The pdb2sql Python Package: Parsing, Manipulation and Analysis of PDB Files Using SQL Queries

Nicolas Renaud\textsuperscript{1} and Cunliang Geng\textsuperscript{1}

\textsuperscript{1}Netherlands eScience Center, Science Park 140 1098 XG Amsterdam, the Netherlands

Summary

The analysis of biomolecular structures is a crucial task for a wide range of applications ranging from drug design to protein engineering. The Protein Data Bank (PDB) file format (Burley et al., 2019) is the most popular format to describe biomolecular structures such as proteins and nucleic acids. In this text-based format, each line represents a given atom and entails its main properties such as atom name and identifier, residue name and identifier, chain identifier, coordinates, etc. Several solutions have been developed to parse PDB files into dedicated objects that facilitate the analysis and manipulation of biomolecular structures. This is, for example, the case for the BioPython parser (Cock et al., 2009,\textsuperscript{1}biopdb) that loads PDB files into a nested dictionary, the structure of which mimics the hierarchical nature of the biomolecular structure. Selecting a given sub-part of the biomolecule can then be done by going through the dictionary and selecting the required atoms. Other packages, such as ProDy (Bakan, Meireles, & Bahar, 2011), BioJava (Lafita, 2019), MMTK (Hinsen, 2000) and MDAnalysis (Gowers et al., 2016) to cite a few, also offer solutions to parse PDB files. However, these parsers are embedded in large codebases that are sometimes difficult to integrate with new applications and are often geared toward the analysis of molecular dynamics simulations. Lightweight applications such as pdb-tools (Rodrigues, Teixeira, Trellet, & Bonvin, 2018) lack the capabilities to manipulate coordinates.

We present here the Python package pdb2sql, which loads individual PDB files into a relational database. Among different solutions, the Structured Query Language (SQL) is a very popular solution to query a given database. However SQL queries are complex and domain scientists such as bioinformaticians are usually not familiar with them. This represents an important barrier to the adoption of SQL technology in bioinformatics. pdb2sql exposes complex SQL queries through simple Python methods that are intuitive for end users. As such, our package leverages the power of SQL queries and removes the barrier that SQL complexity represents. In addition, several advanced modules have also been built, for example, to rotate or translate biomolecular structures, to characterize interface contacts, and to measure structure similarity between two protein complexes. Additional modules can easily be developed following the same scheme. As a consequence, pdb2sql is a lightweight and versatile PDB tool that is easy to extend and to integrate with new applications.

Capabilities of pdb2sql

pdb2sql allows a user to query, manipulate, and process PDB files through a series of dedicated classes. We give an overview of these features and illustrate them with snippets of code. More examples can be found in the documentation (https://pdb2sql.readthedocs.io).
Extracting data from PDB files

`pdb2sql` allows a user to simply query the database using the `get(attr, **kwargs)` method. The attribute `attr` here is a list of or a single column name of the SQL database; see Table 1 for available attributes. The keyword argument `kwargs` can then be used to specify a sub-selection of atoms.

Table 1. Atom attributes and associated definitions in `pdb2sql`

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>serial</td>
<td>Atom serial number</td>
</tr>
<tr>
<td>name</td>
<td>Atom name</td>
</tr>
<tr>
<td>altLoc</td>
<td>Alternate location indicator</td>
</tr>
<tr>
<td>resName</td>
<td>Residue name</td>
</tr>
<tr>
<td>chainID</td>
<td>Chain identifier</td>
</tr>
<tr>
<td>resSeq</td>
<td>Residue sequence number</td>
</tr>
<tr>
<td>iCode</td>
<td>Code for insertion of residues</td>
</tr>
<tr>
<td>x</td>
<td>Orthogonal coordinates for X in Angstroms</td>
</tr>
<tr>
<td>y</td>
<td>Orthogonal coordinates for Y in Angstroms</td>
</tr>
<tr>
<td>z</td>
<td>Orthogonal coordinates for Z in Angstroms</td>
</tr>
<tr>
<td>occ</td>
<td>Occupancy</td>
</tr>
<tr>
<td>temp</td>
<td>Temperature factor</td>
</tr>
<tr>
<td>element</td>
<td>Element symbol</td>
</tr>
<tr>
<td>model</td>
<td>Model serial number</td>
</tr>
</tbody>
</table>

Every attribute name can be used to select specific atoms and multiple conditions can be easily combined. For example, let’s consider the following example:

```python
from pdb2sql import pdb2sql
pdb = pdb2sql('1AK4.pdb')
atoms = pdb.get('x,y,z',
               name=['C', 'H'],
               resName=['VAL', 'LEU'],
               chainID='A')
```

This snippet extracts the coordinates of the carbon and hydrogen atoms that belong to all the valine and leucine residues of the chain labelled A in the PDB file. Atoms can also be excluded from the selection by appending the prefix `no_` to the attribute name. This is the case in the following example:

```python
from pdb2sql import pdb2sql
pdb = pdb2sql('1AK4.pdb')
atoms = pdb.get('name, resName',
                no_resName=['GLY', 'PHE'])
```

This snippet extracts the atom and residue names of all atoms except those belonging to the glycine and phenylalanine residues of the structure. Similar combinations of arguments can be designed to obtain complex selection rules that precisely select the desired atom properties.

Manipulating PDB files

The data contained in the SQL database can also be modified using the `update(attr, vals, **kwargs)` method. The attributes and keyword arguments are identical to those in

the get method. The vals argument should contain a numpy array whose dimension should match the selection criteria. For example:

```python
import numpy as np
from pdb2sql import pdb2sql

pdb = pdb2sql('1AK4.pdb')
xyz = pdb.get('x,y,z', chainID='A', resSeq=1)
xyz = np.array(xyz)
xyz -= np.mean(xyz)
pdb.update('x,y,z', xyz, chainID='A', resSeq=1)
```

This snippet first extracts the coordinates of atoms in the first residue of chain A, then translates this fragment to the origin and updates the coordinate values in the database. pdb2sql also provides a convenient class `transform` to easily translate or rotate structures. For example, to translate the first residue of the structure 5 Å along the Y-axis,

```python
import numpy as np
from pdb2sql import pdb2sql
from pdb2sql import transform

pdb = pdb2sql('1AK4.pdb')
trans_vec = np.array([0, 5, 0])
transform.translation(pdb, trans_vec, resSeq=1, chainID='A')
```

One can also rotate a given selection around a given axis with the `rotate_axis` method:

```python
angle = np.pi
axis = (1., 0., 0.)
transform.rot_axis(pdb, axis, angle, resSeq=1, chainID='A')
```

### Identifying interface

The `interface` class is derived from the `pdb2sql` class and offers functionality to identify contact atoms or residues between two different chains with a given contact distance. It is useful for extracting and analysing the interface of, e.g., protein-protein complexes. The following example snippet returns all the atoms and all the residues of the interface of '1AK4.pdb' defined by a contact distance of 6 Å.

```python
from pdb2sql import interface

pdb = interface('1AK4.pdb')
atoms = pdb.get_contact_atoms(cutoff=6.0)
res = pdb.get_contact_residues(cutoff=6.0)
```

It is also possible to directly create an `interface` instance with a `pdb2sql` instance as input. In this case, all the changes in the `pdb2sql` instance before creating the new interface instance will be kept in the `interface` instance; afterwards, the two instances will be independent, which means changes in one will not affect the other.

```python
from pdb2sql import pdb2sql
from pdb2sql import interface
```
pdb = pdb2sql('1AK4.pdb')

pdbitf = interface(pdb)

atoms = pdbitf.get_contact_atoms(cutoff=6.0)

res = pdbitf.get_contact_residues(cutoff=6.0)

Computing Structure Similarity

The StructureSimilarity class allows a user to compute similarity measures between two protein-protein complexes. Several popular measures used to classify qualities of protein complex structures in the CAPRI (Critical Assessment of PRedicted Interactions) challenges (Méndez, Leplae, Maria, & Wodak, 2003) have been implemented: interface rmsd, ligand rmsd, fraction of native contacts and DockQ (Basu & Wallner, 2016). The approach implemented to compute the interface rmsd and ligand rmsd is identical to the well-known package ProFit (Martin & Porter, 2009). All the methods required to superimpose structures have been implemented in the transform class and therefore this relies on no external dependencies. The following snippet shows how to compute these measures:

```python
from pdb2sql import StructureSimilarity

sim = StructureSimilarity(decoy = '1AK4_model.pdb', ref = '1AK4_xray.pdb')

irmsd = sim.compute_irmsd_fast()
lrmsd = sim.compute_lrmsd_fast()
fnat = sim.compute_fnat_fast()
dockQ = sim.compute_DockQScore(fnat, lrmsd, irmsd)
```

Application

pdb2sql has been used at the Netherlands eScience center for bioinformatics projects. This is, for example, the case of iScore (Geng et al., 2019), which uses graph kernels and support vector machines to rank protein-protein interfaces. We illustrate the use of the package here by computing the interface rmsd and ligand rmsd of a series of structural models using the experimental structure as a reference. This is a common task for protein-protein docking, where a large number of docked conformations are generated and have then to be compared to ground truth to identify the best-generated poses. This calculation is usually done using the ProFit software and we, therefore, compare our results with those obtained with ProFit. The code to compute the similarity measure for different decoys is simple:

```python
from pdb2sql import StructureSimilarity

ref = '1AK4.pdb'
decoys = os.listdir('./decoys')
irmsd = {}

for d in decoys:
sim = StructureSimilarity(d, ref)
irmsd[d] = sim.compute_irmsd_fast(method='svd', izone='1AK4.izone')
```

Note that the method will compute the i-zone, i.e., the zone of the proteins that form the interface in a similar way to ProFit. This is done for the first calculations and the i-zone is
then reused for the subsequent calculations. The comparison of our interface rmsd values to those given by ProFit is shown in Fig 1.

Figure 1. Left - Superimposed model (green) and reference (cyan) structures. Right - comparison of interface rmsd values given by pdb2sql and by ProFit.

Acknowledgements

We acknowledge contributions from Li Xue, Sonja Georgievska, and Lars Ridder.

References


