Discovering Content through Text Mining for a Synthetic Biology Knowledge System

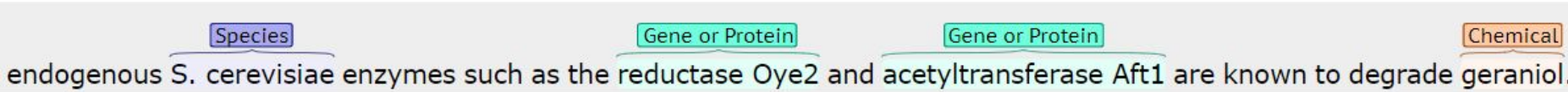
Mai H. Nguyen
Gaurav Nakum
Jaiwei Tang
Xuanyu Wu

Bridget T. McInnes
Nicholas E. Rodriguez

Eric Young
Kevin Keating

This work was funded by the National Science Foundation under Grants NoS. 1939885, 1939951, and 1939860

- **Motivation**
  - Scientific articles contain wealth of information about experimental methods and design results
  - The number of scientific articles published is growing exponentially
  - Automating process of extracting information
    - Extracting information from text is difficult due to ambiguity, variability, and volume
- **Approach**
  - Use Named Entity Recognition (NER) to mine existing literature
  - Goal of NER: Locate and classify entities in text into categories
    - e.g., genes, chemical, species



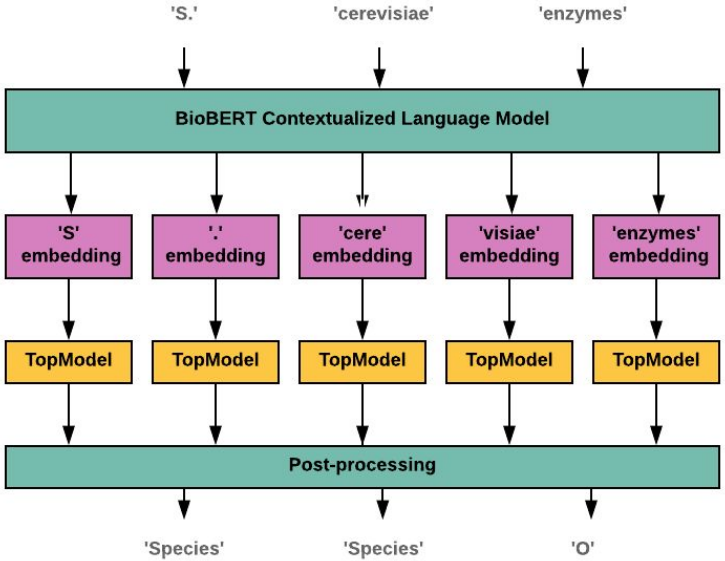
Synthetic Biology Knowledge System: Text Mining Pipeline

- Text mining Pipeline:
  - Parse: XML formatted full articles
  - Extract: Entities using our named entity recognition (NER) system
    - General entities (eg. gene and proteins, chemicals, cell line, and species)
    - Synthetic biology specific entities (eg. Promotors)
  - Validate: NER-discovered annotations are validated by domain experts
  - Refine: Fine tune NER system with validated annotations
  - *Repeat*: Continue until the system obtains a sufficient precision and recall

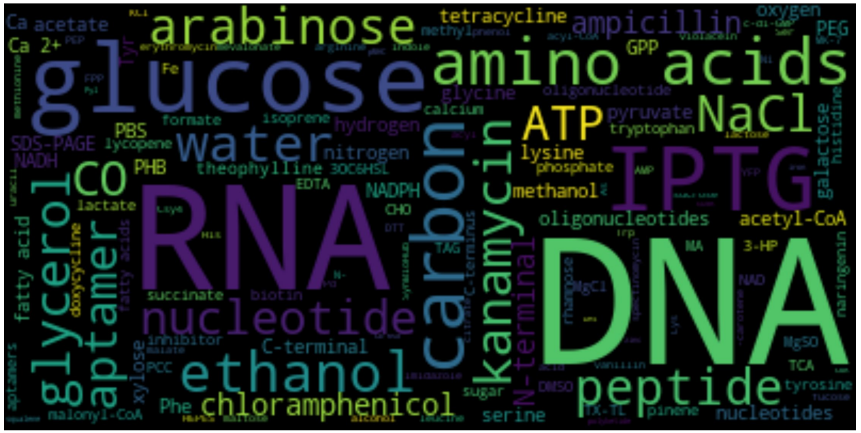
Entities identified are linked to the SBKS open knowledge system

Named Entity Recognition (NER) System

Identifying synthetic biology information from full text journal articles using NER



NER deep learning system



Word cloud of Chemical entity mentions in ACS dataset

Species Gene or Protein Gene or Protein Chemical

endogenous S. cerevisiae enzymes such as the reductase Oye2 and acetyltransferase Aft1 are known to degrade geraniol.

Annotations discovered by NER model in an ACS article