**Disease-Chemical/Drug association from CTD and NDF-RT**

**Comparative Toxicogenomics Database (CTD)**

Download the latest version from CTD <http://ctdbase.org/downloads/#cd>.

Delete the introduction lines.

CTD contains both curated and inferred disease-drug associations. Use the following awk command to filter the curated associations.

*$ cat INPUT\_FILE | awk -F'\t' '{if($6=="therapeutic") print }' > INPUT\_FILE\_curatedOnly*

Retrieve only chemical name (column 1), disease name (column 4) and disease ID (MeSH or OMIM identifier).  *$ cut -f1,4,5 INPUT\_FILE\_curatedOnly > INPUT\_FILE\_selectedColumns*

Identify unique concept identifier (CUI) for chemicals/drugs from the chemicals/drugs lexicon previously compiled from three resources namely UMLS Metathesaurus, DrugBank and PharmGKB (refer Project: DrugDict/DrugsDictionaryCompiler).

Java program /src/edu/uom/med/idMapper/DrugnameToLexicondrugidMapper.java is used to achieve the same and the following are the commands to execute the program file.

*$ javac DrugnamToLexicondrugidMapper.java*

*$ java DrugnamToLexicondrugidMapper INPUT\_FILE\_selectedColumns OUTPUT\_FILE1*

All chemicals/drugs may not have equivalent CUIs. Remove the records without CUI.

*$ awk -F"\t" '$4 != "null" { print $1 "\t" $2 "\t" $3 "\t" $4 }' OUTPUT\_FILE1 > OUTPUT\_FILE2*

Replace Mesh or OMIM ID with CUI for disease mentions.

Java program /src/edu/uom/med/ctd/Conceptname2CUIMapper.java is used to achieve the same and the following are the commands to execute the program file.

*$ javac Conceptname2CUIMapper.java*

*$ java Conceptname2CUIMapper OUTPUT\_FILE2 OUTPUT\_FILE3*

MeSH or OMIM ID may not match with CUI for some disease synonyms. Remove the records without CUI.

*$ awk -F"\t" '$4 != "null" { print $1 "\t" $2 "\t" $3 "\t" $4 }' OUTPUT\_FILE3 >* ***OUTPUT\_FILE\_CTD***

**National Drug File – Reference Terminology (NDF-RT)**

We used MRREL.RRF file within UMLS Metathesaurus to retrieve NDF-RT records.

*$ grep 'NDFRT' MRREL.RRF > OUTPUT\_FILE*

Filter the disease-drug association records with annotation as ‘may treat’.

*$ grep 'may\_treat' OUTPUT\_FILE > OUTPUT\_FILE\_filtered*

Retrieve a unique list of disease-drug associations.

*$ cut -f1,5 -d'|' OUTPUT\_FILE\_filtered > OUTPUT\_FILE\_filtered\_uniquelist*

Retrieve a unique list of CUI and concepts from previously processed file from UMLS Metathesaurus (refer Project: ConceptMap/UMLSMetathesaurusCompiler).

*$ cut -f1,15 -d'|' mrconso.preferredVocabularies > mrconso.preferredVocabularies\_CUIAndConcept*

The associations from NDFRT are represented only with CUI. Map CUI to respective concept name in UMLS Metathesaurus.

Java program /src/edu/uom/med/ndfrt/cui2ConceptnameMapper.java is used to achieve the same and the following are the commands to execute the program file.

*$ javac cui2ConceptnameMapper.java*

*$ java cui2ConceptnameMapper mrconso.preferredVocabularies\_CUIAndConcept OUTPUT\_FILE\_filtered\_uniquelist OUTPUT\_FILE1*

Map the chemicals / drugs to customized ID in the chemicals / drugs lexicon (refer Project: DrugDict/DrugsDictionaryCompiler).

Java program /src/edu/uom/med/idMapper/DrugnameToLexicondrugidMapper.java is used to achieve the same and the following are the commands to execute the program file.

*$ javac DrugnameToLexicondrugidMapper.java*

*$ java DrugnameToLexicondrugidMapper OUTPUT\_FILE1 OUTPUT\_FILE2*

Remove the records that do not map to customized ID in the chemicals / drugs lexicon.

*$ awk -F"\t" '$5 != "null" { print $1 "\t" $2 "\t" $3 "\t" $4 "\t" $5 }' OUTPUT\_FILE2 > OUTPUT\_FILE3*

Reorder the columns in the output file to match with the output file from CTD.

*$ awk -F"\t" ' { print $4 "\t" $2 "\t" $5 "\t" $1 }' OUTPUT\_FILE3 > OUTPUT\_FILE*

**Group records with drug / drug with dosage as drug:**

Sort the drugs based on drug names.

*$ cat OUTPUT\_FILE | sort -k1,1 | uniq > OUTPUT\_FILE\_sorted*

Combine the drug and drug with dosage as one entity and assign the ID for drug. Java program /src/edu/uom/med/ndfrt/DrugnameWithDosageRegexProcessor.java is used to achieve the same and the following are the commands to execute the program file.

*$ javac DrugnameWithDosageRegexProcessor.java*

*$ java DrugnameWithDosageRegexProcessor OUTPUT\_FILE\_sorted* ***OUTPUT\_FILE\_NDFRT***

**Disease-chemical/drug associations from CTD and NDF-RT**

Combine the output from both resources.

*$ cat OUTPUT\_FILE\_CTD OUTPUT\_FILE\_NDFRT >* ***OUTPUT\_FILE\_CTD\_NDFRT***

Retrieve a unique list of disease-chemical/drug associations based on customized chemical/drug ID and disease CUI. Java program /src/edu/uom/med/allResources/DiseaseChemicalORDrugAssociationUniqueList.java is used to achieve the same and the following are the commands to execute the program file.

*$ javac DiseaseChemicalORDrugAssociationUniqueList.java*

*$ java DiseaseChemicalORDrugAssociationUniqueList OUTPUT\_FILE\_CTD\_NDFRT* ***OUTPUT\_FILE\_CTD\_NDFRT\_unique***

**Disease-drugs only associations from CTD and NDF-RT**

Filter a list of disease-drugs only associations by defining drugs as the concepts belonging to UMLS semantic types clinical drugs, antibiotics and Pharmacologically active substances. Java program /src/edu/uom/med/allResources/DrugsOnlyFilterer.java is used to achieve the same and the following are the commands to execute the program file.

*$ javac DrugsOnlyFilterer.java*

*$ java DrugsOnlyFilterer OUTPUT\_FILE\_CTD\_NDFRT\_unique* ***OUTPUT\_FILE\_CTD\_NDFRT\_unique\_drugsOnly***

**Disease-drugs only associations from CTD and NDF-RT restricted to drugs with gene interaction information**

Java program /src/edu/uom/med/allResources/AssociationRetriever.java is used to achieve the same. The program also eliminates the associations involving “Biological Products”. The following the commands are used to execute the program file.

*$ javac AssociationFilterer.java*

*$ java AssociationFilterer* ***DRUG\_GENE\_INTERACTION\_FILE******OUTPUT\_FILE\_CTD\_NDFRT\_unique\_drugsOnly OUTPUT\_FILE\_CTD\_NDFRT\_unique\_drugsOnly\_restrictedToDrugsWithGeneInteractions***

The ***DRUG\_GENE\_INTERACTION\_FILE*** is the final output of Gene2DrugAssociationGoldStandard project.

**Summary (Dec 2017)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **# of associations** | **# of diseases** | **# of chemicals/drugs** | **# of drugs only** |
| Disease-Chemicals/drugs | 44,914 | 2,518 | 8,766 | - |
| Disease-Drugs only | 40,530 | 2,459 | - | 7,105 |
| Disease-Drugs only (drugs with gene interactions) | 27,415 | 2,384 | - | 2,794 |