

Faculty of Bio and Chemical Engineering

Department of Biotechnology

Subject Code: SBI1310

Subject Name: Molecular modeling and Drug designing

UNIT V

Chemical database

A **chemical database** is a database specifically designed to store chemical information. This information is about chemical and crystal structures, spectra, reactions and syntheses, and thermophysical data.

Types of chemical databases

Chemical structures

Chemical structures are traditionally represented using lines indicating chemical bonds between atoms and drawn on paper (2D structural formulae). While these are ideal visual representations for the chemist, they are unsuitable for computational use and especially for search and storage. Small molecules (also called ligands in drug design applications), are usually represented using lists of atoms and their connections. Large molecules such as proteins are however more compactly represented using the sequences of their amino acid building blocks. Large chemical databases for structures are expected to handle the storage and searching of information on millions of molecules taking terabytes of physical memory...

Literature database

Chemical literature databases correlate structures or other chemical information to relevant references such as academic papers or patents. This type of database includes STN, Scifinder, and Reaxys. Links to literature are also included in many databases that focus on chemical characterization.

Crystallographic database

Crystallographic databases store X-ray crystal structure data. Common examples include Protein Data Bank and Cambridge Structural Database.

NMR spectra database

NMR spectra databases correlate chemical structure with NMR data. These databases often include other characterization data such as FTIR and mass spectrometry.

Reactions database

Most chemical databases store information on stable molecules but in databases for reactions also intermediates and temporarily created unstable molecules are stored. Reaction databases contain information about products, educts, and reaction mechanisms.

Thermophysical database

Thermophysical data are information about

- phase equilibria including vapor–liquid equilibrium, solubility of gases in liquids, liquids in solids (SLE), heats of mixing, vaporization, and fusion.
- caloric data like heat capacity, heat of formation and combustion,
- transport properties like viscosity and thermal conductivity

Chemical structure representation

There are two principal techniques for representing chemical structures in digital databases

- As connection tables / adjacency matrices / lists with additional information on bond (edges) and atom attributes (nodes), such as:

MDL Molfile, PDB, CML

- As a linear string notation based on depth first or breadth first traversal, such as:
SMILES/SMARTS, SLN, WLN, InChI

These approaches have been refined to allow representation of stereochemical differences and charges as well as special kinds of bonding such as those seen in organo-metallic compounds. The principal advantage of a computer representation is the possibility for increased storage and fast, flexible search.

Search

Substructure

Chemists can search databases using parts of structures, parts of their IUPAC names as well as based on constraints on properties. Chemical databases are particularly different from other general purpose databases in their support for sub-structure search. This kind of search is achieved by looking for subgraph isomorphism (sometimes also called a monomorphism) and is a widely studied application of Graph theory. The algorithms for searching are computationally intensive, often of $O(n^3)$ or $O(n^4)$ time complexity (where n is the number of atoms involved).

The intensive component of search is called atom-by-atom-searching (ABAS), in which a mapping of the search substructure atoms and bonds with the target molecule is sought. ABAS searching usually makes use of the Ullman algorithm^[1] or variations of it (*i.e.* **SMSD** ^[2]). Speedups are achieved by time amortization, that is, some of the time on search tasks are saved by using precomputed information. This pre-computation typically involves creation of bitstrings representing presence or absence of molecular fragments. By looking at the fragments present in a search structure it is possible to eliminate the need for ABAS comparison with target molecules that do not possess the fragments that are present in the search structure. This elimination is called screening (not to be confused with the screening procedures used in drug-discovery). The bit-strings used for these applications are also called structural-keys. The performance of such keys depends on the choice of the fragments used for constructing the keys and the probability of their presence in the database molecules. Another kind of key makes use of hash-codes based on fragments derived computationally. These are called 'fingerprints' although the term is sometimes used synonymously with structural-keys. The amount of memory needed to store these structural-keys and fingerprints can be reduced by 'folding', which is achieved by combining parts of the key using bitwise-operations and thereby reducing the overall length.^[3]

Conformation

Search by matching 3D conformation of molecules or by specifying spatial constraints is another feature that is particularly of use in drug design. Searches of this kind can be computationally very expensive. Many approximate methods have been proposed, for instance BCUTS, special function representations, moments of inertia, ray-tracing histograms, maximum distance histograms, shape multipoles to name a few.^{[4][5][6][7][8]}

Descriptors

All properties of molecules beyond their structure can be split up into either physico-chemical or pharmacological attributes also called descriptors. On top of that, there exist various artificial and more or less standardized naming systems for molecules that supply more or less ambiguous names and synonyms. The IUPAC name is usually a good choice for representing a molecule's structure in a both human-readable and unique string although it becomes unwieldy for larger molecules. Trivial names on the other hand abound with homonyms and synonyms and are therefore a bad choice as a defining database key. While physico-chemical descriptors like molecular weight, (partial) charge, solubility, etc. can mostly be computed directly based on the molecule's structure, pharmacological descriptors can be derived only indirectly using involved multivariate statistics or experimental (screening, bioassay) results. All of those descriptors can for reasons of computational effort be stored along with the molecule's representation and usually are.

Similarity

There is no single definition of molecular similarity, however the concept may be defined according to the application and is often described as an inverse of a measure of distance in descriptor space. Two molecules might be considered more similar for instance if their difference in molecular weights is lower than when compared with others. A variety of other measures could be combined to produce a multi-variate distance measure. Distance measures are often classified into Euclidean measures and non-Euclidean measures depending on whether the triangle inequality holds. Maximum Common Subgraph (MCS) based substructure search [2] (similarity or distance measure) is also very common. MCS is also used for screening drug like compounds by hitting molecules, which share common subgraph (substructure). [9]

Chemicals in the databases may be clustered into groups of 'similar' molecules based on similarities. Both hierarchical and non-hierarchical clustering approaches can be applied to chemical entities with multiple attributes. These attributes or molecular properties may either be determined empirically or computationally derived descriptors. One of the most popular clustering approaches is the Jarvis-Patrick algorithm. [10]

In pharmacologically oriented chemical repositories, similarity is usually defined in terms of the biological effects of compounds (ADME/tox) that can in turn be semiautomatically inferred from similar combinations of physico-chemical descriptors using QSAR methods.

Registration system

Databases systems for maintaining unique records on chemical compounds are termed as Registration systems. These are often used for chemical indexing, patent systems and industrial databases.

Registration systems usually enforce uniqueness of the chemical represented in the database through the use of unique representations. By applying rules of precedence for the generation of stringified notations, one can obtain unique/'canonical' string representations such as 'canonical SMILES'. Some registration systems such as the CAS system make use of algorithms to generate unique hash codes to achieve the same objective.

A key difference between a registration system and a simple chemical database is the ability to accurately represent that which is known, unknown, and partially known. For example, a chemical database might store a molecule with stereochemistry unspecified, whereas a chemical registry system requires the registrar to specify whether the stereo configuration is unknown, a specific (known) mixture, or racemic. Each of these would be considered a different record in a chemical registry system.

Registration systems also preprocess molecules to avoid considering trivial differences such as differences in halogen ions in chemicals. An example is the Chemical Abstracts Service (CAS) registration system. See also CAS registry number.

PDB-Ligand: a ligand database based on PDB for the automated and customized classification of ligand-binding structures

Jae-Min Shin* and Doo-Ho Cho

Research and Development, IDR Tech. B-3003 Tripolis, 210 KumGok-Dong, BunDang-Ku, SungNam-Shi, KyungKi-Do, Republic of Korea 463-805

Received June 30, 2004; Revised and Accepted October 4, 2004

ABSTRACT

PDB-Ligand (<http://www.idrtech.com/PDB-Ligand/>) is a three-dimensional structure database of small molecular ligands that are bound to larger biomolecules deposited in the Protein Data Bank (PDB). It is also a database tool that allows one to browse, classify, superimpose and visualize these structures. As of May 2004, there are about 4870 types of small molecular ligands, experimentally determined as a complex with protein or DNA in the PDB. The proteins that a given ligand binds are often homologous and present the same binding structure to the ligand. However, there are also many instances wherein a given ligand binds to two or more unrelated proteins, or to the same or homologous protein in different binding environments. PDB-Ligand serves as an interactive structural analysis and clustering tool for all the ligand-binding structures in the PDB. PDB-Ligand also provides an easier way to obtain a number of different structure alignments of many related ligand-binding structures based on a simple and flexible ligand clustering method. PDB-Ligand will be a good resource for both a better interpretation of ligand-binding structures and the development of better scoring functions to be used in many drug discovery applications.

INTRODUCTION

Understanding the interaction between protein and small molecular ligand is very important in post-genomics life science because many important proteins require small molecular ligands or cofactors such as ATP or NAD, in order to function properly. In addition, there is a huge need to design

small molecular inhibitors for new drug discovery, based on the analysis of protein–ligand interaction.

The first step for understanding protein–ligand interaction would be to analyze the known protein–ligand complex structures in the Protein Data Bank (PDB) (1) (<http://www.rcsb.org/>). When analyzing protein–ligand structures, it is often necessary to cluster related ligand-binding structures, according to the ligand conformation, the three-dimensional (3D) ligand-binding structures, and the relative position and orientation of any important residues at the ligand-binding sites.

There are already many protein cluster databases. Protein structure classification databases such as SCOP (2), FSSP (3) and CATH (4) are based on the clustering of the whole 3D structures of protein domains. Other databases such as Pfam (5), Swiss-Model (6) and CDD (7) are primarily based on sequence similarities. With the structural genomics initiatives, these databases have been greatly expanded in size and the structure and function of many experimentally undetermined proteins are now readily inferred using these databases. However, these databases are more focused on the protein structure and function rather than on the structures of ligand or ligand-binding sites. These ligand-binding structures are probably more important in many post-genomics applications such as small molecular inhibitor design for new drug discovery.

There are also many web-based databases of the ligand-binding structure of PDB, including PDBSum (8), Relibase (9) (<http://relibase.ebi.ac.uk/>), Hic-Up (<http://xray.bmc.uu.se/hicup/>) and PLD (10). Although these ligand databases provide very useful information on the ligand–protein binding structures, they cannot easily be used to compare or to classify the ligand-binding structures in 3D. Therefore, there is a need for a convenient tool to analyze and classify the ligand-binding structures based on the clustering of the relevant 3D-structures using all the PDB data.

PDB-Ligand is a 3D ligand-binding structure database, derived from the PDB. It is also a database tool that can be used to build such a database and for conveniently browsing through these databases. One novel feature of PDB-Ligand is

that it allows an interactive clustering of ligand-binding structures based on user-specific clustering criteria such as root-mean-square deviation (RMSD) using flexible combinations of the atoms at the ligand-binding sites.

DATABASE CONTENTS AND FEATURES

Figure 1 shows the scheme used in PDB-Ligand database construction. Currently, PDB-Ligand holds 4870 different types of ligands, extracted from 116,019 ligand-binding structures derived from about 25 000 PDB entries. In PDB-Ligand, a ligand-binding structure is defined by the ligands and all the residues and other atoms that are within 6.5 Å around the ligand. Thus, every ligand-binding structure in PDB-Ligand database is surrounded by the residues of the protein, DNA, RNA, solvent or even other ligands. PDB-Ligand uses Chime Plug-in (<http://www.mdli.com>) as a web-based molecular graphics interface for visualization. It also provides a URL-link to the original PDB file for each ligand-binding structure so that one can easily view the whole ligand-protein structure with other related ligand-binding structures.

One of the most useful features of PDB-Ligand is the interactive clustering of ligand-binding structures, based on the RMSD between different ligand-binding structures. When analyzing the ligand-binding structures for many biologically important ligands such as ATP or FAD, one wants to know how many are in a similar binding environment, and how similar they are in 3D conformation. PDB-Ligand database and its clustering tool allow fast structural classification of the similar ligand-binding structures from all the ligand-binding structures in the PDB. The structure-based clustering feature of PDB-Ligand may be more effectively used with other ligand-binding analysis tools such as LIGPLOT (11) and LPC (12), or with other ligand databases such as Relibase (9) and Ligand-Depot (<http://ligand-depot-i.rutgers.edu/>).

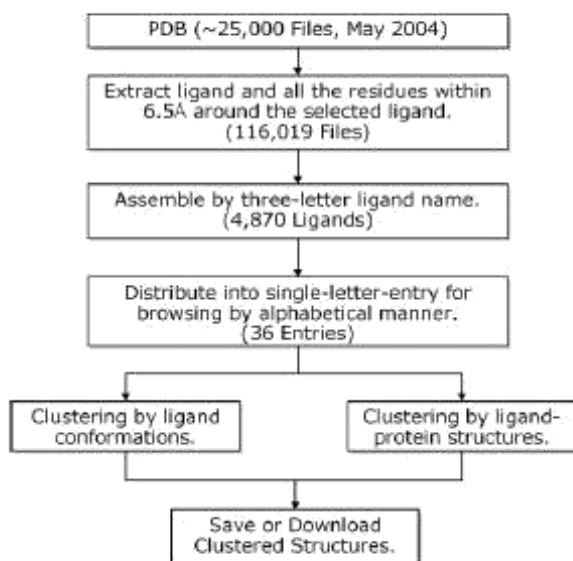


Figure 1. Scheme used in the PDB-Ligand database construction.

In addition, PDB-Ligand allows more flexible clustering based on both the ligand and the protein residues at the ligand-binding sites. This feature is useful, for example, when analyzing the same ligand-binding structures of a structurally related protein family.

CLUSTERING AND STRUCTURE ALIGNMENT

Since PDB-Ligand aims to be a 3D ligand-binding structure database with an interactive clustering feature, it only uses the ligand and the residues within 6.5 Å around the ligand in RMSD-based clustering by structure-structure alignment. Thus, using only the selected number of atoms at the ligand-binding site can greatly speed up the structure-alignment operation for RMSD calculations, while including all the important residues at the ligand-binding sites.

In PDB-Ligand, the clustering of ligand-binding structures is based on the RMSD value between all the corresponding atoms in the ligand after 3D structure superposition by Kabsch method (13). By default, all atoms of the ligand are considered in the superimposition for clustering. Therefore, in this case, every ligand in each cluster will have an overall structural similarity defined by the RMSD cut-off value (default is 0.5 Å).

However, if the ligand shows several different binding modes, it is more important to consider a part of ligand atoms and/or any critical residues at the ligand-binding site in the clustering process. In order to provide users with more convenient atom-selection, PDB-Ligand uses 'copy-and-paste' mechanism, based on chime script utilities (e.g. see E. Martz, <http://www.umass.edu/microbio/chime/>). For example, if a user selects main chain atoms of the residues at the ligand-binding site in the graphics window, these atoms are listed in the chime-log window, then they can be used in the clustering by 'copy-and-paste' into the selected atoms window. The user can copy any set of atoms shown in this window and paste them into the 'Selected Atoms' window. These atoms are then used to compute the superposition matrix. The simplicity and flexibility of the atom selection mechanism allow the users to perform a more precise clustering of ligand-binding structures. Currently, all the atoms in ligand and protein main chain atoms (N, CA, C and O), are allowed in the clustering.

As a clustering method, a simple greedy algorithm similar to that used by Hobohm and Sander (14) is used. In PDB-Ligand, a reference ligand-binding structure is always the one at the top of the list. Based on a given RMSD cut-off, all the structures similar to the reference structure are clustered together and removed from the list. The clustering is complete if no structure remains in the list.

AN EXAMPLE: ATP-BINDING STRUCTURES

In the current release of PDB-Ligand, there are 321 ATP-binding structures derived from 161 PDB entries. The ATP is the 46th most abundant ligand. If these 321 ATP-binding structures are clustered using 0.5 Å RMSD cut-off, we obtain 165 clusters (see Table 1). It means that there are 165 different conformations of ATP, each one of which is different from all

others, at least, by 0.5 Å in RMSD. If 1.0 Å RMSD is used, we obtain 91 different structural clusters for ATP.

Figure 2 shows a sample cluster of ATP-binding structures using 0.5 Å RMSD cut-off. One can easily see in this figure the common 3D structure of the amino acids surrounding the ligand. Interestingly, based on SCOP 1.65 protein family classification (2), the ATP-binding structures shown in Figure 2 are classified as Actin/Hsp70 protein family. This result may be useful for the users who want to investigate further the ATP-binding structures of such protein family.

FUTURE DIRECTIONS AND APPLICATIONS

The ligand-binding structures in PDB-Ligand will be updated, at least, every four months. In addition, the methods and algorithms for ligand-binding structure clustering will be improved for speed and convenience. Substructure search among ligand structures will also be included in the future. This feature will be useful in analyzing binding structures of various functional groups in many important ligands. Protein sequence and structure information based on the clustered ligand-binding structures will be also useful because it provides more complete information about ligand-binding structures. We also believe that the methods and the strategies used in PDB-Ligand, based on the clustering of ligand-binding structures, will be very useful in many applications for new drug discovery. For an example, based on the classification of similar ligand-binding structures, we have a plan to derive more accurate scoring functions for ligand-docking, virtual screening and lead-optimization for specific target proteins.

AVAILABILITY

PDB-Ligand is freely accessible through the URL at <http://www.idrtech.com/PDB-Ligand/>.

ACKNOWLEDGEMENTS

We thank B. K. Lee for useful discussions and for valuable suggestions on the manuscript. We also thank H. C. Shin, S. M. Kim, C. K. Han, J. H. Yoon, Y. H. In and N. D. Kim

for useful discussions and comments. We also thank M. R. Roh and Y. W. Kim for maintaining the website. This study was supported by a grant of Korean Health 21 R&D Project, Ministry of Health and Welfare, Republic of Korea (Grant ID: 03-PJ2-PG4-BD02-0001).

REFERENCES

1. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
2. Andreeva, A., Howorth, D., Bremer, S.E., Hubbard, T.J.P., Chothia, C. and Murzin, A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226–D229.
3. Holm, L. and Sander, C. (1996) Mapping the protein universe. *Science*, **273**, 595–602.
4. Pearl, F.M.G., Lee, D., Bray, J.E., Sillitoe, I., Todd, A.E., Harrison, A.P., Thornton, J.M. and Orengo, C.A. (2000) Assigning genomic sequences to CATH. *Nucleic Acids Res.*, **28**, 277–282.
5. Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
6. Kopp, J. and Schwede, T. (2004) The SWISS-MODEL Repository of annotated three-dimensional protein structure homology models. *Nucleic Acids Res.*, **32**, D230–D234.
7. Marchler-Bauer, A., Anderson, J.B., DeWeese-Scott, C., Fedorova, N.D., Geer, L.Y., He, S., Hurwitz, D.L., Jackson, J.D., Jacobs, A.R., Lanczycki, C.J. *et al.* (2003) CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res.*, **31**, 383–387.
8. Laskowski, R.A. (2001) PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Res.*, **29**, 221–222.
9. Hendlich, M., Bergner, A., Gunther, J. and Klebe, G. (2003) Relibase—design and development of a database for comprehensive analysis of protein–ligand interactions. *J. Mol. Biol.*, **326**, 607–620.
10. Puvanendrapillai, D. and Mitchell, J.B.O. (2003) Protein Ligand Database (PLD): additional understanding of the nature and specificity of protein–ligand complexes. *Bioinformatics*, **19**, 1856–1857.
11. Wallace, A.C., Laskowski, R.A. and Thornton, J.M. (1995) LIGPLOT: a program to generate schematic diagrams of protein–ligand interactions. *Protein Eng.*, **8**, 127–134.
12. Sobolev, V., Sorokine, A., Prilusky, J., Abola, E.E. and Edelman, M. (1999) Automated analysis of interatomic contacts in proteins. *Bioinformatics*, **15**, 327–332.
13. Kabsch, W. (1978) A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr.*, **A34**, 827–828.
14. Hobohm, J. and Sander, C. (1994) Enlarged representative set of protein structures. *Protein Sci.*, **3**, 522–524.

PubChem

PubChem is a database of chemical molecules and their activities against biological assays. The system is maintained by the National Center for Biotechnology Information (NCBI), a component of the National Library of Medicine, which is part of the United States National Institutes of Health (NIH). PubChem can be accessed for free through a web user interface. Millions of compound structures and descriptive datasets can be freely downloaded via FTP. PubChem contains substance descriptions and small molecules with fewer than 1000 atoms and 1000 bonds. More than 80 database vendors contribute to the growing PubChem database.

PubChem

PubChem is designed to provide information on biological activities of small molecules, generally those with molecular weight less than 500 daltons(2). PubChem's integration with NCBI's Entrez (3) information retrieval system provides sub/structure, similarity structure, bioactivity data as well as links to biological property information in PubMed and NCBI's Protein 3D Structure Resource.

PubChem Databases

PubChem is comprised of three linked databases --

PubChem Compound,

PubChem Substance and

PubChem Bioassay

PubChem Compound (unique structures with computed properties)

PubChem Compound (4) is a searchable database of chemical structures with validated chemical depiction information provided to describe substances in PubChem Substance. Structures stored within PubChem Compounds are pre-clustered and cross-referenced by identity and similarity groups. PubChem Compound includes over 5M compounds.

- Molecular Name Searches (e.g., Tylenol, Benzene) allow searching with a variety of chemical synonyms,

- Chemical Property Range Searches (e.g., Molecular Weight between 100 and 200, Hydrogen Bond Acceptor Count between 3 and 5) allow searching for compounds with a variety of physical/chemical properties, and descriptors.
- Simple Elemental Searches (all compounds containing Gallium) allow searching with specific element restrictions.

PubChem Substance (deposited structures)

PubChem Substance (5) is a searchable database containing descriptions of chemical samples, from a variety of sources, and links to PubMed citations, protein 3D structures, and biological screening results available in PubChem BioAssay. PubChem Substance includes over 8M records. Substances with known content are linked to PubChem Compound.

- Molecule Synonym Searches (e.g. all substances with 'deoxythymidine' as a name fragment, or substances that contain 3'-Azido-3'-deoxythymidine).
- Biology Links Search (e.g. substances with tested, active or inactive bioassays).
- Combined Searches (e.g. substances that are 'Active in any BioAssay' and contain the element Ruthenium).

PubChem BioAssay

PubChem BioAssay (6) is a searchable database containing bioactivity screens of chemical substances described in PubChem Substance. PubChem BioAssay includes over 180 bioassays. Searchable descriptions of each bioassay are provided that include descriptions of screening procedural conditions and readouts.

- To Search for BioAssay Data Sets (e.g. HIV growth inhibition).
- To Browse or Download PubChem BioAssay Results (NCI AIDS Antiviral Assay)

Searching PubChem

PubChem Text Search

PubChem Text Search for searching compound name, synonym or ID that defaults to PubChem Compound. The search results page offers a pull down 'databases' menu that allows searching in PubChem Substance, PubChem BioAssay and a variety of other Entrez databases.

PubChem Chemical Structure Search

PubChem Chemical Structure Search (7) has the following options: Search SMILES (including SMARTS or InChI) or Formula which includes a 'Sketch' link to a drawing program that converts structural diagrams to SMILES(exact), SMARTS(substructure) or InChI(exact) strings for searching.

Clicking 'Done' on the 'structure editor' converts the structural diagram to the appropriate string and transfers it to the search box.

Select Structure File allows importation of standard and common chemical file formats (8).

Specify Search Type allows restriction to: same compound, similar compounds (9), formula or substructure.

PubChem Indexes and Index Search

PubChem Indexes and Index Search allows fielded/range searching from either the PubChem homepage or Entrez search page. A extensive list of field aliases and examples of range searching is provided

PubChem Search Results

PubChem Compound

PubChem Compound results are derived from PubChem Substance records that provide structures. Since compounds are structurally unique, one compound may link to multiple substances. The default display is a compound summary with thumbnails with cross links(12) to each PubChem database, other NCBI databases, and depositor's databases.

Limits Preview/Index History Clipboard Details

Display Summary Show 20 Sort by Send to

1: CID: 2244

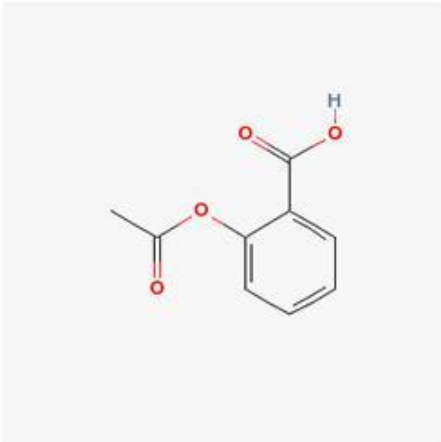
aspirin, Acetosalin ...
IUPAC: 2-acetyloxybenzoic aci
MW: 180.157 | MF: C9H8O4

Links

- ▶ PubChem BioAssay
- ▶ PubChem Inactive BioAssay
- ▶ Same, Connectivity
- ▶ PubChem Substance
- ▶ PubMed via MeSH
- ▶ Protein Structure
- ▶ Similar Compound

Clicking either the structure or SID link gives the full display which includes the compound's property data, description, related substance information, neighboring structures, and cross links.

Compound Summary:



CID: 2244 ⓘ
Substances: ⓘ
All: 51 Links
Same: 10 Links
Mixture: 41 Links

BioActivity: 66 Links ⓘ
Protein Structures: 2 Links ⓘ
Related Compounds: ⓘ
Same, Connectivity: 2 Links
Similar Compounds: 30 Links ⓘ
Structure Search ⓘ

MeSH | **Synonyms** | **Properties** | **Descriptors** | **Exports**

Medical Subject Annotations: (Total:6) ⓘ Display: Next 1 | All

Aspirin [Show MeSH Tree Structure](#)
The prototypical analgesic used in the treatment of mild to moderate pain. It has anti-inflammatory and antipyretic properties and acts as an inhibitor of cyclooxygenase which results in the inhibition of the biosynthesis of prostaglandins. Aspirin also inhibits platelet aggregation and is used in the prevention of arterial and venous thrombosis. (From Martindale, The Extra Pharmacopoeia, 30th ed, p5)

Pharmacological Action:
Anti-Inflammatory Agents, Non-Steroidal
Fibrinolytic Agents
Platelet Aggregation Inhibitors
Cyclooxygenase Inhibitors

PubMed via MeSH Choose by Subheadings:

administration and dosage	adverse effects	analogs and derivatives
analysis	antagonists and inhibitors	blood
contraindications and precautions	drug interactions	therapeutic uses

PubChem Substance

PubChem Substance has unique records if the structure is not known or supplied. For example, Sulfated polymannuroguronate, a novel anti-acquired immune deficiency syndrome (AIDS) drug candidate, and other natural products.

The PubChem Substance Summary Record,

SID: [3724242](#)

[Links](#)



Sulfated poly(2,6-dimethyl-1,4-dioxane-5,8-diol), AIDS218087 ...

Source: [NIAID\(218087\)](#)

is linked to the full record by clicking on the SID number (PubChem's substance identifier). This displays the full substance record, that includes links: to PubMed and the source; the Medical Subject Annotation (MESH Substance Name) and a MESH PubMed search link; and depositor supplied synonyms and comments.

PubChem BioAssay

The PubChem BioAssay Summary Record,

AID: [179](#)

[Links](#)

NCI
Source:

AIDS

Antiviral

Assay
DTP/NCI

15 Readouts, 37678 substances tested

is linked to the full record by clicking on the AID number (PubChem's assay (protocol) identifier). This displays the full bioassay record, that includes: links to the substances tested (all, active, inactive, inconclusive) and related PubMed, Protein, Taxonomy, OMIM and related BioAssay records; and a description of the assay possibly with protocols and comments.

Protein Data Bank

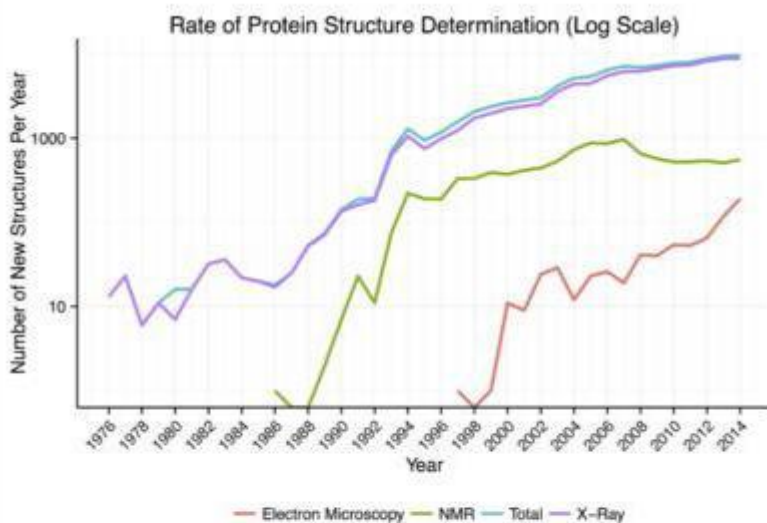
The **Protein Data Bank (PDB)** is a [crystallographic database](#) for the three-dimensional structural data of large biological molecules, such as [proteins](#) and [nucleic acids](#). The data, typically obtained by [X-ray crystallography](#), [NMR spectroscopy](#), or, increasingly, [cryo-electron 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,^[1] PDBj,^[2] and RCSB^[3]). 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.^[6] 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,^[7] the PDB was transferred to the Research Collaboratory for Structural Bioinformatics (RCSB);^[8] 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).^[9] In 2003, with the formation of the wwPDB, the PDB became an international organization. The founding members are PDBe (Europe),^[1] RCSB (USA), and

PDBj (Japan).^[2] The BMRB^[10] 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.^[11] The data are then automatically checked for plausibility (the source code^[12] 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^[13] 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
<i>Total:</i>	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 geometry](#) problem. 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".^{[14][15]}

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.^{[16][17]} 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, [shareware](#) programs include ICM-Browser,^[20] VMD, MDL Chime, UCSF Chimera, Swiss-PDB Viewer,^[21] StarBiochem^[22] (a Java-based interactive molecular viewer with integrated search of protein databank), Sirius, and VisProt3DS^[23] (a tool for Protein Visualization in 3D stereoscopic view in anaglyph 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.^{[1][2][3][4]} The original version of the database was developed around 1995 by Roman Laskowski and collaborators at University College London.^[5] 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.^[6] The RasMol and Jmol molecular graphics software are used to provide a 3D view of molecules and their interactions within PDBsum.^[5]

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.^[6] 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

PDBsum: summaries and analyses of PDB structures

Roman A. Laskowski*

Department of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK

Received August 31, 2000; Accepted October 4, 2000

ABSTRACT

PDBsum is a web-based database providing a largely pictorial summary of the key information on each macromolecular structure deposited at the Protein Data Bank (PDB). It includes images of the structure, annotated plots of each protein chain's secondary structure, detailed structural analyses generated by the PROMOTIF program, summary PROCHECK results and schematic diagrams of protein–ligand and protein–DNA interactions. RasMol scripts highlight key aspects of the structure, such as the protein's domains, PROSITE patterns and protein–ligand interactions, for interactive viewing in 3D. Numerous links take the user to related sites. PDBsum is updated whenever any new structures are released by the PDB and is freely accessible via <http://www.biochem.ucl.ac.uk/bsm/pdbsum>.

INTRODUCTION

To date, the 3D structures of over 13 000 biological macromolecules have been determined experimentally, principally by X-ray crystallography and NMR spectroscopy. The majority of these are protein structures, including protein–DNA and protein–ligand complexes. Together with sequence, physico-chemical and functional annotations they provide a wealth of information crucial for the understanding of biological processes.

Each new structure is deposited in the Protein Data Bank (PDB) (1), which is currently run by the Research Collaboratory in Structural Biology (RCSB) (2). The structures can be downloaded from the RCSB's PDB web server, which also provides additional information about each one. Further information, some of it focusing on specific types of molecules or specific aspects of the molecules, can be obtained from a large number of other structural databases (3) on the Web. One such database is PDBsum, which is the subject of this paper.

DESCRIPTION

The PDBsum database at <http://www.biochem.ucl.ac.uk/bsm/pdbsum> was created in 1995 (4). Its aim was to provide an at-a-glance summary of the molecules contained in each PDB entry (i.e. protein and DNA/RNA chains, small-molecule ligands, metal ions and waters), together with annotations and analyses of their key structural features. Thus, for each PDB

entry there is a corresponding summary web page in PDBsum, accessible by the four-character PDB identifier.

The original PDBsum paper (4) described the basic contents of each entry, namely a block of 'header' information, relating to the entry as a whole, followed by a list of the molecules making up the structure, together with any relevant structural analyses of each. The header details start with a thumbnail image of the molecule(s) in question plus buttons for viewing the whole structure in 3D using RasMol (5) or VRML (Virtual Reality Modelling Language). These are followed by information extracted directly from the header records of the PDB file, summary PROCHECK (6) analyses (including a Ramachandran plot) giving an indication of the stereochemical 'quality' of all the protein chains in the structure, and links to related databases. In the list of molecules that follows, each protein chain is shown schematically by a 'wiring diagram' depicting its secondary structural motifs, primary sequence, structural domains and highlighting active site residues and residues that interact with ligands, metals or DNA/RNA molecules. The secondary structural motifs are computed by the PROMOTIF (7) program, whose detailed outputs are available via hyperlinks, while the domain definitions come from the CATH protein structural classification database (8,9). For each ligand molecule a LIGPLOT (10) diagram gives a schematic depiction of the hydrogen bonds and non-bonded interactions between it and the residues of the protein with which it interacts.

In the time since the original paper was published, a number of new analyses, links and functions have been added, and these are described in the remainder of this paper.

NEW FEATURES

The first of the additions relates only to protein–DNA and DNA–ligand complexes. The interactions between the DNA chains and any other molecules in the complex are shown schematically in a diagram generated by the NUCPLOT (11) program. Like the LIGPLOT diagrams of protein–ligand interactions, the NUCPLOT diagrams show all the hydrogen bonds and non-bonded interactions between the molecules, as calculated by HBPLUS (12). The diagrams are output in PostScript format (see, for example, the PDBsum entry for PDB code 2OR1).

Next, each protein chain now has a direct link to the SAS (Sequence Annotated by Structure) (13) database. Clicking on the link initiates a FASTA search that scans the given chain's sequence of amino acid residues against a database of all sequences in the PDB. The net result is a list of all other chains in the PDB that are similar at the sequence level to the one of interest. The SAS database provides a variety of different

annotations of the resultant multiple-sequence alignment, as well as enabling the user to view the superposed structures in 3D in RasMol.

Also new is the identification of any PROSITE (14) patterns present in each protein chain. These are patterns of residues that are found in regions that are highly conserved across all members of a given protein family and consequently characterise both the family itself and the biologically significant sites in its member proteins. In PDBsum the matching residues are coloured according to their conservation (and hence importance): from red for highly conserved, to blue for highly variable. Not all matching PROSITE patterns are shown; only those that appear to be true positives are included (15). The residues matching the PROSITE pattern can be viewed in RasMol to see where they lie in relation to the rest of the protein structure. A RasMol script renders the residues as thick sticks, coloured as on the PDBsum page, while showing the rest of the protein as a white backbone trace and any nearby ligands in spacefill. This often gives a clear indication of the structural and functional significance of the PROSITE pattern residues. See, for example, the entry for IAAW, an aspartate aminotransferase, which contains the PROSITE pattern AA_TRANSFER_CLASS_1 corresponding to the Class 1 aminotransferases.

The RasMol scripts that display the PROSITE residues are generated on the fly by a program called RomLas (the name being a carefully chosen anagram of RasMol). The program is used throughout PDBsum to generate RasMol scripts for highlighting specific structural features. For example, below each LIGPLOT diagram there is a button for generating a RasMol script that displays the given ligand in the 3D context of the protein residues with which it interacts; the ligand is shown in thick sticks, while the protein residues are shown in wireframe and are labelled with the residue name and number.

Other new features include a simple text search facility on the home page and full listings of all the ligands and hetero groups found in the database. Links to a number of useful new databases have been added.

ACKNOWLEDGEMENTS

PDBsum is maintained at University College, London. The authors of the programs used in generating and running the PDBsum database include David Smith, Gail Hutchinson, Alex Michie, Andrew Martin, Ian McDonald, Andrew

Wallace, Nick Luscombe, Duncan Milburn and Atsushi Kasuya. I would like to thank Martin Jones and John Bouquiere for their contribution to the database's development and running. Thanks also to Frances Pearl, Malcolm MacArthur, Edith Chan and, most of all, Janet Thornton.

REFERENCES

1. Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F.,Jr, Brice,M.D., Rogers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.
2. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242. Updated article in this issue: *Nucleic Acids Res.* (2001), **29**, 214–218.
3. Berman,H.M. (1999) The past and future of structure databases. *Curr. Opin. Struct. Biol.*, **10**, 76–80.
4. Laskowski,R.A., Hutchinson,E.G., Michie,A.D., Wallace,A.C., Jones,M.L. and Thornton,J.M. (1997). PDBsum: a Web-based database of summaries and analyses of all PDB structures. *Trends Biochem. Sci.*, **22**, 488–490.
5. Sayle,R.A. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374–376.
6. Laskowski,R.A., MacArthur,M.W., Moss,D.S. and Thornton,J.M. (1993) PROCHECK - a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.*, **26**, 283–291.
7. Hutchinson,E.G. and Thornton,J.M. (1996) PROMOTIF – a program to identify and analyze structural motifs in proteins. *Protein Sci.*, **5**, 212–220.
8. Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) CATH: a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
9. Pearl,F.M.G., Lee,D., Bray,J.E., Sillitoe,I., Todd,A.E., Harrison,A.P., Thornton,J.M. and Orengo,C.A. (2000) Assigning genomic sequences to CATH. *Nucleic Acids Res.*, **28**, 277–282. Updated article in this issue: *Nucleic Acids Res.* (2001), **29**, 223–227.
10. Wallace,A.C., Laskowski,R.A. and Thornton,J.M. (1995) LIGPLOT: A program to generate schematic diagrams of protein–ligand interactions. *Protein Eng.*, **8**, 127–134.
11. Luscombe,N.M., Laskowski,R.A. and Thornton,J.M. (1997) NUCPLOT: a program to generate schematic diagrams of protein–nucleic acid interactions. *Nucleic Acids Res.*, **25**, 4940–4945.
12. McDonald,I.K. and Thornton,J.M. (1994) Satisfying hydrogen-bonding potential in proteins. *J. Mol. Biol.*, **238**, 777–793.
13. Milburn,D., Laskowski,R.A. and Thornton,J.M. (1998) Sequences annotated by structure: a tool to facilitate the use of structural information in sequence analysis. *Protein Eng.*, **11**, 855–859.
14. Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219.
15. Kasuya,A. and Thornton,J.M. (1999) Three-dimensional structure analysis of PROSITE patterns. *J. Mol. Biol.*, **286**, 1673–1691.

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 in Duluth in the 1980s.[1][2][3] 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." [4] The Environmental Protection Agency funded the initial project to develop SMILES.[5][6]

It has since been modified and extended by others, most notably by Daylight Chemical Information Systems. In 2007, an open 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.[7] 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++++].

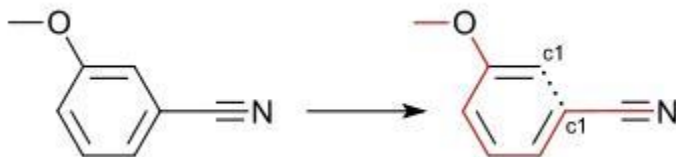
Bonds

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 ethanol can 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 o1ccccc1. 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 deuteriochloroform is [2H]C(Cl)(Cl)Cl.

Examples

Molecule	Structure	SMILES Formula
Dinitrogen	<chem>N#N</chem>	<chem>N#N</chem>
Methyl isocyanate (MIC)	<chem>CH3-N=C=O</chem>	<chem>CN=C=O</chem>
Copper(II) sulfate	<chem>Cu^{2+} SO_4^{2-}</chem>	<chem>[Cu+2].[O-]S(=O)(=O)[O-]</chem>
Oenanthotoxin (C ₁₇ H ₂₂ O ₂)		<chem>CCC[C@@H](O)CC\C=C\C=C\C#CC#C\C=C\CO</chem>

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.