BioCreative VI Precision Medicine Track: creating a training corpus for mining protein-protein interactions affected by mutations

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

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

[2]Institute for Research in Immunology and Cancer, Université de Montréal, Montréal, QC H3C 3J7, Canada

Abstract

The Precision Medicine Track in BioCreative VI aims to bring together the BioNLP community for a novel challenge focused on mining the biomedical literature in search of mutations and protein-protein interactions (PPI). In order to support this track with an effective training dataset with limited curator time, the track organizers carefully reviewed PubMed articles from two different sources: curated public PPI databases, and the results of state-of-the-art public text mining tools. We detail here the data collection, manual review and annotation process and describe this training corpus characteristics. We also describe a corpus performance baseline. 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.

1 Introduction

Genomic technologies now make possible the routine sequencing of individual genomes and such data makes possible to understand how genetic variations are distributed in healthy and sick populations. On the other hand, proteomics and metabolomics approaches are charting the metabolic and interactions maps of the cell. Such data deluge has generated great expectations in the cure of human diseases. Nonetheless, it is still difficult to predict the phenotypic outcome of a specific genome and designing the most appropriate treatment or establishing preventive programs. Linking allelic variation and genomic mutations to protein-protein interactions (PPI) is crucial to understand how cellular networks rewire and to support personalized medicine approaches.

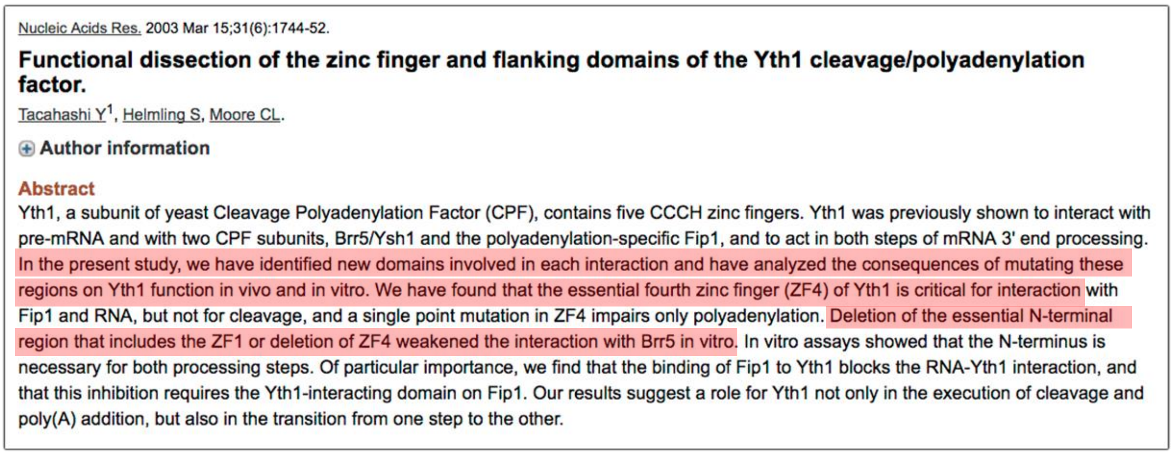
To date, no tool is available to facilitate the specific retrieval of such information that remains buried in the unstructured text within the biomedical literature. Our goal is to foster the development of text mining algorithms that specialize in scanning the published biomedical literature and to extract the reported discoveries of protein interactions changing in nature due to the presence of a genomic variations or artificial mutations.

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
- Relation Extraction: Extraction of experimentally verified PPI pairs affected by the presence of a genetic mutation

Traditionally biological database curators have contributed to the various BioCreative challenges (Hirschman, Yeh et al. 2005, Chatr-aryamontri, Kerrien et al. 2008, Krallinger, Morgan et al. 2008, Lu and Hirschman 2012) supporting the identification of stages in the curation workflow suitable for text mining applications and manually annotating the training and test corpora. Because the manual curation of the current exponentially growing body of biomedical literature is an impossible task, the insertion of robust text mining tools in the curation pipeline represent a feasible and sustainable solution to this problem (Hirschman, Burns et al. 2012).

Proceedings of the BioNLP 2017 workshop, pages 171–175,
Vancouver, Canada, August 4, 2017.



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 A PubMed article describing a protein-protein interaction affected by mutation

As we prepared to create our corpus we faced the common situation of limited reviewer time. We took two steps to maximize this limited, valuable resource: First, we reviewed annotations readily available from manually curated PPI databases (Orchard, Ammari et al. 2014) and marked the relevant publications that could be used for the purposes of this challenge; next, we expanded the training set using a set of publically available text mining tools (Kim, Kwon et al. 2012, Wei, Harris et al. 2013) specifically for the retrieval of literature reporting protein interaction and mutation data.

Both of these sets were manually reviewed and categorized as: 1) Articles describing PPI and mutations affecting those molecular interactions, 2) Articles describing mutations and molecular interactions, with no affect or no relation between the two events, 3) Articles describing PPI, 4) Articles describing mutations or genetic variation, and 5) Articles not relevant for either molecular interaction or mutation information. In addition, the database extracted interactions were carefully reviewed and validated in two important aspects: 1) the annotated PPI were described in the PubMed abstract of the corresponding article, as opposed to the full text, and 2) the extracted interactions were affected by a mutation, and this was stated in the abstract.

All manually selected, categorized and carefully reviewed articles make up a set of 4,082 PubMed abstracts. All of these articles can be used for building machine learning methods and other innovative applications for the Precision Medicine Track in BioCreative VI. Of these, 598 PubMed articles are annotated with specific interactions. This smaller set can be used to develop algorithms for the Relation Extraction task and other similar biomedical text mining problems.

We provide here a detailed description of the assembly of this dataset and report the on-going efforts of building the test corpus.

2 Training Corpus

The Precision Medicine track training corpus was generated as a result of two data selection and validation methods:

- Data repurposing
- Text mining triage and manual validation

These approaches are different and as noticed in the article composition resulting from each of them, they are both important contributors to this dataset. Here we describe the procedure followed in each of these approaches, starting with our annotation guidelines and a detailed view of the corpus characteristics. Figure 1 shows an example article in our dataset.

2.1 Annotation guidelines

All selected articles were manually annotated to answer these questions:

- Does this article describe experimentally verified protein-protein interactions?
- Does this article describe a disease known mutation or a mutational analysis experiment?
- Are the database curated PPI pairs for this article mentioned in the abstract?

Table 1 Training Set annotation and distribution amongst different categories

Annotation Category	Curated database selected articles (PPI set)	Text mining tools selected articles (TM set)	Complete Training Set	
True positives	1079	651	1,730	42%
True Negatives	55	322	377	9%
Negative, Yes PPI, No Mutation	1538	82	1,620	40%
Negative, No PPI, Yes Mutation	136	87	99	2.4%
Negative, No PPI, No Mutation	12	120	256	6.3%
Total	2820	1262	4082	100%

- Is the PPI affected by the mutation?

Then, based on the above annotations, articles are carefully categorized as 1) True Positives, for articles specifically describing PPI influenced by genetic mutations, 2) True Negatives, for articles describing both PPIs and genetic variation analysis with no inference of relation between them, 3) articles containing PPI but no mutations, 4) articles containing mutations but no PPI, and 5) articles mentioning neither.

2.2 Curated Database article selection

The IntAct Molecular Interaction Database (Orchard, Ammari et al. 2014) is a freely available, open source database system and analysis tool for molecular interaction data. It currently lists 14,584 manually annotated PubMed full-text articles with 720,711 molecular interactions for 98,289 different interactors. The curation of these molecular interactions is captured at a required level of detail and frequent updates include mapping to binding regions, point mutations and post-translational modifications to a specified sequence with a reference protein sequence database.

A set of 2,852 articles, containing in-the-abstract information about binding interfaces and mutations influencing the interactions, was retrieved from IntAct and these articles went through a careful review and validation round by an experienced curator. Each one of these articles was carefully considered for their suitability for the precision medicine task.

A second manual validation round was then performed on all positively annotated articles of the first round. As a result, 598 articles were identified as relevant for the Relation Extraction task, with experimentally verified interactions influenced by mutations and with explicit interactors in the abstract. All of these interactors were expressed with both their UniProt ID and Gene Entrez ID. The non-relevant articles were further categorized into the more specific categories as described above.

2.3 Text Mining based article selection

The Text Mining approach used two well-known publically available text mining tools: PIE the search (Kim, Kwon et al. 2012) and tmVar (Wei, Harris et al. 2013). PIE[1] the search is a web service that provides an alternate way of querying PubMed for biologists and database curators. The returned articles are ranked based on their probability of describing protein-protein interactions, using a very competitive algorithm and the winner of BioCreative III ACT competition (Krallinger, Vazquez et al. 2011). tmVar[2] is another text mining tool that is the current gold-standard for recognizing sequence variants in PubMed literature. An article marked by tmVar signals the presence of a sequence variant of a mutation in the title and abstract.

These tools were used as follows:

- Step 1: PIE the search was used to select the top scoring (for PPI) PubMed articles published in the last 10 years. This method selected over 13,000 articles.

- Step 2. tmVar was used on the resulting set of Step 1 to select all articles which had a sequence variant in the title or abstract. This method selected around 1,200 articles.

[1] https://www.ncbi.nlm.nih.gov/CBBresearch/Wilbur/IRET/PIE/

[2] https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/tmTools/#tmVar

Table 2 Document Triage Task results

Methods	Avg. Prec.	Precision	Recall	F1	Positive	Negative	Ratio
10-fold CV (PPI set)	0.7577	0.7184	0.6321	0.6725	1079	1741	38%
Validation (TM set)	0.6551	0.6210	0.6897	0.6536	651	611	52%
10-fold CV (all data)	0.7225	0.6891	0.6260	0.6561	1730	2352	42%

- Step 3. All articles in Step 2 were manually annotated as described in the annotation guidelines.

3 Results and Discussion

3.1 Precision Medicine Task Training Corpus Characteristics

The Precision Medicine Task training corpus contains 4,082 selected PubMed abstracts that come from two different sources: curated databases and text mining tool selection. It is important to see the dataset as a whole and to notice the different composition of classified articles coming from both sources as detailed in Table 1.

In addition, we looked at the PIE score distribution of all articles in the dataset. We noticed that the PubMed articles selected via text mining tools had a higher PIE score average than the articles retrieved from curated databases. In particular, while the PIE scores of the articles selected from the curated databases form a normal distribution, the scores of the text mining selected articles are skewed towards high scores.

On a different experiment, we ran the tmVar tool on all curated database selected articles. Interestingly, only 311 out of 1079 positives articles were marked by tmVar.

Thus, if novel algorithm developers only gave more importance to articles selected via text mining tools, or only the text mining tools used in our experiment, they risk biasing curators to only a particular set of articles. Innovative text mining tools should make use of both sets of articles in order to ensure a better coverage of curatable articles.

3.2 Benchmark results and corpus use

A baseline SVM method was designed using unigram and bigram features from titles and abstracts of the training corpus, as shown in the results in Table 2. A first experiment used articles from the curated database for training in a 10-fold cross validation (CV) setting, and tested on the text mining selected articles. And a second experiment mixed all articles in a 10-fold cross validation setting. Results are detailed in Table 2.

The test dataset for BioCreative VI Precision Medicine Track will be a set selected by database curators. First articles will be retrieved via text mining tools and then each article will be manually evaluated by four experienced curators.

4 Conclusions and Public Availability

A vast amount of precision medicine related information can be found in published literature and extracted by skilled domain expert curators. The BioCreative VI Precision Medicine Track corpus characteristics provide important insights on 1) understanding the structure of biological information and why it is relevant for precision medicine purposes, and 2) the best practices for designing computational automatic methods capable of extracting such information from unstructured text.

By releasing this data we aim to facilitate the curation of precision medicine related information available in published literature. This corpus fosters 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.

This data comes from two realistic, important data sources: 1) articles retrieved from expert curated PPI databases, re-evaluated and found useful for precision medicine purposes, and 2) articles retrieved via state-of-the-art text mining tools trained to identify articles describing PPI and containing identifiable sequence variants. Both sets of data have 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.

The BioCreative VI Precision Medicine training corpus will be available to task participants from the BioCreative website and later to the whole scientific community.

5 References

Chatr-aryamontri, A., S. Kerrien, J. Khadake, S. Orchard, A. Ceol, L. Licata, L. Castagnoli, S. Costa, C. Derow, R. Huntley, B. Aranda, C. Leroy, D. Thorneycroft, R. Apweiler, G. Cesareni and H. Hermjakob (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.

Hirschman, L., G. A. Burns, M. Krallinger, C. Arighi, K. B. Cohen, A. Valencia, C. H. Wu, A. Chatr-Aryamontri, K. G. Dowell, E. Huala, A. Lourenco, R. Nash, A. L. Veuthey, T. Wiegers and A. G. Winter (2012). "Text mining for the biocuration workflow." Database (Oxford) **2012**: bas020.

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

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

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

Krallinger, M., M. Vazquez, F. Leitner, D. Salgado, A. Chatr-aryamontri, A. Winter, L. Perfetto, L. Briganti, L. Licata, M. Iannuccelli, L. Castagnoli, G. Cesareni, M. Tyers, G. Schneider, F. Rinaldi, R. Leaman, G. Gonzalez, S. Matos, S. Kim, W. Wilbur, L. Rocha, H. Shatkay, A. Tendulkar, S. Agarwal, F. Liu, X. Wang, R. Rak, K. Noto, C. Elkan and Z. Lu (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.

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

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

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