# Overview of the BioCreative VI Precision Medicine Track

## Mining protein interactions and mutations for precision medicine

Rezarta Islamaj Doğan[1], Sun Kim[1], Andrew Chatr-aryamontri[2], Chih-Hsuan Wei[1], Donald C. Comeau[1], and Zhiyong Lu[1]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, USA
[2]Institute for Research in Immunology and Cancer, Université de Montréal, Montréal, Canada

***Abstract*— The Precision Medicine Track in BioCreative VI aims to bring together the biomedical text mining community and the biocuration community for a novel challenge composed of two tasks: 1) Triage task focused on identifying scientific articles that describe protein-protein interactions (PPI) being disrupted or significantly affected by the presence of genetic mutations, and 2) Relation extraction task focused on extracting the affected protein pairs. This is a novel challenge for the BioNLP community and, to assist system developers and the assessment of such an endeavor, we created the first large-scale manually annotated corpus of scientific articles that describe such information relevant to precision medicine initiative. The training corpus contained 4,082 articles annotated for triage, of which 598 PubMed articles were annotated for relations. The testing corpus contained 1,464 PubMed articles annotated for both triage and relations. Ten teams worldwide participated in the triage task and sent in results for 22 distinct text mining models. Six teams participated in the relation extraction task and sent in results from 14 different text mining systems. When comparing the text mining system predictions with human annotations, for the triage task, the best F-score was 69.5%, the best precision was 61.6%, the best recall was 97.9% and the best average precision was 72.8%. For the relation extraction task, when we account for similar gene identifiers with HomoloGene database, the best F-score was 37.3%, the best precision was 45.4%, and the best recall was 53.9%. Given the level of participation and team results we find our task to be successful in engaging the text-mining research community, producing a first-of-its-kind, large, manually annotated corpus of scientific articles relevant for precision medicine, and providing the first results of automatically identifying PubMed articles that describe PPI affected by mutations, and extracting the affected relations.***

***Keywords—precision medicine, corpus annotation, relation extraction, protein-protein interaction, mutation, information extraction.***

## I. INTRODUCTION

The goal of the BioCreative challenges (1-8) has been to propose tasks that will bring together text mining community and biology researchers in order to foster the development of systems that can help with biologically relevant problems. One such current research area is precision medicine, an emerging approach for disease treatment and prevention that takes into account variability in genes, environment and lifestyle for each person. Because the intricate network of interactions between genes contributes to control cellular homeostasis, differences in interaction stability, although not resulting in any obvious phenotype, can contribute to the development of diseases in specific contexts. Annotating how gene mutations or variations affect the global behavior of the cellular interaction provides additional support to precision medicine efforts.

Such information can be found in the unstructured text within the scientific articles indexed in PubMed (9-14). Specialized curation databases, such as IntAct and BioGRID have been collecting and cataloging knowledge focused on particular areas of biology so that they may enable insights into conserved networks and pathways that are relevant to human health. Expanding their curation efforts into capturing specific sequence-variant-depended molecular interactions may open up new possibilities and enable insights that pertain to precision medicine. To date, no tool is available to facilitate this kind of specific retrieval. The goal of our track is to foster the development of text mining algorithms that specialize in scanning the published biomedical literature and are capable to extract the reported discoveries of protein interactions changing in nature due to the presence of genomic variations or artificial mutations.

## II. THE PRECISION MEDICINE TRACK

The Precision Medicine Track in BioCreative VI is a community challenge that addresses this problem in the form of two tasks:

- Document Triage: Identification of relevant PubMed citations describing mutations affecting protein-protein interactions (PPI).
- Relation Extraction: Extraction of experimentally verified PPI pairs affected by the presence of a genetic mutation.

### *A. Training and testing datasets*

Our first step was the curation of a manually annotated corpus that could be used for the training, tuning and development of text mining algorithms for such a specialized task. Our research on creating and developing our training corpus (15) showed that biomedical literature is ripe with precision medicine relevant information.

TABLE 1 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 |

We verified this by retrieving articles that were publicly available in expert curated databases, and re-evaluated them for precision medicine purposes. These articles were rich in information of molecular interactions that differ based on the presence of a specific genetic variant, information which could be translated to clinical practice. Moreover, we retrieved articles via state-of-the-art test mining tools that described PPI and contained identifiable sequence variants. Manual curation, again, verified their relevance.

As a result, we released a set 4,082 PubMed abstracts as the Precision Medicine that come from two different sources: curated databases [1] and text mining tool selection. PubMed articles selected from both sources had slightly different, but useful characteristics and as such, novel text mining tools need to use both sources of information for best application in this new domain.

Each article in the precision medicine training dataset was annotated for relevance, and a subset of relevant articles was annotated for relation extraction. Each article annotated for relation extraction contained the relation annotations for the interacting pair of proteins which were affected by mutations identified via their Entrez Gene[2] IDs, and in addition, contained the mention annotations of the interacting genes in the PubMed abstract.

As track participants worked on their text mining models, five BioGRID [3] (16) curators worked on annotating the test dataset. As a result of this effort, 1,464 PubMed articles were annotated by at least two curators for relevance and the interacting genes affected by the presence of a mutation were recorded as interacting pairs in 734 articles. Similarly, to the training set, the relations were described as a pair of Entrez Gene identifiers. Statistics of the dataset are shown in Table 1.

Finally, track participant teams were provided with the raw text of the 1,464 PubMed articles[4] in the test dataset and were asked to return:

- A PubMed article label (relevant/not relevant) for the triage task.
- Pairs of Entrez Gene identifiers, for the relation extraction task.

Each task participant could contribute up to three runs per task and this participation is shown in Table 2.

TABLE 2 PARTICIPATING TEAMS AND THEIR SUBMISSIONS

| Team Number | Triage Task | Relation Task |
|---|---|---|
| 374 | 3 | |
| 375 | 3 | 3 |
| 379 | 1 | 2 |
| 391 | | 3 |
| 405 | 1 | 2 |
| 414 | 3 | |
| 418 | 3 | |
| 419 | 3 | |
| 420 | 1 | 3 |
| 421 | 3 | |
| 433 | 1 | 1 |
| **Total** | **10 teams/22 runs** | **6 teams/14 runs** |

### *B. Evaluation*

For the final evaluation of the participating runs, text mining predictions were compared to manually annotated data using the standard evaluation procedures: precision, recall, F-score, and average precision. To assist the participants, the organizers set up a group e-mail list where information about the task was posted periodically and several discussions were held.

Organizers also set up a PubTator[5] (17) view, so that track participants could visualize the training data annotations. For the evaluation phase, participant teams were provided with the evaluation scripts to use on their results. The evaluation scripts also served as a self-check to ensure that the data was submitted in the correct format for evaluation. Results, in the forms of system output in BioC (18) format (XML/JSON) and a short paragraph description of the applied method, were submitted via e-mail. For each Results Run, organizers asked participants to submit confidence scores for their predictions, which facilitated the ranking results.

For the Relation extraction task, organizers employed a two-level evaluation:

[1] https://www.ebi.ac.uk/intact/
[2] https://www.ncbi.nlm.nih.gov/gene/
[3] https://thebiogrid.org/

[4] 1,500 PubMed articles were initially released as the test set. However, 36 articles that were difficult to assign labels were later removed, and not used for official evaluation.

[5] https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/PubTator/

- **Exact Match**: all system predicted relations were checked against the manual annotated ones for correctness. A PPI relation is not defined as directional in this annotation format, so the order of genes is not considered when checking for exact match.
- **HomoloGene [6] Match**: All gene identifiers in the predicted relations and manually annotated data were mapped to common identifiers representing common HomoloGene classes, then all predicted relations were checked for correctness. If the predicted Gene ID and the annotated Gene ID were homologous genes, they were counted as a match.

### *C. Benchmarking systems*

For comparison purposes, we developed a baseline method for both Triage and Relation Extraction Tasks. For the Triage Task, we designed a baseline SVM classifier using unigram and bigram features from titles and abstracts of the training corpus (15). For the Relation Extraction Task, we implemented a simple co-occurrence baseline method. The Gene entities in the PubMed articles were automatically recognized using our in-house tools (19-23), and a relation was predicted if two gene entities were found in the same sentence, regardless of whether a sequence variant had been predicted for that article or not.

## III. RESULTS

Eleven teams participated in the Precision Medicine Track: ten teams in the document triage task, and six teams in the relation extraction task. Since each team could submit up to 3 runs (i.e. 3 different versions of their tool, or contribute three different methods) for each task, a total of 36 runs were submitted. Participants were from Australia, China, Turkey, Greece, Germany, Portugal and the United States.

For the triage task, we received results of 22 systems (shown in Table 3), 16 of which outperformed our baseline in F-score, 13 on average precision, 2 on precision, and 17 on recall. The best F-score is 69.5%, the best average precision is 72.8%, the best precision is 61.6% and the best recall is 98.0%. The average F-score, average precision, precision and recall were 64.1%, 63.5%, 56.6% and 75.4% respectively.

TABLE 3. TRAGE TASK RESULTS FOR ALL SUBMISSIONS

| Team Number | Submission | Avg Prec | Precision | Recall | F1 | Data Format |
|---|---|---|---|---|---|---|
| 374 | Run 1 | 0.6598 | 0.5916 | 0.8315 | 0.6913 | JSON |
| | Run 2 | 0.6654 | 0.5747 | 0.8699 | 0.6921 | JSON |
| | Run 3 | 0.6930 | 0.6092 | 0.7836 | 0.6854 | JSON |
| 375 | Run 1 | 0.6808 | 0.5821 | 0.7575 | 0.6583 | JSON |
| | Run 2 | 0.6688 | 0.5946 | 0.6973 | 0.6419 | JSON |
| | Run 3 | 0.6750 | 0.5416 | 0.8822 | 0.6712 | JSON |
| 379 | Run 1 | 0.4885 | 0.4622 | 0.3438 | 0.3943 | XML |
| 405 | Run 1 | 0.5877 | 0.5478 | 0.5575 | 0.5526 | JSON |
| 414 | Run 1 | 0.4886 | 0.4792 | 0.5849 | 0.5268 | XML |
| | Run 2 | 0.5055 | 0.4957 | 0.7178 | 0.5865 | XML |
| | Run 3 | 0.5098 | 0.5075 | **0.9795** | 0.6685 | XML |
| 418 | Run 1 | 0.6973 | **0.6164** | 0.7616 | 0.6814 | XML |
| | Run 2 | 0.7083 | 0.5988 | 0.8096 | 0.6884 | XML |
| | Run 3 | 0.7195 | 0.6026 | 0.8205 | **0.6949** | XML |
| 419 | Run 1 | 0.5742 | 0.5718 | 0.8068 | 0.6693 | XML |
| | Run 2 | 0.6010 | 0.5905 | 0.5986 | 0.5946 | XML |
| | Run 3 | 0.6330 | 0.5989 | 0.6096 | 0.6042 | XML |
| 420 | Run 1 | 0.6439 | 0.5473 | 0.8712 | 0.6723 | JSON |
| 421 | Run 1 | 0.6687 | 0.5890 | 0.8068 | 0.6809 | XML |
| | Run 2 | **0.7284** | 0.6112 | 0.7945 | 0.6909 | XML |
| | Run 3 | 0.7103 | 0.5882 | 0.8219 | 0.6857 | XML |
| 433 | Run 1 | 0.6617 | 0.5482 | 0.8877 | 0.6778 | JSON |
| BASELINE | - | 0.6500 | 0.6097 | 0.6356 | 0.6224 | - |

[6] https://www.ncbi.nlm.nih.gov/homologene

TABLE 4. RELATIONS TASK HOMOLOGENE RESULTS FOR ALL SUBMISSIONS

| System | Submission | Precision | Recall | F1 | Data Format |
|---|---|---|---|---|---|
| 375 | Run 1 | 0.3761 | 0.3527 | 0.3640 | XML |
| | Run 2 | 0.3761 | 0.3527 | 0.3640 | XML |
| | Run 3 | 0.4252 | 0.3301 | 0.3717 | XML |
| 379 | Run 1 | 0.3057 | 0.0753 | 0.1208 | XML |
| | Run 2 | 0.1135 | **0.5387** | 0.1875 | XML |
| 391 | Run 1 | 0.2360 | 0.1989 | 0.2159 | XML |
| | Run 2 | 0.2343 | 0.1882 | 0.2087 | XML |
| | Run 3 | 0.2378 | 0.1366 | 0.1735 | XML |
| 405 | Run 1 | 0.0767 | 0.0247 | 0.0374 | JSON |
| | Run 2 | 0.1061 | 0.0376 | 0.0556 | JSON |
| 420 | Run 1 | 0.4321 | 0.2978 | 0.3526 | JSON |
| | Run 2 | **0.4544** | 0.3161 | **0.3729** | JSON |
| | Run 3 | 0.4286 | 0.3097 | 0.3596 | JSON |
| 433 | Run 1 | 0.0803 | 0.2742 | 0.1242 | JSON |
| BASELINE | - | 0.1460 | 0.5215 | 0.2282 | |

For the relations task, we received results from 14 systems, 8 of which outperformed the baseline based on the F-score, 10 on precision, and 1 on recall, when evaluating on Exact Match.

The HomoloGene evaluation showed a slightly different result: 6 systems outperformed the baseline on F-score, 10 on precision, and only one on recall. The average F-score, precision and recall for the HomoloGene evaluation were 23.6%, 27.7% and 24.5%, respectively. The best F-score, precision and recall were 37.3%, 45.4% and 53.9%, respectively. These results are shown in Table 4.

## IV. DISCUSSION AND CONCLUSIONS

Given the level of participation and team results we conclude that the precision medicine track of BioCreative VI was run successfully and is expected to make significant contributions in this novel challenge of mining protein-protein interactions affected by mutations from scientific literature. The training and testing data produced during this effort is novel and substantial in size. Collectively, it consists of 5,546 PubMed articles manually annotated for precision medicine relevance. In addition, the corpus annotations include both text spans and normalized concept identifiers for each of the interacting genes in the mutation-affected PPI relations. We believe that such data will be invaluable in fostering the development of text-mining techniques that increase both precision and recall for such tasks. Another important characteristic is that annotated relations in this corpus are at the abstract level because the majority of such relations are expressed across sentence boundaries.

Participating teams developed systems that specialized in predicting PubMed articles that contain precision-medicine relevant information. Curators at molecular interaction databases will benefit from these text mining systems to select with high accuracy articles relevant for curation. The top achieved recall was 98% and the top achieved precision was 62%. And this is only a first step in this direction. In the future, we plan to build a system that can intelligently merge the results of all individual system submissions with better accuracy.

The relation extraction task on the other hand, showed a somewhat low accuracy. It is to be recognized that this is a very difficult task, as we also showed on the corpus description paper. And, we believe that the accuracy of systems would improve if they were to extract such information from full text. Relation extraction at the abstract level is dependent on accurate entity recognition and correct normalization, as well as the ability to recognize a relation that spans over sentence boundaries, therefore necessitating a system that goes towards abstract-level understanding.

This community effort was designed to foster development of text mining tools that while mining scientific literature could collect information of significant practical value in the clinical practice of precision medicine. The success of the precision medicine endeavor depends on the development of comprehensive knowledge base systems that integrate genomic and sequence variation data, information that lead to tumors and other possible genetic disorders, with clinical response data and outcomes information, as resources for scientists, health care professionals and patients. Leveraging the information already available in scientific literature, and developing automatic text mining methods that facilitate the job of database curators to be able to find and curate such valuable information, is the first step towards this goal.

## ACKNOWLEDGEMENTS

This research was supported by the NIH Intramural Research Program National Library of Medicine. Andrew Chatr-aryamontri is supported by National Institutes of Health Office of Research Infrastructure Programs [R01OD010929 and R24OD011194]. We thank the BioGRID database curators: Rose Oughtred, Jennifer Rust, Christie S. Chang, Lorrie Boucher, and Andrew Chatr-aryamontri for annotating the dataset.

## REFERENCES

1. 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.

2. 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.

3. 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.

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

5. Arighi, C.N., Lu, Z., Krallinger, M., Cohen, K.B., Wilbur, W.J., Valencia, A., Hirschman, L. and Wu, C.H. (2011). Overview of the BioCreative III Workshop. *BMC bioinformatics.* **12 Suppl 8**: S1.

6. 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.

7. Comeau, D.C., Batista-Navarro, R.T., Dai, H.J., Dogan, R.I., Yepes, A.J., Khare, R., Lu, Z., Marques, H., Mattingly, C.J., Neves, M., Peng, Y., Rak, R., Rinaldi, F., Tsai, R.T., Verspoor, K., Wiegers, T.C., Wu, C.H. and Wilbur, W.J. (2014). BioC interoperability track overview. *Database (Oxford).* **2014**.

8. 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.*

9. Singhal, A., Simmons, M. and Lu, Z. (2016). Text Mining Genotype-Phenotype Relationships from Biomedical Literature for Database Curation and Precision Medicine. *PLoS Comput Biol.* **12**(11): e1005017.

10. Caporaso, J.G., Baumgartner, W.A., Jr., Randolph, D.A., Cohen, K.B. and Hunter, L. (2007). MutationFinder: a high-performance system for extracting point mutation mentions from text. *Bioinformatics.* **23**(14): 1862-1865.

11. Cejuela, J.M., Bojchevski, A., Uhlig, C., Bekmukhametov, R., Kumar Karn, S., Mahmuti, S., Baghudana, A., Dubey, A., Satagopam, V.P. and Rost, B. (2017). nala: text mining natural language mutation mentions. *Bioinformatics.* **33**(12): 1852-1858.

12. Horn, F., Lau, A.L. and Cohen, F.E. (2004). Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. *Bioinformatics.* **20**(4): 557-568.

13. Mahmood, A.S., Wu, T.J., Mazumder, R. and Vijay-Shanker, K. (2016). DiMeX: A Text Mining System for Mutation-Disease Association Extraction. *PLoS One.* **11**(4): e0152725.

14. Saunders, R.E. and Perkins, S.J. (2008). CoagMDB: a database analysis of missense mutations within four conserved domains in five vitamin K-dependent coagulation serine proteases using a text-mining tool. *Hum Mutat.* **29**(3): 333-344.

15. 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).*

16. 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.

17. Wei, C.H., Kao, H.Y. and Lu, Z. (2013). PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res.* **41**(Web Server issue): W518-522.

18. Comeau, D.C., Islamaj Dogan, R., 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. and Wilbur, W.J. (2013). BioC: a minimalist approach to interoperability for biomedical text processing. *Database (Oxford).* **2013**: bat064.

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

20. 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.

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

22. Wei, C.H., Kao, H.Y. and Lu, Z. (2015). GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains. *Biomed Res Int.* **2015**: 918710.

23. 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.*