**Drug-Gene interactions from Comparative Toxicogenomics Database (CTD), DrugBank and PharmGKB**

**From CTD:**

Download the latest version of **Chemical-gene interactions** from CTD <http://ctdbase.org/downloads/;jsessionid=02DD8D4FEDB4D115A3884A314F35AE97#cg>

**To remove introduction section in the source file**

Java program Preprocessor.java deletes the introduction lines prior to file processing. The program can be executed with the following command:

*$ javac Preprocessor.java*

*$ java Preprocessor INPUT\_FILE OUTPUT\_FILE\_noIntroSection*

**To extract human related records**

Java program HumanRecordsCollector.java retrieves the records related to human. The program can be executed with the following command:

*$ javac HumanRecordsCollector.java*

*$ java HumanRecordsCollector OUTPUT\_FILE\_noIntroSection OUTPUT\_FILE\_human*

**To retrieve unique list of chemicals/drugs and gene interactions**

Unique list of chemicals/drugs and gene interactions is filtered using Linux command.

*$ cut -f1,4 OUTPUT\_FILE\_human | sort | uniq >* ***OUTPUT\_FILE\_CTD***

**From DrugBank:**

Download the latest version of **drugbank.xml** file from <https://www.drugbank.ca/releases/latest>

**To retrieve unique list of drug-gene interactions**

Java program GeneDrugRelationshipExtractor.java retrieves drug-gene interactions from the downloaded file.

*$ javac GeneDrugRelationshipExtractor.java*

*$ java GeneDrugRelationshipExtractor drugbank.xml OUTPUT\_FILE*

Unique list of drug-gene interactions is collected with the following Linux command.

*$ cut -f2,1 OUTPUT\_FILE | awk -F'\t' '{ print $2 "\t" $1}' >* ***OUTPUT\_FILE\_DrugBank***

**From PharmGKB:**

Download **relationships.tsv** file from <https://www.pharmgkb.org/>

**To retrieve records by excluding ‘LiteratureAnnotation’ only**

Java program GoldStandardGenerator.java retrieves drug-gene interactions by excluding the records annotated only as ‘LiteratureAnnotation’.

*$ javac GoldStandardGenerator.java*

*$ java GoldStandardGenerator relationships.tsv* ***OUTPUT\_FILE\_PharmGKB***

**Drug name normalization**

Chemical/drug name might not be consistent across various resources. In order to identify the chemicals/drugs and their synonyms as the same entity, we assigned drug ID from our chemical/drug lexicon.

CTD / DrugBank / PharmGKB:

*$ javac**DrugnameToLexicondrugidMapper.java*

*$ java DrugnameToLexicondrugidMapper INPUT\_FILE\_resource* ***OUTPUT\_FILE\_resource***

INPUT\_FILE\_resource corresponds to OUTPUT\_FILE\_CTD, OUTPUT\_FILE\_DrugBank and OUTPUT\_FILE\_PharmGKB, based on the resource CTD, DrugBank and PharmGKB respectively.

**Post-processing**

We observed that certain chemical/drug names are not available in our lexicon and returned null value.

CTD / DrugBank / PharmGKB:

Total number of records:

*$ grep 'null' FILE\_NAME | wc*

Total number of drugs

*$ grep 'null' FILE\_NAME | cut -f1 | sort | uniq | wc*

The records without drug ID are subsequently removed.

CTD / DrugBank / PharmGKB:

*$ awk -F"\t" '$3 != "null" { print $1"\t"$2"\t"$3 }' INPUT\_FILE\_NAME > OUTPUT\_FILE\_NAME*

OUTPUT\_FILE\_NAME corresponds to CTD\_FILE, DrugBank\_FILE and PharmGKB\_FILE, based on the resource CTD, DrugBank and PharmGKB respectively.

**Unique list of chemical/drug-gene association from all three resources**

The output from all resources are combined.

*$ cat CTD\_FILE DrugBank\_FILE PharmGKB\_FILE > INTERACTIONS\_OUTPUT\_FILE*

Java Program GeneDrugAssociationGSUniqueList.java retrieves unique list of interactions based on gene symbol (column 1) and lexicon drug ID (column 3).

*$ javac GeneDrugAssociationGSUniqueList.java*

*$ java GeneDrugAssociationGSUniqueList INTERACTIONS\_OUTPUT\_FILE INTERACTIONS\_OUTPUT\_FILE\_uniquelist*

**Unique list of drugsonly-gene association from all three resources**

Java Program DrugsOnlyFilterer.java filters drugs only-gene interactions.

*$ javac DrugsOnlyFilterer.java*

*$ java DrugsOnlyFilterer INTERACTIONS\_OUTPUT\_FILE\_uniquelist INTERACTIONS\_OUTPUT\_FILE\_uniquelist\_drugsonly*

**SUMMARY (Aug 2017)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **# of interactions** | **# of chemicals/drugs** | **# of drugs** | **# of genes** |
| Chemical/drug-gene | 401,133 | 11,156 | - | 23,096 |
| Drug-gene | 240,418 | - | 5,055 | 20,684 |