==========
*
* BioCreative VI
*
* Text mining chemical-protein interactions (CHEMPROT) track
*
*
* Development set - version 1.0 - September 8th
*
*
* 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 development set abstracts and manual annotations.

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. Development set abstracts

- File: *chemprot_development_abstracts.tsv*

This file contains plain-text, UTF8-encoded CHEMPROT development set PubMed record 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 612 PubMed records are included in the ChemProt development set. In this file, each line contains a single PMID, title and abstract separated by tabulators.

Note that the test set abstracts will be provided in the same format!

2. Entity mention annotations

- File: *chemprot_development_entities.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 development 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:

10939639	T10	CHEMICAL	931	940	menadione
10939639	T11	CHEMICAL	1033	1038	Sch B
10939639	T12	CHEMICAL	0	13	Schisandrin B
10939639	T13	CHEMICAL	31	40	menadione
10939639	T14	GENE-N	376	400	alanine aminotransferase
10939639	T15	GENE-Y	649	662	DT-diaphorase
10939639	T16	GENE-Y	664	667	DTD
10939639	T17	GENE-Y	847	850	DTD
10939639	T18	GENE-Y	877	880	DTD
10939639	T19	GENE-Y	1078	1081	DTD
10939639	T1	CHEMICAL	1152	1161	menadione
10939639	T20	GENE-Y	77	90	DT-diaphorase
10939639	T2	CHEMICAL	288	297	menadione
10939639	T3	CHEMICAL	123	136	schisandrin B
10939639	T4	CHEMICAL	428	443	malondialdehyde
10939639	T5	CHEMICAL	138	143	Sch B
10939639	T6	CHEMICAL	480	489	menadione
10939639	T7	CHEMICAL	601	606	Sch B
10939639	T8	CHEMICAL	698	703	Sch B
10939639	T9	CHEMICAL	747	752	Sch B

Important: For the test set only the abstracts and the entity mentions will be released during the prediction phase. Participating teams will asked to submit the automatically predicted ChemProt-relations in the same format as provided for the sample set predictions. This implies that only analogous files to the abstract and entity annotations will by availably during the prediction phase (*chemprot_test_abstracts.tsv* and *chemprot_test_entities.tsv*).

3. CHEMPROT detailed relation annotations

- File: *chemprot_development_relations.tsv*

This file contains the detailed chemical-protein relation annotations prepared for the CHEMPROT development 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:

10939639	CPR:3	Y	ACTIVATOR	Arg1:T11	Arg2:T19
10939639	CPR:3	Y	ACTIVATOR	Arg1:T12	Arg2:T20
10939639	CPR:3	Y	ACTIVATOR	Arg1:T9	Arg2:T17
10939639	CPR:4	Y	INHIBITOR	Arg1:T2	Arg2:T14

4. CHEMPROT task Gold Standard data and predictions

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

- **File: *chemprot_development_gold_standard.tsv***

This file contains the CHEMPROT Gold Standard annotations prepared for the development 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)

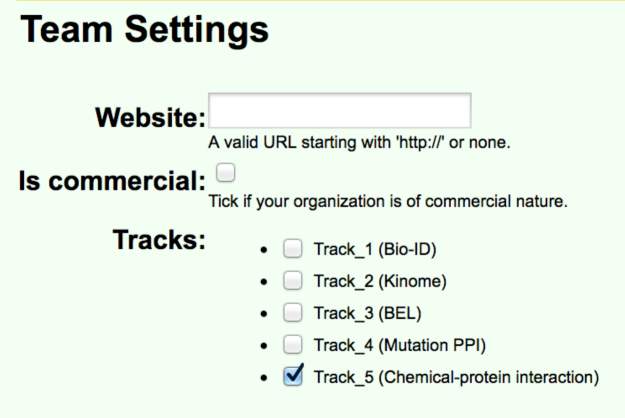
An example illustrating the format of the CHEMPROT Gold Standard annotations is shown below:

10939639	CPR:3	Arg1:T11	Arg2:T19
10939639	CPR:3	Arg1:T12	Arg2:T20
10939639	CPR:3	Arg1:T9	Arg2:T17
10939639	CPR:4	Arg1:T2	Arg2:T14

5. CHEMPROT team registration

In order to participate as a team, you need to register for Track 5 at:

http://www.biocreative.org/events/biocreative-vi/team/



The BioCreative mailing list offers the possibility to discuss-task and workshop related aspects:

https://sourceforge.net/projects/biocreative/lists/biocreative-participant