Identifying genetic interaction evidence passages in biomedical literature

Rezarta Islamaj Doğan[1], Sun Kim[1], Andrew Chatr-Aryamontri[2], Donald C. Comeau[1]
and W. John Wilbur[1]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda MD 20854 USA
{Rezarta.Islamaj, Sun.Kim}@nih.gov
{Comeau, Wilbur}@ncbi.nlm.nih.gov
[2]Institute for Research in Immunology and Cancer, Université de Montréal, Montréal, Quebec
H3C 3J7, Canada
andrew.chatr-aryamontri@umontreal.ca

Abstract. In this work, we report our contributions to the BioC Track of BioCreative V for the task of identifying genetic interaction evidence passages. Text describing genetic interactions is difficult to identify due to no simple definition for these interactions and lack of training data. We prepared two manually annotated datasets containing 1793 PubMed® abstract and 1000 full text sentences, respectively. We also built two classification systems to identify genetic interaction evidence, one based on word and context features, and one based on query features used for genetic evidence information retrieval. Both models gave satisfactory results on our manually annotated datasets and we produced four different runs, which were submitted for inclusion in the complete BioC Track system. Identification of genetic interactions in biomedical text is a challenging problem with much work still needing to be done.

1 Introduction

The understanding of the biological systems that make up the human body and are affected by human disease requires a thorough knowledge of the innumerable biological interactions that take place under various conditions and circumstances. BioGRID[1] comprehensively annotates and compiles biological interaction data in the biomedical literature (1). This data includes genes, proteins and their interactions for all model organism species.

Clearly, the sheer size of the biomedical literature and its growth rate, render the unassisted manual curation of this data simply impossible. The BioC Track (2,3) in BioCreative V consisted of participating teams working in collaboration to develop text mining techniques that could ease the BioGRID curators' daily job.

[1] http://thebiogrid.org

One of the objectives of the BioC Track[2] was the identification of sentences in full text articles claiming genetic interactions and, in particular, sentences that provided evidence for a genetic interaction (GI). The major obstacle was the lack of training data. Because there is no simple definition of genetic interactions (4), there is still some difficulty in defining our task[3]. Here we describe our efforts to address task 7 of BioCreative V Track I. Our contributions are as follows:

- We manually curated two datasets for GI interactions and extracted GI evidence expressions in text.
- We studied and catalogued negation features and speculation features.
- We built two SVM machine learning methods that used a rich combination of features to predict genetic interactions.
  - We built a context-feature SVM classification system with many feature types, and
  - We introduce a novel machine learning classification system with features derived from Lucene search.

2 Methods

2.1 Genetic Interaction Datasets

Here we describe the datasets that we built for GI training and evaluation of our four models.

We started with all the data available from the BioGRID website. We retrieved PubMed IDs of articles that were curated with genetic interactions. From this, we built two datasets. The first dataset, or the abstract dataset, contained 1793 sentences from 819 PubMed abstracts, of which we manually labelled 611 sentences that described genetic interactions (the rest were considered as negative). The second dataset, or the full text dataset, contained 1000 sentences from 39 full text articles, of which we manually labelled 373 sentences that described genetic interactions (the rest were marked as negative).

2.2 Manual curation

For the manual curation we built new tools and used a combination of our available tools, such as the PMID2BioC tool to download the list of abstracts from PubMed and create a BioC collection. Two tools specifically developed for this task are shown in Figures 1 and 2.

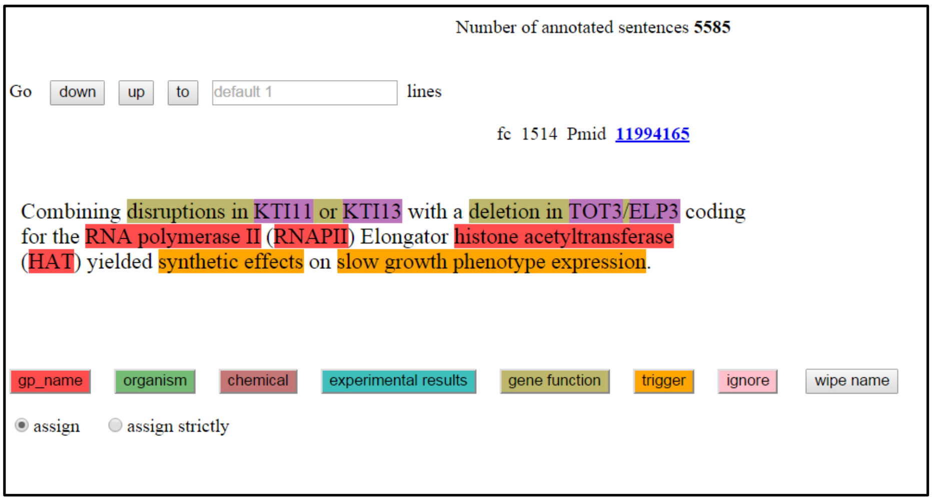
The first tool was used to curate our abstract dataset. For this, we traversed all sentences in the PubMed abstracts that were retrieved from the BioGRID database, and

[2] http://www.biocreative.org/tasks/biocreative-v/track-1-bioc/

[3] http://wiki.thebiogrid.org/doku.php/experimental_systems

automatically pre-annotated all sentences that contained a gene pair from a single entry in BioGRID pertaining to that article. Then, these sentences were loaded into this annotation tool (Figure 1), and the curator annotated with the categories seen in the figure.

One important piece of information that comes with the curation of these sentences is the off-brown and orange marked text. We initially called these segments "gene function" and "trigger", and that is how they are shown in Figure 1. In our later experiments, we found these expressions to be the most useful features in identifying a genetic interaction sentence. We collected a total of 629 unique such patterns when we anonymized the gene names found within.



Number of annotated sentences **5585**

Go down up to default 1 lines

fc 1514 Pmid 11994165

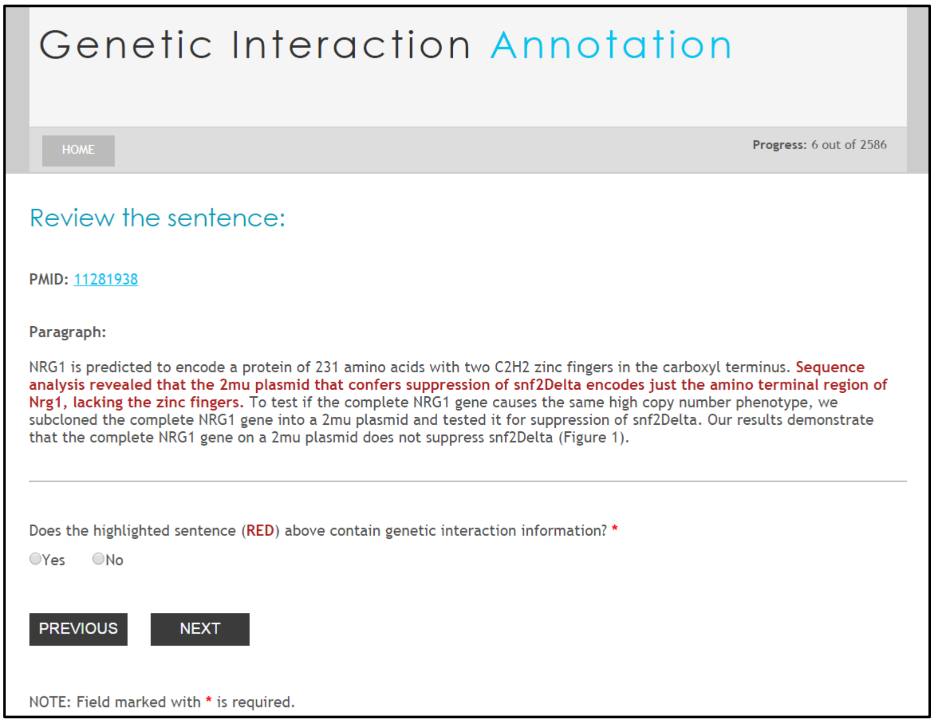
Combining disruptions in KTI11 or KTI13 with a deletion in TOT3/ELP3 coding for the RNA polymerase II (RNAPII) Elongator histone acetyltransferase (HAT) yielded synthetic effects on slow growth phenotype expression.

● assign ○ assign strictly

Fig. 1. A screenshot of the first tool, which allows the curator to traverse one sentence at a time, go to a specific sentence, and mark segments of text, such as Gene (or Protein) Name (gp_name), Organism Name (organism), Chemical Name (chemical), etc., as well as remove a previous annotation (wipe name) or mark a text as not important for our purposes (ignore). In the example shown in the figure, the segments marked in off-brown describe a state of the gene function and the segments marked in orange describe a phenotypic effect of those genetic states, and in turn we believe that the whole sentence describes a genetic interaction.

Since full text is different than abstract text, and the BioC Track aims to provide tools for full text processing for genetic interactions, we built two modules to try and enrich our dataset with full text sentences. The first module found PubMed Central® sentences that were similar to the genetic interaction descriptions on the BioGRID page. The second module used the abstract dataset as a training set, and learned a basic SVM model to distinguish GI sentences from the rest of the sentences. Both modules were applied to full text articles and a set of sentences that scored high for both systems was selected. These sentences were loaded into the second annotation tool, and displayed with their surrounding context as shown in Figure 2. We asked the curator to simply mark if the sentence of interest contained a genetic interaction description. From this tool we collected annotations for 1000 sentences, which constituted our second dataset.

We plan to make both these datasets available to the community. For our automatic genetic interaction identification systems which we describe below, we combined these datasets in a 10 fold cross validation model so that all the abstract sentences and 9/10ths of the full text sentences were used for training, and the remaining 1/10th of the full text sentences were used for testing. All systems were optimized in this manner, and we describe the differences between them below.



Progress: 6 out of 2586

Review the sentence:

PMID: 11281938

Paragraph:

NRG1 is predicted to encode a protein of 231 amino acids with two C2H2 zinc fingers in the carboxyl terminus. **Sequence analysis revealed that the 2mu plasmid that confers suppression of snf2Delta encodes just the amino terminal region of Nrg1, lacking the zinc fingers.** To test if the complete NRG1 gene causes the same high copy number phenotype, we subcloned the complete NRG1 gene into a 2mu plasmid and tested it for suppression of snf2Delta. Our results demonstrate that the complete NRG1 gene on a 2mu plasmid does not suppress snf2Delta (Figure 1).

Does the highlighted sentence (RED) above contain genetic interaction information? *

○Yes ○No

NOTE: Field marked with * is required.

Fig. 2. A screenshot of the second tool which shows a sentence marked in red and the context (the whole paragraph extracted from the full text article, in which the sentence is found). The annotator is asked to label the sentence as containing a genetic interaction or not. The annotator is able to traverse to all sentences using the tool, and can also click on the PMID and view the whole article in a different window if needed.

2.3 Genetic interaction identification system description

In order to classify a given sentence as containing genetic interaction information or not, we built two successful SVM models. We produced two runs using a context- feature SVM model, and two runs using an information-retrieval-based classification model. The four result sets are produced from classification models that differ in the sets of features that they use, and how they combine these features. Here we describe their similarities and their differences.

The context-feature SVM approach used the following process: 1) identify possible gene mentions and anonymize them with a special token, 2) predict negated phrases via NegEx(5) and replace them with a special token, 3) extract unigram, bigram, negation and speculation information features (6), 4) train and test an SVM classifier by using the extracted features. We produced two runs for this rich-feature approach. The difference between these two runs was that the second run used manually created rules for gene mention and GI trigger as extra features.

The basics of the information-retrieval-based classification system (IR-based SVM) are as follows: 1) index the training data using the Lucene model with default settings, 2) identify query phrases (gene function and trigger patterns from the abstract dataset, negation cues (5-7), speculation cues (7) and Interaction Network Ontology[4] literature mining keywords (8,9), 3) query the training data and pair all retrieved sentences with the query and its score. Once we collect the query set and scores for each sentence in the training set, then this interim dataset can serve as the features of a new SVM classification system. The difference between our two submitted runs was that, for the second run, we used aggregate features for each feature type as extra features.

3 Results and Discussion

Table 1 lists the performance results of the baseline and four models that our team submitted for the genetic interaction identification task in the BioC Track of BioCreative V. These results are computed by 10-fold cross-validation on the set of full text sentences with all the abstract sentences included as part of the training set for each fold. Precision, recall and F-score is computed at the score-threshold of 0.

Our team applied these results on the evaluation set of articles for the BioGRID. Initially articles were divided into sentences which were assigned unique IDs. Then each sentence was processed for the feature set that each system required. For the IR-based systems, these new sentences became the new set of indexed documents in which the query search was performed. The learned models then were applied and each sentence was scored. These results were written in the BioC format and submitted for inclusion in the complete system.

While genetic interaction and extraction of genetic networks has seen a lot of research in systems biology and other areas, the text mining methods for genetic identification are by contrast very few, perhaps due to the lack of annotated data for this task and other difficulties arising from not-quite simple definitions. Thus, our efforts were equally divided between developing a dataset that we could use to build a machine learning system, and identifying features that could be useful in such a text mining task. While as a result of this challenge we came out with two original datasets and four

[4] http://bioportal.bioontology.org/ontologies/INO

trained systems, we still feel that the text mining research in the field is barely starting, so there remains much more work to be done.

Table 1. Performance results of our gene interaction identification systems on our datasets (10-fold cross validation results)

Genetic Interaction system	AvePrec	Precision	Recall	F-score
Baseline	0.598	0.571	0.633	0.601
IR-based SVM - run 1	0.785	0.748	0.662	0.703
IR-based SVM - run 2	0.781	0.750	0.684	0.715
Context-feature SVM - run 1	0.819	0.730	0.749	0.739
Context-feature SVM - run 2	0.817	0.727	0.742	0.734

Acknowledgments. 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.

4 REFERENCES

1. Chatr-Aryamontri, A., Breitkreutz, B.J., Oughtred, R., *et al.* (2015) The BioGRID interaction database: 2015 update. *Nucleic Acids Res*, **43**, D470-478.
2. Comeau, D.C., Islamaj Doğan, R., Ciccarese, P., *et al.* (2013) BioC: a minimalist approach to interoperability for biomedical text processing. *Database*, **2013**.
3. Comeau, D.C., Batista-Navarro, R.T., Dai, H.J., *et al.* (2014) BioC interoperability track overview. *Database : the journal of biological databases and curation*, **2014**.
4. Mani, R., St Onge, R.P., Hartman, J.L.t., *et al.* (2008) Defining genetic interaction. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 3461-3466.
5. Chapman, W.W., Bridewell, W., Hanbury, P., *et al.* (2001) A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, **34**, 301-310.
6. Vincze, V., Szarvas, G., Mora, G., *et al.* (2011) Linguistic scope-based and biological event-based speculation and negation annotations in the BioScope and Genia Event corpora. *Journal of biomedical semantics*, **2 Suppl 5**, S8.
7. Vincze, V., Szarvas, G., Farkas, R., *et al.* (2008) The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics*, **9 Suppl 11**, S9.
8. Hur, J., Ozgur, A., Xiang, Z., *et al.* (2012) Identification of fever and vaccine-associated gene interaction networks using ontology-based literature mining. *Journal of biomedical semantics*, **3**, 18.
9. Hur, J., Ozgur, A., Xiang, Z., *et al.* (2015) Development and application of an interaction network ontology for literature mining of vaccine-associated gene-gene interactions. *Journal of biomedical semantics*, **6**, 2.