The BioCreative VI Precision Medicine Track corpus

Selection, annotation and curation of protein-protein interactions affected by mutations in scientific literature

Rezarta Islamaj Doğan[1], Andrew Chatr-aryamontri[2], Chih-Hsuan Wei[1], Christie S. Chang[3], Rose Oughtred[3], Jennifer Rust[3], Lorrie Boucher[4], Sun Kim[1], Donald C. Comeau[1], Zhiyong Lu[1], Kara Dolinski[3], and Mike Tyers[2,4]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD20894, USA

[2]Institute for Research in Immunology and Cancer, Université de Montréal, Montréal, QC Canada

[3]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, USA

[4]The Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, ON, Canada

***Abstract*— The Precision Medicine Track in BioCreative VI aims to bring together the biomedical text mining community for a novel challenge: mining the biomedical literature in search of information of value to precision medicine initiatives such as mutations disrupting/affecting protein-protein interactions (PPI). The Precision Medicine track is organized into two tasks: 1) the triage task – focusing on selection of relevant PubMed articles describing PPI affected by mutations, and 2) the relation extraction task – focusing on extracting the interacting gene pairs for the interactions that are affected by the presence of a mutation.**

To support this track with an effective training dataset and limited curator time, the track organizers used a two-staged approach. First, for the creation of the training dataset, the organizers and curators worked on leveraging the information from expertly curated and publicly available PPI databases, augmenting it with a set of articles selected via publicly available state-of-the-art text mining tools. 4,082 PubMed articles were thus carefully reviewed, annotated and released for system development. They contained 1,729 articles labelled positive for curation, out of which, 597 contained 752 curated relations. The second stage pertained to the creation of the testing dataset, which consisted of 1,464 PubMed articles, previously not curated in any of the known PPI databases. These articles were highly likely to describe PPI and sequence variants according to several text mining tests. Each article in the testing dataset was annotated by at least two curators, for relevance relation extraction. Five BioGRID annotators participated and reviewed more than 600 articles each. The testing set contained 730 articles labelled positive for curation, out of which, 688 articles contained 930 curated relations. We detail here the data collection, manual review and annotation process. We give a report on the precision medicine track corpus characteristics. This analysis will provide useful information to developers and researchers for comparing and developing innovative text mining approaches for the BioCreative VI challenge and other Precision Medicine related applications.

Keywords—corpus creation, manual annotation, protein-protein interaction, mutation, relation extraction, information extraction.

I. INTRODUCTION

Biological knowledgebases, such as BioGRID[1], play an increasingly important role in the scientific community due to the curated, summarized and computable knowledge extracted from the literature by expert curators (1, 2). However, their ability to keep up with the growth of biomedical literature is under scrutiny (3). BioCreative (4-13) has traditionally aimed to bridge the gap between the text mining community and biological database curators by fostering development of text mining tools that have practical applications in extracting with high accuracy biological information from unstructured text.

Precision medicine[2] is the emerging approach of disease-treatment that revolves around the idea that a treatment plan is more effective when it takes into account a patient's individual genetic code and the environment they live in. The practice of precision medicine will only be possible with the establishment of databases that integrate the information of genes and mutations with their corresponding biological function. Such knowledgebases will be available for healthcare providers to reference in order to understand the clinical implications of each patient's genetic makeup. The first step towards this goal calls for development of novel text mining tools that can facilitate such an intricate curation processes, increasing accuracy, coverage, and productivity.

To date, there are no available text mining tools that facilitate the specific retrieval of such information which continues to re-main buried in the unstructured text within the biomedical literature. The goal of the Precision Medicine Track of

[1]https://thebiogrid.org/

[2] https://syndication.nih.gov/multimedia/pmi/infographics/pmi-infographic.pdf

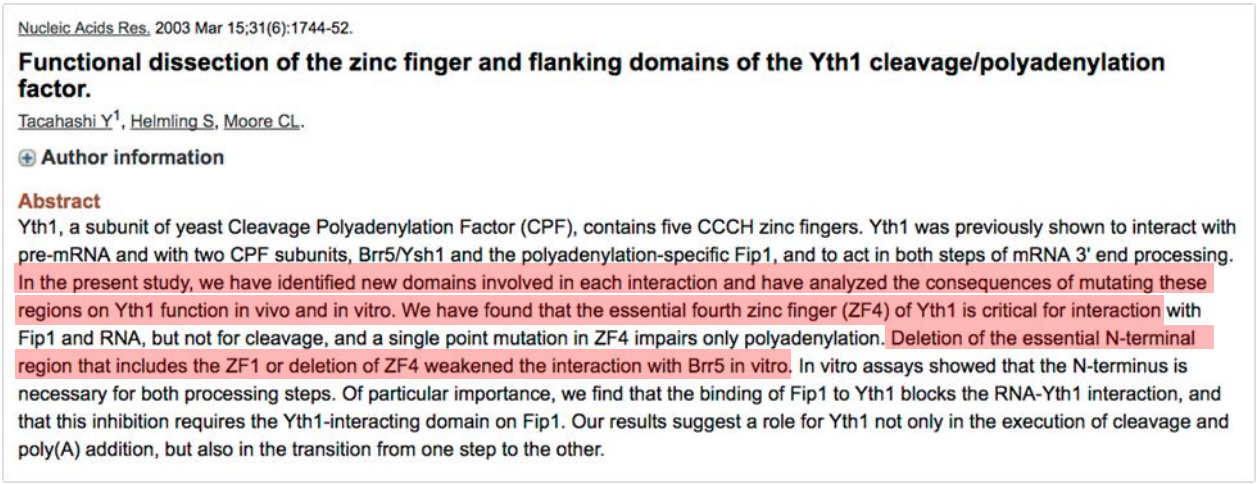
BioCreative VI is to foster the development of text mining algorithms that specialize in scanning the published biomedical literature and to extract the reported discoveries of protein interactions changing in nature due to the presence of genomic variations or artificial mutations. To achieve this goal, we designed the track as a combination of two text mining tasks:

- Document Triage: Identification of relevant PubMed citations describing mutations affecting protein-protein interactions. Figure 1 shows a relevant PubMed article for this purpose, and the highlighted sentences signify to curators that this article describes experimental evidence that the interaction is affected by mutation(s).
- Relation Extraction: Extraction of experimentally verified PPI pairs affected by the presence of a genetic mutation

In order to support this task, we designed and organized an annotation effort that produced a novel corpus. This dataset contains expert human annotations of PPI affected by mutations, as described in the scientific literature. In order to overcome the biggest challenge in building specialized corpora, that of limited reviewer time, we followed several strategies that allowed us to maximize this valuable resource:

First, we built a training dataset consisting of 4,082 PubMed articles, as described in (14). Next, we brought together five BioGRID curators who manually annotated 1,500 PubMed articles for relevance and interacting pairs of proteins that were affected by genetic mutations. Each paper was annotated by at least two curators.

In this manuscript, we describe the process of creating this valuable resource, its manual annotation, annotation guidelines, and inter-annotator agreement. Moreover, we describe how the training and testing datasets complement each other in a rich corpus to test and develop automatic methods for predicting genetically affected protein-protein interactions.



Nucleic Acids Res. 2003 Mar 15;31(6):1744-52.

Functional dissection of the zinc finger and flanking domains of the Yth1 cleavage/polyadenylation factor.

Tacahashi Y[1], Helmling S, Moore CL.

⊕ Author information

Abstract

Yth1, a subunit of yeast Cleavage Polyadenylation Factor (CPF), contains five CCCH zinc fingers. Yth1 was previously shown to interact with pre-mRNA and with two CPF subunits, Brr5/Ysh1 and the polyadenylation-specific Fip1, and to act in both steps of mRNA 3' end processing. In the present study, we have identified new domains involved in each interaction and have analyzed the consequences of mutating these regions on Yth1 function in vivo and in vitro. We have found that the essential fourth zinc finger (ZF4) of Yth1 is critical for interaction with Fip1 and RNA, but not for cleavage, and a single point mutation in ZF4 impairs only polyadenylation. Deletion of the essential N-terminal region that includes the ZF1 or deletion of ZF4 weakened the interaction with Brr5 in vitro. In vitro assays showed that the N-terminus is necessary for both processing steps. Of particular importance, we find that the binding of Fip1 to Yth1 blocks the RNA-Yth1 interaction, and that this inhibition requires the Yth1-interacting domain on Fip1. Our results suggest a role for Yth1 not only in the execution of cleavage and poly(A) addition, but also in the transition from one step to the other.

Figure 1 PubMed article relevant for curation. The abstract describes evidence that a protein pair interaction has been affected by a mutation.

II. CORPUS DEVELOPMENT

The biggest challenge for the organizers of the BioCreative VI Precision Medicine Track was the creation of a high-quality corpus that would serve as a good resource for building automatic algorithms to detect such specialized information.

A possible source for specific PPI information and their related mutations would be the IntAct/Mint database (2), whose curators have had a wide scope when curating protein interactions. Despite the broad coverage and comprehensive curation, such information was not easily retrievable. For this reason, first, our curators selected articles from the IntAct/Mint database that had mutation annotation, and carefully reviewed them and categorized them as relevant/not relevant for the precision medicine track. In addition, we used state-of-the-art text mining methods to select PubMed articles not found in curated databases, that were highly likely to describe protein-protein interactions as well as to contain sequence variations. As a result of this exercise, a set of 4,082 PubMed articles was curated and released as training data to BioCreative VI Precision Medicine Track participants for system development. We described the data repurposing method and text mining triage and manual validation methods that were used to develop this dataset here (14).

The testing dataset, was decided that it would contain previously not annotated articles, and was annotated by five BioGRID curators, with each article being annotated by at least two curators. We describe this process below.

A. Annotation Guidelines

The corpus annotation started with a simple exercise for which every PubMed article was categorized based on these questions:

- Does this article describe experimentally verified protein-protein interactions?

- Does this article describe a known disease mutation or a mutational analysis experiment?
- Are the database curated PPI pairs for this article mentioned in the abstract?
- Is the PPI affected by the mutation?

Then, based on the above annotations, articles were carefully categorized as 1) Positives – articles specifically describing PPI influenced by genetic mutations, or 2) Negatives – a category which comprises articles describing both PPIs and genetic variation analysis with no inference of relation between them, articles containing PPI but no mutations, articles containing mutations but no PPI, and articles mentioning neither.

Another important point of consideration was that the information needed to be present in the abstract. Database curators always look for curatable information in the full text. However, the triage process is often conducted on the article's abstract. Thus, for an article to be labelled positive for curation, the title or abstract had to contain a statement of evidence describing in no ambiguous terms that the interaction between a pair of proteins had been affected by the presence of a genetic mutation. The degree of the effect was not annotated.

For the relation extraction task, the interacting proteins needed to be named in the title or abstract, but the name or description of the specific sequence variant was not required. This degree of specificity is unlikely to be found in the abstract, although the information would be present in the full text. Given the condition that the interacting pair needed to be named, it is possible that an article could be labelled positive for the triage task, but not be eligible for the relation extraction task.

Furthermore, protein-protein interactions could be physical interactions, biochemical reactions, self-interactions and/or aggregations. Examples of molecular interactions which were not considered for the relation extraction task are: protein complexes, cell-organelle interactions, and colocalizations. It was also possible that an abstract could describe experimentally verified PPI, as well as include mutations mentions, but the two events were not related. All such articles were labelled negative.

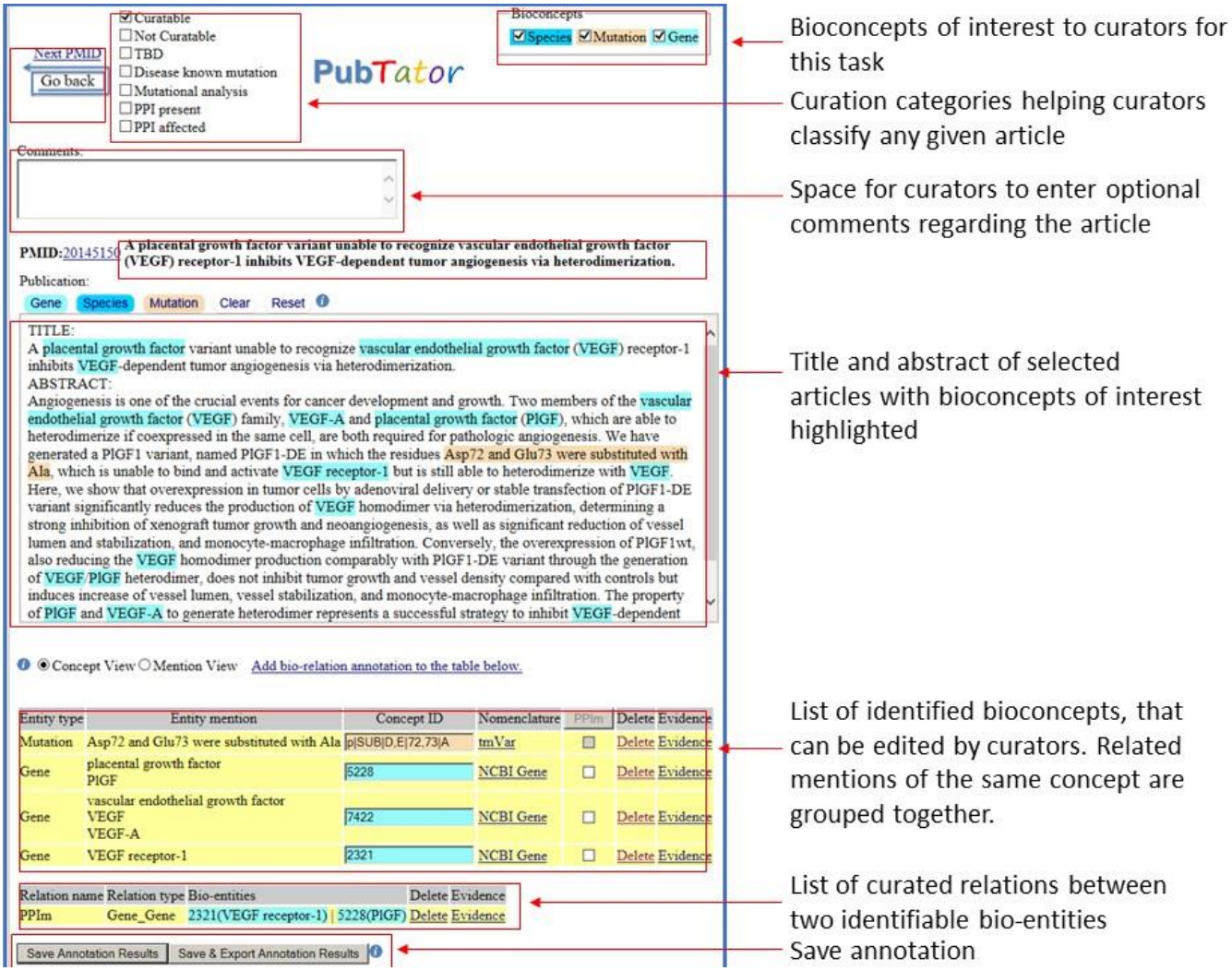


Figure 2 PubTator curation view customized for the BioCreative VI Precision Medicine task. Bioconcepts of interest are: gene names, mutations and species. Automatic detection of these concepts can be turned on and off to help curators.

Regarding mutations, they could be deletions, point mutations and possibly allelic variations. In addition, all mentioned mutations were considered, whether disease mutations or synthetic ones. Often, the mutation was not explicitly mentioned, however if mutational analysis occurred, then the article was curated. Finally, article curation was not limited to any species.

B. Annotation Process

The training data exercise helped collect a set of 4,082 articles which were annotated and distributed for system development. For the testing dataset, we wanted to use PubMed articles that had not been annotated before, and also that could be relevant to curation interests at BioGRID. Starting with a curator-designed comprehensive PubMed query to select articles, which returned 1.4 Million articles, we applied several text mining filters to be able to rank them according to their relevance to the task. Our approach used two well-known publicly available text mining tools: PIE the search (15) and tmVar (16, 17). PIE the search is a web service that ranks PubMed articles based on their probability of describing protein-protein interactions. This algorithm was the winner of BioCreative III ACT competition (5). tmVar is another text mining tool that is used to recognize sequence variants in PubMed literature.

Using these methods, we narrowed down the set of 1.5 Million to 5,000. Given the limited curator time, we randomly selected a set of 1,500 PubMed articles, whose PIE score distribution matched the distribution of scores of the training set, and did two tests of randomly annotating sets of 100 articles to estimate the ratio of relevant to non-relevant articles. After all of these conditions were satisfied, the set of 1,500 PubMed articles was decided upon and the annotation phase of the testing dataset was ready to begin.

All articles in the testing dataset were distributed and randomly assigned to pairs of curators for annotation. The annotation process took place in three phases:

- **Phase 1.** All five curators worked on a set of 20 articles. They spent one week reading and annotating the articles independently, and one week discussing their decisions, for both positively labelled articles and negatively labelled ones. This was the phase where the annotation tool was also adapted to fit curator needs.
- **Phase 2.** Two sets of 100 articles were assigned to three curators at a time. They worked independently for ten days, and then used ten more days to discuss their decisions in groups of three. During this phase, the rules of relation extraction were refined.
- **Phase 3.** The remaining articles, divided into sets of 100, were randomly assigned to pairs of curators. Each pair of curators worked on average for a period of 10 days to curate each set. When both curators were finished with a set, a detailed annotation comparison document was generated, and the curators had independent meetings to review and come to a common agreement. The annotation comparison sheets were used for computing the inter-annotator agreement.

Curators and organizers met weekly to discuss the corpus annotation issues, tool features and report on the progress.

C. Annotation Tool

An annotation portal was built based on PubTator as shown in Figure 2. Testing data was distributed among five curators who accessed the system through private accounts via this system. The system allowed for the organizers to collects multi-annotations for each article and compute annotation comparisons.

When a curator clicks on an article, they view the screen as shown in Figure 2. The tool gives curators the capability to benefit from text mining tools that are specialized in gene/protein, mutation and species identification (mention and normalization). They could easily navigate to the next article, or go to PubMed for more information. They could keep notes on each article. The title and abstract for each article are displayed in one screen, and should the text mining tools be selected, the predicted entities are shown highlighted on the screen. In addition, the predicted list of entities is listed in a table below the abstract. Curators could edit this table to adjust problems. They could annotate from scratch, by highlighting the text mention of interest and selecting the category appearing above the annotation box. Completed annotations could be reviewed, deleted and or edited.

Curators could use the tool to annotate a relation by selecting the entities of interest from the list of bioconcepts in the entity table and clicking on the relation button. The annotated relation then, would get listed in the relations table, shown at the bottom of the screen. Relations could also be edited further as needed. Annotations could be saved, and also exported or downloaded locally.

D. Inter-annotator Agreement

We computed the degree of agreement between pairs of annotators for every set of 100 PubMed articles that was annotated, and then found the average of all sets. For the triage task, on average, our annotators were in agreement for 82% of the articles. The number of articles to be reviewed for classification purposes ranged between 3 and 19 for each set of 100. For each set, on average 2 or 3 articles were difficult to assign a clear label. These were ultimately removed and not used for the official evaluation of the Triage task.

The detailed comparison annotation documents showed that, for each set of 100, on average, 41 articles were marked positive, 42 articles were marked negative and the rest needed to be reviewed to resolve any discrepancies. Of the positive articles, for a typical set of 100 articles, on average, 23 articles needed to be reviewed for the curators to come to a consensus on relation extraction.

TABLE I. STATISTICS OF THE PRECISION MEDICINE TRACK DATASET

Dataset	Articles	Positive	Negative	Articles with relations	Number of relations
Training	4,082	1,729	2,353	597	752
Testing	1,464	730	734	688	930
Total	5,546	2,459	3,087	1,285	1,682

TABLE II. TRIAGE TASK RESULTS OF THE BASELINE SYSTEM

	Avg. Prec.	Precision	Recall	F1
10-fold CV (training data)	0.7225	0.6891	0.6260	0.6561
Testing data	0.6500	0.6097	0.6356	0.6224

TABLE III. RELATION EXTRACTION TASK RESULTS OF THE BASELINE SYSTEM (HOMOLOGENE EVALUATION)

	Precision	Recall	F1
Training data	0.1650	0.4753	0.2449
Testing data	0.1460	0.5215	0.2282

Relations mismatch could be categorized as follows:

- The two curators had picked the same interacting mentions, however, they had normalized them to two different GeneIDs.
- One of the curators had marked additional relations.
- The two curators had marked different interactions, which shared a gene.
- One of the curators, or both, had specifically marked the article for further discussion.

E. Corpus Characteristics

As shown in Table 1, the Precision Medicine Track dataset is a large dataset of 5,546 PubMed articles, manually labelled for triage and relations of PPI affected by mutations. In the training dataset, the relations were repurposed from the previous PPI annotations in the IntAct/Mint databases. The testing dataset was richer in relations, since more curator time was devoted to their extraction. As a collection, the Precision Medicine Track dataset contains 1,285 articles annotated for relations with 1,682 total relations.

F. Benchmark results and corpus use

A baseline SVM method was designed using unigram and bigram features from titles and abstracts of the training corpus. Results are detailed in Table 2. For the Relation Extraction Task, we implemented a simple co-occurrence baseline method, as shown in Table 3. The Gene entities were automatically recognized using our in-house tools (17-19). The co-occurrence method considered every sentence that contained two gene mentions and predicted a relation between them. Predictions of sequence variants were not considered for this baseline. HomoloGene evaluation, considered whether the curator's annotated gene in the relation and the predicted gene were homologous genes.

This dataset was used for the BioCreative VI Precision Medicine Task. The results of twenty-two systems were submitted for the Triage Task and the results of fourteen systems were submitted for the Relation Extraction Task. This is an indication of the necessity of developing this dataset. We anticipate that more systems will use the released corpus in the future.

III. CONCLUSIONS AND PUBLIC AVAILABILITY

Scientific articles indexed in PubMed contain a vast amount of precision medicine related information, because they often detail experimentally verified protein-protein interactions, which in some cases are affected by differences in sequence variation. Currently, such information can only be extracted by skilled domain expert curators.

The BioCreative VI Precision Medicine Track corpus contains 5,546 PubMed articles and is of high quality. It was curated by five BioGRID curators and each article was annotated by at least two curators, with an inter-annotator agreement of 82%.

By releasing the BioCreative VI Precision Medicine Track corpus, we aim to facilitate the curation of precision-medicine-related information available in published literature. This corpus fosters the development of innovative text mining algorithms that may help database curators in identifying molecular

interactions that differ based on the presence of a specific genetic variant, information which could be translated to clinical practice.

In addition, this dataset may provide important insights on 1) understanding the specific biological information in the unstructured text that may be relevant for precision medicine purposes, and 2) the best practices for designing automatic computational methods that can extract such information.

The BioCreative VI Precision Medicine training corpus is available from the BioCreative website for the scientific community.

ACKNOWLEDGEMENT

This research was supported by the NIH Intramural Research Program National Library of Medicine and by grants from the National Institutes of Health R01OD010929 (to M.T. and K.D.) and 3OT3TR002026-01S1 (sub-award to M.T., Principal Investigator, S. Huang, Institute for Systems Biology, Seattle WA).

REFERENCES

1. Chatr-Aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N.K., O'Donnell, L., Oster, S., Theesfeld, C., Sellam, A., Stark, C., Breitkreutz, B.J., Dolinski, K. and Tyers, M. (2017). The BioGRID interaction database: 2017 update. *Nucleic Acids Res.* **45**(D1): D369-D379.

2. Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., del-Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., Jagannathan, S., Jimenez, R., Khadake, J., Lagreid, A., Licata, L., Lovering, R.C., Meldal, B., Melidoni, A.N., Milagros, M., Peluso, D., Perfetto, L., Porras, P., Raghunath, A., Ricard-Blum, S., Roechert, B., Stutz, A., Tognolli, M., van Roey, K., Cesareni, G. and Hermjakob, H. (2014). The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42**(Database issue): D358-363.

3. Hirschman, J., Berardini, T.Z., Drabkin, H.J. and Howe, D. (2010). A MOD(ern) perspective on literature curation. *Mol Genet Genomics.* **283**(5): 415-425.

4. Wu, C.H., Arighi, C.N., Cohen, K.B., Hirschman, L., Krallinger, M., Lu, Z., Mattingly, C., Valencia, A., Wiegers, T.C. and John Wilbur, W. (2012). BioCreative-2012 virtual issue. *Database (Oxford).* **2012**: bas049.

5. Krallinger, M., Vazquez, M., Leitner, F., Salgado, D., Chatr-aryamontri, A., Winter, A., Perfetto, L., Briganti, L., Licata, L., Iannuccelli, M., Castagnoli, L., Cesareni, G., Tyers, M., Schneider, G., Rinaldi, F., Leaman, R., Gonzalez, G., Matos, S., Kim, S., Wilbur, W., Rocha, L., Shatkay, H., Tendulkar, A., Agarwal, S., Liu, F., Wang, X., Rak, R., Noto, K., Elkan, C. and Lu, Z. (2011). The Protein-Protein Interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text. *BMC bioinformatics.* **12**(Suppl 8): S3.

6. Lu, Z. and Wilbur, J. (2010). Overview of BioCreative III Gene Normalization. Proceedings of the BioCreative III workshop, Bethesda, MD.

7. Leitner, F., Mardis, S.A., Krallinger, M., Cesareni, G., Hirschman, L.A. and Valencia, A. (2010). An Overview of BioCreative II.5. *IEEE/ACM Trans Comput Biol Bioinform.* **7**(3): 385-399.

8. Krallinger, M., Morgan, A., Smith, L., Leitner, F., Tanabe, L., Wilbur, J., Hirschman, L. and Valencia, A. (2008). Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. *Genome Biol.* **9 Suppl 2**: S1.

9. Chatr-aryamontri, A., Kerrien, S., Khadake, J., Orchard, S., Ceol, A., Licata, L., Castagnoli, L., Costa, S., Derow, C., Huntley, R., Aranda, B., Leroy, C., Thorneycroft, D., Apweiler, R., Cesareni, G. and Hermjakob, H. (2008). MINT and IntAct contribute to the Second BioCreative challenge: serving the text-mining community with high quality molecular interaction data. *Genome Biol.* **9 Suppl 2**: S5.

10. Hirschman, L., Yeh, A., Blaschke, C. and Valencia, A. (2005). Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC bioinformatics.* **6 Suppl 1**: S1.

11. Kim, S., Islamaj Doğan, R., Chatr-aryamontri, A., Chang, C.S., Oughtred, R., Rust, J., Batista-Navarro, R., Carter, J., Ananiadou, S., Matos, S., Santos, A., Campos, D., Oliveira, J.L., Singh, O., Jonnagaddala, J., Dai, H.-J., Su, E.C.-Y., Chang, Y.-C., Su, Y.-C., Chu, C.-H., Chen, C.C., Hsu, W.-L., Peng, Y., Arighi, C., Wu, C.H., Vijay-Shanker, K., Aydın, F., Hüsünbeyi, Z.M., Özgür, A., Shin, S.-Y., Kwon, D., Tyers, M., Dolinski, K., Wilbur, W.J. and Comeau, D.C. (2016). BioCreative V BioC Track Overview: Collaborative Biocurator Assistant Task for BioGRID. *Database.*

12. Doğan, R.I., Kim, S., Chatr-Aryamontri, A., Comeau, D.C. and Wilbur, W.J. (2015). Identifying genetic interaction evidence passages in biomedical literature. BioCreative V Workshop. Seville, Spain. 36-41.

13. Lu, Z. and Hirschman, L. (2012). Biocuration workflows and text mining: overview of the BioCreative 2012 Workshop Track II. *Database : the journal of biological databases and curation.* bas043.

14. Islamaj Dogan, R., Chatr-aryamontri, A., Kim, S., Wei, C.-H., Peng, Y., Comeau, D.C. and Lu, Z. (2017). BioCreative VI Precision Medicine Track: creating a training corpus for mining protein-protein interactions affected by mutations *Proceedings of the 2017 ACL Workshop on Biomedical Natural Language Processing (BioNLP).*

15. Kim, S., Kwon, D., Shin, S.Y. and Wilbur, W.J. (2012). PIE the search: searching PubMed literature for protein interaction information. *Bioinformatics.* **28**(4): 597-598.

16. Wei, C.H., Phan, L., Feltz, J., Maiti, R., Hefferon, T. and Lu, Z. (2017). tmVar 2.0: Integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine. *Bioinformatics.*

17. Wei, C.H., Harris, B.R., Kao, H.Y. and Lu, Z. (2013). tmVar: a text mining approach for extracting sequence variants in biomedical literature. *Bioinformatics.* **29**(11): 1433-1439.

18. Wei, C.H., Kao, H.Y. and Lu, Z. (2012). SR4GN: a species recognition software tool for gene normalization. *PLoS One.* **7**(6): e38460.

19. Wei, C.H., Leaman, R. and Lu, Z. (2016). Beyond accuracy: creating interoperable and scalable text-mining web services. *Bioinformatics.* **32**(12): 1907-1910.