Overview of the BioCreative VI chemical-protein interaction Track

Martin Krallinger[1§], Obdulia Rabal[2], Saber A. Akhondi[3], Martín Pérez Pérez[4], Jesús Santamaría[1], Gael Pérez Rodríguez[4], Georgios Tsatsaronis[3], Ander Intxaurrondo[1], José Antonio López[1] Umesh Nandal[3], Erin Van Buel[3], Akileshwari Chandrasekhar[3], Marleen Rodenburg[3], Astrid Laegreid[5,] Marius Doornenbal[3], Julen Oyarzabal[2], Analia Lourenço[4], Alfonso Valencia[6]

[1]Spanish National Cancer Research Centre, Madrid, Spain
[2]Center for Applied Medical Research (CIMA), University of Navarra, Spain
[3]Content and Innovation, Elsevier BV, Netherlands
[4]University of Vigo, Spain
[5]Norwegian University of Science and Technology (NTNU), Norway
[6]Barcelona Supercomputing Center, Spain

§ Corresponding author

Abstract—**Despite the considerable number of available systems that recognize automatically mentions of genes/proteins and chemicals in text, only a limited number of attempts were made so far to extract interactions between them. Most biomedical relation extraction systems focus on the extraction of protein-protein or gene/chemical-disease relations. The detection of interactions between drugs and proteins/genes is of key relevance for pharmacological and clinical research, playing an important role for drug discovery, understanding of molecular mechanism of adverse drug reactions, describing drug metabolism or drawing regulatory networks of importance for systems pharmacology. The BioCreative VI - ChemProt track represents the first attempt to promote the development of systems for extracting chemical-protein interactions (CPIs), of relevance for precision medicine as well as for drug discovery and basic biomedical research. The novel ChemProt corpus consists of text exhaustively annotated by hand with mentions of chemical compounds/drugs and genes/proteins, as well as 22 different types of compound-protein relations. To focus on a subset of important relations, 5 relation classes were used for evaluation purposes, including agonist, antagonist, inhibitor, activator and substrate/product relations. A total of 13 participating teams returned 45 runs for this track. Despite the biological complexity of the considered relation types, top-scoring teams could obtain an F-measure across relation classes of 64.10%. Performance varied depending on the relation class: for the antagonist relation class the best team obtained an F-measure of 72.56% (precision of 80.75%, recall of 65.87%) while for inhibition/down-regulation the best value was of 71.48% (with a precision of 76.51% and a recall of 67.07%).**

Keywords—text mining; chemical compound; drug; protein; drug target; agonist; antagonist, inhibitor; activator; gene regulation; chemical-protein relation

I. INTRODUCTION

A growing amount of scientific articles, medicinal chemistry patents and other biomedical documents provide descriptions of interactions between chemical compounds and gene products. Compared to the extraction of protein-protein (1) or gene/chemical-disease relations (2), the detection of associations between chemical entities (e.g. drugs or active pharmaceutical ingredients) and proteins/genes is an underexplored biomedical text mining research area.

There is an increasing interest in the integration of chemical and biomedical data understood as the 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. Certain types of chemical-protein/gene interactions are of key relevance for biomedicine, including metabolic relations (e.g. substrates, products), antagonist, agonist, inhibitor or activator associations.

Despite the existence of competitive named entity recognition tools for tagging chemicals and genes/proteins, the retrieval of certain relationships between these two types of entities using text mining and information extraction approaches has only be attempted by a limited number of systems. A comprehensive review of previously published text mining systems for different types of chemical-protein relations is summarized in Krallinger et al. 2017 (3). For instance, Rindflesch et al. introduced an early system called EDGAR (4) that extracted drug-gene relations (drugs altering gene expression) and gene-drug relations (gene/protein altering drug activity) by exploiting syntactic information and relational vocabulary. The drug-target interaction resource SuperTarget (5) applied the EBIMed tool to select text passages potentially describing drug-target associations (6). Tari et al. applied syntactic dependencies to recover gene-drug relations (7) while Czarnecki et al. (8) published a pattern-matching and rule-based approach for detecting metabolic reaction relations. The LimTox online server incorporates both pattern matching and machine learning techniques to retrieve gene expression induction, inhibition and metabolism relations between cytochromes P-450 (CYPs) and drugs (9). The relation patterns used by LimTox, including also protein-protein interactions

patterns, were released as part of the LimTox resource collection. The reasons for missing publicly available text mining resources for chemical-protein interaction extraction lay mainly in: (a) the difficulty to define chemical-protein interactions in the first place, despite some attempts to define the underlying relation concepts, (b) the lack of large corpora of manually curated text bound annotations of mentions of chemical entities and genes/proteins and their interactions, done by domain experts and (c) the lack of publicly shared annotation guidelines and rules for labeling mentions of chemicals, genes and proteins relationships.

Due to the critical importance of CPI we have posed a track specially devoted to this kind of biological relation at the BioCreative VI community evaluation effort.

II. THE CHEMPROT TRACK AND CORPUS

The ChemProt track aimed to promote the development of systems able to extract chemical-protein interactions that are relevant for precision medicine, drug discovery and basic biomedical research. To carry out the ChemProt track it was necessary to create a novel manually annotated corpus that annotates exhaustively all the mentions of the underlying entity types, namely chemical compounds and drugs as well as genes and proteins. In order to make this feasible in practice, and to focus on types of text that are easily accessible, we restricted the annotation process to PubMed abstracts essentially published between 2005 and 2014. Domain experts with experience in text annotation and database curation tasks annotated by hand all abstracts. The manual labeling of chemicals and genes was done separately to avoid cross-influence during the annotation process. The labeling of mentions of chemical entities, genes and proteins was done following the annotation process and guidelines previously used for several BioCreative tasks (10–12). Gene/protein mentions were manually normalized to their corresponding database identifiers whenever possible and classified as either normalizable to databases (tag: GENE-Y) or non normalizable mentions (GENE-N). Participanting teams were only provided with this classification of gene mentions and not the actual database identifier to avoid usage of external knowledgebases for producing their predictions.

After the completion of chemical and gene annotations, all mention annotations were merged in order to carry out the relation annotation process. To enable the annotation of chemical-protein interactions, the ChemProt track organizers constructed very granular relation annotation rules described in a 33 pages annotation guidelines document. These guidelines were refined during an iterative process based on the annotation of sample documents.

The guidelines provided the basic details of the chemical-protein interaction annotation task and the conventions that had to be followed during the corpus construction process. They incorporated suggestions made by curators as well as observations of annotation inconsistencies encountered when comparing results from different human curators.

In brief, the annotated ChemProt interactions included direct interactions (when a physical contact exists between a chemical and a gene/protein) as well as indirect regulatory interactions that alter either the function or the quantity of the gene/gene product. The aim of the iterative manual annotation cycles was to improve the quality and consistency of the guidelines, in order to make them more intuitive and easier to follow. During the preparation of the guidelines some rules had to be reformulated to make them more explicit and additional rules were added wherever necessary to better cover the practical annotation scenario and for being more complete.

The manual annotation task basically consisted of labeling or marking up manually through a customized web-interface the mentions of CHEMPROT interactions in text. Figure 1 summarizes the ChemProt relation types included in the annotation guidelines and ChemProt corpus.

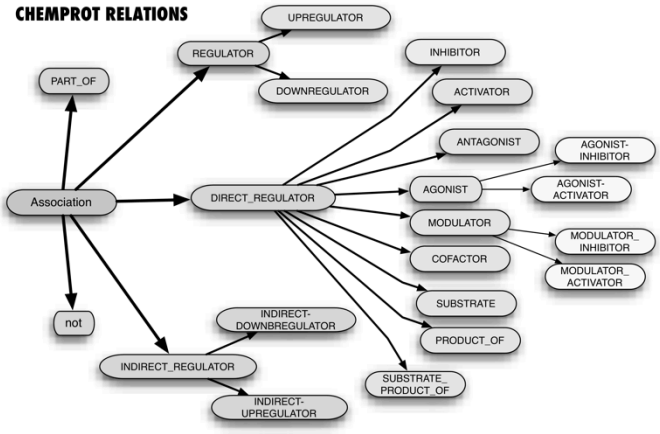

Fig. 1. Overview of the ChemProt track relation types and classification.

The annotation carried out for the ChemProt track was exhaustive for the types of interactions previously specified. This implied that mentions of other relationships between chemicals and genes (e.g. phenotypic and biological responses) were not annotated. The ChemProt relations are directed in the sense that only relations of "what a chemical does to a gene/protein" (chemical → gene/protein direction) were annotated, and not vice versa.

In order to establish a homogenous nomenclature and avoid redundant class definitions, we reviewed several chemical repositories that incorporate chemical – biology information. We thus inspected DrugBank (13), the Therapeutic Targets Database (TTD) (14) and ChEMBL (15), assay normalization ontologies (BAO) (16) and previously existing formalizations for the annotation of relationships: the Biological Expression Language (BEL) (17,18), curation guidelines for transcription regulation interactions (DNA-binding transcription factor – target gene interaction) and SIGNOR, a database of causal relationships between biological entities (19).

Each of these resources inspired the definition of the subclasses DIRECT REGULATOR (e.g. DrugBank, ChEMBL, BAO and SIGNOR) and the INDIRECT REGULATOR (e.g. BEL, curation guidelines for transcription

regulation interactions and SIGNOR). For example, DrugBank relationships for drugs included a total of 22 definitions, some of them overlapping with CHEMPROT subclasses (e.g. "Inhibitor", "Antagonist", "Agonist",…), some of them being regarded as highly specific for the purpose of this task (e.g. "intercalation", "cross-linking/alkylation") or referring to biological roles (e.g. "Antibody", "Incorporation into and Destabilization") and others, partially overlapping between them (e.g. "Binder" and "Ligand"), that were merged into a single class. Concerning indirect regulatory aspects, the five classes of casual relationships between a subject and an object term defined by BEL ("decreases", "directlyDecreases", "increases", "directlyIncreases" and "causesNoChange") were highly inspiring. Subclasses definitions of pharmacological modes of action were defined according to the UPHAR/BPS Guide to Pharmacology in 2016 (20,21).

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 Figure 2. For evaluation purposes only five groups labeled with 'Y' were used, that is: CPR:3, CPR:4, CPR:5, CPR:6, CPR:9.

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

Fig. 2. Overview of the ChemProt track relation groups (CPRs).

Four different randomly sampled data subsets (Table 1) were released for the ChemProt track:

- An initial *sample set* of 50 abstracts to illustrate the type of chemical-protein relations and annotations that were used for the track, together with annotation guidelines for labeling the chemicals, genes/proteins and chemical-protein interaction relations. The sample set also contains an illustrative prediction example in the chemprot track format.
- A training set of 1,020 abstracts annotated exhaustively with chemicals, genes/proteins and chemical-gene interactions.
- A development set of 612 abstracts annotated exhaustively with chemicals, genes/proteins and chemical-gene interactions.
- A test set of 800 abstracts annotated exhaustively with chemical, genes/proteins and blinded chemical-gene interactions. Also, 2,599 additional abstracts were included in order to avoid manual corrections of team submissions and assure that systems could process larger datasets. Teams had to return results for the entire set of 3,399 records.

TABLE I. OVERVIEW OF CHEMPROT DATA SETS

Dataset	**Annotations**			
	Chemicals	***Genes***	***All CPI * relations***	***Evaluated CPI***
Sample	683	606	339	239
Training	13,017	12,735	6,437	4,157
Development	8,004	7,563	3,558	2,416
Test	10,810	10,018	5,744	3,458

* Chemical-protein interaction (CPI)

In practice, chemical-protein relation annotations prepared for the ChemProt track consisted of simple 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 (chemical entity followed by the interactor term identifier)
6. Interactor argument 2 (gene/protein entity followed by the interactor term identifier)

Track participants had to return for the collection of test set document identifiers the detected pairs of entities (one corresponding to a chemical entity and another to a gene/protein) together with the corresponding CPR group of the predicted relation. Only relations between a chemical and a gene/protein were allowed. Relations between a chemical and another chemical or between a gene/protein and another gene/protein were not permitted. Moreover participants were allowed to return for a given entity pair multiple relation groups. A total of 5 runs were accepted per team.

In addition to the ChemProt track data sets, a special evaluation script was available at the track webpage (22). For evaluation purposes we considered the micro-averaged precision, recall and, in particular, balanced micro F1-score.

III. RESULTS

The Markyt evaluation platform was used to register and upload the team submissions (23). A total of 13 teams returned overall 45 submissions for the ChemProt track. A detailed description of the underlying strategy used by each of the

participating teams can be found in the systems description papers published in the BioCreative VI workshop proceedings. Table 2 shows a summary of the participating teams.

TABLE II. OVERVIEW OF CHEMPROT PARTICIPATING TEAMS

Team Id	Details		
	Team Leader	*Institution*	*Nr. runs*
374	Sérgio Matos	Universidade de Aveiro	5
379	Sijia Liu	Mayo Clinic	4
394	Neha Warikoo	Academia Sinica	5
397	Atakan Yüksel	Boğaziçi University	1
403	Peter Corbett	Royal Society of Chemistry	1
404	Ignacio Tripodi	University of Colorado	5
417	Farrokh Mehryary	University of Turku	5
421	Cong Sun	DaLian University of Technology	1
424	Sangrak Lim	Korea University	2
427	Wei Wang	National University of Defense Technology	5
430	Yifan Peng	NCBI, NLM, NIH	5
432	Pat Verga	UMass Amherst	4
433	Pei-Yau Lung	Florida State University	2

In addition to evaluate all the team predictions, we prepared two simple baseline predictions. These consisted of predictions of all CPR classes for all the co-occurrence of chemical entities and genes, either within the entire abstracts, or within individual sentences. These baseline predictions can be considered as a sort of upper boundaries in terms of recall and lower boundaries in terms of precision when using simple entity co-mention. Table 3 illustrates the results obtained for each of the evaluated runs of all participating teams. Taking into account the complexity of the chemical-protein relations examined for this task, the results obtained by participating teams are very promising. The best F-measure, across all CPI relations, was reached by team 430 (run 5) with a score of 64.10%. Overall, the results were better in terms of precision when compared to recall, in particular for the top performing teams. Team 430 could obtain particularly high precision values for all runs, reaching up to 74.37% for run 1. The best recall values were obtained by team 403 (run 1) with a score of 67.84% followed by run 5 of team 427 (66.63%).

TABLE III. LIST OF CHEMPROT RESULTS PER TEAM AND RUN

Team Id	Run	Precision	Recall	F-Score
Co-mention	Abstract	0.0050	1.0000	0.0099
Co-mention	Sentence	0.0437	0.9803	0.0837
TEAM_374	RUN_1	0.6419	0.2577	0.3677
TEAM_374	RUN_2	0.5156	0.4670	0.4901
TEAM_374	RUN_3	0.5919	0.2403	0.3418
TEAM_374	RUN_4	0.4024	0.4193	0.4107
TEAM_374	RUN_5	0.5738	0.4722	0.5181
TEAM_379	RUN_1	0.4773	0.4375	0.4565
TEAM_379	RUN_2	0.4849	0.4913	0.4881
TEAM_379	RUN_4	0.5072	0.4306	0.4657
TEAM_379	RUN_5	0.5301	0.4639	0.4948
TEAM_394	RUN_1	0.2446	0.3407	0.2847
TEAM_394	RUN_2	0.2563	0.3456	0.2943
TEAM_394	RUN_3	0.2932	0.3271	0.3092
TEAM_394	RUN_4	0.0729	0.0150	0.0249
TEAM_394	RUN_5	0.2587	0.3456	0.2959
TEAM_397	RUN_1	0.6057	0.1102	0.1864
TEAM_403	RUN_1	0.5610	**0.6784**	0.6141
TEAM_404	RUN_1	0.3460	0.3913	0.3673
TEAM_404	RUN_2	0.3387	0.4078	0.3700
TEAM_404	RUN_3	0.3305	0.1666	0.2215
TEAM_404	RUN_4	0.3307	0.3641	0.3466
TEAM_404	RUN_5	0.3058	0.3603	0.3309
TEAM_417	RUN_1	0.6373	0.4462	0.5249
TEAM_417	RUN_2	0.6337	0.4387	0.5185
TEAM_417	RUN_3	0.6608	0.5662	0.6099
TEAM_417	RUN_4	0.6105	0.6006	0.6055
TEAM_417	RUN_5	0.6088	0.5989	0.6038
TEAM_421	RUN_1	0.1618	0.3409	0.2195
TEAM_424	RUN_1	0.6760	0.5159	0.5852
TEAM_424	RUN_2	0.6704	0.5194	0.5853
TEAM_427	RUN_1	0.2496	0.6417	0.3594
TEAM_427	RUN_2	0.2535	0.6478	0.3643
TEAM_427	RUN_3	0.2634	0.6622	0.3769
TEAM_427	RUN_4	0.2674	0.6602	0.3806
TEAM_427	RUN_5	0.2696	0.6663	0.3839
TEAM_430	RUN_1	**0.7437**	0.5529	0.6343
TEAM_430	RUN_2	0.7283	0.5503	0.6269
TEAM_430	RUN_3	0.7426	0.5382	0.6241
TEAM_430	RUN_4	0.7311	0.5685	0.6397
TEAM_430	RUN_5	0.7266	0.5735	**0.6410**
TEAM_432	RUN_1	0.2211	0.2024	0.2114
TEAM_432	RUN_2	0.5491	0.2021	0.2955
TEAM_432	RUN_3	0.4073	0.4783	0.4400
TEAM_432	RUN_4	0.4718	0.4453	0.4582
TEAM_433	RUN_1	0.6276	0.4858	0.5477
TEAM_433	RUN_2	0.6352	0.5121	0.5671

Performance varied also depending on the particular class of chemical-protein relations. For the antagonist relation class the best team obtained an F-measure of 72.56% (precision of 80.75%, recall of 65.87%) while for inhibition/down-regulation it was of 71.48% (with a precision of 76.51% and a recall of 67.07%).

IV. Discussion and conclusions

The ChemProt track could engage a considerable number of teams, opening up further research on this topic. Systems resulting from this track can result in valuable contributions to improve the curation of chemical and biological data and promote the extraction of various types of chemical-protein interactions demanded by drug-discovery, metabolic reaction, gene regulatory network and systems biology analysis pipelines.

The ChemProt corpus comprises a large collection of manually annotated mentions of chemical compounds and genes/proteins that can serve to improve and validate bio-entity recognition tools, while the granular interaction types annotated for this task can foster more sophisticated bio-entity relation extraction pipelines. Only five general CPI classes were tested for this track, implying that the remaining relation classes and the actual granular relation types remain largely unexplored. We plan to promote additional efforts focusing on these more granular relation types, and provide the annotation of gene mention normalizations, i.e. biological database identifiers for the ChemProt corpus together with inter-annotator agreement results. When examining the underlying methodologies tested by participating teams, it becomes clear that machine learning techniques, and especially the use of artificial neural network approaches represent highly competitive strategies for chemical-gene interaction extraction tasks.

Acknowledgment

We acknowledge the encomienda MINETAD-CNIO/OTG Sanidad Plan TL and OpenMinted (654021) H2020 project for funding. We thank the Foundation for Applied Medical Research (FIMA) of the University of Navarra for financial support.

References

1. Krallinger, M., Leitner, F., Vazquez, M., Valencia, A. (2014) Text Mining. *Comprehensive Biomedical Physics,* **6**, 51-66.

2. Wei, C. H., Peng, Y., Leaman, R., Davis, A. P., Mattingly, C. J., Li, J., et al. (2016). Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. *Database*, *2016*.

3. Krallinger, M., Rabal, O., Lourenço, A., Oyarzabal, J., and Valencia, A. (2017). Information Retrieval and Text Mining Technologies for Chemistry. *Chem. Rev.*, **117** (12), 7673–7761

4. Rindflesch, T. C., Tanabe, L., Weinstein, J. N., and Hunter, L. (2000). EDGAR: extraction of drugs, genes and relations from the biomedical literature. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 517–528.

5. Günther, S., Kuhn, M., Dunkel, M., Campillos, M., Senger, C., Petsalaki, E., et. al. (2007). SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic acids research*, ***36***(suppl_1), D919-D922.

6. Rebholz-Schuhmann, D., Kirsch, H., Arregui, M., Gaudan, S., Riethoven, M., and Stoehr, P. (2007). EBIMed—text crunching to gather facts for proteins from Medline. *Bioinformatics*, ***23***(2), e237-e244.

7. Tari, L., Hakenberg, J., Gonzalez, G., and Baral, C. (2009). Querying parse tree database of Medline text to synthesize user-specific biomolecular networks. In *Pacific symposium on biocomputing*, **14**, 87-98.

8. Czarnecki, J., Nobeli, I., Smith, A. M., and Shepherd, A. J. (2012). A text-mining system for extracting metabolic reactions from full-text articles. *BMC bioinformatics*, ***13***(1), 172.

9. Cañada, A., Capella-Gutierrez, S., Rabal, O., Oyarzabal, J., Valencia, A., & Krallinger, M. (2017). LimTox: a web tool for applied text mining of adverse event and toxicity associations of compounds, drugs and genes. *Nucleic Acids Research*. **45**, Issue W1, W484–W489.

10. Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., et al. (2015). The CHEMDNER corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, **7**(S1), S2.

11. Perez, M. P., Rabal, O., Rodriguez, G. P., Vazquez, M., Fdez-Riverola, F., Oyarzabal, J., ... & Krallinger, M. (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.

12. Krallinger, M., Rabal, O., Lourenço, A., Perez, M. P., Rodriguez, G. P., Vazquez, M., et al. (2015). Overview of the CHEMDNER patents task. In Proceedings of the fifth BioCreative challenge evaluation workshop (pp. 63-75).

13. Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., et. al. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, ***34***(suppl_1), D668-D672.

14. Yang, H., Qin, C., Li, Y. H., Tao, L., Zhou, J., Yu, C. Y., ... & Chen, Y. Z. (2015). Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information. *Nucleic Acids Research*, **44**(D1), D1069-D1074.

15. Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., et al. (2014). The ChEMBL bioactivity database: an update. *Nucleic Acids Research*, **42**(D1), D1083-D1090.

16. Visser, U., Abeyruwan, S., Vempati, U., Smith, R. P., Lemmon, V., and Schürer, S. C. (2011). BioAssay Ontology (BAO): a semantic description of bioassays and high-throughput screening results. *BMC bioinformatics*, **12**(1), 257.

17. BEL Relationships. https://wiki.openbel.org/display/BLD/BEL+Relationships (accessed Mar 30, 2017).

18. BEL Causal Relationships. https://wiki.openbel.org/display/BLD/Causal+Relationships (accessed Mar 30, 2017)

19. Perfetto, L., Briganti, L., Calderone, A., Cerquone Perpetuini, A., Iannuccelli, M., Langone, F., ... & Peluso, D. (2016). SIGNOR: a database of causal relationships between biological entities. Nucleic acids research, 44(D1), D548-D554.

20. Southan, C., Sharman, J. L., Benson, H. E., Faccenda, E., Pawson, A. J., Alexander, S. P., et al. (2015). The IUPHAR/BPS Guide to PHARMACOLOGY in 2016: towards curated quantitative interactions between 1300 protein targets and 6000 ligands. *Nucleic Acids Research*, **44**(D1), D1054-D1068.

21. IUPHAR/BPS Guide to Pharmacology.

http://www.guidetopharmacology.org/helpPage.jsp#ligand (accessed Mar 30, 2017).

22. http://www.biocreative.org/tasks/biocreative-vi/track-5/

23. Pérez-Pérez, M., Pérez-Rodríguez, G., Rabal, O., Vazquez, M., Oyarzabal, J., Fdez-Riverola, F., et al. (2016). The Markyt visualisation, prediction and benchmark platform for chemical and gene entity recognition at BioCreative/CHEMDNER challenge. *Database*, *2016*.