Bayesian Efficient Multiple Kernel Learning (BEMKL):

In this study [1], we used the state-of-the-art BEMKL model [2] for drug response prediction from multi-omics datasets, integrated in a biologically meaningful way. BEMKL was the top-performing model in the NCI/DREAM7 drug sensitivity prediction challenge [3], among various classes of machine learning model.

The implementation of BEMKL model in Matlab is provided in 'BEMKL' folder, where following files can be found:

1. bemkl_supervised_multitask_regression_variational_train.m
2. bemkl_supervised_multitask_regression_variational_test.m
3. bemkl_loocv.m: loads data, performs analysis by calling bemkl model and applies Leave-One-Out Cross-Validation.
4. bemkl_demo.m: uses bemkl_loocv and selects cytotoxic or targeted agents and view combinations of choice. It also contains code for performance measurement using Spearman's correlation, root-mean-square value and concordance index.
5. civalue.m: calculates concordance index between true and predicted values.

NB: If you use BEMKL in your study, remember to cite the original sources of the model [2, 3] in addition to this study [1].

Rule Based Protein Selection (RBPS):

In order to identify combinations of protein abundances that best explain the drug sensitivity profiles (and which eventually could be used as predictive biomarkers for clinical translation), we used RBPS model [1].

The implementation of RBPS in R is provided in 'RBPS' folder, where following files can be found:

1. rbps.R: contains implementation of RBPS as well as demo code to identify protein combinations for selected seven targeted compounds (macbecin II, selumetinib, tamoxifen, tanespimycin, alvespimycin alvocidib and lapatinib), as mentioned in the study [1].
2. ProtViews.RData: is a list and contains three objects
   (a) prot: the input data matrix containing aggregated set of 55 proteins from RPPA and MS data sets for NCI60 cell lines, as discussed in the study [1], for GMP6 view combination for targeted agents.
   (b) prot_bin: the binarized form of prot matrix. In the absence of a ground truth, we used scores above mean as up-regulated and below mean as down-regulated.
   (c) prot_names: contains names of the proteins (denoting to the columns of prot)
3. DrugResponseTargeted.RData: is a list and contains two objects
   (a) targeted_response: the input data matrix contains drug responses, as -pGI$_{50}$

($-\log_{10}GI_{50}$), of 24 selected targeted agents with known targeted mechanism of action across NCI60 cell lines, as discussed in the study [1].

(b) targeted_agents: contains names of the targeted compounds (denoting to the columns of targeted_response)

4. CellTypes.csv: contains list of cell lines and cell types of NCI60 data set.

NB: If you use RBPS in your study, remember to cite this study [1]. If you use NCI60 data set in your study, remember to cite the original sources of the data [4, 5] in addition to this study [1].

NCI60 Data set:

The primary data set used in this study [1] comprised of genomic, molecular and proteomics profiles of 58 human pan-cancer cell lines from the National Cancer Institute (NCI), also referred to as NCI-60 cell lines panel [5], along with their drug responses reported as $-pGI_{50}$ ($-\log_{10}GI_{50}$) values. Further details on this data used, can be found in supplementary Table 1 of study [1].

Frequent feature-wise data missingness was observed in MS-based proteomics dataset (on average 55% partially-measured proteins across NCI-60 cell-lines panel). To avoid data-sparsity-induced noise issues, only the completely available MS features were considered.

NCI60 data set used in the study [1], is provided in in Matlab format in 'Processed NCI60 Data' folder, where in the following data objects can be found:

1. DataViews.mat is a list and contains 22 objects consisting of names of NCI60 cell lines, data matrices and gene/protein names of 6 omics profiles (mRNA, miRNA, CNV, MUT, RPPA and MS), across 58 NCI60 cell lines. Subset of all these matrices for genes overlapping Catalogue Of Somatic Mutations In Cancer (COSMIC, v77, http://cancer.sanger.ac.uk/cosmic) is also provided in objects ending with 'cc'.
2. ViewCombinations.mat contains object view_combinations. It consists of individual as well combinations of omics profiles (column 1) used in study [1] and kernel types (column 2) for calculating their respective kernel matrices.
3. DrugResponse.mat: is a list and contains four objects
   (a) cytotoxic_response: the input data matrix contains drug responses of 47 FDA-approved cytotoxic agents across NCI60 cell lines, as discussed in the study [1].
   (b) cytotoxic_agents: contains NSC number - the NCI's internal ID number and names of the cytotoxic compounds (denoting to the columns of cytotoxic_response).
   (c) targeted_response: the input data matrix contains drug responses of 24 selected targeted agents with known targeted mechanism of action across NCI60 cell lines, as discussed in the study [1].
   (d) targeted_agents: contains NSC number - the NCI's internal ID number and names of the targeted compounds (denoting to the columns of targeted_response).

NB: If you use NCI60 data set in your study, remember to cite the original sources of the data [4, 5] in addition to this study [1].

References

[1] M. Ali, et al. Global proteomics profiling improves drug sensitivity prediction: results from a multi-omics, pan-cancer modeling approach. (Manuscript under review)

[2] Costello, J. C. et al. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. Nat. Biotechnology, 32, 1202-1212.

[3] Gönen, M. (2012). Bayesian efficient multiple kernel learning. In J. Langford, and J. Pineau (Ed.), 29th International Conference on Machine Learning (ICML-12) (pp. 1-8). New York, USA: ACM.

[4] Shankavaram, U. T. et al. (2009). CellMiner: a relational database and query tool for the NCI-60 cancer cell lines. BMC Genomics, 10, 1.

[5] Shoemaker, R. H. (2006). The NCI60 human tumour cell line anticancer drug screen. Nat. Reviews Cancer, 6, 813-823.
