Drug repurposing for SARS-CoV-2

In this homework, you will predict the potential drug candidates targeting the SARS-CoV-2 proteins, using the information from the given heterogeneous networks.

1. Develop and evaluate your model:

You need to develop an algorithm to predict drug-target interactions based on the following three types of networks: drug-target interaction networks, protein-protein interaction networks (target=protein) and similarity networks. The interaction networks are binary and the similarity networks are real values. There are three types of nodes in these networks: drugs, virus targets and human targets. The goal is to predict the interactions in the drug-virus target interaction network.

Please use a cross-validation procedure to evaluate performance of your algorithm. That is, randomly split the drug-virus target interaction network according to the target dimension into 10 folds. This new-target setting is close to the real situation, in which the test targets are not seen in the training data. **Please report the AUC score on all the test data**, and (optional) AUPR scores under 1:1 and 1:10 positive-negative ratios. Note that there are much more negative samples than positive samples in the label matrix, as all the unknown drug-target pairs are regarded as negative samples. If you would like to calculate the AUPR scores, please use all the positive samples in the test data and sample a corresponding amount of negative samples. In the training process, you can either use all the negative training samples and sample some negative samples according to a fixed ratio.

Here is the detailed description of the attached data:
Network data:
(1) drug - **virus protein target** interaction matrix: **VDTI_net.npy** (This is the label matrix)
(2) drug - **human protein target** interaction matrix: **HDTI_net.npy**
(3) **drug similarity** matrix: **Drug_simi_net.npy**
(4) **human protein similarity** matrix: **human.npy**
(5) **virus protein similarity** matrix: **virusseq_add_ncov.npy**
(6) **human protein - protein interaction** matrix: **PPI_net.npy**
(7) **virus protein - human protein interaction** matrix: **VHI_net.npy**

Indices of the matrices:
(8) **virus protein index** dictionary (key: ID, value: index): **virusseq_add_ncov.pkl**
(9) **human protein index** dictionary (key: ID, value: index): **human_seq_iddict.pkl**
(10) **drug index** dictionary (key: InChI of drug, value: index): **drug_iddict**
For example, the corresponding drugs and targets in the rows and columns of (1) drug - virus protein target interaction matrix are stored in (10) and (8). The rows and columns of (5) virus protein similarity matrix also follow the same order in (8), and so do other matrices.

Other information in case you need:
(11) **virus protein sequence dictionary** (key: ID, value: protein sequence): **virusseq_add_ncovseqdict.pkl**
(12) **human protein sequence dictionary** (key: ID, value: protein sequence): **human_seq_dict.pkl**
(13) The drug - virus protein target interactions were extracted from file **final_Virus_DTI_add_organism.tsv**.
(14) InChIs, DrugBank IDs and names of the drugs - **drug_info.tsv**

2. Apply your model to predict drugs for SARS-CoV-2 proteins

The drug candidates should be predicted from all the 6255 drugs provided. Please choose to predict for 2-3 viral proteins, from the following list. In the virus protein ID dictionary, those proteins whose IDs cannot be converted to a number belong to SARS-CoV-2. The corresponding proteins are listed in the table below.

**Please list top 10 predicted drugs for each of the selected 2-3 proteins.** It would be nice if you can provide more information about the drugs. The DrugBank IDs of the drugs can be found in file "drug_info.tsv", and the chemical structures (images), original indications and targets of these drugs can be found in https://www.drugbank.ca/.

| **Protein ID** | **Protein name** |
|---|---|
| sp\|P0DTC1\|R1A_WCPV | Replicase polyprotein 1a |
| sp\|P0DTC2\|SPIKE_WCPV | Spike glycoprotein |
| sp\|P0DTC3\|AP3A_WCPV | Protein 3a |
| sp\|P0DTC4\|VEMP_WCPV | Envelope small membrane protein |
| sp\|P0DTC5\|VME1_WCPV | Membrane protein |
| sp\|P0DTC6\|NS6_WCPV | Non-structural protein 6 |
| sp\|P0DTC7\|NS7A_WCPV | Protein 7a |
| sp\|P0DTC8\|NS8_WCPV | Non-structural protein 8 |
| sp\|P0DTC9\|NCAP_WCPV | Nucleocapsid protein |
| sp\|P0DTD1\|R1AB_WCPV | Replicase polyprotein 1ab |
| sp\|P0DTD2\|ORF9B_WCPV | Protein 9b |
| sp\|P0DTD3\|Y14_WCPV | Uncharacterized protein 14 |
| sp\|P0DTD8\|NS7B_WCPV | Protein non-structural 7b |
| tr\|A0A663DJA2\|A0A663DJA2_9BETC | ORF10 |

Note: As these data are unpublished, please DO NOT distribute these data outside this course.