```
========================================================
*
*                    BioCreative VI
*
*        Text mining chemical-protein interactions (CHEMPROT) track
*
*
*        Gold Standard CHEMPROT test set - version 1.0 - November 21th
*
*
*   URL: http://www.biocreative.org/tasks/biocreative-vi/track-5/
*
*
*          contact e-mail: krallinger.martin@gmail.com
*
========================================================
```

This directory contains the BioCreative VI CHEMPROT track Gold Standard test set, including the set abstracts, the manual annotations of entity mentions and the manually annotated ChemProt relations.

Important: Do revise the ChemProt Sample set for additional details on the used annotation guidelines and example predictions/format. It is available at:

http://www.biocreative.org/media/store/files/2017/chemprot_sample.zip

1. Gold Standard test set abstracts

- File: *chemprot_test_abstracts_gs.tsv*

This file contains the plain-text UTF8-encoded CHEMPROT Gold Standard test set PubMed records. These are distributed in a tab-separated format with the following three columns:

1- Article identifier (PMID, PubMed identifier)
2- Title of the article
3- Abstract of the article

In total 800 PubMed records are included in the ChemProt Gold Standard test set.

2. Entity mention annotations

- File: *chemprot_test_entities_gs.tsv*

This file contains the manually labeled mention annotations of chemical compounds and genes/proteins (so-called gene and protein related objects –

GPRO as defined during BioCreative V) generated for the Gold Standard test set records.

This file consists of tab-separated fields containing:

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. (See ChemProt sample set for additional details).

Example CHEMPROT entity mention annotations:

10076535	T42	CHEMICAL	943	951	androgen
10076535	T43	CHEMICAL	1004	1012	androgen
10076535	T44	CHEMICAL	0	8	Androgen
10076535	T45	CHEMICAL	32	54	estramustine phosphate
10076535	T46	CHEMICAL	56	59	EMP
10076535	T47	CHEMICAL	98	106	androgen
10076535	T48	GENE-Y	1220	1237	androgen receptor
10076535	T49	GENE-Y	1827	1852	prostate-specific antigen
10076535	T4	CHEMICAL	1178	1200	estramustine phosphate
10076535	T50	GENE-Y	1854	1857	PSA
10076535	T51	GENE-Y	1874	1876	AR
10076535	T52	GENE-Y	1970	1987	androgen receptor

3. CHEMPROT detailed relation annotations

- File: *chemprot_test_relations_gs.tsv*

This file contains the detailed chemical-protein relation annotations prepared for the CHEMPROT Gold Standard test set. It consists of tab-separated columns containing:

1- Article identifier (PMID)
2- Chemical-Protein relation (CPR) group*
3- Evaluation type (Y: group evaluated, N: group not evaluated – extra annotation).
4- CHEMPROT relation (CPR)
5- interactor argument 1 (Arg1: followed by the interactor term identifier)
6- interactor argument 2 (Arg2: followed by the interactor term identifier)

For the CHEMPROT 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 perspective.

Nevertheless, to simplify the CHEMPROT track, and to focus mainly on a subset of key relevant relation types, all the annotated CHEMPROT relations (CPRs) were grouped into 10 semantically related classes that do share some underlying biological properties.

Those groups were labeled as [CPR:1, CPR:2, … CPR:10] ; and are detailed in the table below:

Group	*Eval.*	*CHEMPROT relations belonging to this group*
CPR:1	N	PART_OF
CPR:2	N	REGULATOR|DIRECT_REGULATOR|INDIRECT_REGULATOR
CPR:3	Y	UPREGULATOR|ACTIVATOR|INDIRECT_UPREGULATOR
CPR:4	Y	DOWNREGULATOR|INHIBITOR|INDIRECT_DOWNREGULATOR
CPR:5	Y	AGONIST|AGONIST-ACTIVATOR|AGONIST-INHIBITOR
CPR:6	Y	ANTAGONIST
CPR:7	N	MODULATOR|MODULATOR-ACTIVATOR|MODULATOR-INHIBITOR
CPR:8	N	COFACTOR
CPR:9	Y	SUBSTRATE|PRODUCT_OF|SUBSTRATE_PRODUCT_OF
CPR:10	N	NOT

Important: For evaluation purposes only five groups labeled with 'Y' will be used, that is: **CPR:3, CPR:4, CPR:5, CPR:6, CPR:9.**

Example CHEMPROT entity relation annotations:

```
0076535     CPR:2  N    DIRECT-REGULATOR     Arg1:T23     Arg2:T55
10076535     CPR:2  N    DIRECT-REGULATOR     Arg1:T2 Arg2:T48
10076535     CPR:2  N    DIRECT-REGULATOR     Arg1:T3 Arg2:T48
10076535     CPR:2  N    DIRECT-REGULATOR     Arg1:T4 Arg2:T48
10076535     CPR:3  Y    INDIRECT-UPREGULATOR   Arg1:T23     Arg2:T56
10076535     CPR:4  Y    INDIRECT-DOWNREGULATOR  Arg1:T18     Arg2:T49
10076535     CPR:4  Y    INDIRECT-DOWNREGULATOR  Arg1:T18     Arg2:T50
10076535     CPR:4  Y    INDIRECT-DOWNREGULATOR  Arg1:T18     Arg2:T51
10076535     CPR:4  Y    INDIRECT-DOWNREGULATOR  Arg1:T19     Arg2:T49
10076535     CPR:4  Y    INDIRECT-DOWNREGULATOR  Arg1:T19     Arg2:T50
10076535     CPR:4  Y    INDIRECT-DOWNREGULATOR  Arg1:T19     Arg2:T51
10076535     CPR:4  Y    INDIRECT-DOWNREGULATOR  Arg1:T22     Arg2:T54
10076535     CPR:4  Y    INHIBITOR    Arg1:T18     Arg2:T52
10076535     CPR:4  Y    INHIBITOR    Arg1:T19     Arg2:T52
10076535     CPR:4  Y    INHIBITOR    Arg1:T21     Arg2:T53
10076535     CPR:5  Y    AGONIST Arg1:T24     Arg2:T57
10076535     CPR:6  Y    ANTAGONIST    Arg1:T26     Arg2:T58
10076535     CPR:6  Y    ANTAGONIST    Arg1:T27     Arg2:T58
```

4. CHEMPROT task Gold Standard data

The CHEMPROT task requires the correct recognition of relations between chemicals and proteins. Participants have to return pairs of entities (one corresponding to a chemical entity and another to a gene/protein) together with the corresponding CPR group of the detected relation.

Please notice that:
1. Only relations between a chemical and a genes/protein were allowed. Relations between a chemical and another chemical or between a genes/protein and another gene/protein were not allowed.
2. Only relations of the following classes were considered for evaluation purposes: CPR:3, CPR:4, CPR:5, CPR:6, CPR:9.
3. Participants were allowed to return for a given entity pair multiple relation groups.

- **File: *chemprot_test_gold_standard.tsv***

This file contains the CHEMPROT Gold Standard annotations prepared for the Gold Standard test set. It corresponds essentially to a subset of the relation annotation file.

It consists of tab-separated columns containing:

1- Article identifier (PMID)
2- Manually annotated Chemical-Protein relation (CPR) group*
3- interactor argument 1 (Arg1: followed by the interactor term identifier)
4- interactor argument 2 (Arg2: followed by the interactor term identifier)

An example illustrating the format of the CHEMPROT Gold Standard annotations is shown below:

```
10076535    CPR:3   Arg1:T23    Arg2:T56
10076535    CPR:4   Arg1:T18    Arg2:T49
10076535    CPR:4   Arg1:T18    Arg2:T50
10076535    CPR:4   Arg1:T18    Arg2:T51
10076535    CPR:4   Arg1:T18    Arg2:T52
10076535    CPR:4   Arg1:T19    Arg2:T49
10076535    CPR:4   Arg1:T19    Arg2:T50
```

5. CHEMPROT Track BioCreative VI workshop proceedings

- The overall results of the ChemProt track together with short technical papers summarizing the techniques used for this track by each participating team can be found in the BioCreative VI workshop proceedings, available at:

http://www.biocreative.org/media/store/files/2017/ProceedingsBCVI_v2.pdf

- For a more general background review article covering both the recognition of chemical entities, genes as well as the relation extraction of chemical entities with other entities including genes and proteins, please refer to:

- Krallinger, M., Rabal, O., Lourenço, A., Oyarzabal, J., & Valencia, A. (2017). Information Retrieval and Text Mining Technologies for Chemistry. *Chemical Reviews*, 2017, 117 (12), pp 7673–7761

  URL: http://pubs.acs.org/doi/abs/10.1021/acs.chemrev.6b00851

Note that a Special issue covering all BioCreative VI tracks, including the CHEMPROT task , will be published in the Journal “*Database*: *The Journal of Biological Databases and Curation”.* Outlinks and details related to the Special issue will be posted on the biocreative.org webpage.