Bioinformatics, 31(8), 2015, 1274–1278
doi: 10.1093/bioinformatics/btu789
Advance Access Publication Date: 2 December 2014
Original Paper

OXFORD

Databases and ontologies

The chemical component dictionary: complete descriptions of constituent molecules in experimentally determined 3D macromolecules in the Protein Data Bank

John D. Westbrook[1,*], Chenghua Shao[1], Zukang Feng[1], Marina Zhuravleva[1], Sameer Velankar[2] and Jasmine Young[1]

[1]RCSB PDB, Department of Chemistry and Chemical Biology and Center for Integrative Proteomics Research, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA and [2]Protein Data Bank in Europe, European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

*To whom correspondence should be addressed.

Associate Editor: Anna Tramontano

Received on August 7, 2014; revised on October 31, 2014; accepted on November 22, 2014

Abstract

Summary: The Chemical Component Dictionary (CCD) is a chemical reference data resource that describes all residue and small molecule components found in Protein Data Bank (PDB) entries. The CCD contains detailed chemical descriptions for standard and modified amino acids/nucleotides, small molecule ligands and solvent molecules. Each chemical definition includes descriptions of chemical properties such as stereochemical assignments, chemical descriptors, systematic chemical names and idealized coordinates. The content, preparation, validation and distribution of this CCD chemical reference dataset are described.

Availability and implementation: The CCD is updated regularly in conjunction with the scheduled weekly release of new PDB structure data. The CCD and amino acid variant reference datasets are hosted in the public PDB ftp repository at ftp://ftp.wwpdb.org/pub/pdb/data/monomers/components.cif.gz, ftp://ftp.wwpdb.org/pub/pdb/data/monomers/aa-variants-v1.cif.gz, and its mirror sites, and can be accessed from http://wwpdb.org.

Contact: jwest@rcsb.rutgers.edu.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The Protein Data Bank (PDB) is the single worldwide repository of information about the 3D structures of biological macromolecules (Berman *et al.*, 2003). PDB structures of macromolecules and macromolecular–ligand complexes provide direct experimental insights into protein function. These structures provide essential information for the understanding of biochemical processes and are critical data for structure-based drug design studies. PDB structure data are also central to the understanding the energetics of intermolecular interactions among proteins and nucleic acids, structure classification and structure prediction. The knowledge obtained through the study of 3D structure data is key to deducing the role of a protein or nucleic acid in human health and disease. To enable this rich interpretation of structural data requires accurate and consistent annotation of the chemical identities and nomenclatures for all of the constituent molecules in the PDB archive.

The individual monomers within a polymer, individual ligands and solvent molecules form the principal building blocks or chemical components within a PDB entry. These chemical components form the basis for the high-level macromolecular descriptions of

polymers and complexes in the PDB archive. Over the years, molecular descriptions that fully characterize all unique chemical components in the PDB structures are maintained and made available as part of the PDB archive. These molecular definitions are assembled in a reference data resource called the PDB chemical component dictionary (CCD). The current release of the CCD contains over 18 500 molecular definitions.

The CCD is maintained by the worldwide Protein Data Bank (wwPDB), which consists of organizations that act as deposition, data processing and distribution centers for PDB data. The wwPDB partners are the RCSB PDB (Berman *et al.*, 2000); PDB in Europe, PDBe (Gutmanas *et al.*, 2014); PDB in Japan, PDBj (Kinjo *et al.*, 2012) and BioMagResBank, BMRB (Markley *et al.*, 2008).

2 Methods

2.1 Data representation and organization

The detailed representation of the chemical components maintained in the PDB chemical dictionary has evolved over the 40-year history of the PDB archive. The early PDB chemical component representation is captured in the PDB format file as shown in Figure 1. The information in a structure entry in the original PDB file format includes an identifier code (1-3 alphanumeric characters), atom connection table, atom count, formula, molecular name and synonyms. The identifier codes for standard amino acids (e.g. ALA, GLY, TRP) and a small number of common ligands (e.g. adenosine triphosphate) are represented using the well-known three-character codes; however, for the majority of components, CCD identifier (ID) codes have no chemical significance. For each atom, a connection table provides the count and a list of adjacent atoms. This connection table also defines a unique name for each atom in the chemical component. Atom names in the original PDB file format contain up to four characters. Corresponding atoms for all instances of the chemical component in the archive share the same atom naming. These early chemical descriptors did not define stereochemistry or bond orders and lacked additional information such as formal charge, molecular weight or SMILES and InChI representation of the molecule.

Following a major nomenclature standardization effort in 2007 (Henrick *et al.*, 2008), the PDB introduced a dictionary of chemical definitions with much richer content. An example of a current molecular definition is illustrated in Figure 2(a-d).

The definition is encoded using the macromolecular crystallographic information framework (mmCIF) syntax (Fitzgerald *et al.*, 2005). This is also the primary archival syntax used to store PDB structure entries. This syntax provides for two simple organizational styles: key-value pairs (Fig. 2a) and tabular format (Fig. 2b-d). A typical chemical definition contains between 50 and 60 unique items of data.

The data names used in the chemical definition example are described separately in the PDB exchange data dictionary (PDBx; Westbrook *et al.*, 2005). This metadata resource defines all data items used in PDB data entries and reference data. Just as a conventional dictionary, the PDBx dictionary provides definitions and examples for each item of PDB data. In addition to these semantics, the PDBx dictionary defines properties such as data type, controlled vocabularies, boundary values and parent-child relationships. Collectively, these properties enable the software to verify the data integrity within individual chemical definitions and maintain data consistency between chemical reference data and PDB data entries. The PDBx dictionary is maintained as an electronic metadata resource using the syntax and technology of mmCIF at http://mmcif.wwpdb.org.

Data categories describing chemical components are defined in the CHEM_COMP_GROUP category group. High-level molecular descriptions of each component are presented in CHEM_COMP data category that includes data items describing the general features of the chemical component as a whole as shown in Figure 2a and Supplementary Table S1a.

This category provides identifying information including: the unique identifier (1-3 alphanumeric characters) assigned by the PDB, any one-letter-code monomer abbreviation, molecular name and synonyms. The molecular name is typically assigned a computed systematic name [ACD/Name (Advanced Chemistry Development, 2007)]. Well-known common names, trade names or other alternative names may be included as synonyms. The formal charge and molecular formula are also provided. The molecular formula is presented using the Hill system convention (Hill, 1900).

A component type distinguishes polymer, non-polymer and solvent components. For peptide and nucleic acid polymer components, the component type further describes polymer linking behavior. For instance, for most standard amino acids the linking type is 'L-peptide linking'.

For polymer components related to a standard amino acid or nucleotide by a simple chemical modification, the parent standard residue is identified. For components that have recognizable internal substructure, the list of constituent component identifiers is included. For instance, Lisinopril (CCD ID LPR) has the subcomponent sequence 'CLT LYS PRO' for the substituent components 4-phenyl-butanoic acid (CLT), lysine (LYS) and proline (PRO).

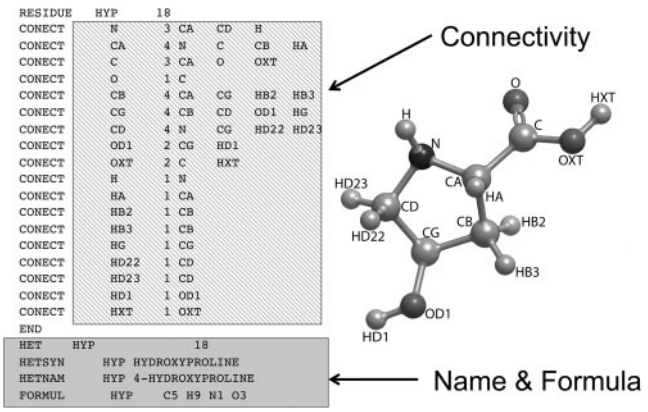

Fig. 1. An example of an early PDB molecular description for 4-hydroxyproline including atom nomenclature, connectivity, formula and molecular names

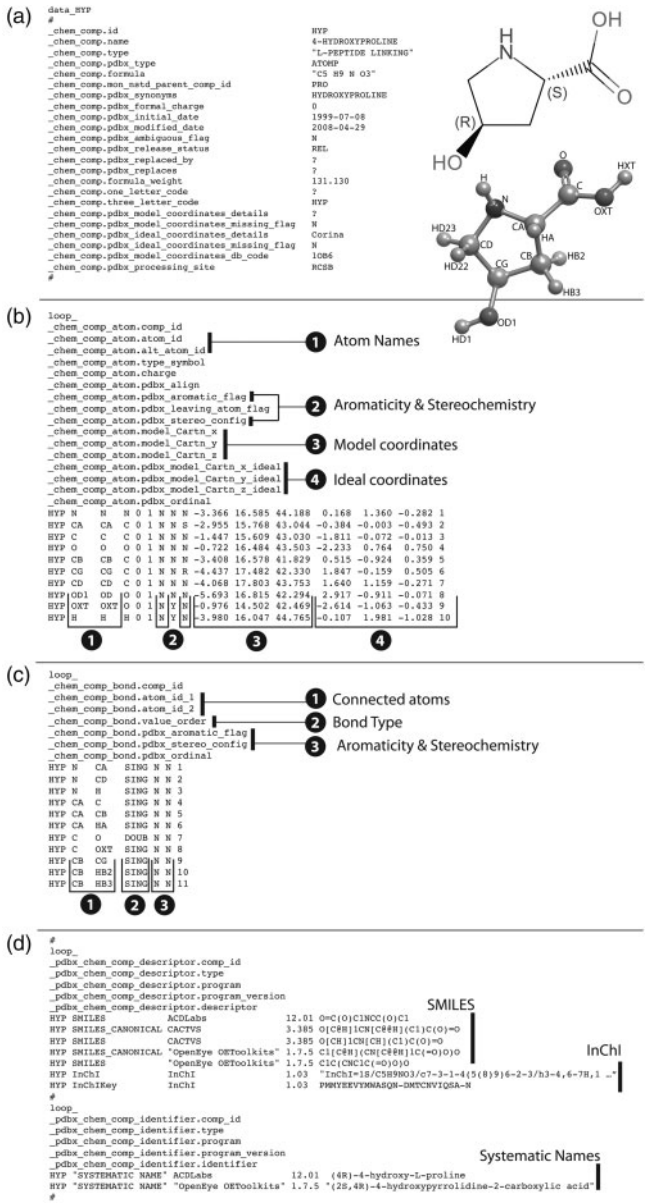

Fig. 2. (**a**) An example CCD definition for 4-hydroxyproline, including molecular-level information: accession details (e.g. identifiers and status information), molecular names, molecular type, formula and formula weight; (**b**) An abbreviated example CCD definition for 4-hydroxyproline illustrating the atom-level features including atom names, stereochemical and aromatic molecular identifiers, experimental model coordinates, computed ideal coordinates; (**c**) An abbreviated example CCD definition for 4-hydroxyproline illustrating the bond-level features including bonding atom names, bond type, stereochemical and aromatic molecular identifiers; (**d**) An example CCD definition for 4-hydroxyproline illustrating SMILES and InChI descriptors, systematic chemical name identifiers and the programs and versions used to compute each name and descriptor

Accession details are also given in this data category. In addition to the CCD ID assigned to the component, the PDB ID for the entry containing a representative example of the molecule may be provided. If the chemical component definition has been replaced, the CCD ID of the superseding chemical component is provided. Other special details concerning the chemical definition or the associated representative model atomic coordinates are also included in this definition category.

The CHEM_COMP_ATOM data category describes the atom-level features of the chemical component (Fig. 2b; Supplementary Table S1b). This includes the atom type symbol and atom nomenclature in current use in the PDB archive. Atoms names for standard amino acids and nucleotides follow IUPAC recommendations (IUPAC Commission on Macromolecular Nomenclature, 1979) with the exception of the well-established convention for C-terminal atoms OXT and HXT. In early PDB entries, an alternative atom nomenclature was used and this prior atom nomenclature was also included in definitions where the nomenclature has changed (Henrick *et al.*, 2008). For standard amino acids, additional molecular definitions have been created to specify common protonation variants. This provides a naming reference for experimental methods such as nuclear magnetic resonance (NMR), which typically report these additional hydrogen atoms.

A set of model coordinates from a single experimental entry and a computed set of ideal coordinates are provided. Hydrogen atoms

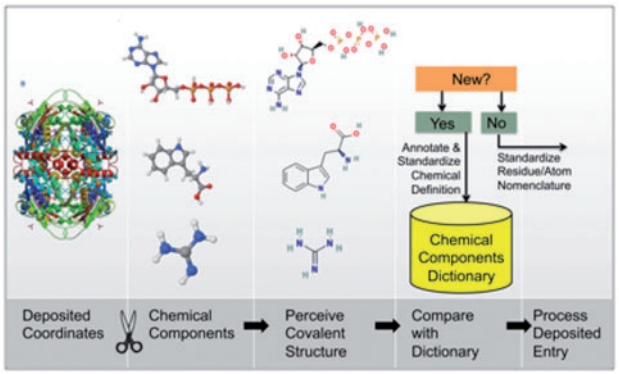

Fig. 3. The chemical component data preparation pipeline

are computationally added to the experimental coordinates if not present in the modeled structure. Computed coordinates are obtained from the programs CORINA (Sadowski and Gasteiger, 1994) or OpenEye/Omega (Bostrom *et al.*, 2003). Cahn–Ingold–Prelog (CIP) stereochemical assignments (Cahn *et al.*, 1966) and aromatic annotations are recorded for each atom. For chemical components with recognizable subcomponent architecture, the subcomponent CCD ID code is specified for each atom. Leaving atom positions are identified for components in which covalent linkages are observed in PDB entries. Leaving atoms identify atoms that are cleaved in the formation of covalent bonds with other chemical components. This data item identifies those atoms that are cleaved in covalent interactions observed in PDB entries.

Bonding information is provided in CHEM_COMP_BOND data category (Fig. 2c; Supplementary Table S1c). Each bond is specified as a pair of connected atoms where the atoms are identified by the atom names declared in the CHEM_COMP_ATOM data category. Bond types (i.e. single, double, triple), CIP stereochemical assignments and aromatic assignments are recorded for each bond.

Figure 2d and Supplementary Tables S1d and e illustrate additional details describing identifiers and descriptors related to the chemical definition. Descriptors and identifiers are accompanied by the provenance details such as software application name and version used to compute the quantity. Each chemical definition commonly includes SMILES, Standard InChI and InChI key descriptors (Heller *et al.*, 2013; Weininger, 1988). SMILES descriptors are calculated and reported using multiple programs (ACD/Name), CACTVS (Ihlenfeldt *et al.*, 1992), OpenEye OEChemTk (OpenEye Scientific Software Inc., 2007) owing to differences in the approach to canonicalization implemented by these programs. Systematic chemical names are also computed for each component (ACD/Name) and these and other synonyms are included as chemical identifiers.

Audit and revision history details (Supplementary Table 1f) report the details of any revisions to the existing definitions. The type of revision in each audit record is reported using a precise controlled vocabulary. This makes it possible to distinguish changes involving the addition of molecular synonyms or more substantive corrections to nomenclature or chemical structure.

2.2 Technical validation

A variety of software tools are used to maintain the CCD and to provide for consistent chemical representation across the PDB archive.

Referential integrity and data consistency within chemical component definitions are maintained using the constraints such as parent-child relationships, boundary values and controlled vocabularies defined in the underlying metadata definitions from the PDBx exchange dictionary. Chemical dictionary maintenance software tools provide checks for consistency of chemical formula and formula weight, atom name and type consistency, unusual chemical valences and non-bonded atoms.

Multiple software libraries are used to calculate key annotations. In cases where coordinate data are required to define the chemical structure of a component, the perception of covalent bond types from the 3D experimental coordinates is based on a consensus of results obtained from OpenBabel (O'Boyle *et al.*, 2011) and OEChemTK (Stahl and Mauser, 2005) libraries. Both the CACTVS and OpenEye OEChemTK libraries are used to compute CIP rule-based stereochemical and aromatic annotations from model coordinated data. When a consensus between these methods is not obtained, the results are subject to a manual review. Differences are typically due to unusual or strained geometries.

Chemical substructure search tools are used for matching and aligning covalent chemical structures in molecular definitions (Cordella *et al.*, 2004; Young *et al.*, 2013). Searching new chemical structures against the existing collection of definitions avoids duplication of molecular definitions.

Audit records are maintained to track changes in each definition. Revisions of individual chemical components definitions are archived in the CVS version control system.

2.3 Usage and maintenance of the CCD

The CCD is an important validation resource in the processing of new PDB entries as it serves as a reference for nomenclature, stereochemistry and structure.

As wwPDB members curate new experimental structure depositions, a principal task is to determine the accurate chemical descriptions of all of the molecular components within the entry (Fig. 3; Young *et al.*, 2013). A deposited macromolecular structure is decomposed into a set of constituent molecules. Within the experimental resolution of a typical macromolecular structure determination, there may be considerable latitude in precise details of any observed chemical moiety. Where possible, chemical components are chosen to correspond to realizable molecules such as neutral off-the shelf reagents. To the extent possible, the molecular identity is verified with the contributing scientist as part of the deposition and annotation of PDB entries. Hydrogen atom positions and charge details may be reported for NMR and some high-resolution X-ray and neutron diffraction experiments. In such cases, the protonation variant dictionary is used as a naming reference for protonated hydrogen atoms on standard amino acids. Separate chemical definitions are created for ligands with a reported charged state.

Atomic coordinate data or other chemical data provided with a deposited entry are used to establish the covalent structure of each chemical component. The covalent structures are compared with known chemical components within entries in the current archive. Molecules corresponding to known components are standardized for nomenclature while new components are added to the CCD containing all observed components.

The CCD is updated regularly in conjunction with the scheduled weekly release of new PDB structure data. The CCD and amino acid variant reference datasets are hosted in the public PDB ftp repository ftp://ftp.wwpdb.org/pub/pdb/data/monomers/components.cif.gz, ftp://ftp.wwpdb.org/pub/pdb/data/monomers/aa-variants-v1.cif.gz, its mirror sites, and accessed from http://wwpdb.org.

Acknowledgements

We thank Kim Henrick (formerly PDBe) and Dimitris Dimitropoulos (formerly PDBe and RCSB PDB) for their efforts in nomenclature standardization and software tool development that contributed significantly to the development of CCD. We thank Helen Berman (RCSB PDB) for her leadership, long-standing support for developing data standards and assistance in the preparation of this manuscript.

Funding

RCSB PDB is supported by NSF [DBI-1338415], NIH, DOE; PDBe by EMBL-EBI, Wellcome Trust [088944], BBSRC [BB/J007471/1, BB/K016970/1, BB/K020013/1, BB/M013146/1], NIGMS [1RO1 GM079429-01A1], EU [284209] and MRC [MR/L007835/1]; PDBj by JST-NBDC and BMRB by NLM P41 LM05799.

Conflict of Interest: none declared.

References

Advanced Chemistry Development,I. (2007) ACD/Name Batch, version 9.0.

Berman,H.M. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Berman,H.M. *et al.* (2003) Announcing the worldwide protein data bank. *Nat. Struct. Biol.*, **10**, 980.

Bostrom,J. *et al.* (2003) Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J. Mol. Graph Model*, **21**, 449–462.

Cahn,R.S. *et al.* (1966) Specification of molecular chirality. *Angew. Chem. Int. Edition*, **5**, 385–415.

Cordella,L.P.F. *et al.* (2004) A (sub)graph isomorphism algorithm for matching large graphs. *IEEE Trans. Pattern Anal. Mach. Intell.*, **26**, 1367–1372.

Fitzgerald,P.M.D. *et al.* (2005) 3.6 classification and use of macromolecular data. In: Hall,S.R. and McMahon,B. (eds.) *International Tables for Crystallography*. Springer: Dordrecht, The Netherlands, pp. 144–198.

Gutmanas,A. *et al.* (2014) PDBe: protein data bank in Europe. *Nucleic Acids Res.*, **42**, D285–D291.

Heller,S. *et al.* (2013) InChI: the worldwide chemical structure identifier standard. *J Cheminform.*, **5**, 7.

Henrick,K. *et al.* (2008) Remediation of the protein data bank archive. *Nucleic Acids Res.*, **36**, D426–D433.

Hill,E.A. (1900) On a system of indexing chemical literature; adopted by the classification division of the U.S. patent office. *J. Am. Chem. Soc.*, **22**, 478–494.

Ihlenfeldt,W.D. *et al.* (1992) CACTVS: a chemistry algorithm development environment. In: Machida,K. and Nishioka,T. (eds.) *Daijuukagakutouronkai Dainijuukai Kouzoukasseisoukan Shinpojiumu Kouenyoushishuu*. Kyoto University Press, Kyoto, pp. 102–105.

IUPAC Commission on Macromolecular Nomenclature. (1979) Stereochemical definitions and notations relating to polymers. *Pure Appl. Chem.*, **51**, 1101–1121.

Kinjo,A.R. *et al.* (2012) Protein data bank Japan (PDBj): maintaining a structural data archive and resource description framework format. *Nucleic Acids Res.*, **40**, D453–D460.

Markley,J.L. *et al.* (2008) BioMagResBank (BMRB) as a partner in the worldwide protein data bank (wwPDB): new policies affecting biomolecular NMR depositions. *J. Biomol. NMR*, **40**, 153–155.

O'Boyle,N.M. *et al.* (2011) Open babel: an open chemical toolbox. *J. Cheminform.*, **3**, 33.

OpenEye Scientific Software Inc. (2007) OpenEye OEChem, version 1.5.

Sadowski,J. and Gasteiger,J. (1994) Comparison of automatic three-dimensional model builders using 639 X-ray structures. *J. Chem. Inj. Comput. Sci.*, **34**, 1000–1008.

Stahl,M. and Mauser,H. (2005) Database clustering with a combination of fingerprint and maximum common substructure methods. *J. Chem. Inf. Model*, **45**, 449–462.

Weininger,D. (1988) SMILES 1: introduction and encoding rules. *J. Chem. Inf. Comput. Sci.*, **28**, 31.

Westbrook,J. *et al.* (2005) 3.6.2 the protein data bank exchange data dictionary. In: Hall,S.R. and McMahon,B. (eds.) *International Tables for Crystallography*. Springer, Dordrecht, The Netherlands, pp. 195–198.

Young,J.Y. *et al.* (2013) Chemical annotation of small and peptide-like molecules at the protein data bank. *Database*, **2013**, bat079.