INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

PDB$_{cle}$ – An online tool for extracting chain structure and sequence of macromolecule and small molecule structure from the Protein Data Bank

T Ashok Kumar[1]*

1 Department of Plant Biotechnology, Kerala Agricultural University, Vellayani, 695522, Kerala, Tel.: +91-9655307178

OPEN ACCESS

Received: 14.02.2021

Accepted: 02.05.2021

Published: 10.05.2021

Citation: Ashok Kumar T (2021) PDB$_{cle}$ – An online tool for extracting chain structure and sequence of macromolecule and small molecule structure from the Protein Data Bank. Indian Journal of Science and Technology 14(16): 1296-1301. https://doi.org/10.17485/IJST/v14i16.282

* **Corresponding author**.

Tel: +91-9655307178
ashok.bioinformatics@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Background: Protein Data Bank (PDB) is the most popular structure database that contains experimentally determined three-dimensional (3D) structures of biological macromolecules and small molecules. The rich features of PDB are keyword assisted advanced text search, structure search by sequence alignment, sequence motif search, ligand to target-ligand complex search through SMILES substructure search, JSON API query search, structure alignment, structure quality assessment, genome viewer, and 3D structure viewer. It is widely used in molecular modelling and computer-aided drug design. PDB$_{cle}$ is a simple tool to extract chain sequence of protein/nucleotide and 3D structure of protein/ nucleotide/ ligand from the PDB. **Objectives:** To construct an online tool for separating molecule-wise chain sequence and structure of polymers and non-polymer structures in a macromolecule. Moreover, the separated sequences and structures are produced to molecule-specific standard file format. **Methods**: The graphical web-interface of PDB$_{cle}$ tool has been designed using PHP, CSS, and pure JavaScript. Parsing the atomic coordinate records and sequence records from the PDBML/XML file and/or PDBx/mmJSON file through the API of PDB was done through PHP server script. **Findings**: The PDB$_{cle}$ tool retrieves and generates separate structure/sequence files for each amino acid/RNA chain, and pair of chains for DNA base pairs with/without ligand complex from the PDB. The ligand molecules are separated and sorted from the chains and produced to an SDF file. **Applications**: PDB$_{cle}$ tool is publicly accessible at https://www.biogem.org/tool/pdbcle/.

Keywords: PDBcle; PDB; PDB Chain and Ligand Extractor; PDB Sequence Extractor; PDBML

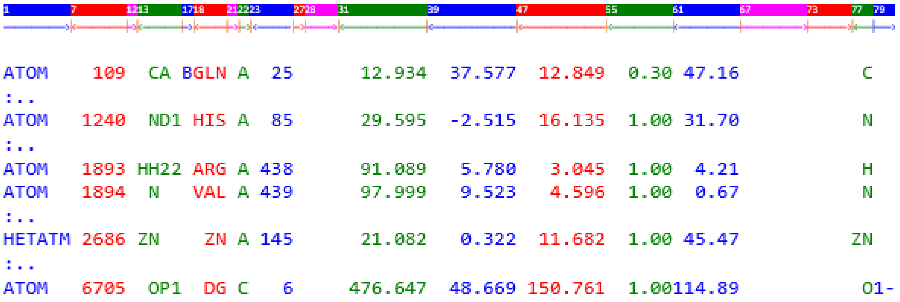
1 Introduction

Theoretical and computational molecular modelling are the most promising approaches used in CADD to reduce the experimental costs and duration. The design and quality of the molecular model rely on the available 3D structure of biomolecules. In homology modelling, the templates are threaded with the target sequence and optimized to build a 3D structure model. The templates are fragments of 3D experimental structures retrieved from the structure database. PDB is the worldwide repository for the 3D structure of biomolecules determined using X-ray crystallography, NMR spectroscopy, and 3D electron microscopy. The collaborative members of wwPDB are RCSB PDB, PDBe, PDBj, and BMRB[(1)].

The PDB releases macromolecular structure data in three types of file formats namely PDB (.pdb), PDBx/mmCIF (.cif), or PDBML/XML (.xml) for various purpose. PDB file format is the standard file format released for displaying/analyzing macromolecules using 3D molecular visualization and modelling tools. Due to certain limitations in the PDB file format, PDBx/mmCIF and PDBML/XML file formats are introduced to extend the accessibility. Moreover, the wwPDB has stopped modifying or extending the data dictionary of the PDB file format. The PDBx/mmCIF file format is considered as standard PDB archive format and PDBML/XML file format for programming purposes[(1–3)]. PDB$_{cle}$ tool was designed to generate a 3D template structure (biomolecule chains) in PDB file format and retrieve compounds in SDF file format from PDB sever through parsing the PDBML/XML file.

2 Methods

The atomic coordinate records in the PDB file format starts with ATOM/HETATM identifier (Figure 1). The standard residues such as proteins and nucleic acids start with ATOM record, and HETATM record for non-standard residues such as ions, solvent, cofactors, and inhibitors.



```
ATOM    109  CA BGLN A  25      12.934  37.577  12.849  0.30 47.16           C
:..
ATOM   1240  ND1 HIS A  85      29.595  -2.515  16.135  1.00 31.70           N
:..
ATOM   1893 HH22 ARG A 438      91.089   5.780   3.045  1.00  4.21           H
ATOM   1894  N   VAL A 439      97.999   9.523   4.596  1.00  0.67           N
:..
HETATM 2686 ZN    ZN A 145      21.082   0.322  11.682  1.00 45.47          ZN
:..
ATOM   6705  OP1  DG C   6     476.647  48.669 150.761  1.00114.89           O1-
```

Fig 1. Example for atomic coordinate records in the PDB file.

The information of atomic coordinates in the PDB file is arranged in tabular format. Each row represents atom record and column represents data field. There are a total of 20 columns in a record with a fixed width for each field. The maximum length to display the atom serial number is limited to 99999, due to the width size of the field. A brief explanation of fields in the atom records is given in the tables below (Table 1). The entry format and data description are given according to the current specification of PDB DDL (version 3.3). Some of the fields are either empty or ignored due to depreciation from the older version[(4,5)].

Table 1. Definition of columns in atomic coordinates section of standard PDB file format.

Column	Data Description	Data Type	Alignment	Width
1-6	Record type identifier	String	Left	6
7-11	Atom serial number	Integer	Right	5
12	Empty space	Null	Left	1
13-16	Atom name	String	Left*	4
17	Alternate location indicator	String	Left	1
18-20	Residue name	String	Right	3
21	Empty space	Null	Left	1
22	Chain identifier	String	Left	1
23-26	Residue sequence number	Integer	Right	4
27	Code for insertions of residues	String	Left	1
28-30	Empty space	Null	Left	3
31-38	X orthogonal coordinate (Å)	Decimal	Right	8.3
39-46	Y orthogonal coordinate (Å)	Decimal	Right	8.3
47-54	Z orthogonal coordinate (Å)	Decimal	Right	8.3
55-60	Occupancy	Decimal	Right	6.2

Continued on next page

Table 1 continued

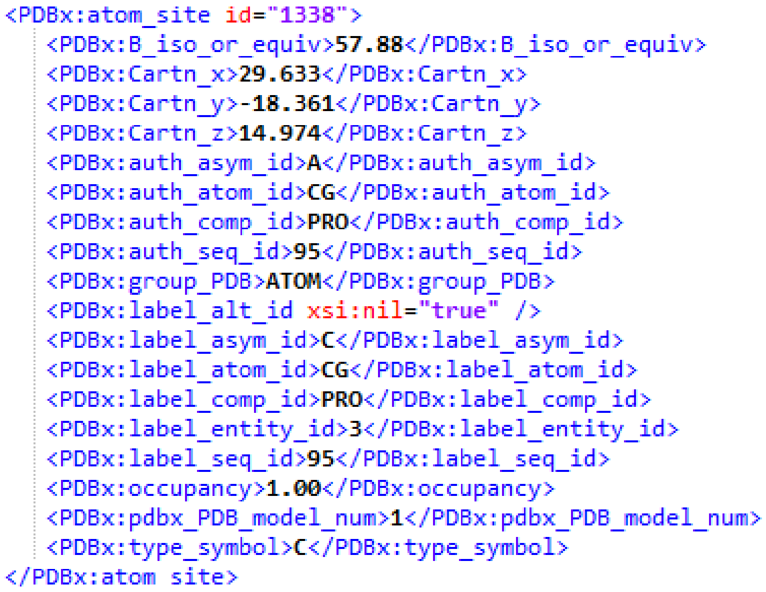
61-66	Temperature factor	Decimal	Right	6.2
67-72	Empty space	Null	Left	6
73-76	Segment identifier (deprecated)	String	Left	4
77-78	Element symbol	String	Right	2
79-80	Charge	String	Left	2

* The column 13-16 in alignment varies according to the length of an atom symbol name.

The PDB$_{cle}$ tool retrieves the PDBML/XML and/or PDBx/mmJSON file from the PDB server and generates a standard PDB file for macromolecules, FASTA file for sequence, and SDF file for small molecules. Separate PDB files are generated for DNA molecule by combining the atom records of two chains in pair or group according to the molecule type. The molecules in the PDB repository are categorized as a polypeptide, polydeoxyribonucleotide, cyclic-pseudo-peptide, polysaccharide, polydeoxyribonucleotide/ polyribonucleotide hybrid, polyribonucleotide, and small molecules[2–4]. The PDB$_{cle}$ tool splits the chains in the molecule under three major categories, (1) macromolecule, (2) macromolecule and small molecule complex, and (3) small molecule (Table 2). Based on the category, the PDB file is generated for each chain by parsing the <PDBx:atom_site> elements and attributes/objects in the PDBML/XML file (Figure 2).

Table 2. Classification of separation of sequence/structure file using PDB$_{cle}$ tool.

Molecule Type		Separated Files and Formats	
		3D Structure	Sequence
Macromolecule	Protein	Chains and complex structures (.pdb)	Chains (.fasta)
	DNA	Chains, pairs, and complex structures (.pdb)	Chains (.fasta)
	RNA	Chains and complex structures (.pdb)	Chains (.fasta)
	Hybrid	Chains and complex structures (.pdb)	Chains (.fasta)
Small molecule	Ions	Chain-wise sorted structures (.sdf)	NA
	Solvent	Chain-wise sorted structures (.sdf)	NA
	Cofactors	Chain-wise sorted structures (.sdf)	NA
	Inhibitors	Chain-wise sorted structures (.sdf)	NA



```
<PDBx:atom_site id="1338">
   <PDBx:B_iso_or_equiv>57.88</PDBx:B_iso_or_equiv>
   <PDBx:Cartn_x>29.633</PDBx:Cartn_x>
   <PDBx:Cartn_y>-18.361</PDBx:Cartn_y>
   <PDBx:Cartn_z>14.974</PDBx:Cartn_z>
   <PDBx:auth_asym_id>A</PDBx:auth_asym_id>
   <PDBx:auth_atom_id>CG</PDBx:auth_atom_id>
   <PDBx:auth_comp_id>PRO</PDBx:auth_comp_id>
   <PDBx:auth_seq_id>95</PDBx:auth_seq_id>
   <PDBx:group_PDB>ATOM</PDBx:group_PDB>
   <PDBx:label_alt_id xsi:nil="true" />
   <PDBx:label_asym_id>C</PDBx:label_asym_id>
   <PDBx:label_atom_id>CG</PDBx:label_atom_id>
   <PDBx:label_comp_id>PRO</PDBx:label_comp_id>
   <PDBx:label_entity_id>3</PDBx:label_entity_id>
   <PDBx:label_seq_id>95</PDBx:label_seq_id>
   <PDBx:occupancy>1.00</PDBx:occupancy>
   <PDBx:pdbx_PDB_model_num>1</PDBx:pdbx_PDB_model_num>
   <PDBx:type_symbol>C</PDBx:type_symbol>
</PDBx:atom_site>
```

Fig 2. Example of a single PDBML/XML atomic coordinate record that is equivalent to PDBx/mmCIF content.

The sequence files are generated to FASTA file for each chain by parsing the <PDBx:pdbx_seq_one_letter_code_can>, <PDBx:pdbx_gene_src_ncbi_taxonomy_id>, and <PDBx:pdbx_gene_src_scientific_name> elements and attributes/objects in the PDBML/XML file[5]. Whereas, small molecules are retrieved from the PDB server through the API interface, according to each HETATM records.

3 Results and Discussion

PDBML is a document markup language defined by PDB that consist of a set of DTDs framed according to SGML protocol. The current stable release of PDBML schema is PDBx v50 (https://pdbml.pdb.org/schema/pdbx-v50.xsd) for atomic coordinates data and wwPDB Validation v004 (https://www.wwpdb.org/validation/schema/wwpdb_validation_v004.xsd) for structure validation data. Each PDBML/XML files consist of different schema referenced by the PDB exchange data object (xmlns:PDBx, xmlns:xsi, and xsi:schemaLocation) for atomic coordinates[5]. The PDB$_{cle}$ tool acts as a client that submits the query (PDB ID) to the PDB server through RESTful service and retrieves the PDBML/XML archive data. Moreover, the retrieved data are parsed and converted to separate PDB files, FASTA files, and SDF files according to the molecule type and chains. The result hits of chains and ligands are properly annotated by mouse move-over tool-tip text containing a short title of the molecule and hyperlinks to the original resource.

A sample query search for the macromolecule (PDB ID: 6PL7) from the Protein Data Bank was done (Figure 3). The Figure 4 represents the result of separated 3D structure of human DNA polymerase eta (Pol η) complexed with DNA, 1FZ, magnesium, glycerol, and water extracted using PDB$_{cle}$ tool. It consists of protein chain A, nucleotide chain P and T, nucleotide base pair group (P, T), protein chain A complex with 1FZ, magnesium, glycerol, and water, nucleotide chain P complex with water, nucleotide chain T complex with water, 1FZ, magnesium, glycerol, and water. The retrieved sequence of protein chain A, nucleotide chain P, nucleotide chain T in FASTA file format is given bellow[6].

>6PL7, Chain A|DNA polymerase eta|Homo sapiens (9606)
GPHMATGQDRVVALVDMDCFFVQVEQRQNPHLRNKPCAVVQYKSWKGGGIIAVSYEARAFGVTRSMWADDAKKLCPDLLL
AQVRESRGKANLTKYREASVEVMEIMSRFAVIERASIDEAYVDLTSAVQERLQKLQGQPISADLLPSTYIEGLPQGPTTA
EETVQKEGMRKQGLFQWLDSLQIDNLTSPDLQLTVGAVIVEEMRAAIERETGFQCSAGISHNKVLAKLACGLNKPNRQTL
VSHGSVPQLFSQMPIRKIRSLGGKLGASVIEILGIEYMGELTQFTESQLQSHFGEKNGSWLYAMCRGIEHDPVKPRQLPK
TIGCSKNFPGKTALATREQVQWWLLQLAQELEERLTKDRNDNDRVATQLVVSIRVQGDKRLSSLRRCCALTRYDAHKMSH
DAFTVIKNCNTSGIQTEWSPPLTMLFLCATKFSAS

>6PL7, Chain P|DNA (5'-D(*AP*GP*CP*GP*TP*CP*AP*T)-3')|Homo sapiens (9606) AGCGTCAT

>6PL7, Chain T|DNA (5'-D(*CP*AP*TP*AP*AP*TP*GP*AP*CP*GP*CP*T)-3')|Homo sapiens (9606) CATAATGACGCT

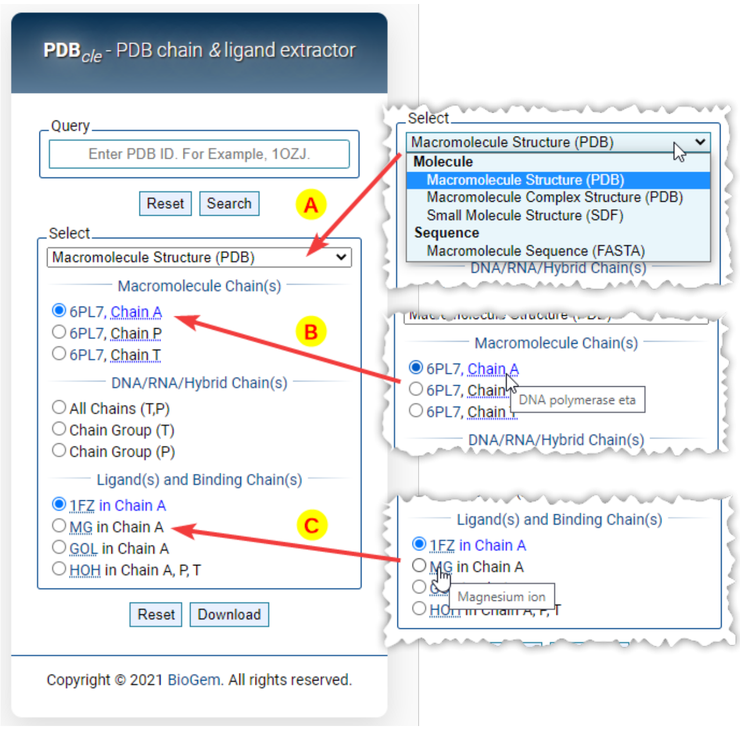


Fig 3. Screenshot of the PDB$_{cle}$ tool result page for a query search '6PL7'. (A) Dropdown list displaying category of molecule for splitting structure and sequence, (B) tool-tip text displaying short title of the polymer chain, and (C) tool-tip text and hyperlink to the original resource.

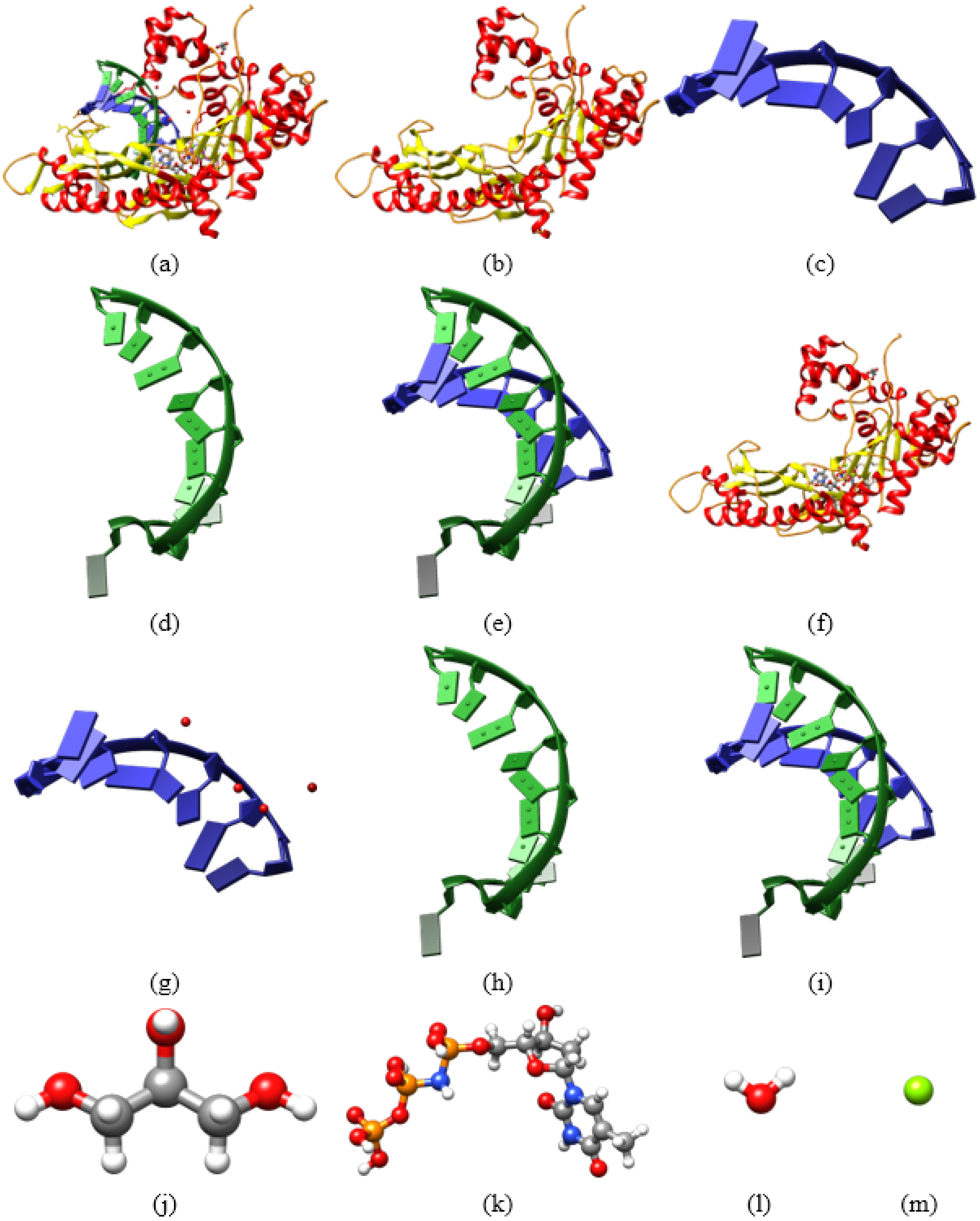


Fig 4. The macromolecule (a) human DNA polymerase eta complex (PDB ID: 6PL 7) is split into (b) protein chain A, (c) DNA chain P, (d) DNA chain T, (e) DNA base pairs (chains P and T), (f) protein chain A complex with thymidine-5’-[α,β-imide]triphosphoric acid (1FZ), magnesium ion (MG), glycerol (GOL), and water (HOH), (g) DNA chain P complex with HOH, (h) DNA chain T complex with HOH, (i) DNA chain T and P complex with HOH, (j) GOL, (k) 1FZ, (l) HOH, and (m) MG using PDB$_{cle}$ tool.

The PDB$_{cle}$ tool is tested with the various complex of molecules from the PDB which include protein, DNA, RNA, hybrid, ions, solvent, co-factors, and inhibitors (Table 3). Moreover, the splits of chains and ligands are downloaded in standard file format.

Table 3. Summary of some splits of macromolecule extracted using PDB$_{cle}$ tool.

PDB ID	Chains	Chain Complexes	Ligands
1OZJ	A, B, C, D	Zn^{+2} (**A, B**), H_2O (**A, B, C, D**)	H_2O, Zn^{+2}
2B00	A	Ca^{+2} (**A**), Glycocholic acid (**A**), H_2O (**A**)	Ca^{+2}, Glycocholic acid, H_2O
3A01	A, B, C, D, E, F, G, H	-	-
4CA1	A, B	Zn^{+2} (**A, B**), SO_4 (**A, B**), Cl^{-} (**A, B**), Glycerol (**A, B**)	Zn^{+2}, SO_4, Cl^{-}, Glycerol
5E21	A	H_2O (**A**)	H_2O
6GL7	A, B, C, D, E, F	-	-
7CY7	A, C, D	1,2-Ethanediol (**A**); Isopropyl (**A**); Fe^{+2} (**A**); H_2O (**A, C, D**)	1,2-Ethanediol; Isopropyl alcohol, Fe^{+2}, H_2O
8BNA	A, B	2’-(4-Hydroxyphenyl)-5-(4-Methyl-1-Piperazinyl)-2,5’-Bi-Benzimidazole (**A**); Mg^{+2} (**A**); H_2O (**A**)	2’-(4-Hydroxyphenyl)-5-(4-Methyl-1-Piperazinyl)-2,5’-Bi-Benzimidazole; Mg^{+2}; H_2O
9EST	A	SO_4 (**A**); Ca^{+2} (**A**); (2-Bromoethyl)(2-‘Formyl-4’-Aminophenyl) Acetate (**A**); H_2O (**A**)	SO_4; Ca^{+2}; (2-Bromoethyl)(2-‘Formyl-4’-Aminophenyl) Acetate; H_2O
101D	A, B	Mg^{+2} (**A**), H_2O (**A, B**), Netropsin (**B**)	Mg^{+2}, H_2O, Netropsin

4 Conclusion

The PDB$_{cle}$ tool finds structural and sequence insights of a macromolecule retrieved from the PDB. It allows downloading chain-wise molecule-specific 3D structure and sequence from the PDB. The PDB, SDF, and FASTA files are generated by PDB$_{cle}$ that mimic the standard biological database file format. In the future, service to download macromolecule in standard CIF file format will be available.

Abbreviations

1FZ – Thymidine-5’-[α,β-imide]triphosphoric acid, API – Application Programming Interface, BMRB – Biological Magnetic Resonance Data Bank, CADD – Computer-Aided Drug Design, CSS – Cascading Style Sheets, DDL – Dictionary Description Language, DTD – Document Type Definition, DNA – Deoxyribonucleic Acid, GOL – Glycerol, HOH – Water, JSON – JavaScript Object Notation, MG – Magnesium, mmCIF – macromolecular Crystallographic Information File, PDB – Protein Data Bank, PDB$_{cle}$ – PDB chain & ligand extractor, PDBe – Protein Data Bank in Europe, PDBj – Protein Data Bank Japan, PDBML – Protein Data Bank Markup Language, PDBx – PDB Exchange, PHP – PHP: Hypertext Preprocessor, RCSB – Research Collaboratory for Structural Bioinformatics, REST – REpresentational State Transfer, RNA – Ribonucleic acid, SMILES – Simplified Molecular Input Line Entry System, wwPDB – worldwide Protein Data Bank, XML – eXtensible Markup Language, XMLNS – XML Namespace, XSD – XML Schema Definition, XSI – XML Schema Instance

References

1) Berman HM. The Protein Data Bank. *Nucleic Acids Research*. 2000;28(1):235–242. Available from: https://dx.doi.org/10.1093/nar/28.1.235.
2) Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, Rodgers JR. The protein data bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology*. 1977;112(3):535–542. Available from: https://dx.doi.org/10.1016/s0022-2836(77)80200-3.
3) Burley SK, Bhikadiya C, Bi C, Bittrich S, Chen L, Crichlow GV, et al. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Research*. 2021;49(D1):D437–D451. Available from: https://dx.doi.org/10.1093/nar/gkaa1038.
4) Bourne PE, Berman HM, Mcmahon B, Watenpaugh KD, Westbrook JD, Fitzgerald P. Macromolecular Crystallographic Information File. *Methods in Enzymology*. 1997;277:571–590. Available from: https://doi.org/10.1016/S0076-6879(97)77032-0.
5) Westbrook J, Ito N, Nakamura H, Henrick K, Berman HM. PDBML: the representation of archival macromolecular structure data in XML. *Bioinformatics*. 2005;21(7):988–992. Available from: https://dx.doi.org/10.1093/bioinformatics/bti082.
6) Koag MC, Lee S. Structure of human DNA polymerase eta complexed with A in the template base paired with incoming non-hydrolyzable TTP. 2020. doi:10.2210/pdb6PL7/pdb.