ACCESS BIOINFORMATICS DATABASES WITH BIO-PYTHON

This project is aimed to deploy python-based programming pipelines and scripts to automate biological data retrieval and analysis.

3. PDB

This section fetches protein structures from PDB (Protein Data Bank), using the PDB module from Biopython to fetch, parse, and filter details of protein sequences from the PDB database.

Import Modules

In [45]:
```
from Bio.PDB import PDBParser,PDBList
```

The structure of the 7BYR protein is downloaded and stored:

In [48]:
```
pdbl=PDBList()
pdbl.retrieve_pdb_file("7BYR", file_format="pdb", pdir="dir")
```

```
Structure exists: 'dir\pdb7byr.ent'
```

Out[48]:

```
'dir\\pdb7byr.ent'
```

Then, the structure of the protein is read by PDBParser.

```
In [49]: parser = PDBParser()
         structure = parser.get_structure("7BYR","dir/pdb7byr.ent")
```

```
C:\Users\fxy40\anaconda3\lib\site-packages\Bio\PDB\StructureBuilder.py:89: PDBCo
nstructionWarning: WARNING: Chain A is discontinuous at line 26237.
  warnings.warn(
C:\Users\fxy40\anaconda3\lib\site-packages\Bio\PDB\StructureBuilder.py:89: PDBCo
nstructionWarning: WARNING: Chain B is discontinuous at line 26405.
  warnings.warn(
C:\Users\fxy40\anaconda3\lib\site-packages\Bio\PDB\StructureBuilder.py:89: PDBCo
nstructionWarning: WARNING: Chain C is discontinuous at line 26545.
  warnings.warn(
```

I then used a *for* loop to identify the number of chains in the protein sequence:

```
In [50]: for chain in structure[0]:
             print(f"chainid:{chain.id}")
```

```
chainid:A
chainid:B
chainid:C
chainid:H
chainid:L
chainid:D
chainid:E
chainid:F
chainid:G
chainid:I
chainid:J
```

As shown, the protein structure has a total of 11 chains named in alphabetical order.

The resolution of the protein is 3.84 angstroms:

In [51]:
```
resolution= structure.header["resolution"]
resolution
```

Out[51]:

3.84

Using a keyboard variable, I passed in the same structure variable containing the protein structure details. From here, I fetched the header, which has the keywords that are associated with the proteins.

```
In [52]: keywords = structure.header["keywords"]
         keywords
```

Out[52]:

```
'sars-cov-2, antigen, rbd, neutralizing antibody, viral protein'
```

As shown, the protein is a SARS-CoV-2 protein, and is associated with the keywords antigen, RBD, neutralizing antibody, and viral protein.