Overview of BioCreative V BioC Track

Sun Kim[1], Rezarta Islamaj Doğan[1], Andrew Chatr-aryamontri[2], Mike Tyers[3], W. John Wilbur[1], and Donald C. Comeau[*1]

[1] National Center for Biotechnology Information (NCBI), NLM, NIH, USA
{sun.kim,islamaj,wilbur,comeau}@ncbi.nlm.nih.gov
[2] Institute for Research in Immunology and Cancer, Université de Montréal, Canada
andrew.chatr-aryamontri@umontreal.ca
[3] The Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Canada
md.tyers@umontreal.ca

Abstract. BioC is a simple XML format for text, annotations and relations, and was developed to achieve interoperability for biomedical text processing. Following the success of BioC in BioCreative IV, the BioCreative V BioC track addressed a collaborative task to build an assistant tool for BioGRID curation. For this track, we divided the whole task into 8 different subtopics including gene/protein/organism named entity recognition and protein-protein/genetic interaction passage identification. A total of 9 teams participated in 7 tasks and the submitted runs were merged by using a machine learning classifier to produce optimized output. The curation tool was evaluated by 4 BioGRID curators in terms of practical usability. The feedback given by curators was positive overall due to the user-friendly design and an easy-to-use gene/protein curation tool. In this paper, we describe the framework of the collaborative BioC task and discuss preliminary findings based on the user survey.

1 Introduction

Molecular interaction (MI) information is of great importance both in the field of experimental biology as well as from the perspective of systems biology and bioinformatics [8]. Thus, many efforts have been made to capture this information in databases such as BioGRID[1] [3], IntAct [6] and DIP [11]. Text mining techniques have been tried to ease the curation burden. However, there have been few successes for improving biocuration throughput using text mining [5].

The purpose of the BioC track in BioCreative V is to create BioC-compatible modules [4] which complement each other and integrate into a system that assists BioGRID curators. In previous BioCreative workshops, great emphasis was given to the identification of protein-protein interactions (PPI). The PPI track [8, 7, 9] was divided into subcategories and each task was addressed independently, i.e. article classification, interaction pair extraction, interaction sentence classification and experimental method identification. The user interaction track (IAT) [2, 1, 10] promoted the development of annotation systems that can assist

[1] http://thebiogrid.org

in biocuration tasks by bringing text mining tool developers and database curators together. But, no attempt has been made to integrate text mining modules developed in the formal BioCreative PPI track into one annotation tool. This may be due to interoperability and data exchange problems, or performance is not good enough for certain extraction modules.

While the previous BioC track focused on releasing the BioC resources such as datasets and biomedical NLP tools, the BioCreative V BioC track addresses a more practical issue by setting up a biocurator assistant tool in a collaborative way, in particular, for the BioGRID database. The main goals of the BioC track are as follows.

- To define a collaborative task for MI information extraction, so each team can develop a module independently, but can also use other modules' outputs.
- To develop practical MI tools by combining or improving existing methods for full-text articles.
- To improve interoperability by developing BioC-compatible MI extraction modules.
- To implement an annotation assistant tool by closely working with biocurators in BioGRID.
- To produce a full-text benchmark set while evaluating the new biocurator assistant tool.

In this respect, we first divided the BioC track into 8 different tasks. Based on this division, 8 teams created modules which were cooperatively used for annotating MI information from full-text documents.

2 BioC Track Tasks

One distinctive feature of the BioC track was the lack of competition among participating teams. The organizers promoted a collaborative framework and helped each team to collaborate with the others for building an integrated annotation system. The tasks defined for the BioGRID assistant tool are as follows:

- Task 1, "Gene/protein named entity recognition (NER)": This task is to identify gene/protein mentions. Participating teams combine results from existing tools or develop their own methods to improve NER performance.
- Task 2, "Species/organism NER": This task is to identify and normalize species/ organism names. Participating teams either combine results from existing techniques or proposes a new way for identifying species/organisms.
- Task 3, "Normalization of gene/protein names": This task is to determine gene/protein IDs based on gene/protein names and species/organisms mentioned in surrounding text. Previous BioCreative datasets may be used for system development. The system can alternatively use prediction results from Tasks 1 and 2.

- Task 4, "Passages with PPIs": This task is to find passages describing physical PPIs. Physical interactions may appear in single or several sentences. Participating team(s) may use the PPI corpora[2] such as BioCreative, BioNLP Shared Task, AIMed and LLL for training, but they also can develop additional training data.
- Task 5, "Passages with genetic interactions (GIs)": This task is to find passages claiming GIs. GIs may appear in single or several sentences. The BioGRID set may be used for creating a training set.
- Task 6, "Passages with experimental methods for physical interactions": This task is to search for passages describing experimental methods used for finding physical interactions. There are 17 experimental methods defined in BioGRID. For this task, BioGRID, MINT, and/or IntAct may be used for training data.
- Task 7, "Passages with GI types": This task is to search for passages describing GI types. These passages may overlap with the ones from Task 5. However, for Task 7, a type of GI should be clearly shown. There are 11 interaction types defined in BioGRID. The BioGRID set may be used for training data.
- Task 8, "Visual tool for displaying various annotations": This task is to develop a visualization tool for highlighting annotation results from other tasks above. The tool should allow easy navigation and display user-selected annotations. A participating team should work closely with biocurators in BioGRID, in order to develop a visualization tool that curators find most useful.

3 Methods

Unlike other BioCreative tasks, no official training/test set was released for the BioC track. Participating teams proposed a method for each task and defined a training set they were going to use. A total of 10 teams submitted their task proposals in March, however one team later withdrew. As a result, we received 24 runs from 8 teams plus the visual tool by the July submission deadline. Task 5 was not performed and it was considered covered by Task 7 because no team submitted a proposal. Table 1 shows submitted runs for each task for each team. The number of runs varied from 3 to 6 except for the visual interface task. Task 8 is to implement a visual tool for BioGRID curation, hence there are no runs for the task. A submitted run contains predicted text, e.g. gene/protein/organism names or PPI/GI passages, optionally with normalized IDs for gene/protein/organism names.

For merging multiple outputs from teams and evaluating the BioGRID curation tool, we recruited 4 curators from BioGRID to build a gold-annotation set. Since most of the information in the BioGRID database is from the yeast or human domain, we randomly chose 60 full-text PubMed Central® (PMC) articles for each of these organisms. For the selected documents in the human set, there

[2] `http://corpora.informatik.hu-berlin.de`

Team	Task 1	Task 2	Task 3	Task 4	Task 6	Task 7	Task 8
T1	1						
T2	1	1	1				
T3	1	1	1				
T4	1	1	1				
T5				4			
T6				1			
T7					2		
T8				1	2	4	
T9							1
Total	4	3	3	6	4	4	1

Table 1. Submitted runs from 9 participating teams. To boost the synergy effect of using multiple runs, we (T8) produced additional results for Tasks 4 and 6. Only one team was selected for Task 8 as it was to develop a user interface.

Organisms	Documents	MI information
Yeast	60	PPI and GI
Human	38	PPI and GI
Human	17	PPI
Human	5	GI

Table 2. Evaluation set used for merging submitted runs and for testing the BioGRID curation tool. Documents were randomly selected from PMC articles relevant to either yeasts or humans. Of these, 98 documents contained both PPI and GI information, the remaining 22 documents contained either PPI or GI.

were 38 OAPMCs with entries in BioGRID containing both PPI and GI information, 17 OAPMCs with PPIs and 5 OAPMCs with GIs. Table 2 summarizes the newly created annotation set for this merging and evaluation process.

To evaluate the BioGRID curation tool, we first assigned 10 articles to each curator (i.e. the test set). The remaining 110 articles were used for training to optimize parameters for merging the multiple runs on the test set. For gene/protein/orgasim tasks (Tasks 1, 2 and 3), we measured the performance of individual runs by precisions and overlaps between submitted runs. The merging process for Tasks 1, 2 and 3 did not include machine learning and simply took the union of selected submitted runs to maximize recall. This is a reasonable strategy for NER tasks because curators prefer high recall.

The PPI/GI tasks are somewhat different than NER tasks. Users expect high recall in general, meanwhile precision should not be ignored. To address the PPI/GI issue, we used the following process for merging and optimizing results.

- We removed uninformative sections such as acknowledgements and references.
- We converted all the remaining paragraphs into sentences.
- For each run of each team, we did the following[3].
  - ⋆ We treated a submission's predicted sentences as a gold standard.
  - ⋆ We used unigrams and bigrams from text as features.
  - ⋆ We performed a 10-fold cross-validation using an SVM classifier [12]. By doing this, all sentences whether positive or negative in the original submission, received scores.
- We learned an SVM classifier using the training set with the obtained scores above as feature weights. If there are four runs, the number of features is exactly four. This process prioritizes submitted runs while maximizing the prediction performance.
- We made predictions on the test set using the combined results.

4 Results

After testing the system, BioGRID curators were asked to rate the usefulness of the system and each functionality on a scale of 1 (bad) to 5 (good). Moreover, they were encouraged to give feedback regarding what they thought hinders usefulness and what might improve usefulness of the system and its functions. Table 3 presents the questionnaire used for the user feedback. The questionnaire consists of 7 categories: "Overall reaction", "Overall comparison to similar systems", "System's ability to help complete tasks", "Prediction performance", "Design of BioC Viewer", "Learning to use BioC Viewer" and "Usability". In the table, average ratings from 4 curators are shown for each question. The rating with 'N/A' (not available) was not used for calculating average rates. From the table, the curators were positive overall for the design and the learnability of the curation tool. Only two curators had experience on other text mining tools and their responses were positive as well. However, passage predictions still need improvement in accuracy to significantly benefit the curation process. Other comments noted that the functionalities for curation were limited. The interface was designed as a viewer, but this can easily be changed by incorporating the curators' comments and suggestions in the future.

Figure 1 depicts the detailed ratings from curators for "Prediction performance". All four curators were satisfied with gene/protein NER and normalization (Tasks 1 and 3), whereas they showed less favorable views for the organism NER and normalization task (Task 2). This may be partly because their goal was to curate PPI and GI pairs, not organism mentions. The PPI/GI passage tasks received rather mixed ratings, but the reactions for finding passages with PPIs and GI types (Tasks 4 and 7) were slightly better than finding passages with PPI experimental methods (Task 6). Curators' comments suggest this could

[3] The procedure in the nested items was performed for assigning scores to all sentences. This was done due to lack of scores for negative predictions. Applying this process also showed better classification performance on the training data.

Questions	Rates
Overall reaction	
Please rate your experience with BioC Viewer.	3.3
Overall, I am satisfied with BioC Viewer.	3.0
I would recommend BioC Viewer to other PPI/GI curators.	2.8
Overall comparison to similar systems	
It is easy to use BioC Viewer.	5.0
I am satisfied with using BioC Viewer.	4.0
BioC Viewer is powerful enough to complete the task.	3.0
System's ability to help complete tasks	
Speed: the system would reduce annotation time to reach my curation goal.	3.5
Effectiveness: the system would help me get closer to my curation goal.	3.0
Efficiency: I can be both fast and effective with the system.	2.8
Prediction performance	
Task 1 (gene/protein NER)	4.3
Task 2 (organism NER)	2.7
Task 3 (gene/protein name normalization)	3.8
Task 4 (Passages with PPIs)	3.3
Task 6 (Passages with PPI experimental systems)	2.5
Task 7 (Passages with GI types)	3.0
Design of BioC Viewer	
It was easy to find and read information.	4.0
Highlights were adequate and helpful.	3.5
Information was well organized.	3.5
Learning to use BioC Viewer	
It was easy to learn how to operate the interface.	4.3
It was easy to remember features in BioC Viewer.	4.3
It was straightforward to use the interface.	4.3
Usability	
The interface was fast enough to do my job.	3.5
The interface was performed consistently.	4.0
The interface provided a means to easily correct mistakes.	3.0

Table 3. Questionnaire used for user feedback. For each question, BioGRID curators rated on a 1 (bad) to 5 (good) scale. The scores shown are the average rates from 4 curators.

be a matter of personal display preference, i.e. some preferred higher recall, but others preferred higher precision.

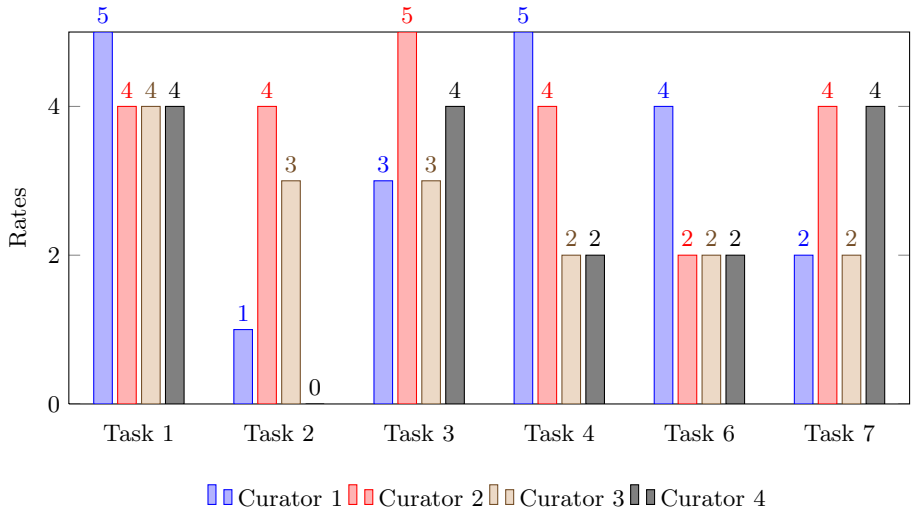

Fig. 1. Curators' ratings for prediction performance for each task. Tasks 1 and 3 received positive responses overall, however ratings were mixed for other tasks depending on curators' preferences. '0' means 'Not Available'.

5 Conclusion

BioC is an XML format for text, annotations and relations that is easy to learn and provides interoperability in biomedical text processing. For BioCreative V, the BioC track focused on developing BioC modules for assisting BioGRID curators. This task was unique since participating teams had to produce independent, but collaborative modules for the BioGRID curation tool. Eight tasks were defined to achieve this goal and a machine learning process was utilized to merge 24 runs from 8 teams. Through the evaluation process, 4 BioGRID curators judged the integrated curation tool in terms of its practical usability. The feedback from curators indicates that the performance of the curation tool and text mining results for gene/protein NER and normalization are reasonable and adequate to support the BioGRID curation task. However, much work needs to be done for suggesting PPI/GI-related passages. The future work includes revising the curation tool based on curators' feedback, releasing BioC modules and the gold-standard set.

Acknowledgments. The authors would like to thank Cecilia Arighi for providing resources for the user questionnaire. This research was supported by the Intramural Research Program of the NIH, National Library of Medicine. Andrew Chatr-aryamontri was supported by National Institutes of Health [R01OD010929 and R24OD011194 to Mike Tyers] and Genome Quebec International Recruitment Award to Mike Tyers.

References

1. Arighi, C.N., Carterette, B., Cohen, K.B., Krallinger, M., Wilbur, W.J., Fey, P., Dodson, R., Cooper, L., Slyke, C.E.V., Dahdul, W., Mabee, P., Li, D., Harris, B., Gillespie, M., Jimenez, S., Roberts, P., Matthews, L., Becker, K., Drabkin, H., Bello, S., Licata, L., Chatr-aryamontri, A., Schaeffer, M.L., Park, J., Haendel, M., Auken, K.V., Li, Y., Chan, J., Muller, H.M., Cui, H., Balhoff, J.P., Wu, J.C.Y., Lu, Z., Wei, C.H., Tudor, C.O., Raja, K., Subramani, S., Natarajan, J., Cejuela, J.M., Dubey, P., Wu, C.: An overview of the BioCreative 2012 Workshop Track III: interactive text mining task. Database 2013, bas056 (2013)
2. Arighi, C.N., Roberts, P.M., Agarwal, S., Bhattacharya, S., Cesareni, G., Chatr-aryamontri, A., Clematide, S., Gaudet, P., Giglio, M.G., Harrow, I., Huala, E., Krallinger, M., Leser, U., Li, D., Liu, F., Lu, Z., Maltais, L.J., Okazaki, N., Perfetto, L., Rinaldi, F., Sætre, R., Salgado, D., Srinivasan, P., Thomas, P.E., Toldo, L., Hirschman, L., Wu, C.H.: BioCreative III interactive task: an overview. BMC Bioinformatics 12(Suppl 8), S4 (2011)
3. Chatr-aryamontri, A., Breitkreutz, B.J., Heinicke, S., Boucher, L., Winter, A., Stark, C., Nixon, J., Ramage, L., Kolas, N., O'Donnell, L., Reguly, T., Breitkreutz, A., Sellam, A., Chen, D., Chang, C., Rust, J., Livstone, M., Oughtred, R., Dolinski, K., Tyers, M.: The BioGRID interaction database: 2013 update. Nucleic Acids Research 41, D816–D823 (2013)
4. Comeau, D.C., Doğan, R.I., Ciccarese, P., Cohen, K.B., Krallinger, M., Leitner, F., Lu, Z., Peng, Y., Rinaldi, F., Torii, M., Valencia, A., Verspoor, K., Wiegers, T.C., Wu, C.H., Wilbur, W.J.: BioC: a minimalist approach to interoperability for biomedical text processing. Database 2013, bat064 (2013)
5. Hirschman, L., Burns, G.A.P.C., Krallinger, M., Arighi, C., Cohen, K.B., Valencia, A., Wu, C.H., Chatr-Aryamontri, A., Dowell, K.G., Huala, E., Lourenco, A., Nash, R., Veuthey, A.L., Wiegers, T., Winter, A.G.: Text mining for the biocuration workflow. Database 2012, bas020 (2012)
6. Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., Jandrasits, C., Jimenez, R.C., Khadake, J., Mahadevan, U., Masson, P., Pedruzzi, I., Pfeiffenberger, E., Porras, P., Raghunath, A., Roechert, B., Orchard, S., Hermjakob, H.: The IntAct molecular interaction database in 2012. Nucleic Acids Research 40, D841–D846 (2012)
7. Krallinger, M., Leitner, F., Rodriguez-Penagos, C., Valencia, A.: Overview of the protein-protein interaction annotation extraction task of BioCreative II. Genome Biology 9(Suppl 2), S4 (2008)
8. 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.J., Rocha, L., Shatkay, H., Tendulkar, A.V., Agarwal, S., Liu, F., Wang, X., Rak, R., Noto, K., Elkan, C., Lu, Z., Dogan, R.I., Fontaine, J.F., Andrade-Navarro, M.A., Valencia, A.: 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 (2011)
9. Leitner, F., Mardis, S.A., Krallinger, M., Cesareni, G., Hirschman, L.A., Valencia, A.: An overview of BioCreative II.5. IEEE/ACM Transactions on Computational Biology and Bioinformatics 7(3), 385–399 (2010)

10. Matis-Mitchell, S., Roberts, P., Tudor, C.O., Arighi, C.N.: BioCreative IV Interactive Task. In: Proceedings of the BioCreative IV Workshop. pp. 190–203 (2013)
11. Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., Eisenberg, D.: The Database of Interacting Proteins: 2004 update. Nucleic Acids Research 32, D449–D451 (2004)
12. Smith, L.H., Wilbur, W.J.: Finding related sentence pairs in MEDLINE. Information Retrieval 13(6), 601–617 (2010)