Automated knowledge harvesting from literature text, tables, and figures using natural language processing and machine learning

Shinjae Yoo,[1,*] (sjyoo@bnl.gov), Ian Blaby[2], Sean McCorkle[1], Gilchan Park[1], and Carlos Soto[1]

[1] Computational Science Initiative, Brookhaven National Laboratory, Upton, NY; [2]Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA

Project Goals:

The overall goal of this project is to develop tools and techniques enabling the efficient extraction of protein-associated information from large volumes of literature. This high-level goal is divided into three core objectives: i) develop machine learning (ML) methods to visually process documents which are not in machine-readable formats; ii) use natural language processing (NLP) to identify protein relations described in text; iii) leverage machine learning techniques to demonstrate automatic extraction of structure and information from tables and figure embedded in documents.

Abstract:

A significant knowledge gap currently exists between sequenced genomes and the cellular function of the encoded proteins. This gap is growing as sequencing techniques accelerate while gene function-validating experiments continue at a slower pace. Since the cost (financial and time) of investigations seeking to capitalize on genome-enabled organisms by biological redesign to meet BER goals, the automated, and up to date with the current literature, annotation of target genes is essential. Current techniques for managing this resource are inadequate: keyword-based search is largely limited to hand-picked terms or at best the contents of the abstract, and reference crawling helps to expand a query, but not to refine it. Consequently, at present the most reliable functional annotations in databases are manually curated, which clearly cannot keep pace with the ever-growing body of literature. Moreover, much of the scientific contents of a publication are found within tables and figures, which are all but ignored by current literature search techniques. In this work, we use machine learning (ML) and natural language processing (NLP) techniques to move past these limitations and develop new tools to harvest knowledge from the literature at scale.

We identified several challenges to the goal of scalable scientific literature mining for functional genomics: full-text document processing; non-machine-readable formats; inconsistent gene and protein identifiers; semantic ambiguity and complex relationship ontologies; scale and diversity of table and figure structures and contents; and extensible knowledge representations. Here, **we focus on three subproblems**, their associated challenges, and our approaches, methods, and results in addressing them in this work: 1) ML for **processing non-machine-readable documents**, 2) NLP for **identifying protein entities and relationships between them in the text**, and 3) ML for **automated information extraction from tables and figures**.

Most published scientific works are available as PDFs – either as scans of old printed manuscripts or as digitally-sourced documents. These are readily accessible by human readers, but unfortunately cannot be processed automatically by computers. The publications' titles, authors, and abstracts may be indexed for digital search, but it remains relatively uncommon in most scientific fields to publish full-text articles in machine-readable formats. To alleviate this limitation, we developed a document processing pipeline that leverages ML techniques originally designed for object detection to visually segment the salient regions of a PDF article. The ML method was trained to recognize and isolate figures, tables, captions, main body text, and other document components for downstream processing by further specialized techniques. Our method achieved 80% - 94% detection accuracy on major region classes after training on a relatively small 100-document novel annotated dataset.

Due to broad inconsistencies in the in-text gene/protein identifiers found within the literature, a simple dictionary approach would not suffice for seeking textual evidence of relationships between these entities. We therefore used NLP techniques to train a named entity recognition (NER) model specialized in identifying mentioned of genes and proteins in the main-body text of biology articles. We then built upon this NER model to develop and train an entity-relationship model that identifies a refined set of relationships from the semantics of the textual evidence surrounding identified gene/protein entities. This effort includes ongoing annotation of a novel dataset for this purpose, which currently has over 400 entries and which we expect to quickly grow to over 1000. Our model currently achieves over 85% accuracy in identifying protein-protein interactions in the text.

Finally, although tables and figures often contain much of the scientific contents in research publication, the information contained in these has largely remained opaque to automated information extraction techniques. To address this opening, we are adapting existing ML techniques as well as developing new ones to identify and isolate tables and figures of relevance, as well as to extract their structure and contents. We are building upon semantic segmentation ML methods to accurately capture the structure and contents of document-embedded tables, after which we may apply NLP techniques to process the text contents. We identified bar charts as a case study for demonstrating the ability to identify data plots of interest and automatically extract the data values they contain. For this purpose, we are developing a two-stage detection model and a novel value extraction model. Both of these efforts are in early to intermediate stages but we have already demonstrated up to 88% accuracy in table identification, 45% raw table structure recognition accuracy (before post-processing), and 73% sub-figure detection accuracy.

This project aims to provide biologists with new tools to accelerate their work and to discover promising new directions of research informed by the wealth of knowledge buried in the published literature.