Manual: Link Prediction for Drug Sensitivity

Zachary Stanfield[1], Mustafa Coskun[2], and Mehmet Koyuturk[1,2]

[1]Center for Proteomics and Bioinformatics, School of Medicine
[2]Department of Electrical Engineering and Computer Science, School of Engineering, Case Western Reserve University, Cleveland, OH

1 Description

This user manual provides instructions for implementing the network-based approach described in "Drug response prediction as a link prediction problem" published in *Scientific Reports* (2017). This approach results in a simple formulation through representation of multiple data types as a heterogeneous network that integrates the relationships between drugs, cell lines, molecular aberrations, and functional associations among biomolecules. We compute "network profiles" for drugs and cell lines, which represent their location in this network. The associations between these profiles are then used to predict links between drugs and cell lines. Through leave-one-out cross validation (LOOCV) on the Genomics of Drug Sensitivity in Cancer (GDSC) data set, we achieve and accuracy of 88% for the classification of sensitive and resistant cell line-drug pairs. We also performed cross-classification studies by using profiles derived from the GDSC network to predict links in the Cancer Cell Line Encyclopedia (CCLE) network, and observed 85% prediction accuracy. Finally, we determined that our approach is highly accurate for the problem of ranking drugs by their effectiveness against a new cell line/sample.

2 Implementation

This method is implemented in **MATLAB**.

2.1 GDSC LOOCV

This section describes how to perform leave-one-out cross validation on the GDSC dataset as discussed in the paper. To obtain the predicted scores and assess performance, run the script *RunGDSCLOOCV.m* from the MATLAB command line. The script will prompt for the following inputs:

- Number of desired MATLAB workers (for running in parallel). Using only 1 process, runtime for this script is roughly 5 hours, so run in parallel if results are needed more quickly.
- A value for the restart probability α (should be in the interval $(0, 1)$). The smaller this parameter, the greater the influence of the network topology on the calculation of cell line and drug profiles.
- A value for the sparsity parameter ϵ (should be in the interval $[0, 1)$). Decreasing this parameter increases the dimensionality of the profiles to be correlated by allowing nodes more distant from the seed node of the RWR to affect the final predicted scores.

Upon completion, *RunGDSCLOOCV.m* will output four files to a directory named *GDSCLOOCVoutput*.

Output Files

- *GDSCprocessed_α_ϵ.csv*: A matrix containing the sensitivity and resistance correlation scores for each cell line-drug pair. There are eight columns in this matrix and contain the following information:
  - 1 and 2 contain the drug and cell line index of each pair (referencing the variable *NBPMnodes_GDSCstr_red.mat*)
  - 3 and 4 contain the sensitive and resistant normalized edge weights derived from the original GDSC IC50 values
  - 5 and 7 contain the sensitive and resistant correlation scores from the calculated network profiles
  - 6 and 8 are carried over from processing previous matrices
- *GDSCevalDrugs_α_ϵ.csv*: A matrix containing multiple accuracy measures for each drug. These measures correspond to the columns and are correlation and concordance index with the normalized IC50 values, AUC, accuracy, sensitivity, specificity, precision, and f_measure.
- *GDSCevalCLs_α_ϵ.csv*: Performance measures for cell lines. Same format as the above file.
- *GDSCLOOCVresults_α_ϵ.mat*: - A MATLAB file containing the three above matrices as three separate MATLAB data objects.

2.2 CCLE Validation

This section describes how to perform the cross-study classification using the CCLE dataset as described in the paper. The response prediction scores and performance values can be obtained by running the script *RunCCLEvalidation.m* from the MATLAB command line. Similarly to the GDSC script, a prompt for the following inputs will appear:

- A value for the restart probability α (should be in the interval $(0, 1)$). See GDSC LOOCV for description.
- A value for the sparsity parameter ϵ (should be in the interval $[0, 1)$). See GDSC LOOCV for description.

Runtime for this script is about one hour on a standard laptop. The script will generate five output files that will be placed in a directory with the name *CCLEoutput*.

Output Files

- *CCLEprocessed_α_ϵ.csv*: A matrix containing the sensitivity and resistance correlation scores for each cell line-drug pair. There are eight columns in this matrix (see GDSC LOOCV for a description). Note here that all pair weights (columns 3 and 4) are zero as these cell lines do not appear in the GDSC dataset. Since the cell lines here are new, their index (column 2) starts after the last node index of the GDSC network.
- *CCLEevalDrugs_α_ϵ.csv*: A matrix containing AUC values for each drug. The seven columns represent different cutoffs for calling sensitive/resistant cell line-drug pairs. Unlike the GDSC, the CCLE performs each drug screening assay with the same dosage profile for all drugs (1-8 μM). The seven columns correspond to denoting pairs as sensitive if their IC50 value is less than 1, 2, 3,... 7.
- *CCLEevalCLs_α_ϵ.csv*: Performance measures for cell lines (using a class cutoff of 2; can be changed in main script). Same format as the GDSC cell line file.
- *CCLEevalOverallAUC_α_ϵ.csv*: AUC values when assessing accuracy of all cell line-drug pairs at the seven CCLE sensitivity cutoffs.
- *CCLEresults_α_ϵ.mat*: - A MATLAB file containing the three above matrices as three separate MATLAB data objects. The MATLAB variable *CCLE_drug_allpairs_eval* is a structural variable containing the drug AUCs, other accuracy measures, and overall AUC values.

2.3 Use Case: Predicting Sensitivity for a Provided Sample

A script named *RunNewSamplePred.m* was written in order to allow users to apply our network-based method to their cell lines or samples of interest. Upon running this script, the user will be provided with the same three prompts as for the GDSC LOOCV pipeline (number of processes, α, and ϵ). Runtime for each cell line is roughly 3 minutes, so parallelization may be desired for analyzing multiple cell lines. The user must also have a mutation file in their MATLAB path (it can be added directly to the unzipped folder provided on the download website before running the script). This file will be loaded by the script and used to rank drugs by their effectiveness against each provided cell line.

Input File

- File should be named *MutationFile.csv*
- Must be a numeric matrix containing genes (Entrez IDs only) with no row or column headers. Each column should represent a cell line. Genes in each column are not required to have any particular order.
- Cell lines have differing number of mutations. Make sure that the matrix is complete (i.e. add zeros to the end of any column such that all columns have the same number of elements/rows).
- The file *example_MutationFile.csv* is provided as a template

Output Files

The output files will be placed in a directory with the name *UserCaseOutput*.

- *UserProcessed_α_ϵ_date.csv*: A matrix with the same format as *CCLEprocessed*.
- *UserResults_α_ϵ_date.csv*: A matrix containing the final predicted scores (resistance score - sensitivity score) for each cell line-drug pair. Rows represent the 138 GDSC drugs (order is that of the data file GDSCuniquedrugs_orig_red.mat) and columns represent the user-provided cell lines.
- *UserRankedDrugs_α_ϵ_date.csv*: A matrix the same size as UserResults where now the drug names are listed and have been sorted, for each cell line, by their final score such that the most effective drug is in row one. Again, the user-provided cell lines represent the columns.
- *UserRankedDrugScores_α_ϵ_date.csv*: A matrix corresponding to the previous file. Drug scores for each cell line in the same order as in UserRankedDrugs (i.e. the drug in row 2 column 4 is the second most effective drug for the fourth input cell line; the name of the drug is in UserRankedDrugs[2,4] and its sensitivity score is in UserRankedDrugScores[2,4]). The user-provided cell lines represent the columns. Note that in our framework, a negative score indicates sensitivity and a positive score indicates resistance.
- *UserResults_α_ϵ_date.mat*: - A MATLAB file containing the first two of the above matrices as two separate MATLAB data objects.

3 Raw Data

The original data is all publicly available for download. The protein interactions were obtained from Biogrid, GDSC cell line mutations from COSMIC (*CosmicCLP_MutantExport* file at cancer.sanger.ac.uk/cosmic), GDSC drugs and IC50 values from cancerrxgene.org/downloads, and CCLE data from broadinstitute.org/ccle/home. Original files for the COSMIC mutations and GDSC IC50 values were quite large and required extensive, sometimes partially manual, processing before they were able to be used in a meaningful way for our algorithm. For these reasons, all original, raw data is not provided at our download page.