Disease-target association prediction

1 Problem Description

In this problem, your goal is to develop an algorithm to predict disease-target associations based on the following two types of data:

(1) The interactions or associations between nodes in a heterogeneous network depicting the relationships among drugs, targets, diseases and drug side-effects. In particular, we have linkages between drugs, proteins, diseases and side effects. The specific graph depicting the relationship is shown in Fig. 1.

(1) The similarity scores between drugs or targets. Specifically, for chemicals (drugs), the scores are calculated through Tanimoto coefficient using the product-graph of these two structures [1]. For targets, the scores are computed using the simple Smith-Waterman scores [2] on genome sequences.

There is one thing that needs to be specified about the interaction network data: the all-zero lines are very common in the drug-target interaction matrix, that is, one drug (target) may have no interaction with any one of the existed targets (drugs).

Implement your algorithm in any programming language that you are familiar with, such as Java, C/C++, Matlab, etc., and then test your algorithm on the given data. You are allowed to call any other available public package in your program. If so, please include the library in your final submission.

Use a cross-validation procedure to evaluate performance of your algorithm. More details about cross-validation can be found in wikipedia or other references. You can use the *area under receiver operating characteristic* (AUC) curve, the *area under the precision-recall* (AUPR) curve, or other measures to assess the performance of your prediction method.

After running the cross-validation procedure, use the whole dataset as training data to predict the new disease-target associations. For each new association in your top prediction list (e.g., top 10 predictions), use "google" or "google scholar" to search the literature and check whether there exists any evidence to support your prediction. If so, describe them in your report.

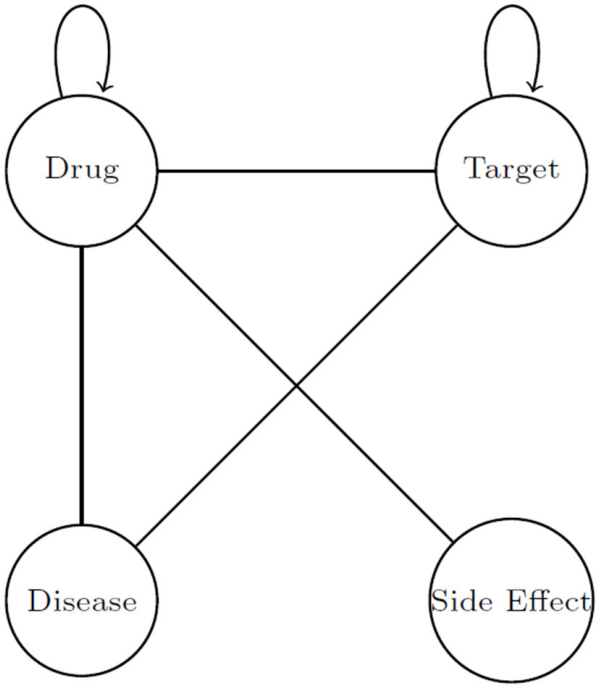

Figure 1: The schema of the heterogeneous network.

2 Data

In the the root folder, `drug_dict_map` and `protein_dict_map` suggest the (complete) real drug and protein names corresponding to each name label used in subfolder `./InteractionData`'s specification files.

All the interaction and similarity data are organized in a normal matrix-like format. And the specification data are just names separated by carriage return `\n`.

In the folder `./InteractionData`, there are in total 6 kinds of interactions between different entities, which are listed as follows:

`mat_drug_se.txt`	:	Drug-SideEffect interaction matrix
`mat_protein_protein.txt`	:	Protein-Protein interaction matrix
`mat_protein_drug.txt`	:	Protein-Drug interaction matrix
`mat_drug_drug.txt`	:	Drug-Drug interaction matrix
`mat_protein_disease.txt`	:	Protein-Disease interaction matrix
`mat_drug_disease.txt`	:	Drug-Disease interaction matrix

Note that these matrices have been already preprocessed to be perfectly matched to each other (e.g., the rows of drug-disease interaction matrix, which are drugs, corresponds to exactly the same

drugs listed in the rows of protein-drug interaction matrix). The corresponding list of the entities, i.e., the list of diseases, drugs, proteins and side-effects are listed in the following files:

`disease.txt`	:	Disease names
`drug.txt`	:	Drug names
`protein.txt`	:	Protein names
`se.txt`	:	Side Effects

In the folder `./SimilarityData`, the two matrices denote the similarity scores within drugs and targets (proteins):

`Similarity_Matrix_Drugs.txt`	:	Drug similarity scores
`Similarity_Matrix_Proteins.txt`	:	Protein similarity scores

Note: (1) As these data are unpublished, please DO NOT distribute these data outside this course. (2) These data are very raw data, you can preprocess the data according to some principles. If so, please describe the principles that you use.

3 Requirement of Report

In your final report, you should address the following points:

(1) Details of your algorithm, such as overview, pseudo-code (or flow chart), etc.

(2) Performance evaluation of your algorithm.

(3) Discussion about strength and limitation of your algorithm.

(4) Description of your top prediction results using the whole dataset as training data, and validation results from the literature.

4 Final Submission

For final submission, you need to provide: (1) report; (2) source code and binary executable file of your program, and a short readme file that describes how to compile and run your program.

References

[1] Masahiro Hattori, Yasushi Okuno, Susumu Goto, and Minoru Kanehisa. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *Journal of the American Chemical Society*, 125(39):11853–11865, 2003.

[2] Temple F Smith and Michael S Waterman. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.