Readme

**Prediction of Adverse Drug Effect for drugs in combination using transcriptomic data.**

Computational analysis of drugs in combination to predict Adverse drug reaction is a promising solution to reduce the fatality and economic loss reported in post-marketing surveillance. Recent works suggest that transcriptomic data are potentially informative about the phenotypic properties of drugs, including Adverse Drug Reaction by a single drug. In this study, we aim to build predictive models for Adverse Drug Reactions caused by drugs in combinations using transcriptomic data. Our model provides a scalable approach to derive up to 243 Adverse Drug Reactions with an average accuracy of 84%. We suggest 11 Adverse Drug Reactions that can be predicted without any inclusion of negative samples using our model. Our model can be used to fill in the gaps that exist when it comes to extracting drug-drug-ADR association.

**Getting stated:**

These instructions will get to you a copy of the research work up on your local machine for testing purpose.

**Prerequisite:**

R environment – version: 3.6.3

* Link to install: Follow the steps in <https://www.datacamp.com/community/tutorials/installing-R-windows-mac-ubuntu>
* Packages: tidyr. For installing, use command: install.packages(“tidyr”)

Python – version: 3.7.3

Jupyter notebook

* Link to install: <https://jupyter.org/install>
* Packages: pandas, numpy, ast, sklearn

**Description of data**

* **Input Data**
  + *LINCS\_Gene\_Experssion\_signatures\_CD.csv:* Contains differential Gene Expression for 978 landmark genes for 20.338 drug. Dimension: 20,338 x 978
  + *GO\_transformed\_signatures\_log.csv:*Expression values for the 20,338 drug for 4,438 Gene ontology terms. Dimension: 20,338 x 4,438
  + *MACCS\_bitmatrix.csv:*Chemical Fingerprints representing drugs. Dimension: 20,338 x 166
  + *Label\_two\_side.*csv:Dataset for Adverse Drug Reaction caused by drugs in combination as in PharmGKB dataset
* **Processed File**
  + *represesntation\_of\_single\_drug.csv :* Drug represented by their DGE, expression for Gene Ontology and MACCS bit
  + *ADR\_list.csv:* list of Adverse Drug Reaction in categorical text format corresponding to the drugs in combination as extracted from PharmGKB TwoSide table
  + *dataset\_subset.csv:* drugs in combination represented as concatenation of each of the drug vectors
  + *ADR\_combine\_hot\_encoded.csv:* One hot encoded vector of Adverse Drug Reaction corresponding to each drug combination
  + *required\_ADR.csv:* list of 243 ADRs after processing as per the literature
  + *ADR\_combined.csv:* one hot encoded vector of 243 ADRs after processing as per the literature
* **Output**
  + *pca\_dataset\_for\_5000\_features.csv:* dataset after PCA reduction
  + *pca\_loading\_for\_model.csv:* Eigen vectors after PCA reduction
  + *pca\_center\_for\_model.csv:* moved centers after PCA reduction

**Process for deployment:**

1. Download the input files for the research work:
   1. Folder: Input Dataset
   2. Files:
      1. label\_two\_side.csv: Dataset for Adverse Drug Reactions caused by drugs in combination as in PharmGKB dataset
      2. GO\_transformed\_signatures\_log.csv : Expression value of Gene Ontology terms for 20k drugs.
      3. LINCS\_Gene\_Experssion\_signatures\_CD.csv : Differential Gene Expression (DGE) of 978 landmark genes for 20k drugs.
      4. MACCS\_bitmatrix.csv: Chemical structure of drugs in MACCS format.

The datasets - GO\_transformed\_signatures\_log.csv, LINCS\_Gene\_Experssion\_signatures\_CD.csv and MACCS\_bitmatrix.csv can be downloaded from  <http://maayanlab.net/SEP-L1000/> . For further information about the dataset, please refer to the publication.

1. Preprocessing of the input data
   1. Code name: data\_preparation\_for\_model.R
   2. Environment: R
   3. Input: files downloaded in Step 1
   4. Outputs: (Folder: Processed\_Files)
      1. represesntation\_of\_single\_drug.csv : Drug represented by their DGE, expression for Gene Ontology and MACCS bit
      2. ADR\_list.csv: list of Adverse Drug Reaction in categorical text format corresponding to the drugs in combination as extracted from PharmGKB TwoSide table
      3. dataset\_subset.csv: drugs in combination represented as concatenation of each of the drug vectors
2. Convert the ADR list from Step 2 to one hot encoded format
   1. Code name: one\_hot\_code.py
   2. Environment: python
   3. Input: ADR\_list.csv
   4. Output: (Folder: Processed\_Files)
      1. ADR\_combine\_hot\_encoded.csv: One hot encoded vector of Adverse Drug Reaction corresponding to each drug combination
3. Preprocess the one hot encoded Adverse drug Reaction list
   1. Code name: code\_for\_required\_ADR.R
   2. Environment: R
   3. Input: ADR\_combine\_hot\_encoded.csv
   4. Output: (Folder: Processed Files)
      1. required\_ADR.csv: list of 243 ADRs after processing as per the literature
      2. ADR\_combined.csv: one hot encoded vector of 243 ADRs after processing as per the literature

Write up below has been explained for one subset. Similar steps can lead to results for all 5 subsets. However, final output datasets for all the subsets are present in the folder

1. Prepare the division of sets
   1. Code name: subset\_formation.R
      1. Environment: R
      2. Input: dataset\_subset.csv,ADR\_combined.csv
      3. Output: (Folder: Processed Files)
         1. ADR\_subset\_1.csv
   2. Code\_name:code\_for\_validation\_and\_pca.R
      1. Environment: R
      2. Input: ADR\_subset\_1.csv, dataset\_subset.csv
      3. Output: (Folder: Processed Files)
         1. dataset\_subset\_1.csv
         2. dataset\_validation\_1.csv
         3. dataset\_for\_validation\_after\_pca.csv
         4. dataset\_for\_validation\_after\_pca\_2000\_features.csv
         5. dataset\_for\_validation\_after\_pca\_3000\_features.csv
         6. dataset\_for\_validation\_after\_pca\_2250\_features.csv
         7. dataset\_for\_validation\_after\_pca\_2500\_features.csv
         8. dataset\_for\_validation\_after\_pca\_2750\_features.csv
         9. dataset\_for\_validation\_after\_pca\_500\_features.csv
         10. dataset\_for\_validation\_after\_pca\_750\_features.csv
         11. dataset\_for\_validation\_after\_pca\_1000\_features.csv
         12. dataset\_for\_validation\_after\_pca\_1250\_features.csv
         13. dataset\_for\_validation\_after\_pca\_1500\_features.csv
         14. dataset\_for\_validation\_after\_pca\_1750\_features.csv
         15. dataset\_after\_pca.csv
         16. dataset\_after\_pca\_2000\_features.csv
         17. dataset\_after\_pca\_3000\_features.csv
         18. dataset\_after\_pca\_2250\_features.csv
         19. dataset\_after\_pca\_2500\_features.csv
         20. dataset\_after\_pca\_2750\_features.csv
         21. dataset\_after\_pca\_500\_features.csv
         22. dataset\_after\_pca\_750\_features.csv
         23. dataset\_after\_pca\_1000\_features.csv
         24. dataset\_after\_pca\_1250\_features.csv
         25. dataset\_after\_pca\_1500\_features.csv
         26. dataset\_after\_pca\_1750\_features.csv
         27. pca\_for\_subset\_1\_for\_all\_features.csv
         28. center\_for\_pca\_subset\_1\_all\_features.csv
         29. rotation\_for\_pca\_subset\_1\_all\_features.csv
         30. ADR\_dataset\_for\_training\_subset\_1.csv
         31. ADR\_validation\_for\_validation\_subset\_1.csv
2. Run ANN for complete dataset
   1. Code name: subset\_formation.R
      1. Environment:Python
      2. Input: dataset\_subset.csv, ADR\_dataset\_for\_training\_subset\_1.csv, ADR\_validation\_for\_validation\_subset\_1.csv
      3. Output: (Folder: output\_files)
         1. model-1.json
         2. model-1.h5
         3. predict-1.py
3. Run machine learning algorithms on the PCA reduced data
   1. Code name: extraTree.py,KNN.py,Logistic\_Regression.py,Naïve\_Bayes.py,ANN.py
      1. Environment: python (preferably jupyter)
      2. Input:
         1. dataset\_after\_pca\_1000\_features.csv
         2. ADR\_dataset\_for\_training\_subset\_1.csv
         3. dataset\_for\_validation\_after\_pca\_1000\_features.csv
         4. ADR\_validation\_for\_validation\_subset\_1.csv
      3. Output (Folder: output): 486 files – 243 for ADRs and 243 for evaluation metrics of ADR
   2. Code name: compilation.py
      1. Environment: python (preferably jupyter)
      2. Input:
         1. 243 files with listed evaluation metrics for ADRs. Files end with “\_acc.csv”
      3. Output (Folder: output):
         1. Performance\_subset\_1.csv

Changing the value of variable “sample” would result with different number of pca components. For example, if sample =1000, number of PCA components taken is 1000.

**All the codes are present in the Codes folder**

**Data Selection**

1. **Input Data:** 
   1. *dataset\_subset\_1*: Training dataset for subset 1
   2. *dataset\_validation\_1.csv* : Testing data for subset 1
   3. *dataset\_subset\_2*: Training dataset for subset 2
   4. *dataset\_validation\_2.csv* : Testing data for subset 2
   5. *dataset\_subset3*: Training dataset for subset 3
   6. *dataset\_validation\_3.csv* : Testing data for subset 3
   7. *dataset\_subset\_4*: Training dataset for subset 4
   8. *dataset\_validation\_4.csv* : Testing data for subset 4
   9. *thirty\_percent.csv*: ADR’s for 30% cut-off
   10. *fifty\_percent.csv*: ADR’s for 50% cut-off
   11. *seventy\_percent.csv*: ADR’s for 70% cut-off
   12. *ninety\_percent.csv*: ADR’s for 90% cut-off
2. **Output Data**
   1. *ten.csv* : ANN Results for 10% cut-off
   2. *thirty.csv* : ANN Results for 30% cut-off
   3. *fifty.csv* : ANN Results for 50% cut-off
   4. *seventy.csv* : ANN Results for 70% cut-off
   5. *ninety.csv* : ANN Results for 90% cut-off
   6. *Training Predictions*: Predictions for 4 training subsets for each of the 5 cut offs
   7. *Testing Predictions*: Predictions for 4 validation subsets for each of the 5 cut offs

1. **Training the model:**
   1. **Code name:**  training.ipynb, comparison.ipynb
   2. **Environment:** Python

**Model Tuning**

1. **Input Data:** 
   1. *dataset\_subset\_1*: Training dataset for subset 1
   2. *dataset\_validation\_1.csv* : Testing data for subset 1
   3. *dataset\_subset\_2*: Training dataset for subset 2
   4. *dataset\_validation\_2.csv* : Testing data for subset 2
   5. *dataset\_subset3*: Training dataset for subset 3
   6. *dataset\_validation\_3.csv* : Testing data for subset 3
   7. *dataset\_subset\_4*: Training dataset for subset 4
   8. *dataset\_validation\_4.csv* : Testing data for subset 4
2. **Output Data**
   1. *Training Predictions*: Predictions for 4 training subsets for different parameters
   2. *Testing Predictions*: Predictions for 4 validation subsets for different parameters
3. **Training the model:**
   1. **Code name:**  training.ipynb, tuning.ipynb
   2. **Environment:** Python

**Selecting Thresholds**

1. **Input Data:** 
   1. *Training Predictions*: Predictions for 4 training subsets for different parameters
   2. *Testing Predictions*: Predictions for 4 validation subsets for different parameters
2. **Output Data**
   1. *selectedThresholds.csv*
3. **Training the model:**
   1. **Code name:**  thresholdSelection.ipynb, comparision.ipynb
   2. **Environment:** Python

**Authors:**

Susmitha Shankar, Ishita Bhandari, David Okou, Gowri Srinivasa, Prashanth Athri

**License**

This project is licensed under the MIT License - see the [LICENSE.md](https://gist.github.com/PurpleBooth/LICENSE.md) file for details