==========
*
* BioCreative VI
*
* Text mining chemical-protein interactions (CHEMPROT) track
*
*
* Training set - version 1.0 - August 30th
*
*
* 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 training set abstracts and manual annotations.

Important: Do revise the ChemProt Training 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_training.zip

1. Training set abstracts

- File: *chemprot_training_abstracts.tsv*

This file contains plain-text, UTF8-encoded CHEMPROT training 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 1020 records are provided in this training set, where each line contains a single PMID, title and abstract separated by tabulators.

Note that the training, development and test set abstracts will be provided in the same format.

Important: For the test set only the abstracts and the entity mentions will be provided. Participating teams have to return the automatically predicted ChemProt-relations in the same format as provided for the sample set predictions.

2. Entity mention annotations

- File: *chemprot_training_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 training 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:

11319232	T1	CHEMICAL	242	251	acyl-CoAs
11319232	T2	CHEMICAL	1193	1201	triacsin
11319232	T3	CHEMICAL	1441	1448	sucrose
11319232	T4	CHEMICAL	1637	1652	triacylglycerol
11319232	T5	CHEMICAL	1702	1711	acyl-CoAs
11319232	T6	CHEMICAL	176	184	acyl-CoA
11319232	T7	CHEMICAL	790	806	N-ethylmaleimide
11319232	T8	CHEMICAL	898	910	Troglitazone
11319232	T9	CHEMICAL	1012	1022	Triacsin C
11319232	T10	CHEMICAL	0	8	Acyl-CoA
11319232	T11	GENE-N	1212	1215	ACS
11319232	T12	GENE-Y	1244	1248	ACS1
11319232	T13	GENE-Y	1250	1254	ACS4

3. CHEMPROT detailed relation annotations

- File: *chemprot_training_relations.tsv*

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

```
10047461    CPR:3  Y    ACTIVATOR     Arg1:T13      Arg2:T57
10047461    CPR:3  Y    ACTIVATOR     Arg1:T7 Arg2:T39
10047461    CPR:3  Y    ACTIVATOR     Arg1:T7 Arg2:T40
10047461    CPR:3  Y    ACTIVATOR     Arg1:T7 Arg2:T41
10047461    CPR:3  Y    INDIRECT-UPREGULATOR   Arg1:T13      Arg2:T55
10047461    CPR:3  Y    INDIRECT-UPREGULATOR   Arg1:T13      Arg2:T56
10047461    CPR:3  Y    INDIRECT-UPREGULATOR   Arg1:T2 Arg2:T30
10047461    CPR:3  Y    INDIRECT-UPREGULATOR   Arg1:T2 Arg2:T31
10047461    CPR:3  Y    INDIRECT-UPREGULATOR   Arg1:T2 Arg2:T32
10047461    CPR:3  Y    INDIRECT-UPREGULATOR   Arg1:T2 Arg2:T33
10047461    CPR:4  Y    INDIRECT-DOWNREGULATOR Arg1:T13      Arg2:T54
10047461    CPR:4  Y    INDIRECT-DOWNREGULATOR Arg1:T7 Arg2:T38
10047461    CPR:4  Y    INHIBITOR     Arg1:T10      Arg2:T46
10047461    CPR:4  Y    INHIBITOR     Arg1:T16      Arg2:T62
```

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_training_gold_standard.tsv***

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

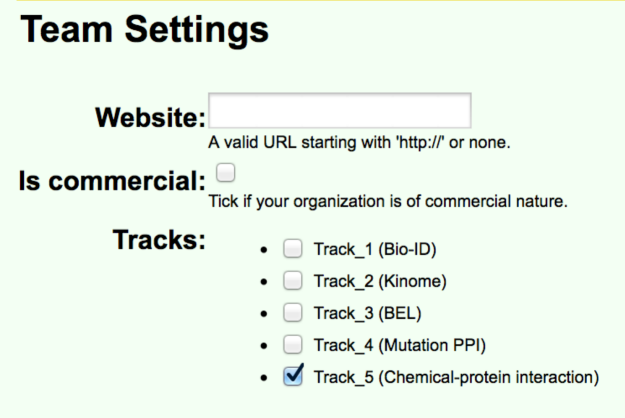
An example illustrating the format of the CHEMPROT Gold Standard annotations is shown below:

```
10047461    CPR:3  Arg1:T13    Arg2:T55
10047461    CPR:3  Arg1:T13    Arg2:T56
10047461    CPR:3  Arg1:T13    Arg2:T57
10047461    CPR:3  Arg1:T2 Arg2:T30
10047461    CPR:3  Arg1:T2 Arg2:T31
10047461    CPR:3  Arg1:T2 Arg2:T32
10047461    CPR:3  Arg1:T2 Arg2:T33
10047461    CPR:3  Arg1:T7 Arg2:T39
10047461    CPR:3  Arg1:T7 Arg2:T40
10047461    CPR:3  Arg1:T7 Arg2:T41
10047461    CPR:4  Arg1:T10    Arg2:T46
10047461    CPR:4  Arg1:T13    Arg2:T54
```

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