Title: Automated Knowledge Harvesting from Literature Text, Tables, and Figures

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

Institutions: [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. There is substantial cost (financial and time) in investigations seeking to capitalize on genome-enabled organisms by biological redesign to meet BER goals. Therefore, 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; 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. For this year, **we focus on two major subproblems**, their associated challenges, and our approaches, methods, and results in addressing them in this work: 1) NLP for **identifying protein entities and relationships (i.e., regulation) between them in the text**, and 2) ML for **automated information extraction from tables and figures**.

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. Based on our last success of PPI (protein-protein interaction) inference from literature, we leverage multisource gene databases and refine their labels to train natural language processing (NLP) models for predicting ***gene regulatory relations***. Our best model achieves 90.04 F1 performance and we have used it to identify over 200 promising regulation candidates for our demonstration organism of Pseudomonas putida. We are currently working to combine other evidence sources (i.e., expression profiles) to further improve regulatory relation detection.

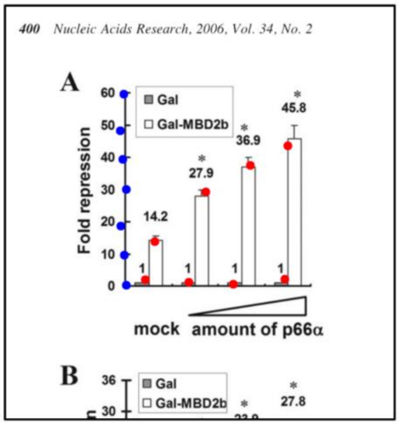

Figure 1 the predictions are bar peaks (red), and data axis values (blue). The results show the robust Figure detection results on over-cropped images.

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 novel ability to identify data plots of interest and automatically extract the data values they contain**. For this purpose, we developed a novel point proposal network (PPN) for automatically identifying data points in charts and graphs. Our PPN model efficiently leverages synthetic data during training and achieves 87.05 F1 on real scientific bar charts from unseen literature; we are working on extending our PPN algorithm to other chart types such as pie chart. We also significantly improved Table structure detection accuracies, compared to the last year by leveraging direct cell structure inference instead of regular cell structure prediction, in order to better handle irregular table structures.

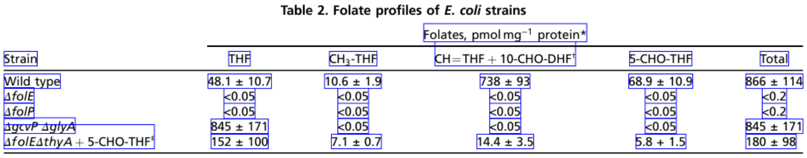

Table 2. Folate profiles of *E. coli* strains

	Folates, pmol mg^{-1} protein*				
Strain	THF	CH_3-THF	CH=THF + 10-CHO-DHF†	5-CHO-THF	Total
Wild type	48.1 ± 10.7	10.6 ± 1.9	738 ± 93	68.9 ± 10.9	866 ± 114
Δ*folE*	<0.05	<0.05	<0.05	<0.05	<0.2
Δ*folP*	<0.05	<0.05	<0.05	<0.05	<0.2
Δ*gcvP* Δ*glyA*	845 ± 171	<0.05	<0.05	<0.05	845 ± 171
Δ*folE*Δ*thyA* + 5-CHO-THF‡	152 ± 100	7.1 ± 0.7	14.4 ± 3.5	5.8 + 1.5	180 ± 98

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.

Funding Statement: This research was supported by the DOE Office of Science, Office of Biological and Environmental Research (BER), grant no. DE-SC0012704.