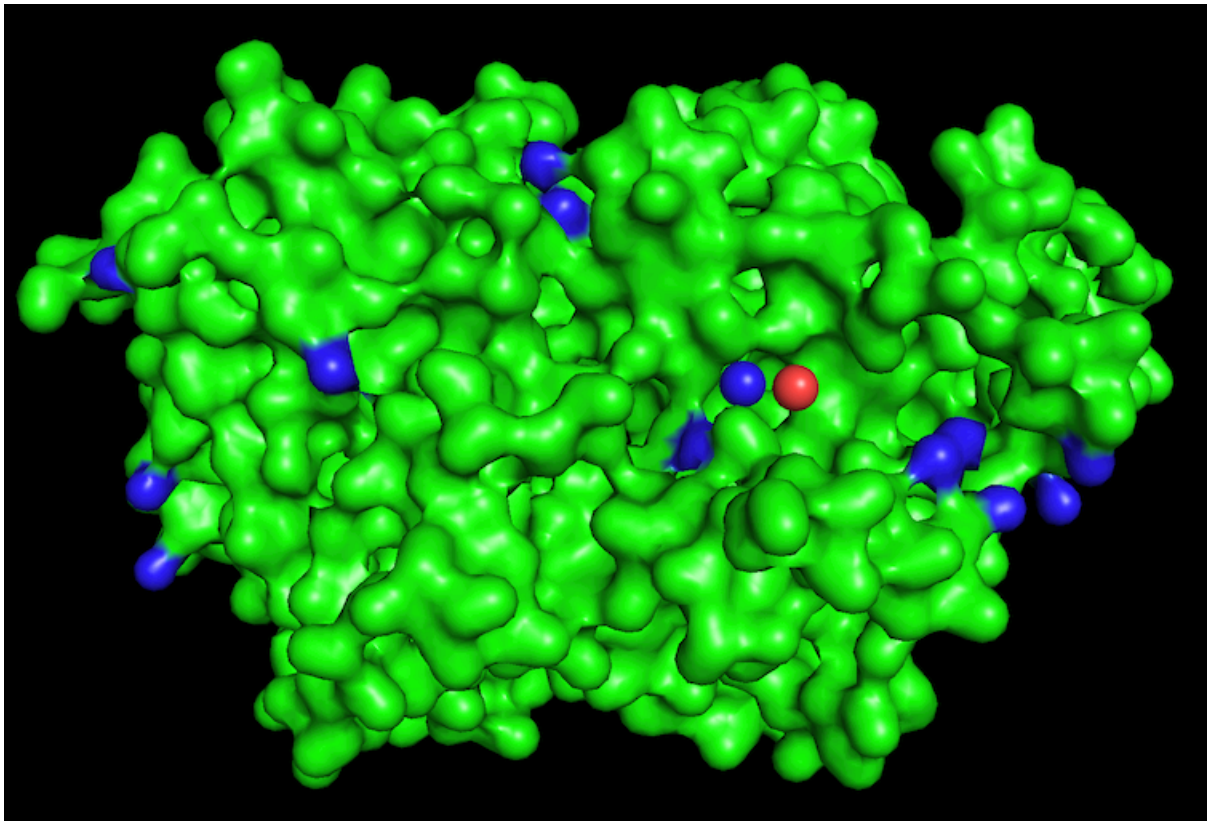

CS5242
Neural Networks and Deep Learning

2018 – 2019 Semester 1

Predicting Protein – Ligand Interaction by using Deep Learning Models

(Term Project)

Background

Molecular recognition between proteins and ligands plays an important role in many biological processes, such as membrane receptor signaling and enzyme catalysis. Predicting the structures of protein-ligand complexes and and finding ligands by virtual screening of small molecule databases are two long-standing goals in molecular biophysics and medicinal chemistry [1, 2]. Knowledge-based statistical potentials have been developed for modeling protein-ligand interactions. They are based on distributions of intermolecular features in large databases of protein-ligand complexes.

Over the past decade, deep learning has achieved remarkable success in various artificial intelligence research areas. Evolved from the previous research on artificial neural networks, this technology has shown superior performance to other machine learning algorithms in areas such as image and voice recognition, natural language processing, among others. The first wave of applications of deep learning in pharmaceutical research has emerged in recent years, and its utility has gone beyond bioactivity predictions and has shown promise in addressing diverse problems in drug discovery [3].

Training Dataset

In our dataset, there are ~3000 protein-ligand complexes that were determined experimentally with 3D structures available (For now, we are providing 100 of them). Each protein (xxxx_pro_cg_.pdb) and its ligand (xxxx_lig_cg_.pdb) are of one-to-one correspondence, i.e. they can bind to each other and make protein-ligand complex.

For each protein-ligand complex, protein and ligand data are provided in xxxx_pro_cg_.pdb and xxxx_lig_cg_.pdb files, respectively [4]. Due to one-to-one correspondence, only matching pairs of protein and ligand can make protein-ligand complex. In other words, xxxx_pro_cg_.pdb and yyyy_lig_cg_.pdb cannot make protein-ligand complex. Proteins and ligands are consisting of atoms and these atoms forming the structure of them (characterize them) by coming together. Protein and ligand data files contain (x, y, z) coordinates of each atom of the proteins and ligands. There is a 4th feature containing atom type. In Figure 1 and Figure 2, data structure of these files are shown for protein and ligand data files. All data fields are also shown in Figure 3 in protein and ligand files. For the detailed description of each field, please refer to pp.180-182 and pp.187-188 of [4].

0001_pro_cg.pdb

						X	Y	Z			TYPE
1	ATOM	2	CA	HIS A	0	17.186	-28.155	-12.495	1.00	26.12	C
2	ATOM	5	CB	HIS A	0	15.862	-28.669	-13.037	1.00	26.47	C
3	ATOM	12	CA	MET A	1	16.156	-26.144	-9.429	1.00	28.80	C
4	ATOM	15	CB	MET A	1	15.469	-24.766	-9.530	1.00	32.87	C
5	ATOM	20	CA	ASN A	2	15.018	-27.739	-6.188	1.00	22.61	C
6	ATOM	23	CB	ASN A	2	15.903	-27.912	-4.946	1.00	21.54	C
7	ATOM	28	CA	PRO A	3	11.654	-26.110	-5.652	1.00	21.30	C
8	ATOM	31	CB	PRO A	3	10.353	-26.736	-6.196	1.00	20.53	C
9	ATOM	35	CA	ILE A	4	10.653	-23.899	-2.732	1.00	20.12	C
10	ATOM	38	CB	ILE A	4	11.944	-23.314	-2.084	1.00	20.76	C
11	ATOM	43	CA	VAL A	5	7.516	-22.209	-1.354	1.00	21.45	C
12	ATOM	46	CB	VAL A	5	6.127	-22.882	-1.483	1.00	20.70	C
13	ATOM	50	CA	VAL A	6	7.351	-19.607	1.413	1.00	23.08	C
14	ATOM	53	CB	VAL A	6	8.432	-18.519	1.571	1.00	23.11	C
15	ATOM	57	CA	VAL A	7	4.032	-18.435	2.799	1.00	22.22	C
16	ATOM	60	CB	VAL A	7	2.994	-19.564	3.031	1.00	22.77	C
17	ATOM	64	CA	HIS A	8	2.839	-15.691	5.155	1.00	22.40	C
18	ATOM	67	CB	HIS A	8	3.714	-14.408	5.123	1.00	20.60	C
19	ATOM	75	CA	AGLY A	9	-0.090	-14.108	6.817	0.50	29.08	C
20	ATOM	76	CA	BGLY A	9	-0.201	-14.020	6.722	0.50	27.76	C

Figure 1: Data structure in protein data files

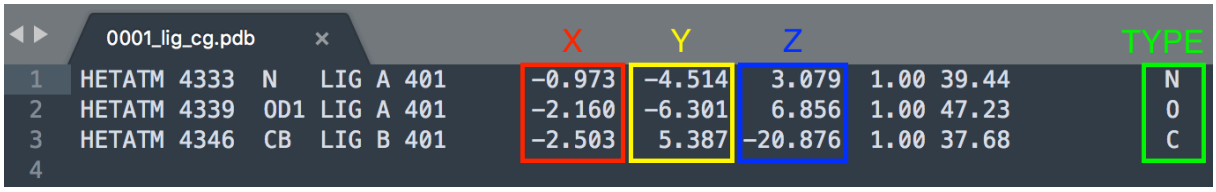

Figure 2: Data structure in ligand data files

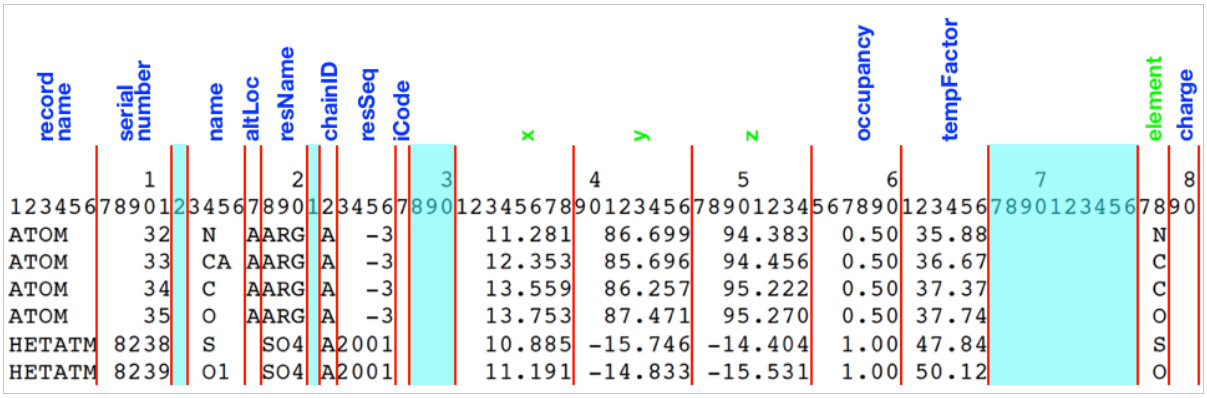

Figure 3: Data fields in a protein and ligand data files

In this project, we are solely interested in coordinates and types of the atoms (which are indicated in Figure 1 and Figure 2) constructing the structure of the proteins and ligands. In the "TYPE" field, atom types are given as: 'C - Carbon', 'O – Oxygen', 'N – Nitrogen', etc. In this project, we will not use atom types directly; instead we will treat them either as hydrophobic or polar. 'C' is interpreted as hydrophobic and the rest is interpreted as polar. The reason why we haven't changed them to hydrophobic ('h') and polar ('p') in our dataset is that you may visualize the structure of proteins and ligands by using some open-source software like PyMOL, which requires actual atom types [5, 6]. Visualization of protein1 and ligand1 with PyMOL software is shown in Figure 4. An example python script (read_pdb_file.py) is also provided in course website for reading pdb files to extract atom coordinates and atom types.

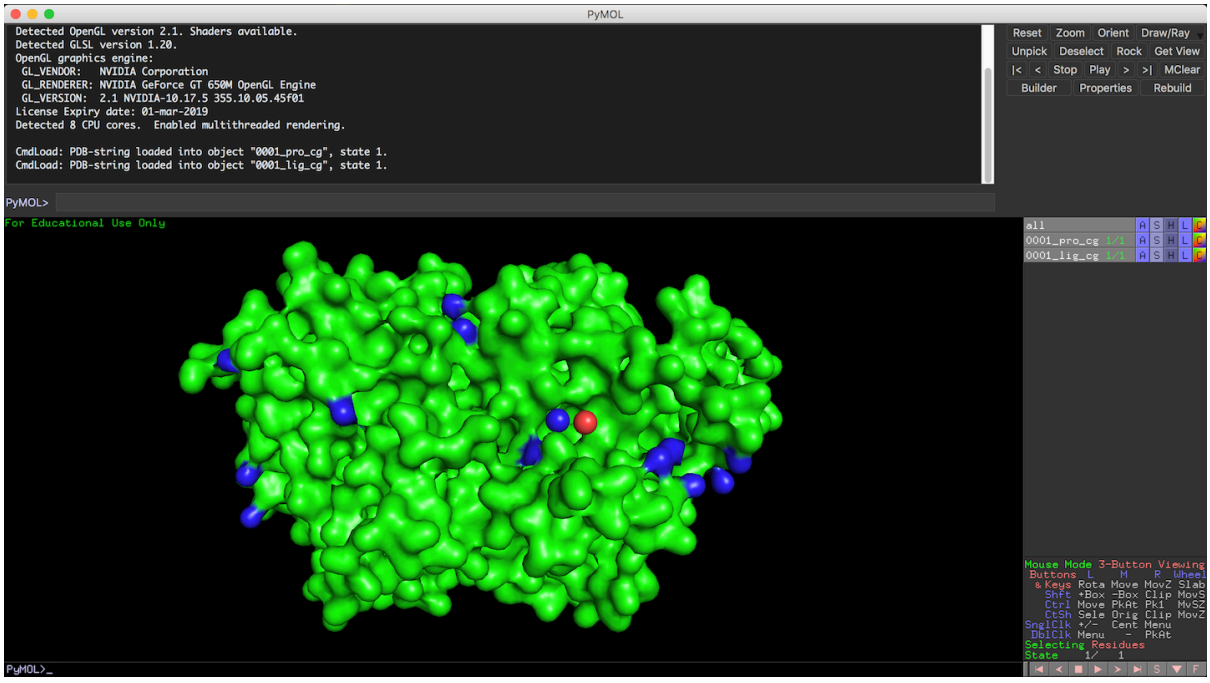

Figure 4: Visualizing protein and ligand structure in PyMOL

Project Task and Testing Data

You are required to train a neural network that takes in as input, the (x, y, z) coordinates of atoms and atom type of a pair of protein and the ligand, and then predict if they bind at the output of the network.

For the metric of grading, for each protein in the test data set, you are required to identify 10 ligands that are predicted to bind the protein. Prediction is considered to be correct if true ligand, which will bind to protein, is among the 10 ligands predicted to bind. For each protein, only one ligand will bind to it. You are allowed to suggest 10 candidates and hope that among your 10 candidates, you identify the correct one. Final score of the project is the number of proteins with a correct prediction for binding.

The testing dataset will consist of a set of protein .pdb files and ligand .pdb files. Each file will be named as rrrr_pro.cg.pdb and tttt_lig_cg.pdb where “rrrr” and “tttt” will be randomized indices. We will not give you which protein and ligand files are matching. For your project output, you need to provide a .csv file specifying which protein random index files match to which of ligand random index files.

File format for submissions will be given in later weeks. For now, you have been given enough information to start the project.

References

1) https://en.wikipedia.org/wiki/Protein
2) https://en.wikipedia.org/wiki/Protein_structure
3) https://en.wikipedia.org/wiki/Drug_design
4) ftp://ftp.wwpdb.org/pub/pdb/doc/format_descriptions/Format_v33_A4.pdf
5) https://pymol.org/2/#download
6) http://pymol.sourceforge.net/newman/userman.pdf