The Synthetic Biology Knowledge System

Chris Myers
Jeanet Mante
Eric Yu
Logan Terry
University of Utah
Salt Lake City, UT, USA
myers@ece.utah.edu,jv@mante.net
ejyu99@gmail.com,randoom97@live.com

Mai H. Nguyen
Gaurav Nakum
Jiawei Tang
Xuanyu Wu
University of California, San Diego
La Jolla, CA, USA
{mhnguyen,gnakum,jit072,xuw057}@ucsd.edu

Kevin Keating
Eric Young
Worcester Polytechnic Institute
Worcester, MA, USA
kwkeating@wpi.edu,emyoung@wpi.edu

Bridget T. McInnes
Nicholas E. Rodriguez
Virginia Commonwealth University
Richmond, USA
btmcinnes@vcu.edu,rodrigueazne2@vcu.edu

Jacob Jett
J. Stephen Downie
University of Illinois at Urbana-Champaign
Champaign, Illinois, USA
jjett2@illinois.edu,jdownie@illinois.edu

Brandon Sepulvado
NORC at the University of Chicago
Bethesda, MD, USA
sepulvado-brandon@norc.org

1 INTRODUCTION

Synthetic biology has transformative potential in a variety of application areas including agriculture, energy, materials, and health. While much of the research in this field has been in *E. Coli*, many applications require yeast and other bacteria. Researchers often use trial-and-error, since information can be difficult to locate. The goal of the *Synthetic Biology Knowledge System* (SBKS) is to create an an open and integrated resource that harnesses disparate, heterogeneous data sources to accelerate scientific exploration and discovery. This abstract gives an overview of the SBKS project, while several other abstracts submitted to this workshop explain different aspects in more detail.

2 SBKS CURATION PIPELINE

The core of SBKS is a curation pipeline (see Figure 1) that integrates knowledge found in both text and data sources. The knowledge once harvested is encoded into the *Synthetic Biology Open Language* (SBOL) [1], a *rich data format* (RDF) data standard for genetic design. This SBOL representation is then uploaded to the SBKS instance of the SynBioHub [2] data repository. Once deposited, it can be searched and accessed using either a graphical user interface (GUI) or programmatically by its application programmers interface (API).

Text Mining Pipeline: Our initial text data set is all the articles that have been published in ACS Synthetic Biology. These articles are provided in richly annotated JATS XML markup, which includes both a rich set of metadata and the full article text. The metadata and citation elements of the structured article file are harvested and converted into SBOL-compliant RDF/XML with Dublin Core annotations suitable for ingestion into SynBioHub. Among the steps taken during

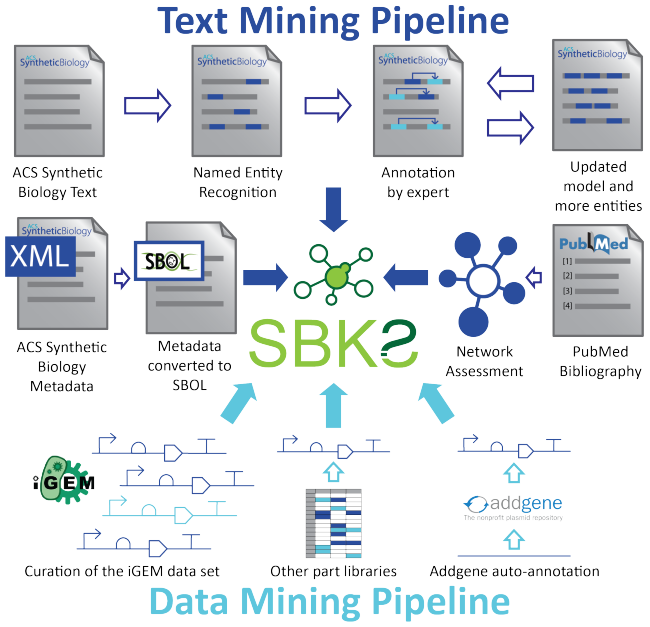

Figure 1: SBKS curation pipeline.

this process is the employment of python scripts to match article DOI's to corresponding PubMed ID's.

The XML files are also parsed to extract the full article text. The article text is then processed using techniques for *named entity recognition* (NER), which is a sub-task of text mining. The goal of NER is to locate and classify named entities present in text into pre-defined categories. We use deep neural network models to perform NER on these articles. For the initial round of NER, standard biological entity categories (e.g., genes and chemicals) are used since there

is no labeled dataset for synthetic biology entities. Results from this initial round are reviewed and corrected by domain experts to create a more refined dataset with entities more specific to synthetic biology that can be used to fine tune the NER models. Named entities expected to be detected within synthetic biology articles are also added to the articles as suggested annotations, to be confirmed by expert annotators in order to facilitate the creation of gold-standard synthetic biology-specific training data.

Another component of the text mining pipeline is the mapping of the social and conceptual structure of synthetic biology ethics, which is accomplished with network analysis and topic modeling. Building upon established bibliometric techniques to identify the synthetic biology literature, we located all 15,152 publications in the Web of Science pertaining to synthetic biology and then derived from this set of publications a smaller corpus of 562 ethical texts. Although synthetic biology literature began to increase exponentially around 2000, not much attention was devoted to ethics until roughly 2010. Ethical discourse in this field is currently dominated by small set of institutions, and scholars tend to collaborate only with a few others. The next stage of the ethics component will build upon this knowledge in order to return known ethical concerns and relevant literature related to SBKS users' queries.

Data Mining Pipeline: Our initial data set for the data mining pipeline are the synthetic biology parts and designs found in the *International Genetically Engineered Machine* (iGEM) registry of parts. Whilst the iGEM registry is large and expanding, it is not easy to extract information from parts to encourage reuse. As such a more standardized form of the iGEM library was created by converting it into SBOL. As part of the validation process we realised that there are many spurious records either due to improper completion of the record, making the information undecipherable, or as records were simply created as test exercises. Furthermore, many sequences have multiple records—one for each use case, with different use cases sometimes using the sequence in different ways. As such, we propose creating an iGEM library of sequences with different 'experiences' or uses of each sequence being linked to a core sequence. To help in this process, we are using both simple filtering and hand annotations as well as machine learning methods to create 'useful unique entity' records that are fully annotated.

In addition to the iGEM registry, parts are commonly reported in the literature in the form of "toolkit" papers. These papers almost always include part name, type, host organism and characterization data. Sequence information is often included as well, typically in the form of a table in supplemental information. Transferring parts from primary literature into an SBOL database will help to bridge the gap between highly-characterized parts reported in the literature and design tools which require data in a standardized format.

Finally, we plan to link part use to articles using the Addgene data set. Addgene is a company that stores plasmids typically created for published research studies. Once we have a good library of parts, we can use this library to annotate the sequences of the plasmids being stored by Addgene, and thus link parts to their uses in published research papers connecting the data and text pipeline results.

User Interface

One final aspect of the SBKS project is the development of a user interface that can access the information stored in SBKS to assist a designer of a genetic circuit. SBOLCanvas is a web application for creation and editing of genetic constructs using the SBOL data and visual standard. SBOLCanvas allows a user to create a genetic design from start to finish, with the option to incorporate existing SBOL data from a SynBioHub repository, such as SBKS. While SBOLCanvas is currently able to efficiently create genetic designs for parts selected via searches on SynBioHub, the end goal will be to have a design tool that provides a synthetic biology designer a seamless connection to knowledge about the parts that they are or could use in their designs.

3 DISCUSSION

The SBKS project began less than a year ago, and it is being executed by a team that met only a few months before that. While the scope is ambitious, the progress so far is very promising. We look forward to feedback from the community about the needs and potential applications for SBKS.

ACKNOWLEDGEMENTS

This work was funded by the National Science Foundation under Grants No. 1939892, 1939929, 1939885, 1939887, 1939951, and 1939860.

REFERENCES

[1] Beal, J., Nguyen, T., Gorochowski, T. E., Goñi-Moreno, A., Scott-Brown, J., McLaughlin, J. A., Madsen, C., Aleritsch, B., Bartley, B., Bhakta, S., Bissell, M., Castillo Hair, S., Clancy, K., Luna, A., Le Novère, N., Palchick, Z., Pocock, M., Sauro, H., Sexton, J. T., Tabor, J. J., Voigt, C. A., Zundel, Z., Myers, C., and Wipat, A. Communicating structure and function in synthetic biology diagrams. *ACS Synthetic Biology 8*, 8 (2019), 1818–1825. PMID: 31348656.

[2] McLaughlin, J. A., Myers, C. J., Zundel, Z., Misirli, G., Zhang, M., Ofiteru, I. D., Goñi-Moreno, A., and Wipat, A. SynBioHub: A standards-enabled design repository for synthetic biology. *ACS Synthetic Biology 7*, 2 (2018), 682–688. PMID: 29316788.