==========
*
* BioCreative VI
*
* Text mining chemical-protein interactions (CHEMPROT) track
*
*
* Sample set - version 1.0 - July 31st
*
*
* URL: http://www.biocreative.org/tasks/biocreative-vi/track-5/
*
*
* contact e-mil: krallinger.martin@gmail.com
*
==========

This directory contains the BioCreative VI CHEMPROT track sample set data: abstracts, manual annotations of entity mentions, manual annotations of chemical-protein relations, annotation guidelines and example ChemProt relation (CPR) predictions.

A considerable number of approaches have been implemented to detect automatically mentions of chemical compounds and genes/proteins in running text, while far less attempts have been made to recognize automatically relations between them [1]. The aim of the CHEMPROT track is to promote the development of systems able to extract chemical-protein interactions of relevance for precision medicine, drug discovery as well as basic biomedical research.

1. Annotation Guidelines

- File: *guidelines/CHEMPROT_guidelines_v6.pdf*

This document contains the manual data annotation guidelines/rules that are used for labeling CHEMPROT relations. In order to annotate these relations, chemical entities, genes and proteins were previously tagged by hand by domain experts according to the entity mention annotation guidelines summarized below.

- File: *guidelines/GPRO_guidelines.pdf*

This document contains the annotation guidelines used for the manual annotation of gene and protein related objects, which essentially consists of genes and gene product mentions. These guidelines were used previously to prepare the CHEMDNER patents task corpus for BioCreative V and V.5 [2,3,4]

- File: *guidelines/CEM_guidelines.pdf*

This document describes the annotation guidelines used for the manual annotation of chemical entity mentions in PubMed abstracts. These guidelines were used previously to prepare the CHEMDNER task corpus for BioCreative IV [1,5,6].

2. Sample set abstracts

- File: *chemprot_sample_abstracts.tsv*

This file contains plain-text, UTF8-encoded CHEMPROT sample 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 50 records are provided in this sample 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. Essentially the same simple input format was used previously for the CHEMDNER tracks of BioCreative IV, V and V.5 [1,5,6].

3. Entity mention annotations

- File: *chemprot_sample_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 sample 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 (see document *GPRO_guidelines.pdf* description of GPRO entity mention type 1); GENE-N: gene/protein mention type that cannot be normalized to a database identifier (see document *GPRO_guidelines.pdf* description of GPRO entity mention type 2).

Example CHEMPROT entity mention annotations:

23538162	T1	CHEMICAL	1305	1308	Rg1
23538162	T2	CHEMICAL	291	306	Ginsenoside Rg1

23538162	T3	CHEMICAL	308	311	Rg1
23538162	T4	CHEMICAL	549	552	Rg1
23538162	T5	CHEMICAL	581	584	Rg1
23538162	T6	CHEMICAL	730	738	nitrogen
23538162	T7	CHEMICAL	873	878	RU486
23538162	T8	CHEMICAL	898	903	U0126
23538162	T9	CHEMICAL	916	924	estrogen
23538162	T10	CHEMICAL	947	957	ICI 82,780
23538162	T11	CHEMICAL	1002	1005	Rg1
23538162	T12	CHEMICAL	72	87	ginsenoside Rg1
23538162	T13	GENE-Y 1330	1337	Aβ25-35	
23538162	T14	GENE-Y 1391	1393	GR	
23538162	T15	GENE-N 1394	1397	ERK	

4. CHEMPROT detailed relation annotations

- File: *chemprot_sample_relations.tsv*

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

23538162	CPR:4	Y	DOWNREGULATOR	Arg1:T5	Arg2:T19
23538162	CPR:4	Y	INDIRECT-DOWNREGULATOR	Arg1:T5	Arg2:T20
23538162	CPR:6	Y	ANTAGONIST	Arg1:T7	Arg2:T21
23538162	CPR:6	Y	ANTAGONIST	Arg1:T7	Arg2:T22
23538162	CPR:4	Y	INHIBITOR	Arg1:T8	Arg2:T23
23538162	CPR:6	Y	ANTAGONIST	Arg1:T10	Arg2:T24
23538162	CPR:2	N	REGULATOR	Arg1:T11	Arg2:T26
23538162	CPR:2	N	REGULATOR	Arg1:T11	Arg2:T27
23538162	CPR:4	Y	DOWNREGULATOR	Arg1:T11	Arg2:T28
23538162	CPR:2	N	REGULATOR	Arg1:T1	Arg2:T14
23538162	CPR:2	N	REGULATOR	Arg1:T1	Arg2:T15
12453616	CPR:3	Y	INDIRECT-UPREGULATOR	Arg1:T6	Arg2:T17
12453616	CPR:3	Y	INDIRECT-UPREGULATOR	Arg1:T6	Arg2:T18

5. 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 fro 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_sample_gold_standard.tsv*

This file contains the CHEMPROT Gold Standard annotations prepared for the sample set. 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:

23538162	CPR:4	Arg1:T5	Arg2:T19
23538162	CPR:4	Arg1:T5	Arg2:T20
23538162	CPR:6	Arg1:T7	Arg2:T21
23538162	CPR:6	Arg1:T7	Arg2:T22
23538162	CPR:4	Arg1:T8	Arg2:T23
23538162	CPR:6	Arg1:T10	Arg2:T24
23538162	CPR:4	Arg1:T11	Arg2:T28
12453616	CPR:3	Arg1:T6	Arg2:T17
12453616	CPR:3	Arg1:T6	Arg2:T18
12453616	CPR:3	Arg1:T1	Arg2:T10
12453616	CPR:3	Arg1:T1	Arg2:T15

- File: *chemprot_sample_predictions.tsv*

This file contains CHEMPROT task example predictions for the sample set. It consists of tab-separated columns containing:

1- Article identifier (PMID)
2- Predicted 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)

Teams that participate in the CHEMPROT relation track have to return predictions in the same format as the Gold Standard manual annotations for this subtask, corresponding to a plain text file with tab separated columns containing the PubMed identifier (PMID), a single CHEMPROT relation group [CPR:3, CPR:4, CPR:5, CPR:6 or CPR:9], the interactor argument term 1 and the interactor argument term 2.

Please notice that:
1. CHEMPROT interactor terms (columns 3 and 4 of the prediction file) have to be sorted in ascending order according to their corresponding term number. Correct order: Arg1:T10 Arg2:T45 ; Wrong: order: Arg1:T10 Arg2:T5.
2. No duplicate predictions (several times the same prediction) are allowed:

10403635	CPR:3	Arg1:T10	Arg2:T45	
10403635	*CPR:3*	*Arg1:T10*	*Arg2:T45*	*[No duplicates predictions!]*
10403635	CPR:3	Arg1:T8	Arg2:T43	

An example illustrating the format of the CHEMPROT task prediction format is shown below:

10403635	CPR:3	Arg1:T10	Arg2:T45
10403635	CPR:3	Arg1:T8	Arg2:T43
10403635	CPR:4	Arg1:T11	Arg2:T45

10403635	CPR:4	Arg1:T20	Arg2:T40
10403635	CPR:4	Arg1:T20	Arg2:T42
10403635	CPR:4	Arg1:T35	Arg2:T40
10403635	CPR:4	Arg1:T35	Arg2:T42
10403635	CPR:4	Arg1:T55	Arg2:T40
10403635	CPR:4	Arg1:T9	Arg2:T43
10403635	CPR:9	Arg1:T16	Arg2:T40
10403635	CPR:9	Arg1:T16	Arg2:T42
10403635	CPR:9	Arg1:T24	Arg2:T49
10403635	CPR:9	Arg1:T24	Arg2:T50

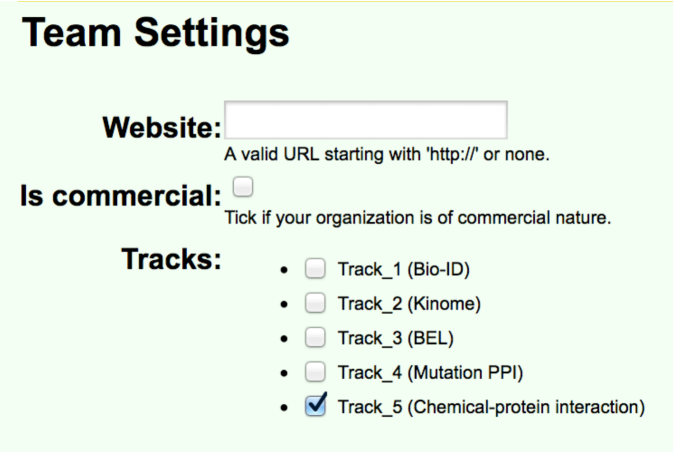
- File: *chemprot_sample_predictions_eval.txt*

This file contains the CHEMPROT task evaluation of the predictions done for the sample set. An evaluation script will be distributed together with the training data. The main evaluation metric for this task will be the micro-averaged F1-score.

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

References

[1] Krallinger, M., Rabal, O., Lourenço, A., et al. (2017). Information Retrieval and Text Mining Technologies for Chemistry. *Chemical Reviews*.
[2] http://www.biocreative.org/tasks/biocreative-v/track-2-chemdner/

[3] Perez, M. P., Rabal, O., Rodriguez, G. P., et al. (2017). Evaluation of chemical and gene/protein entity recognition systems at BioCreative V. 5: the CEMP and GPRO patents tracks. *Proceedings of the BioCreative*, *5*, 3-11.
[4] Krallinger, M., Rabal, O., Lourenço, A., et al. (2015). Overview of the CHEMDNER patents task. In *Proceedings of the fifth BioCreative challenge evaluation workshop* (pp. 63-75).
[5] Krallinger, M., Rabal, O., Leitner, F., et al. (2015). The CHEMDNER corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, *7*(S1), S2.
[6] Krallinger, M., Leitner, F., Rabal, O., et al. (2015). CHEMDNER: The drugs and chemical names extraction challenge. *Journal of cheminformatics*, *7*(S1), S1.