Towards Evidence Extraction : Analysis of Scientific Figures from Studies of Molecular Interactions

Gully BURNS [a,1], Xiangyang SHI [a], Yue WU [a], Huaigu CAO [a] and
Prem NATARAJAN [a]

[a] *USC Information Sciences Institute, 4676 Admiralty Way, Suite 1001, Marina del Rey CA 90292, United States of America*

Abstract. Scientific figures, captions and accompanying text provide a valuable resource that comprise the evidence generated by a published scientific study. Extracting information pertaining to that evidence requires a pipeline made up of several intermediate steps. We describe machine reading analysis applied to papers that had been curated into the European Bioinformatics Institute's INTACT database describing molecular interactions. We unpack multiple steps in an extraction pipeline that ultimately attempts to identify the type of experiments being performed automatically. We apply machine vision and natural language processing to classify figures and their associated text based on the type of methods used in the experiment to a level of accuracy that can likely support future biocuration tasks.

Keywords. Information Extraction, Molecular Interactions, Biomedical Informatics, Image Analysis

1. Introduction

Figures in the results sections of experimental research articles papers serve as the primary representation of evidence in scientific publications. They anchor the narrative flow of a paper in data by showcasing relevant aspects that illuminate points in papers' arguments. As scientists mature, they tend to focus more on the methods and results of papers, and find figures easier to understand when reading the literature [1]. Although experimental findings shown in figures are the most informative and valuable semantic elements of scientific papers, in-depth knowledge of the domain may be required to interpret the data correctly. This may make developing semantic representation of figures' scientific content a less attractive target for information extraction (IE) researchers. Most existing IE systems work with text and extract information from all claims available in the text (not just those derived from evidence presented in the paper). Our goal in this

[1] Corresponding Author: Gully A Burns, USC Information Sciences Institute, 4676 Admiralty Way, Suite 1001, Marina del Rey CA 90292, United States of America; E-mail: burns@isi.edu.

This work will be published as part of the book "Emerging Topics in Semantic Technologies. ISWC 2018 Satellite Events. E. Demidova, A.J. Zaveri, E. Simperl (Eds.), ISBN: 978-3-89838-736-1, 2018, AKA Verlag Berlin".

paper is to describe preliminary results from deep learning classification and extraction work based on text and images pertaining to figures in a well-defined experimental domain.

Molecular interactions are binding events where two molecules join to form a single, larger “molecular complex”. The European Bioinformatics Institute’s (EBI) INTACT database provides an open-access, high-quality repository of molecular interactions that have been manually-curated from primary research papers. INTACT links subfigure references (i.e., 1a, 2b, 5f, etc.) of experiments that describe interactions directly to their database records [2]. INTACT, therefore, provides a high-quality resource for IE in this domain. We previously developed methods to link ‘evidence fragments’ (i.e., text from the main narrative of papers pertaining to figures) [3]. We report initial efforts to develop evidence extraction infrastructure. This involves extracting images from PDF files, breaking them into subfigures, and classifying each based on the type of image. Each of the various pieces described here should be considered preliminary and will be described in subsequent technical papers. Here, we focus on synthesis of these multiple steps into a workflow.

2. Related Work

Detecting and processing scientific figures in biomedical papers using machine vision techniques is a well-established area of research [4]. This work includes [7], who extracted vector images from PDF files to analyze their substructure. FigSearch [8] classified the text of figure captions to identify ‘schematic representations of protein interactions and signaling events’ with an F-score of 0.77. The Yale Imagefinder system searched and examined data from scientific figures, based on OCR analysis of text in the figures [5]. More focused extraction work from the same team was then centered on molecular gel images given their ubiquity and regular structure [6]. Our long-term goal follows their example by applying deep learning to gel-based images to reconstruct primary measurements made with gels in molecular interaction experiments (see Discussion).

There are a few methods for segmentation of multipanel figures in the literature. In [9], panels are located by a line segment detection algorithm followed by a line vectorization process that connects broken line segments on the boundary of the panel. As a useful step to analyze and understand figures in biomedical papers, a caption localization and recognition algorithm is presented in [10]. It is worth noting that ImageCLEF competition [11] is an evaluation campaign with several image related tasks. Prediction of condensed textual descriptions for biomedical images has become one of the ImageCLEF tasks since 2017.

The YOLO (“You Only Look Once”) method is a high-performance approach to object detection in computer vision [12]. YOLO is designed for real-time object detection, and can be trained with user-provided training data and deployed to a customized set of objects. This method has been used for medical image analysis [13], but not yet (to our knowledge) for literature-based IE.

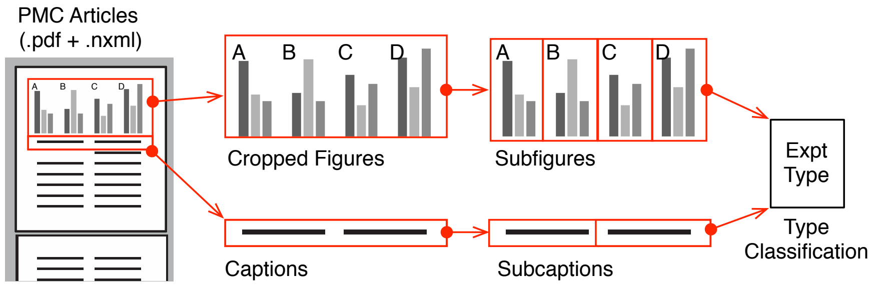


Figure 1. The data processing pipeline.

3. Methods

3.1. INTACT Data

Our INTACT data contains 20,065 papers of which 2,254 were available as part of the open access subset of Pubmed Central's (PMC) online digital collection. We downloaded bundled `.tar.gz` files from the PMC `ftp` service (available at `ftp://ftp.ncbi.nlm.nih.gov/pub/pmc/`), which provided access to both the `.nxml` and `.pdf` formatted versions of each article. We downloaded access to the original INTACT data records for each paper in PSI-MI25 format [15] from `https://www.ebi.ac.uk/intact/downloads`.

3.2. Preprocessing Figure-Based Image and Text Data

Preprocessing of the text of each paper was performed using the UIMA-BIOC library (`https://github.com/SciKnowEngine/UimaBioC`) using regular expressions to identify and standardize subfigure references within the captions of papers. Subfigure references are provided in the body of `.nxml` formatted papers and may be read easily.

We used LAPDFText (`https://github.com/SciKnowEngine/lapdftext`), providing a new figure extraction capability based on finding captions in PDF files (i.e., blocks that start with the word 'Figure') and identifying a nearby region with very low word density over the page. Caption text was painted out by masking individual words with whitespace and the region cropped from the PDF to provide a bitmap version of the figure image.

3.3. Data Pipeline

Figure 1 shows the organization of the data processing pipeline for this work. We started with INTACT papers in the PMC open access set. We then performed extraction experiments to retrieve subfigure panels from figures. Finally, we performed classification experiments in order to identify the types of experiment being performed.

3.4. Figure Subpanel Extraction

3.4.1. Simple Baseline: A Heuristic Connected Component Approach

We developed a heuristic approach for subpanel extraction based on detecting the uppercase letters that denote each subfigure ('A', 'B', etc.) and then use a greedy tiling mechanism that places the letter in the top left corner of panels to construct a rectangular layout for each panel in a figure. Letters are detected using connected component analysis. A figure is cut into multiple su-panels by straight lines that go along the top or left side of each detected letter. As a baseline, this is designed to be an easy-to-implement solution that we use for comparison with more sophisticated methods.

3.4.2. Applying and Modifying Convolutional Neural Networks for Subpanel Detection

We applied the YOLO algorithm [12] to detect subpanels in scientific figures (rather than objects in photo-quality images). The multipanel figure was resized to obtain 1:1 aspect ratio and fed into the input layer of YOLO. All subpanels were considered the same type of object. The YOLO network produces an output indicating the locations of at most 13 by 13 (i.e., 169) subpanels from the input figure. In the architecture, images are horizontally and vertically split into finely-divided, regular grids. The system then uses this grid structure to predict the existence of bounding-boxes. Since YOLO finds the bounding-boxes based on regression, these delineations are sensitive to the center position of each box and more finely grained grids are more likely to ensure that each subgraph has a more accurate center point. Thus, YOLO tends to split subfigures, causing errors by splitting the image too finely. We implemented a variant of YOLO to act more flexibly with irregularly distributed grids by introducing constraints on the generated layout of the figure. This work is ongoing and is reported here in a preliminary form.

3.5. Image Type Detection

We applied the LeNet image classification algorithm [14] directly to subfigures labeled as “gel” (for images of gel data), “graph” (for data visualizations with axes such as bar and line charts), “histology” (for photographic images of tissue), and “diagram” (for any conceptual diagrams). We hand-annotated figures extracted from the INTACT database and created binary classifiers for each of the four types of image.

3.6. Text Classification of Experimental Type

We processed open access INTACT papers with pattern-based extraction to identify individual sentences from figure captions that refer to specific subfigures (concatenating them with captions for the figure as a whole). We matched these caption documents to INTACT records to yield 3,366 entries with an associated annotation for the types of methods used to detect molecules and their interactions [15]. There were 122 separate codes for “interaction detection method” which we grouped into to 18 higher-level codes. Similarly, the INTACT set used 48 separate codes for detecting molecular participants in interactions, which we simplified to 6 higher-level codes. We then applied document classification tools based on one-dimensional convolution neural networks (CNN), and Long-Short Term Memory networks (LSTM). Source code for each classifier (with complete configuration details) may be found through the paper’s accompanying research object descriptor: `http://purl.org/ske/ro/semsci18`.

4. Results

We describe multiple stages of analyses that together provide the initial stages of a full information extraction pipeline for evidence from scientific figures. They do not themselves provide a complete solution but each contributes a step towards the construction of such a system. We provide access to code and data for this work as a research object [16]: `http://purl.org/ske/ro/semsci18`.

4.1. Sub-panel Extraction

Within the preliminary INTACT evidence extraction pipeline, the augmented YOLO method yields an accuracy of 0.87. This stands in comparison to the use of our heuristic baseline (accuracy=0.78) and the use of plain YOLO without our modifications (accuracy=0.76). This is an essential part of the pipeline for constructing the basic data record pertaining to each individual piece of evidence in a paper and will be a focus of continued improvement going forward.

4.2. Image Type Detection

Table 1. Subfigure type detection performance.

Figure Type	**N(train)**	**N(test)**	**Tagging Accuracy**
Chart	980	315	0.92
Diagram	819	197	0.40
Gel	1402	404	0.83
Histology	1299	398	**0.97**

We performed machine learning experiments on manually-tagged subfigure images from INTACT. Table 1 shows very good performance even with simple, off-the-shelf image classification technology. In our sample, we were able to detect histological images with near-perfect accuracy (0.97), charts with an accuracy of 0.92, and gel images with an accuracy of 0.83. Tagging accuracy for general conceptual diagrams was only 0.40. Given the variety of visual design that these diagrams can have, this is unsurprising and perhaps requires a more finely-divided classification scheme. Table 1 also shows the number of training and testing examples we performed our experiments on.

4.3. Text Classification of Experimental Type

Table 2 shows how text source (from evidence fragments[3] or subcaptions), number of classes, and neural network model affected the accuracy of experimental type classifications. We investigated multi-way classification of the PSI-MI2.5 experimental 'participant detection method' codes (marked 'Participant' in Table 2) or 'interaction detection method' (marked 'Interaction') for each curated data record in our corpus at two levels of granularity for both CNN and LSTM classifiers.

First, we attempted to reconstruct the INTACT record classification, involving a large number of target categories (48 for participant methods and 122 for interaction methods). Our systems generally had quite poor performance for this data. We then

Table 2. Accuracy for experimental type classification from text.

	Evidence Fragment		Sub-Caption	
Detection Method	**LSTM**	**CNN**	**LSTM**	**CNN**
Participant (48 types)	0.37	0.48	0.48	0.59
Participant (6 types)	0.58	0.70	0.72	**0.75**
Interaction (122 types)	0.26	0.50	0.56	0.62
Interaction (18 types)	0.71	0.73	0.77	**0.83**
Interaction(Co-IP tagging)	0.79	0.84	0.87	**0.90**
Participant(WB tagging)	0.71	0.79	0.76	**0.85**

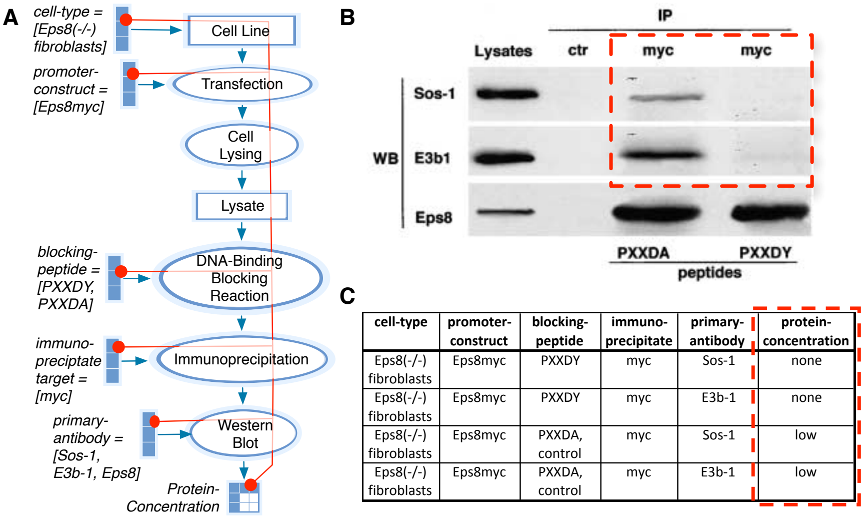
grouped together more finely delineated records into more general categories. For example, we replaced the low-level category for ‘anti-tag coimmunoprecipitation’ (MI:0007) with the higher-level category ‘affinity chromatography technology’ (MI:0004) to provide a coarser classification target. We reduced the number of classification categories from 48 to 6 for participant detection methods and from 122 to 18 for interaction detection methods. This improve prediction accuracy for interaction detection methods to 0.83 (using a CNN document classifier) and 0.75 for participant detection methods. Finally, we performed a binary tagging classification to identify specific subtypes of method: Coimmunoprecipitation (‘Co-IP’) as the most common interaction detection method and Western Blot (‘WB’) as the most common method for participant detection. The classification accuracy for tagging coimmunoprecipitation experiments was 0.90 and western blots was 0.85. We found that prediction performance was consistently better based on caption text rather than text from evidence fragments. This is consistent with findings from previous work [17].

5. Discussion

Ultimately, we seek to isolate, model, and extract scientific evidence as a distinct class of entity from interpreted ‘facts’ in scientific papers. Scientists spend the majority of their effort on creating evidence to support mechanistic explanations through experimentation. Yet, informatics systems rarely support the complete chain of reasoning that supports a given assertion. Typically coding schemes, such as the Evidence Code Ontology (ECO), designate the *type* of evidence for a given claim (i.e., inferred from data, asserted by curator, etc.), but do not deal with detailed representations of the evidence itself [18]. Similarly, the PSI-MI25 codes for interaction and participant detection methods [15] provide a human-generated classification scheme for methods but do not provide any structures to help understand and interpret data acting as evidence.

An important use case is ‘document triage’ where biocurators need to prioritize studies. Typically, this is viewed as a whole-document task [22], but being able to identify types of individual experiments could provide a powerful, lower-level set of features for triage.

Figure 2 illustrates the desired outcome of what an evidence extraction system should be able to do: given a scientific publication where experimental work is described in text and figures, we envisage a system that can (A) identify a semantic model of the experiment being performed; and (B) populate a tabular representation of the experiment’s results through the execution of IE technology. We seek to use ontology-based semantic models to accomplish this goal [19,20,21].


A
cell-type = [Eps8(-/-) fibroblasts]
Cell Line
promoter-construct = [Eps8myc]
Transfection
Cell Lysing
Lysate
blocking-peptide = [PXXDY, PXXDA]
DNA-Binding Blocking Reaction
immuno-preciptate target = [myc]
Immunoprecipitation
primary-antibody = [Sos-1, E3b-1, Eps8]
Western Blot
Protein-Concentration
B
IP
Lysates
ctr
myc
myc
WB
Sos-1
E3b1
Eps8
PXXDA
PXXDY
peptides
C

cell-type	promoter-construct	blocking-peptide	immuno-precipitate	primary-antibody	protein-concentration
Eps8(-/-) fibroblasts	Eps8myc	PXXDY	myc	Sos-1	none
Eps8(-/-) fibroblasts	Eps8myc	PXXDY	myc	E3b-1	none
Eps8(-/-) fibroblasts	Eps8myc	PXXDA, control	myc	Sos-1	low
Eps8(-/-) fibroblasts	Eps8myc	PXXDA, control	myc	E3b-1	low

Figure 2. A manually-curated example of ‘evidence extraction’. **A**. Flowchart of the protocol for experiment 1C from [23]. Dependency relations between independent and dependent variables are shown as red line. **B**. Original gel image showing gel-based measurements of protein concentration indexed by values of independent variables (‘WB’, ‘IP’, ‘peptides’). **C**. Desired extracted data table showing four values corresponding to values enclosed by a red dotted line in B.

As is often the case with work in eScience and biomedical informatics, developing useful tools for scientists must initially pass through several intermediate data-science steps. The contribution of this paper is preliminary but provides clear demonstration of feasibility to define a framework for extracting and classifying types of evidence pertaining to specific types of experiment.

Acknowledgments.

This work was funded by DARPA Big Mechanism program under ARO contract W911NF-14-1-0436, by DARPA MEDIFOR Project and by NIH grant R01 LM012592-01. This material is based on research sponsored by the Air Force Research Laboratory and the Defense Advanced Research Projects Agency under agreement number FA8750-16-2-0204. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory and the Defense Advanced Research Projects Agency or the U.S. Government. Additionally, we thank Dr. Nanyung Peng for her generous advice and guidance. Thanks too to Deepthi Devaraj.

July 2018

References

[1] K. E. Hubbard, S. D. Dunbar, Perceptions of scientific research literature and strategies for reading papers depend on academic career stage. *PLoS One* **12**, (2017), e0189753.

[2] S. Orchard, M. Ammari, B. Aranda, L. Breuza, L. Briganti, et al., The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* **42**, (2014) D358-363.

[3] G. A. Burns, P. Dasigi and E. H. Hovy. Extracting Evidence Fragments for Distant Supervision of Molecular Interactions. *SemSci 2017 Workshop, ISWC* (2017).

[4] T. Kuhn, T. Luong. and M. Krauthammer, Finding and accessing diagrams in biomedical publications. *AMIA Annu Symp Proc* **2012**, (2012) 468-474.

[5] S. Xu, J. McCusker, and M. Krauthammer Yale Image Finder (YIF): a new search engine for retrieving biomedical images. *Bioinformatics* **24**, (2008) 1968-1970.

[6] T. Kuhn, M. L. Nagy, T. Luong, and M. Krauthammer, Mining images in biomedical publications: Detection and analysis of gel diagrams. *J Biomed Semantics* **5**, (2014) 10.

[7] M. Shao and R. P. Futrelle, Recognition and Classification of Figures in PDF Documents. *Graphics Recognition. Ten Years Review and Future Perspectives* (eds. Liu, W. and Llads, J.) 231-242, Springer Berlin Heidelberg, 2006.

[8] F. Liu, T.-K. Jenssen, V. Nygaard, J. Sack, and E. Hovig, FigSearch: a figure legend indexing and classification system. *Bioinformatics* **20** (2004) 2880-2882

[9] K. C. Santosh, A. Aafaque, S. K. Antani, G. R. Thoma, Line Segment-Based Stitched Multipanel Figure Separation for Effective Biomedical CBIR. *IJPRAI* **31** (2017) 1-18

[10] J. Zou, S. K. Antani, G. R. Thoma: Localizing and Recognizing Labels for Multi-Panel Figures in Biomedical Journals. *ICDAR* **14** (2017) 753-758.

[11] A. García Seco de Herrera, R. Schaer, S. Bromuri, and H. Müller, Overview of the ImageCLEF 2016 medical task *Working Notes of CLEF 2016 (Cross Language Evaluation Forum)*, (2016).

[12] J. Redmon, S. K. Divvala, R.B. Girshick, and A. Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016) 779-788

[13] M. van Rijthoven, Z. Swiderska-Chadaj, K. Seeliger, J. van der Laak, and F. Ciompi. You Only Look on Lymphocytes Once. in *Medical Imaging with Deep Learning 2018* (2018) Amsterdam.

[14] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86, (1998) 2278-2324

[15] S. Kerrien, et al. Broadening the horizon–level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol* **5**, (2007) 44

[16] K. Belhajjame, J. Zhao, D. Garijo, K. M. Hettne, R. Palma, R., O. Corcho, J. M. Gomez-Perez, S. Bechhofer, G. Klyne, and C. A. Goble, The Research Object Suite of Ontologies: Sharing and Exchanging Research Data and Methods on the Open Web. *CoRR* **abs/1401.4307**, (2014).

[17] G. A. Burns, P. Dasigi, A. de Waard, and E. H. Hovy, Automated detection of discourse segment and experimental types from the text of cancer pathway results sections. *Database (Oxford)* (2016) baw122.

[18] M. C. Chibucos, C. J. Mungall, R. Balakrishnan, K. R. Christie, R. P Huntley, O. White, J. A. Blake, S. E. Lewis, and M Giglio, Standardized description of scientific evidence using the Evidence Ontology (ECO). *Database (Oxford)* **2014**, (2014) bau075

[19] T. Russ, C. Ramakrishnan, C., E. H. Hovy, M. Bota, and G. A. Burns, Knowledge Engineering Tools for Reasoning with Scientific Observations and Interpretations: a Neural Connectivity Use Case. *BMC Bioinformatics* **12** (2011) 351

[20] A. Bandrowski, et al. The Ontology for Biomedical Investigations. *PLoS One* **11**, (2016) e0154556

[21] G. A. Burns, and H. Chalupsky, 'Its All Made Up' - Why we should stop building representations based on interpretive models and focus on experimental evidence instead. in *Discovery Informatics: Scientific Discoveries Enabled by AI* (2014) Quebec City.

[22] A. M. Cohen, and W. R. Hersh, The TREC 2004 genomics track categorization task: classifying full text biomedical documents. *J Biomed Discov Collab* **1**, (2006) 4.

[23] M. Innocenti, E. Frittoli, I. Ponzanelli, J. R. Falck, S. M. Brachmann, P. P. Di Fiore, and G. Scita, Phosphoinositide 3-kinase activates Rac by entering in a complex with Eps8, Abi1, and Sos-1. *J Cell Biol* **160** (2003)17-23