```
==============================================================
*
* BioCreative VII
*
* Track 1 - Text mining drug and chemical-protein interactions (DrugProt)
*
*
* Training set - version 1.1 - June 28th
* Development set - version 1.1 - June 28th
*
*
* URL: https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-1/
*
*
* contact e - mail: krallinger.martin@gmail.com
*
==============================================================
```

This directory contains the BioCreative VII DrugProt track training and development set abstracts and manual annotations.

- **Abstracts**
    - drugprot_training_abstracts.tsv
    - drugprot_development_abstracts.tsv

These files contain plain-text, UTF8-encoded, NFC normalized DrugProt PubMed records in a tab - separated format with the following three columns:

1. Article identifier (PMID, PubMedidentifier)
2. Title of the article
3. Abstract of the article

In total 3500 training set and 750 development set records are provided, where each line in the fails contains a single PMID, title and abstract separated by tabulators.

- **Entity mention annotations**
    - drugprot_training_entities.tsv
    - drugprot_development_entities.tsv

These files contain the manually labeled mention annotations of chemical compounds and genes/proteins (so-called gene and protein-related objects as defined during BioCreative V) generated for the training (and development) set records. Tab-separated format:

1. Article identifier (PMID)
2. Entity or term number (for this record)
3. Type of entity mention (CHEMICAL, GENE-Y, GENE-N)*
4. Start character offset of the entity mention**
5. End character offset of the entity mention**

6. Text string of the entity mention

*CHEMICAL: Chemical entity mention type; GENE-Y: gene/protein mention type that can be normalized or associated to a biological database identifier; GENE-N: gene/protein mention type that cannot be normalized to a database identifier.
***IMPORTANT:** development set GENE entities are not split into GENE-Y and GENE-N. All gene/protein mentions are tagged as GENE. This will be the test set format as well.

****IMPORTANT**: Character offsets are in relation to the complete PubMed record. That is, the string composed of: title, a single blankspace and abstract body. The equivalent Python one-liner to obtain it would be:

```
complete = ‘\t‘.join(title, abstract)
```

Example DrugProt *training* entity mention annotations:

11808879	T12	GENE-Y	1860	1866	KIR6.2
11808879	T13	GENE-N	1993	2016	glutamate dehydrogenase
11808879	T14	GENE-Y	2242	2253	glucokinase
23017395	T1	CHEMICAL	216	223	HMG-CoA
23017395	T2	CHEMICAL	258	261	EPA

Example DrugProt *development* entity mention annotations:

11808879	T12	GENE	1860	1866	KIR6.2
11808879	T13	GENE	1993	2016	glutamate dehydrogenase
11808879	T14	GENE	2242	2253	glucokinase
23017395	T1	CHEMICAL	216	223	HMG-CoA
23017395	T2	CHEMICAL	258	261	EPA

- **DRUGPROT relation annotations**
  - drugprot_training_relations.tsv
  - drugprot_development_relations.tsv

These files contain the detailed chemical-protein relation annotations prepared for the DrugProt training and development set. It consists of tab-separated columns containing:

1. Article identifier (PMID)
2. DrugProt relation
3. Interactor argument 1 (Arg1: followed by the interactor term identifier)
4. Interactor argument 2 (Arg2: followed by the interactor term identifier)

For the DrugProt track, a very granular chemical-protein relation annotation was carried out, with the aim to cover most of the relations that are of importance from the point of view of biochemical and pharmacological/biomedical perspectives.

Example DrugProt entity relation annotations:

12488248	INHIBITOR	Arg1:T1	Arg2:T52

12488248	INHIBITOR	Arg1:T2	Arg2:T52
23220562	ACTIVATOR	Arg1:T12	Arg2:T42
23220562	ACTIVATOR	Arg1:T12	Arg2:T43
23220562	INDIRECT-DOWNREGULATOR	Arg1:T1	Arg2:T14

IMPORTANT: For the test set only the abstracts and the entity mentions will be provided. Participating teams have to return the automatically predicted DrugProt **relations** in the same format as provided for the training set predictions.