a structure factor file.

8,531 structures have an NMR restraint file.

2,289 structures in the PDB have a chemical shifts file.

901 structures in the PDB have a 3DEM map file deposited in EM Data Bank

These data show that most structures are determined by X-ray diffraction, but about 10% of structures are now determined by protein NMR. When using X-ray diffraction, approximations of the coordinates of the atoms of the protein are obtained, whereas estimations of the distances between pairs of atoms of the protein are found through NMR experiments. Therefore, the final conformation of the protein is obtained, in the latter case, by solving a distance geometryproblem. A few proteins are determined by cryo-electron microscopy. (Clicking on the numbers in the original table will bring up examples of structures determined by that method.)

The significance of the structure factor files, mentioned above, is that, for PDB structures determined by X-ray diffraction that have a structure file, the electron density map may be viewed. The data of such structures is stored on the "electron density server".

In the past, the number of structures in the PDB has grown at an approximately exponential rate, passing the 100 registered structures milestone in 1982, the 1,000 in 1993, the 10,000 in 1999 and the 100,000 in 2014 However, since 2007, the rate of accumulation of new protein structures appears to have plateaued.

Viewing the data

The structure files may be viewed using one of several free and open source computer programs, including Jmol, Pymol, and Rasmol. Other non-free, sharewareprograms include ICM-Browser,[20] VMD, MDL Chime, UCSF Chimera, Swiss-PDB Viewer, StarBiochem (a Java-based interactive molecular viewer with integrated search of protein databank), Sirius, and VisProt3DS (a tool for Protein Visualization in 3D stereoscopic view in anaglyth and other modes), and Discovery Studio. The RCSB PDB website contains an extensive list of both free and commercial molecule visualization programs and web browser plugins.

PDBsum

PDBsum is database that provides an overview of the contents of each 3D macromolecular structure deposited in the Protein Data Bank. The original version of the database was developed around 1995 by Roman Laskowski and collaborators at University College London. As of 2014, PDBsum is maintained by Laskowski and collaborators in the laboratory of Janet Thornton at the European Bioinformatics Institute (EBI).

Each structure in the PDBsum database includes an image of structure (main view, Bottom view and right view), molecular components contained in the complex(structure), enzyme reaction diagram if appropriate, Gene Ontology functional assignments, a 1D sequence annotated by Pfam and InterPro domain assignments, description of bound molecules and graphic showing interactions between protein and secondary structure, schematic diagrams of protein-protein interactions, analysis of clefts contained within the structure and links to external databases. The RasMol and Jmol molecular graphics software are used to provide a 3D view of molecules and their interactions within PDBsum.

Since the release of the 1000 Genomes Project in October 2012, all single amino acid variants identified by the project have been mapped to the corresponding protein sequences in the Protein Data Bank. These variants are also displayed within PDBsum, cross-referenced to the relevant UniProt identifier. PDBsum contains a number of protein structures which may be of interest in structure-based drug design. One branch of PDBsum, known as DrugPort, focuses on these models and is linked with the DrugBank drug target database

SMILES

The simplified molecular-input line-entry system (SMILES) is a specification in form of a line notation for describing the structure of chemical species using short ASCII strings. SMILES strings can be imported by most molecule editors for conversion back into two-dimensional drawings or three-dimensional models of the molecules.

The original SMILES specification was initiated by David Weininger at the USEPA Mid-Continent Ecology Division Laboratory inDuluth in the 1980s. Acknowledged for their parts in the early development were "Gilman Veith and Rose Russo (USEPA) and Albert Leo and Corwin Hansch (Pomona College) for supporting the work, and Arthur Weininger (Pomona;

Daylight CIS) and Jeremy Scofield (Cedar River Software, Renton, WA) for assistance in programming the system." The Environmental Protection Agency funded the initial project to develop SMILES.

It has since been modified and extended by others, most notably by Daylight Chemical Information Systems. In 2007, anopen standard called "OpenSMILES" was developed by the Blue Obelisk open-source chemistry community. Other 'linear' notations include the Wiswesser Line Notation (WLN), ROSDAL and SLN (Tripos Inc).

In July 2006, the IUPAC introduced the InChI as a standard for formula representation. SMILES is generally considered to have the advantage of being slightly more human-readable than InChI; it also has a wide base of software support with extensive theoretical (e.g., graph theory) backing

Terminology

The term SMILES refers to a line notation for encoding molecular structures and specific instances should strictly be called SMILES strings. However, the term SMILES is also commonly used to refer to both a single SMILES string and a number of SMILES strings; the exact meaning is usually apparent from the context. The terms "canonical" and "isomeric" can lead to some confusion when applied to SMILES. The terms describe different attributes of SMILES strings and are not mutually exclusive.

Typically, a number of equally valid SMILES strings can be written for a molecule. For example, CCO, OCC and C(O)C all specify the structure of ethanol. Algorithms have been developed to generate the same SMILES string for a given molecule; of the many possible strings, these algorithms choose only one of them. This SMILES is unique for each structure, although dependent on the canonicalization algorithm used to generate it, and is termed the canonical SMILES. These algorithms first convert the SMILES to an internal representation of the molecular structure; an algorithm then examines that structure and produces a unique SMILES string. Various algorithms for generating canonical SMILES have been developed and

include those by Daylight Chemical Information Systems, OpenEye Scientific Software, MEDIT, Chemical Computing Group, MolSoft LLC, and the Chemistry Development Kit. A common application of canonical SMILES is indexing and ensuring uniqueness of molecules in a database.

The original paper that described the CANGEN[2] algorithm claimed to generate unique SMILES strings for graphs representing molecules, but the algorithm fails for a number of simple cases (e.g. cuneane, 1,2-dicyclopropylethane) and cannot be considered a correct method for representing a graph canonically. There is currently no systematic comparison across commercial software to test if such flaws exist in those packages.

SMILES notation allows the specification of configuration at tetrahedral centers, and double bond geometry. These are structural features that cannot be specified by connectivity alone and SMILES which encode this information are termed isomeric SMILES. A notable feature of these rules is that they allow rigorous partial specification of chirality. The term isomeric SMILES is also applied to SMILES in which isotopes are specified.

Graph-based definition

In terms of a graph-based computational procedure, SMILES is a string obtained by printing the symbol nodes encountered in a depth-first tree traversal of a chemical graph. The chemical graph is first trimmed to remove hydrogen atoms and cycles are broken to turn it into a spanning tree. Where cycles have been broken, numeric suffix labels are included to indicate the connected nodes. Parentheses are used to indicate points of branching on the tree.

Examples

Atoms

Atoms are represented by the standard abbreviation of the chemical elements, in square brackets, such as [Au] for gold. Brackets can be omitted for the "organic subset" of B, C, N, O, P, S, F, Cl, Br, and I. All other elements must be enclosed in brackets. If the brackets are omitted, the proper number of implicit hydrogen atoms is assumed; for instance the SMILES for water is simply O.

An atom holding one or more electrical charges is enclosed in brackets, followed by the symbol H if it is bonded to one or more atoms of hydrogen, followed by the number of hydrogen atoms (as usual one is omitted example: NH4 for ammonium), then by the sign '+' for a positive charge or by '-' for a negative charge. The number of charges is specified after the sign (except if there is one only); however, it is also possible write the sign as many times as the ion has charges: instead of "Ti+4", one can also write "Ti++++" (Titanium IV, Ti4+). Thus, the hydroxide anion is represented by [OH-], the oxonium cation is [OH3+] and the cobalt III cation (Co3+) is either [Co+3] or [Co+++].

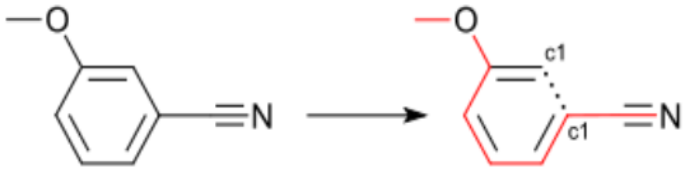
Bonds

Bonds between aliphatic atoms are assumed to be single unless specified otherwise and are implied by adjacency in the SMILES string. For example, the SMILES for ethanolcan be written as CCO. Ring closure labels are used to indicate connectivity between non-adjacent atoms in the SMILES string, which for cyclohexane and dioxane can be written as C1CCCCC1 and O1CCOCC1 respectively. For a second ring, the label will be 2 (naphthalene: c1cccc2c1cccc2 (note the lower case for aromatic compounds)), and so on. After reaching 9, the label must be preceded by a '%', in order to differentiate it from two different labels bonded to the same atom (~C12~ will mean the atom of carbon holds the ring closure labels 1 and 2, whereas ~C%12~ will indicate one label only, 12). Double, triple, and quadruple bonds are represented by the symbols '=', '#', and '$' respectively as illustrated by the SMILES O=C=O (carbon dioxide), C#N (hydrogen cyanide) and [Ga-]$[As+] (gallium arsenide).

Aromaticity

Aromatic C, O, S and N atoms are shown in their lower case 'c', 'o', 's' and 'n' respectively. Benzene, pyridine and furan can be represented respectively by the SMILES c1ccccc1, n1ccccc1 and o1cccc1. Bonds between aromatic atoms are, by default, aromatic although these can be specified explicitly using the ':' symbol. Aromatic atoms can be singly bonded to each other and biphenyl can be represented by c1ccccc1-c2ccccc2. Aromatic nitrogen bonded to hydrogen, as found in pyrrole must be represented as [nH] and imidazole is written in SMILES notation as n1c[nH]cc1.

The Daylight and OpenEye algorithms for generating canonical SMILES differ in their treatment of aromaticity.



Visualization of 3-cyanoanisole as COc(c1)cccc1C#N.

Branching

Branches are described with parentheses, as in CCC(=O)O for propionic acid and C(F)(F)F for fluoroform. Substituted rings can be written with the branching point in the ring as illustrated by the SMILES COc(c1)cccc1C#N (see depiction) and COc(cc1)ccc1C#N (see depiction) which encode the 3 and 4-cyanoanisole isomers. Writing SMILES for substituted rings in this way can make them more human-readable.

Stereochemistry

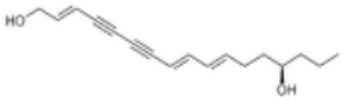
Configuration around double bonds is specified using the characters "/" and "\". For example, F/C=C/F (see depiction) is one representation of trans-difluoroethene, in which the fluorine atoms are on opposite sides of the double bond, whereas F/C=C\F (see depiction) is one possible representation of cis-difluoroethene, in which the Fs are on the same side of the double bond, as shown in the figure.

Configuration at tetrahedral carbon is specified by @ or @@. L-Alanine, the more common enantiomer of the amino acid alanine can be written as N[C@@H](C)C(=O)O (see depiction). The @@ specifier indicates that, when viewed from nitrogen along the bond to the chiral center, the sequence of substituents hydrogen (H), methyl (C) and carboxylate (C(=O)O) appear clockwise. D-Alanine can be written as N[C@H](C)C(=O)O (see depiction). The order of the substituents in the SMILES string is very important and D-alanine can also be encoded as N[C@@H](C(=O)O)C (see depiction).

Isotopes

Isotopes are specified with a number equal to the integer isotopic mass preceding the atomic symbol. Benzene in which one atom is carbon-14 is written as [14c]1ccccc1 and deuterochloroform is [2H]C(Cl)(Cl)Cl.

Examples

Molecule	Structure	SMILES Formula
Dinitrogen	N≡N	N#N
Methyl isocyanate (MIC)	CH_3–N=C=O	CN=C=O
Copper(II) sulfate	Cu^{2+} SO_4^{2-}	[Cu+2].[O-]S(=O)(=O)[O-]
Oenanthotoxin ($C_{17}H_{22}O_2$)		CCC[C@@H](O)CC\C=C\C=C\C#CC#C\C=C\CO

SMILES can be converted back to 2-dimensional representations using Structure Diagram Generation algorithms (Helson, 1999). This conversion is not always unambiguous. Conversion to 3-dimensional representation is achieved by energy minimization approaches. There are many downloadable and web-based conversion utilities.