Retrieving Passages Describing Experimental Methods using Ontology and Term Relevance based Query Matching

Ferhat Aydın*, Zehra Melce Hüsünbeyi*, and Arzucan Özgür*

Department of Computer Engineering
Boğaziçi University
TR-34342, Bebek, Istanbul, Turkey
{ferhat.aydin,zehra.husunbeyi,arzucan.ozgur}@boun.edu.tr
http://www.cmpe.boun.edu.tr

Abstract. With the increasing amount of published articles in the biomedical domain, text mining has emerged as a significant research area to extract biologically crucial information such as protein-protein interactions from the scientific literature. The reliability of the extracted protein-protein interactions depends on the methods used to experimentally verify them. We participated in the Collaborative Biocurator Assistant Task (BioC) of the BioCreative V challenge assessment by developing an experimental method detection module as part of the collaborative BioC-compatible text mining system to assist biocurators. Unlike most previous studies, besides identifying the experimental methods in an article, we also target identifying the passages where they are described. Our approach is based on query matching, where the queries are generated using the terms in the PSI-MI ontology and expanded with the most salient terms for each experimental method using the term frequency-relevance frequency (tf.rf) metric over our manually annotated data set.

Key words: interaction method detection, text mining, information retrieval, PSI-MI ontology, term frequency-relevance frequency (tf.rf)

1 Introduction

Protein-protein interactions (PPIs) are central for a variety of biological processes including DNA replication, transcription, translation, cell cycle control, signal transduction, intermediary metabolism, and so forth. A large number of manually curated databases including *BioGrid [5], IntAct [8], DIP [19], MINT [6], and BIND [3]* are created to store information about PPIs in structured format. Manual curation is becoming harder due to the rapidly growing biomedical literature. Therefore, automatically extracting biologically useful information through text mining techniques has become an essential research topic in the bio-text mining community. The community-wide shared tasks including the

* Registered to BioCreative V - Track 1 (BioC) as Team-330.

BioCreative Challenge [9, 2] and the BioNLP Shared Tasks [11] further boosted research in this area.

Experimental methods such as affinity capture, two-hybrid, and coimmunoprecipitation, which are used to detect protein-protein physical interaction, have their own limitations in terms of metrics such as cost, time, and certainty. Therefore, besides extracting the interactions among bio-molecules, identifying the experimental context is also essential for the characterization and biological affirmation of the extracted interactions. The problem of extracting experimental methods for physical protein interactions has been addressed in the BioCreative (Critical Assessment of Information Extraction systems in Biology) Challenge Evaluations [9, 2]. Most previous studies on experimental method detection are based on pattern matching (e.g., [16]) and/or machine learning (e.g., [18, 1]) approaches. A dictionary of experimental method names and their synonyms is typically used in pattern matching approaches to perform exact or approximate string matching. Experimental methods are not identifiable when their canonical names and synonyms do not trivially appear in the text. Such cases require the deduction of the experimental methods from the descriptions of the experimental procedures presented in the articles.

Approaches based on machine learning usually perform a text classification task, where the entirety of the articles are classified as containing a particular experimental method or not [17]. The identification of the experimental methods in the articles is possible even when their canonical names and synonyms are not used [10]; however, the issue of identifying the positions of the experimental method descriptions in the article remains unaddressed. The identification of the position of an experimental method description is especially essential for articles in which multiple PPIs and experimental methods are mentioned. The position information can be used for the mapping of the PPIs to their related experimental methods.

In this paper we describe our participation in the Collaborative Biocurator Assistant Task (BioC) of the BioCreative V Challenge. The goal of the BioC Task is to develop a text mining system consisting of BioC-compatible [7] modules integrated together to assist biocurators. We contributed to the system by developing a module for identifying the passages (i.e., sequences of sentences) that describe experimental methods for physical PPIs. Our approach, which is detailed in the following sections, is based on traditional information retrieval techniques.

2 Systems description and methods

2.1 Data set

To the best of our knowledge, there does not exist a data set annotated for experimental methods at the passage level. The available data sets for experimental methods are either annotated at the article level (e.g. BioCreative III [2]) or PPI level (e.g. BioCreative II [9]). Therefore, we manually annotated a data set of

full text papers for the passages (i.e., sequences of sentences) that describe an experimental method and for the specific method that each passage describes.

We selected the full text articles to annotate from the BioCreative III Interaction Method Task (IMT) [2]. The BioCreative III IMT data set consists of 2003 training, 587 development, and 305 test articles. The data set contains the pdf, text and xml versions of each article. The pdf versions are full text and the text versions are generated from the pdf versions using the *pdftotext* program. The xml versions are prepared from the abstracts of the articles, which were obtained from *Pubmed*. The BioC versions of the full text articles of the BioCreative III IMT data set were requested from the BioC Task organizers and were kindly provided to us by converting the text versions into BioC format. However, since the passages were not well-formatted in the text versions (due to the automatic conversion from pdf to text), these articles were not of sufficient quality to be useful for further processing. Therefore, we decided to use the well formatted BioC versions of the articles which are accessible from *PMC Open Access [14]* as full text in xml format. The final distribution of the articles of the initial data set selected for annotation is shown in the first row of Table 1.

Table 1. Overview of Biocreative III IMT data set and its open access available subset

	Training	Development	Test
Open Access available	145	2	48
BioCreative-III IMT	2003	587	305

Table 2. List of experimental interaction detection methods in the annotated data set

Id	Name	Articles	Passages
MI:0018	two hybrid	6	26
MI:0019	coimmunoprecipitation	11	47
MI:0040	electron microscopy	1	3
MI:0055	fluorescent resonance energy transfer	2	6
MI:0081	peptide array	1	5
MI:0096	pull down	7	28
MI:0114	x-ray crystallography	2	12
MI:0402	chromatin immunoprecipitation assay	3	6
MI:0416	fluorescence microscopy	3	11
MI:0424	protein kinase assay	3	15
MI:0676	tandem affinity purification	1	4

Due to the difficulty of manual annotation for creating a data set that contains 195 full text articles within a limited time, initially a small set of ar-

ticles was decided to be selected for annotation. We first identified the most frequent five experimental methods in the data set, which are anti bait co-immunoprecipitation, anti tag co-immunoprecipitation, two hybrid, pull down, and fluorescence microscopy. Then, we randomly selected 10 articles for each method to be annotated at the passage level. Out of the 50 selected articles, 13 could be annotated for the BioC Task system development. The articles were annotated by considering all experimental methods in the PSI-MI ontology, not only the most common five methods. The experimental methods annotated manually in the data set of 13 articles are shown in Table 2. The PSI-MI identifiers of the methods, their canonical names in the PSI-MI ontology, the number of articles each method occurs in, as well as the total number of passages annotated for each method are presented in the table.

```
<passage>
  <infon key="type">paragraph</infon>
  <offset>9525</offset>
  <text>To help identify factors that might be shuttled from the cytosol to the ER by the GET system,
    we performed a yeast two-hybrid (Y2H) screen for polypeptides that can interact with Get3. Y2H
    analysis, which reports on weak interactions occurring within the nucleus of assayed strains,
    is well suited for identifying Get3 binding proteins, as it can detect transient interactions that
    are independent of the presence of Get1 and Get2. We used yeast expressing Get3 as bait to screen
    a genomic library encoding prey proteins (James et al., 1996). Physical interactions caused activation
    of the Gal4-driven HIS3 reporter gene, allowing growth on plates lacking histidine. The strongest
    hit from the screen was a fragment of Sed5 (amino acid 197 to the C terminus) (Figure 2A), a TA protein
    that acts as a SNARE in vesicular traffic within the Golgi and between the Golgi and the ER
    (Hardwick and Pelham, 1992). The Get3-Sed5 interaction was dependent on the presence of the C-terminal
    TMD (Figure 2A).</text>
  <annotation id="0">
    <infon key="type">ExperimentalMethod</infon>
    <infon key="PSIMI">0018</infon>
    <location offset="9525" length="542"/>
    <text>To help identify factors that might be shuttled from the cytosol to the ER by the GET system,
        we performed a yeast two-hybrid (Y2H) screen for polypeptides that can interact with Get3.
        Y2H analysis, which reports on weak interactions occurring within the nucleus of assayed strains,
        is well suited for identifying Get3 binding proteins, as it can detect transient interactions that
        are independent of the presence of Get1 and Get2. We used yeast expressing Get3 as bait to screen
        a genomic library encoding prey proteins (James et al., 1996).</text>
  </annotation>
</passage>
```

Fig. 1. Manual annotation example

A sample annotation from a portion of an article in the data set is shown in Fig. 1. Besides the canonical names and synonyms of the experimental methods, the method definitions from the PSI-MI ontology, as well as the hierarchies from the *Ontology Look-up Service [4]* were used to facilitate the manual annotation of the passages for the experimental methods. Each annotation has an identifier that is incremented by one throughout the article. Moreover, the value of the 'type' infon is static and set to 'ExperimentalMethod' for all annotations. The value of the 'PSIMI' infon is set to the PSI-MI identifier of the interaction detection method. The 'text' tag holds the annotated sentence(s). The 'location' tag holds the position of the annotated portion in the article with the 'offset' and 'length' attributes.

2.2 Methodology

We developed an information retrieval based system, where a query using the PSI-MI ontology and the term frequency - relevance frequency (tf.rf) term weighting metric [13] is generated for each experimental method (see below) and matched against the passages in the articles. The overall workflow of the system is shown in Fig. 2. The system pipeline takes a BioC article as input, processes it, and returns the article with the annotated passages for experimental methods in BioC format as output. *The BioC Java library [7]* is used to read, modify, and re-create the BioC files.

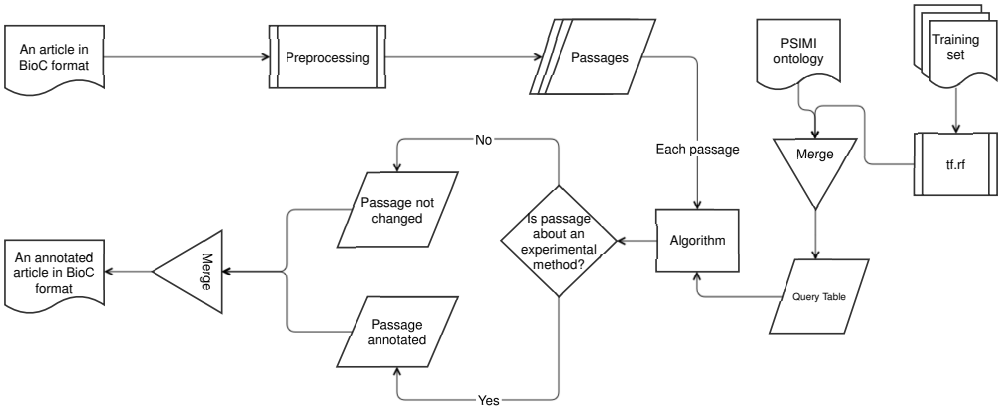

Fig. 2. Overall system workflow

In the preprocessing step a rule-based sentence splitting method, which we developed based on the period followed by a space pattern, is used. The infon types such as 'title', 'table caption', 'table', 'ref', 'footnote', and 'front' are excluded, since the text of some of these infon types are not sentences, but may contain experimental method relevant keywords. In order to reduce the number of false positives, these infon types are not used for query matching. The *Stanford CoreNLP toolkit [15]* is used to tokenize the sentences. At the tokenization phase, the punctuation marks, braces, left and right parenthesis, brackets, digits, floats etc. are removed from the sentences.

In order to determine whether a passage has sentences related to an experimental method or not, a query table is generated for each method. The query tables of the experimental methods are composed of three term lists. The first list is constructed from the canonical names and synonyms that are taken from the PSI-MI ontology entries of the experimental methods. The second and the third lists are constructed using tf.rf. The manually annotated BioC articles in our data set were used to extract the most relevant words for each experimental method. The texts under the 'annotation' tags in the passages (see Fig. 1) were

filtered according to each experimental method, split into sentences, and tokenized. The frequency of each token was calculated and token-frequency tuples were prepared. These tuples were used to calculate the weight of each token with the tf.rf method as follows.

$$tf.rf = tf * \log(2 + \frac{a}{max(1, c)})$$

tf is the number of times the token occurs in the passages annotated for the given experimental method (i.e., passages in the positive category), a is the number of passages in the positive category that contain the token, and c is the number of passages in the negative category (i.e., passages annotated with other experimental methods) that contain the token. The intuition behind relevance frequency (rf) is that a term that occurs more in the positive category compared to the negative category has more discriminating power. For each experimental method the terms are ranked by their tf.rf weights and manually examined to create the first tier tf.rf and second tier tf.rf term lists in the query table. The first tier tf.rf list consists of high scored relevant tf.rf terms, whereas the second tier tf.rf list consists of lower scored, yet still relevant terms. An example query table for the two-hybrid experimental method is shown in Table 3. The names and synonyms of the experimental methods are not included to the first and second tier lists even if they have high tf.rf weights, since they are already included in the name and synonyms list in the query table.

Table 3. Query table for the two-hybrid experimental method. The canonical names and synonyms are extracted from the PSI-MI ontology. The Tier 1 and Tier 2 terms are extracted based on tf.rf weights

Name and Synonyms	Tier 1 Terms	Tier 2 Terms
two hybrid	yeast	bait
two-hybrid	hybrid	cdna
yeast two hybrid	y-2h	gal4
2 hybrid		gal
2-hybrid		galactosidase
2h		
y2h		
classical two hybrid		
gal4 transcription regeneration		

The sentences in the passages are matched against the created queries for each experimental method. First, the name and synonyms list of the query table is used. The name and synonyms list contains terms which can be unigrams, bigram, or trigram. On the other hand, the first and second tier lists only consist of unigrams. If the term that is being searched in a sentence is a bigram or trigram, the sentence is converted to the corresponding language model. The

number of occurrences of the canonical names and synonyms found in the sentence is multiplied with the weight of 0.5. Then, the terms in the first and second tier lists are searched in the sentences and the number of their occurrences are multiplied with the weights of 0.25 and 0.125, respectively. These weights have been determined manually by trial. The threshold for selecting a sentence as relevant to an experimental method is set as 0.5. If the sum of the three scores for a sentence is greater than or equal to 0.5, the sentence is annotated with the experimental method for which it scored highest. The previous and next sentences of the selected sentence are also processed to check whether they are relevant to the same experimental method or not. If the previous and next sentences of the annotated sentence obtain the highest score for the same experimental method and if this score is greater than or equal to 0.25, they are annotated with the same experimental method. All the successive sentences with the same annotation are concatenated under one annotation tag. As a result, sentences or groups of sentences in passages are annotated for experimental methods.

3 Evaluation and Results

For evaluation, the raw versions of the 13 articles were given to the developed system, and the output articles of the system were compared to the manually annotated versions. The validation of a passage annotation (i.e., passage retrieval) by the system was done as follows; If a portion of a passage is annotated with the same experimental method and contains at least one common sentence in both the manually annotated article and the output article of the system, this annotation is considered as relevant (i.e., correct) and it is considered as not relevant (i.e., incorrect) otherwise. The summary of the results are shown in Table 4.

Table 4. Summary of the results

	Relevant	Not Relevant
Retrieved	105	87
Not Retrieved	38	493

Precision	Recall	F-Score	Accuracy
0.547	0.734	0.627	0.830

4 Discussion

We described the experimental method detection module that we developed for the BioC Task of Biocreative V. The goal of this module is to identify passages

in full text articles describing experimental methods. There does not exist a data set annotated for experimental methods at the passage level. Therefore, we manually annotated a small subset consisting of 13 full text articles from the open access portion of the Biocreative III IMT data set.

We developed a pattern-matching based approach utilizing traditional information retrieval methods. For each experimental method queries were created consisting of weighted terms identified using the PSI-MI ontology and the tf.rf weights of the terms in the annotated data set. Promising results are obtained over the manually annotated data set (62.7% F-score). However, the size of the data set and its coverage for different types of experimental methods is very limited.

As future work we plan to extend the data set by manually annotating more full text articles including a wider variety of experimental methods. Once we have a larger data set for training, we will investigate using supervised machine learning methods for identifying passages that describe an experimental method. Each sentence will be classified as the beginning, inside, or outside of an experimental method (experimental procedure) description by utilizing sequence labeling algorithms such as *Conditional Random Fields (CRF) [12].*

Acknowledgments. This work has been supported by Marie Curie FP7-Reintegration-Grants within the 7th European Community Framework Programme. We would like to thank the BioC Task organizers for organizing the shared task and for their help with the data preparation and the questions.

References

1. Agarwal, S., Liu, F., Yu, H.: Simple and Efficient Machine Learning Frameworks for Identifying Protein-Protein Interaction Relevant Articles and Experimental Methods Used to Study the Interactions. BMC Bioinformatics (2011)
2. Arighi, C., Cohen, K., Hirschman, L., Krallinger, M., Lu, Z., Valencia, A., Wilbur, J., Wu, C.: The Third Biocreative - Critical Assessment of Information Extraction in Biology Challenge. BMC Bioinformatics (2010)
3. Bader, G.D., Betel, D., Hogue, C.W.V.: BIND: The Biomolecular Interaction Network Database. Nucleic acids research 31(1), 248–250 (2003)
4. Barsnes, H., Ct, R.G., Eidhammer, I., Martens, L.: Ols Dialog: An Open-Source Front End to The Ontology Lookup Service. BMC Bioinformatics (2010)
5. Chartaryamontri, A., Breitkreutz, B.J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., Stark, C., Breitkreutz, A., Kolas, N., O'Donnell, L., Reguly, T., Nixon, J., Ramage, L., Winter, A., Sellam, A., Chang, C., Hirschman, J., Theesfeld, C., Rust, J., Livstone, M.S., Dolinski, K., Tyers, M.: The BioGRID Interaction Database: 2015 Update. Oxford University (2014)
6. Chatr-Aryamontri, A., Ceol, A., Palazzi, L.M., Nardelli, G., Schneider, M.V., Castagnoli, L., Cesareni, G.: MINT: The Molecular INTeraction Database. Nucleid Acids Research D572-D574 (2007)
7. Comeau, D.C., Dogan, 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)
8. Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D., Apweiler, R.: IntAct: An Open Source Molecular Interaction Database. Nucleic acids research 32(suppl 1), D452–D455 (2004)
9. Hirschman, L., Krallinger, M., Wilbur, J., Valencia, A.: The Biocreative II - Critical Assessment for Information Extraction in Biology Challenge. Genome Biology (2007)
10. Kappeler, T., Clematide, S., Kaljurand, K., Schneider, G., Rinaldi, F.: Towards Automatic Detection of Experimental Methods from Biomedical Literature. Third International Symposium on Semantic Mining in Biomedicine (SMBM) (2015)
11. Kim, J.D., Ohta, T., Pyysalo, S., Kano, Y., Tsujii, J.: Overview of Bionlp'09 Shared Task on Event Extraction. In: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task. pp. 1–9. BioNLP '09, Association for Computational Linguistics, Stroudsburg, PA, USA (2009)
12. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Morgan Kaufmann pp. 282–289 (2001)
13. Lan, M., C.L., T., Su, J., Lu, Y.: Supervised And Traditional Term Weighting Methods for Automatic Text Categorization. Pattern Analysis and Machine Intelligence, IEEE Transactions 31(4), 721–735 (2008)
14. Maloney, C., Sequeira, E., Kelly, C., Orris, R., Beck, J.: Pubmed Central. National Center for Biotechnology Information (US), Bethesda (MD) (2002)
15. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP Natural Language Processing Toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 55–60 (2014)
16. Rinaldi, F., Kappeler, T., Kaljurand, K., Schneider, G., Klenner, M., Clematide, S., Hess, M., von Allmen, J.M., Parisot, P., Romacker, M., Vachon, T.: Ontogene in Biocreative II. Genome Biology (2008)
17. Schneider, G., Clematide, S., Rinaldi, F.: Detection of Interaction Articles and Experimental Methods in Biomedical Literature. BMC Bioinformatics (2011)
18. Wang, X., Rak, R., Restificar, A., Nobata, C., Rupp, C., Batista-Navarro, R.T.B., Nawaz, R., Ananiadou, S.: Detecting Experimental Techniques and Selecting Relevant Documents for Protein-Protein Interactions from Biomedical Literature. BMC Bioinformatics (2011)
19. Xenarios, I., Rice, D.W., Salwinski, L., Baron, M.K., Marcotte, E.M., Eisenberg, D.: DIP: The Database of Interacting Proteins. Nucleic acids research 30(1), 303–305 (2002)