BioCreative VI

Track 5: Text mining chemical-protein interactions [2017-11-21]

Task 5: Text mining chemical-protein interactions (CHEMPROT)

The aim of the CHEMPROT task of BioCreative VI is to promote the development and evaluation of systems that are able to automatically detect in running text (PubMed abstracts) relations between chemical compounds/drug and genes/proteins. We will therefore release a manually annotated corpus, the *CHEMPROT corpus*, where domain experts have exhaustively labeled:(a) all chemical and gene mentions, and (b) all binary relationships between them corresponding to a specific set of biologically relevant relation types (*CHEMPROT relation classes*). 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. Compared to the extraction of protein-protein or gene/chemical-disease relations, the detection of associations between chemical entities, in particular drugs and active pharmaceutical ingredients, with proteins/genes has resulted in a considerably lower number of text mining systems coping with this relation type. Moreover, despite the existence of competitive named entity recognition tools for tagging chemicals and genes/proteins, the retrieval of certain relationships between these two entities using text mining and information extraction approaches has only be attempted by a limited number of systems. In an early work by Craven and Kumlien published in 1999 (2) the automatic detection of interactions between drugs and protein targets from text was already proposed, while Rindflesch et al. published a system called EDGAR (3) that extracted several relation types including drug-gene relations (drugs affecting gene expression) and gene-drug relations (gene/protein affecting drug activity). There is also an increasing interested in the integration of chemical and biomedical data understood as curation of relationships between biological and chemical entities from text and storing such information in form of structured annotation databases. Such databases are of key relevance not only for biological but also for pharmacological and clinical research. A range of different types chemical-protein/gene interactions are of key relevance for biology, including metabolic relations (e.g. substrates, products) inhibition, binding or induction associations.

The ChemProt track aims to address these needs and to promote the development of systems able to extract chemical-protein interactions that might be of relevance for precision medicine as well as for drug discovery and basic biomedical research.

The ChemProt track in BioCreative VI (BC VI) will explore recognition of chemical-protein entity relations from abstracts. To do support this task we will provide a set of manually annotated chemical and protein/gene entity mentions adapting the annotation processes used for the BioCreative V CHEMNDER task together with the manual annotation of the chemical-protein relation types.

Teams participating in this track will be provided with:

- PubMed abstracts

- Manually annotated chemical compound mentions
- Manually annotated gene/protein mentions
- Manually annotated chemical compound-protein relations

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.
4 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.

The entire corpus prepared for this track is available at: **ChemProt corpus**.

Timeline:

1. Sample set release and annotation guidelines, task details: 10th of August
2. Training set release - 30 August: ChemProt evaluation script and Training set
3. Development set release - 8th of September: Development set
4. Test set release- 12th of September: Test set abstracts and entity annotations
5. Test set prediction submission instructions and CHEMPROT-Elsevier Prize info: 27th September (updated)
6. Test set prediction submission due: 7th October
7. Test set evaluation returned to participants: 11th October
8. Short technical systems description paper due 13th October
9. Paper acceptance and review returned 15th October
10. Test set Gold Standard annotations: Test set abstracts, entity annotations and relations

CHEMPROT Test set prediction submission instructions

We have opened the submission system of the test set predictions for the BioCreative VI ChemProt track.
Step 1. biocreative.org team id
In order to upload predictions you need to have registered at a team on the BioCreative website at:
http://www.biocreative.org/events/biocreative-vi/team/
After registration you will have team number (the team number is your official team ID).
If you are not sure what your team id was, just log into the biocreative.org webpage and select the on the page header the field Team Page.

Step 2. Markyt submission account
The submission of test set runs for the ChemProt track will be done through the Markyt system. You need to register your team also at Markyt in order to have access to the submission platform.

Teams can register at:
http://www.markyt.org/mmes/users/register

Step 3. Login into Markyt

Then, you need to log into the Markyt system (click on the red 'Sign In' icon on the top right corner). While you log in select the CHEMPROT track
Login URL:
http://www.markyt.org/mmes/

4. Upload predictions

Once you are logged in you can upload your test set runs for the CHEMPROT track.

Important:

1. You are allowed to upload up to 5 runs per team
2. The prediction file must 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, corresponds to the chemical entity)
4- interactor argument 2 (Arg2: followed by the interactor term identifier, corresponds to the gene entity)

* Has to be one of these groups: CPR:3, CPR:4, CPR:5, CPR:6 or CPR:9.

Step 4. Final Submission

After uploading the predictions you have to select the option: 'Final task submission?' In case you want to consider the prediction to be evaluated by the track organizers (to be included in your final submissions).
You have to upload each of your up to five runs separately, file by file.

Elsevier Prize

The Top scoring team will be awared a prize of 500 euro sponsored by Elsevier

Data sets:

We will provide a training consisting of a collection of manually annotated chemical and gene/protein mentions as well as their relation types.
The input files for the CHEMPROT track will be plain-text, UTF8-encoded PubMed records 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

For the CHEMPROT track all entity mention labels will be provided both for the training/development data sets as well as for the test set. This implies that participants will only need to focus on the relation detection aspect for this task.

CHEMPROT entity mention annotation files do contain manually labeled mention annotations of chemical compounds and genes/proteins (so-called gene and protein related objects – GPRO as defined during BioCreative V). Such files consist of tab-separated fields containing: with the following three columns:

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

* Note that: 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

CHEMPROT relation annotations will be distributed as a file that contains the detailed chemical-protein relation annotations prepared for the CHEMPROT track. 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)

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
```

Evaluation:

We will provide another data set as a blind test set for which we will calculate precision, recall and F-measure.

This CHEMPROT team prediction file will consist 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 its format, 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.

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

CHEMPROT team registration

In order to participate as a team, you need to **register** for Track 5.

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

BioCreative VI workshop and ChemProt track

Note that all teams sending a test set prediction submission will be invited to send a workshop proceedings paper on their system (systems description paper), similarly to previous BioCreative events. It is not mandatory to assist to the workshop, but we encourage very much attendance. You can send both submissions and the workshop proceedings paper without attending the workshop.
We will invite all the track 5 participants to give a short flash presentation of their system at the workshop. Additionally, top scoring teams will have the opportunity to give a longer presentation at the workshop.
We will invite selected works for full publication in the Database Journal Special Issue for BioCreative. Invitation to the special issue will consider multiple aspects such as performance, novelty of the system, availability of the underlying system (software/web-service) as well as the workshop presentation.
All workshop info is in http://www.biocreative.org/news/biocreative-vi/workshop/
1-Registration is open and the early fee applies before September 10
2-Funds are available for US participants for the amount up to $700 to participate in the BioCreative workshop. To apply complete the application by September 1st (application in the workshop info website). Women, under-represented minorities, students, and post-doctoral fellows are encouraged to apply.

The task overview and result summary presentation can be found **here**

Paper submission

Submit a paper (max. 4 pages) describing your system and track results for your work to be included in the conference Proceedings, be considered for a talk in the workshop and be considered for publication in Database virtual issue.

- Deadline for paper submissions is October 13th
- Instructions and paper template
- Submission link: https://easychair.org/conferences/?conf=bc6

CHEMPROT Organizers

- Martin Krallinger, Spanish National Cancer Research Centre, Spain

- Analia Lourenço, University of Vigo, Spain
- Obdulia Rabal, Center for Applied Medical Research (CIMA), University of Navarra, Spain
- Julen Oyarzabal, , Center for Applied Medical Research (CIMA), University of Navarra, Spain
- Georgios Tsatsaronis, Content and innovation, Elsevier BV
- Saber A. Akhondi, Content and innovation, Elsevier BV
- Alfonso Valencia, Barcelona Supercomputing Center, Spain

CHEMPROT contact/info

- Martin Krallinger: krallinger.martin@gmail.com

References

[1] Krallinger, M., Rabal, O., Lourenço, A., et al. (2017). Information Retrieval and Text Mining Technologies for Chemistry. Chemical Reviews.

[2] Craven, M., & Kumlien, J. (1999, August). Constructing biological knowledge bases by extracting information from text sources. In ISMB (Vol. 1999, pp. 77-86).

[3] Rindflesch, T. C., Tanabe, L., Weinstein, J. N., & Hunter, L. (2000). EDGAR: extraction of drugs, genes and relations from the biomedical literature. In Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing (p. 517). NIH Public Access.

Back to top

Downloads

- ChemProt overview talk at BioCreative VI workshop
- CHEMPROT test set Gold Standard annotations
- CHEMPROT test set abstracts and entities
- CHEMPROT development set
- CHEMPROT training set
- ChemProt evaluation kit
- CHEMPROT sample set and description