

The Synthetic Biology Knowledge System (SBKS)

C. Myers, J. Mante, E. Yu, L. Terry, M. Nguyen,
G. Nakum, J. Tang, X. Wu, K. Keating, E. Young,
B. McInnes, N. Rodriguez, J. Jett, J. Downie,
B. Sepulvado



Introduction

- Synthetic biology has a transformative potential.
- Many parts come from or are initially tested in *E. coli*.
- Many applications require different bacteria or higher-level organisms.
- Researchers currently use trial-and-error.
- Access to a wide range of data is essential to scale.
- The SBKS project is developing an open and integrated knowledge system.



Synthetic biology has a transformative potential in applications from energy, agriculture, materials, and health.

Many parts come from or are initially tested in *E. coli*.

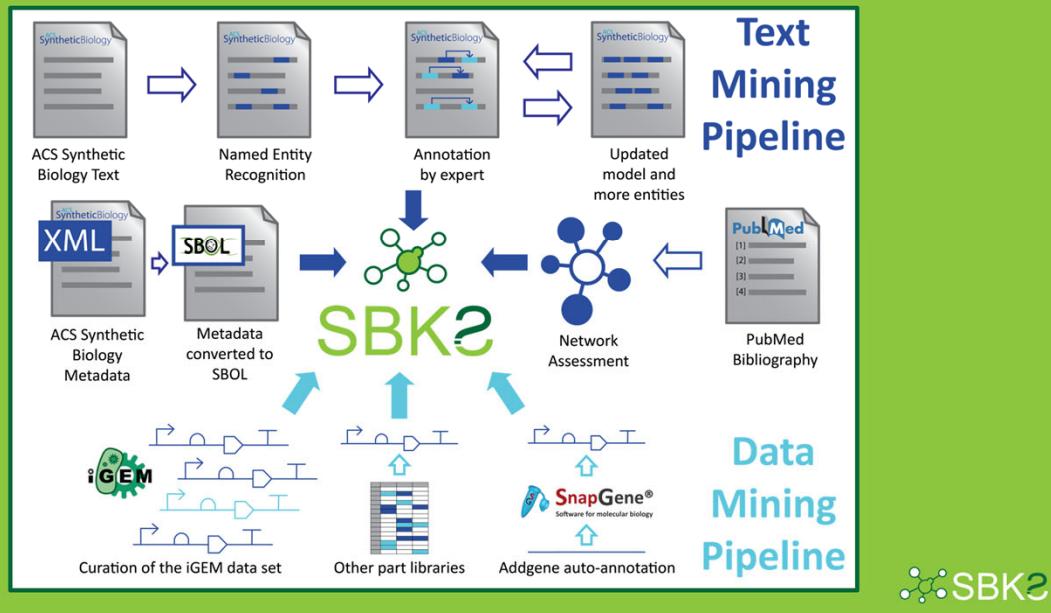
Many applications require different bacteria or higher-level organisms (i.e., yeast and other eukaryotic cells).

Researchers currently use trial-and-error, since reliable information about prior attempts is difficult to locate.

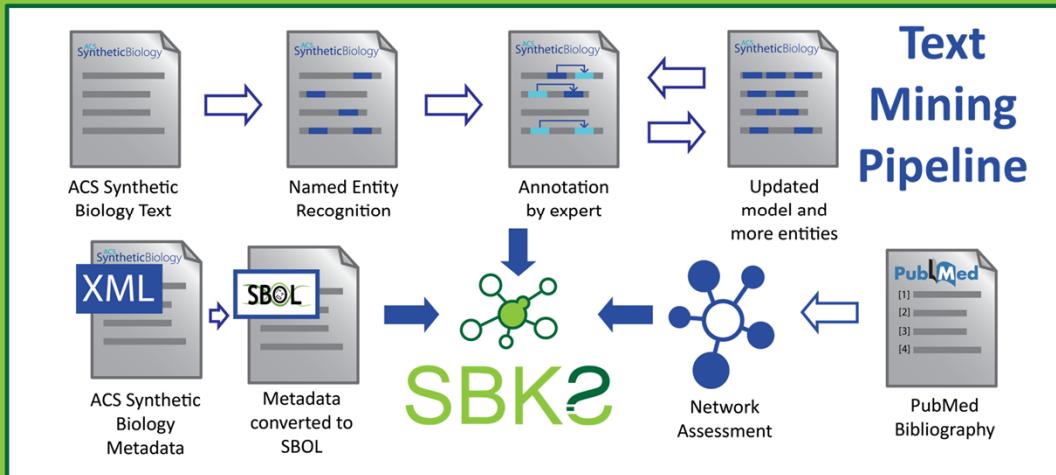
Access to a wide range of data is essential to scale.

The SBKS project is developing an open and integrated knowledge system to harness disparate, heterogeneous data sources to accelerate scientific exploration and discovery.

SBKS Curation Pipeline



Text Mining Pipeline (Overview)



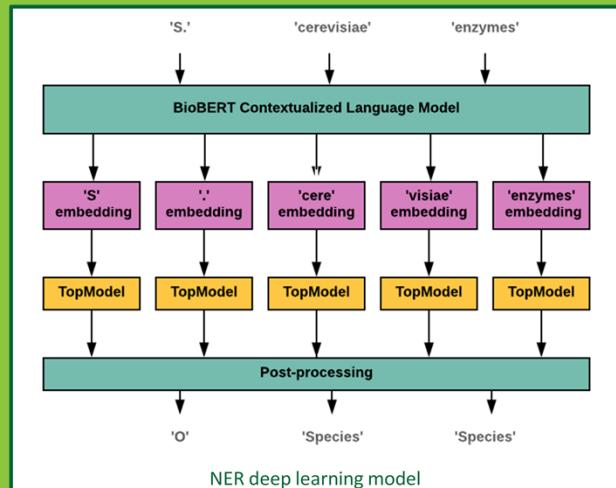
ACS Synthetic Biology provided full text in Journal Article Tag Suite (JATS) XML format. Text extracted and annotated using Named Entity Recognition (NER) and validated by domain experts.

Metadata extracted and encoded into the Synthetic Biology Open Language (SBOL). In parallel, full synthetic biology corpus found in PubMed is searched for articles on synthetic biology ethics.

Results are shared via the SBK2 instance of SynBioHub.

Named Entity Recognition

- Parse XML formatted full articles into text.
- Process text with NER using deep learning models to identify synthetic biology related entities (e.g., gene names).

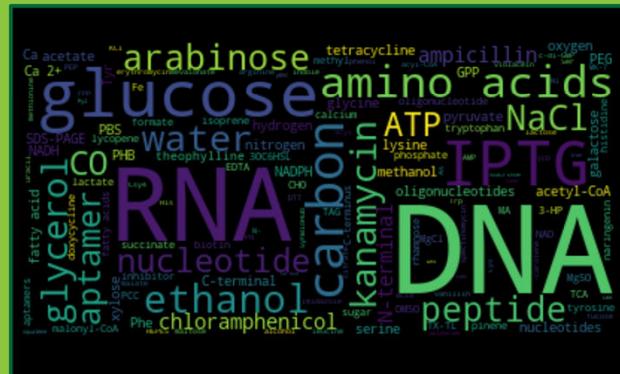


We use a deep learning model for NER, specifically a BioBERT model. The figure gives a schematic view of the model. The model inputs text, tokenizes it, and generates embeddings for the tokenized text. The token embeddings are then classified into pre-specified entity types (e.g., Species), and additional processing is performed to assign a label to an entire entity (e.g., Species for S. cerevisiae).
[Note: An embedding is a numeric vector representation of a token.]

Named Entity Recognition

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

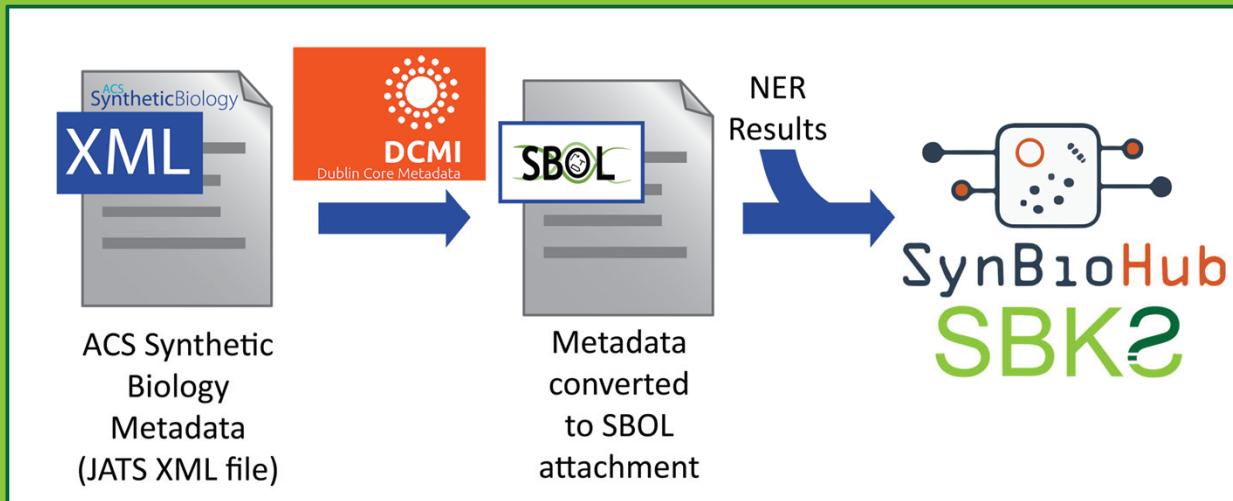
- NER-discovered annotations are validated by domain experts and corrected, as necessary.
 - Refined annotations are used to fine tune models.
 - Entities identified through this process are used to tag articles linked in SBKS.
 - Analysis can also be carried out on tag frequency.



The lower figure is a word cloud showing how often the Chemical entity type was found in the ACS dataset. The larger the term is in the word cloud, the more often it was identified by the NER model in the set of ACS articles.

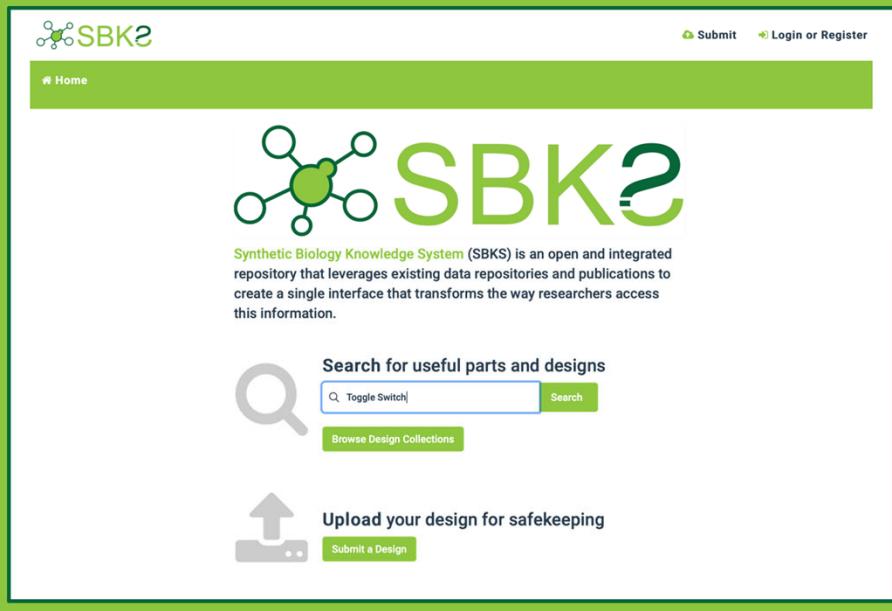
The top figure shows entity mentions found by the NER model in an ACS article. Each mention is associated with a single entity type. Note that when the NER-discovered terms are validated by the domain experts, "reductase Oye2" is split into two entities, "reductase" and "Oye2."

Metadata Harvesting



Create an SBOL Attachment object for each article in ACS Synthetic Biology.
Parse the metadata and reference sections from the JATS XML file, and map to SBOL and Dublin Core.
Results of NER added as keywords to these objects.
Results uploaded to the SBK2 instance of SynBioHub.

SBKS (<https://synbioks.org>)



The screenshot shows the homepage of the Synthetic Biology Knowledge System (SBKS). At the top left is the SBKS logo, which consists of a green molecular structure icon followed by the text "SBKS". At the top right are two buttons: "Submit" and "Login or Register". Below the header is a green navigation bar with the text "Home". The main content area features a large green "SBKS" logo with a molecular structure icon to its left. Below this, a text block explains that SBKS is an open and integrated repository that leverages existing data repositories and publications to create a single interface that transforms the way researchers access this information. There is a search bar with the placeholder "Toggle Switch" and a "Search" button. Below the search bar is a green button labeled "Browse Design Collections". Further down, there is a section for uploading designs, featuring a grey arrow icon pointing up, the text "Upload your design for safekeeping", and a green "Submit a Design" button. The bottom right corner of the page also features the SBKS logo.

SBKS (<https://synbioks.org>)

The screenshot shows the SBKS website interface. At the top, there is a navigation bar with a logo, a search bar, and links for 'Submit', 'Login or Register', and 'Logout'. Below the navigation bar, a search bar contains the query 'Toggle Switch'. To the right of the search bar is a green 'Search' button. Underneath the search bar, there are links for 'Sequence Search | Advanced Search | SPARQL' and a message indicating 'Showing 1 - 8 of 8 result(s)'. The main content area displays two search results, each with a title, a brief description, and a 'PUBLIC' badge.

General Applicability of Synthetic Gene-Overexpression for Cell-Type Ratio Control via Reprogramming
a160870518 Version 1 (Attachment)
Control of the cell-type ratio in multistable systems requires wide-range control of the initial states of cells. Here, using a synthetic circuit in *E. coli*, we describe the use of a simple gene-overexpression system combined with a bistable toggle switch, for the purposes of enabling the wide-range control of cellular states and thus generating arbitrary cell-type ratios. Theoretically, overexpression induction temporarily alters the bistable system to a monostable system, in which the location of the single steady state of cells can be manipulated over a wide range by regulating the overexpression levels. This induced cellular state becomes the initial state of the basal bistable system upon overexpression cessation, which restores the original bistable system. We experimentally demonstrated that the overexpression induced a monomodal cell distribution, and subsequent overexpression withdrawal generated a bimodal distribution. Furthermore, as designed theoretically, regulating the overexpression levels by adjusting the concentrations of small molecules generated arbitrary cell-type ratios.

Automated Design of Genetic Toggle Switches with Predetermined Bistability
a310159261 Version 1 (Attachment)
Synthetic biology aims to rationally construct biological devices with required functionalities. Methods that automate the design of genetic devices without post-hoc adjustment are therefore highly desired. Here we provide a method to predictably design genetic toggle switches with predetermined bistability. To accomplish this task, a biophysical model that links ribosome binding site (RBS) DNA sequence to toggle switch bistability was first developed by integrating a stochastic model with RBS design method. Then, to parametrize the model, a library of genetic toggle switch mutants was experimentally built, followed by establishing the equivalence between RBS DNA sequences and switch bistability. To test this equivalence, RBS nucleotide sequences for different specified bistabilities were *in silico* designed and experimentally verified. Results show that the deciphered equivalence is highly predictive for the toggle switch design with predetermined bistability. This method can be generalized to quantitative design of other probabilistic genetic devices in synthetic biology.



SBKS (<https://synbioks.org>)

The screenshot shows a detailed view of a model entry on the SBKS platform. The title of the entry is "Efficient Analysis of Systems Biology Markup Language Models of Cellular Populations Using Arrays". The identifier is c947433925, and it is labeled as Version 1. The entry was created by Leandro Watanabe and Chris J. Myers on 2016-02-25. The main text discusses the limitations of the standard SBML language in representing large complex regular systems like whole-cell and cellular population models, and how the proposed SBML arrays package can handle such models more efficiently. It includes examples of repressor and genetic toggle switch circuits. At the bottom of the page, there are links for "Download", "Search", and "Back".



SBKS (<https://synbioks.org>)

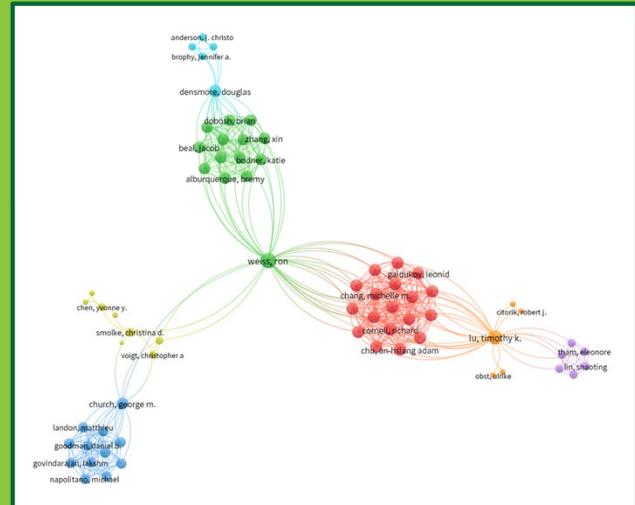
The screenshot shows the SBKS website interface. At the top, there is a navigation bar with a logo, a search bar, and links for 'Submit', 'Login or Register', and 'Logout'. Below the navigation bar, there is a search bar with a placeholder 'Q dc:creator=Chris J. Myers&'. To the right of the search bar is a green 'Search' button. Underneath the search bar, there are links for 'Sequence Search | Advanced Search | SPARQL' and a message stating 'Showing 1 - 17 of 17 result(s)'. The main content area displays three search results, each with a title, a small image, and a 'PUBLIC' badge:

- Generating Systems Biology Markup Language Models from the Synthetic Biology Open Language**
a011001475 Version 1 (Attachment)
In the context of synthetic biology, model generation is the automated process of constructing biochemical models based on genetic designs. This paper discusses the use cases for model generation in genetic design automation (GDA) software tools and introduces the foundational concepts of standards and model annotation that make this process useful. Finally, this paper presents an implementation of model generation in the GDA software tool BioSim and provides an example of generating a Systems Biology Markup Language (SBML) model from a design of a 4-input AND sensor written in the Synthetic Biology Open Language (SBOL).
- Proposed Data Model for the Next Version of the Synthetic Biology Open Language**
a420043052 Version 1 (Attachment)
While the first version of the Synthetic Biology Open Language (SBOL) has been adopted by several academic and commercial genetic design automation (GDA) software tools, it only covers a limited number of the requirements for a standardized exchange format for synthetic biology. In particular, SBOL Version 1.1 is capable of representing DNA components and their hierarchical composition via sequence annotations. This proposal revises SBOL Version 1.1, enabling the representation of a wider range of components with and without sequences, including RNA components, protein components, small molecules, and molecular complexes. It also introduces modules to instantiate groups of components on the basis of their shared function and assert molecular interactions between components. By increasing the range of structural and functional descriptions in SBOL and allowing for their composition, the proposed improvements enable SBOL to represent and facilitate the exchange of a broader class of genetic designs.
- SBOLDesigner 2: An Intuitive Tool for Structural Genetic Design**
a548410277 Version 1 (Attachment)
As the Synthetic Biology Open Language (SBOL) data and visual standards gain acceptance for describing genetic designs in a detailed and



Synthetic Biology Ethics Literature

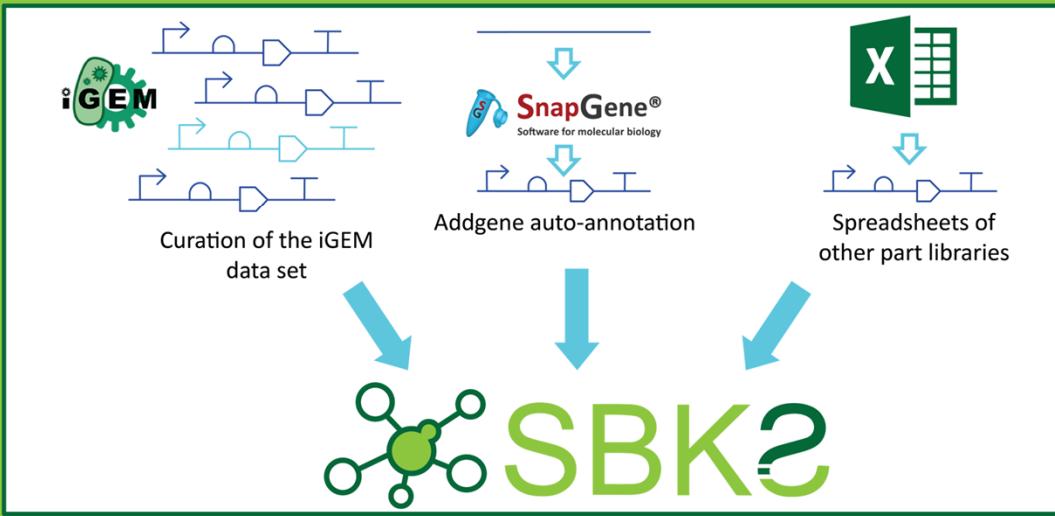
- Only ~4% of synthetic biology texts discuss ethical issues.
- Ethics gets little attention until 2010, 10 years after field began.
- Social networks among ethically-concerned researchers differ greatly from the overall synthetic biology collaboration network.
- Topics discussed range from concrete (e.g., food technology) to abstract (e.g., moral principles, defn of life).
- Analytical results from NLP will link SBKS queries with ethical issues.



- Overview
 - Innovations within synthetic biology offer a lot of promises (e.g., medicine, food technologies), but they nonetheless bring about many ethical questions. CRISPR is probably the most well-known example, but actually most research deals in some way with issues that concern the public. Pew Research Center pretty consistently publishes public opinion data showing this.
 - Our SBKS doesn't attempt to resolve these questions but rather seeks to inform researchers about known ethical concerns surrounding their research/query. Researchers will then be able to look into the concerns on their own or to turn to ethical experts identified by SBKS results.
 - In order to do this well, we need to know what is being discussed and who is discussing each topic.
 - Without knowing the communities in which certain concerns tend to focus, information retrieval systems can easily collect a biased sample of the ethical literature. This can happen because such systems do not consistently identify the ways in which certain communities discuss ethical issues (and thus fail to learn new ways they should be querying literature).
- Findings
 - Ethics currently represents a very small proportion of synthetic biology literature.
 - Synthetic biology literature started to grow exponentially around 2000, but it wasn't until closer to 2010 that people started to publish consistently on ethics.

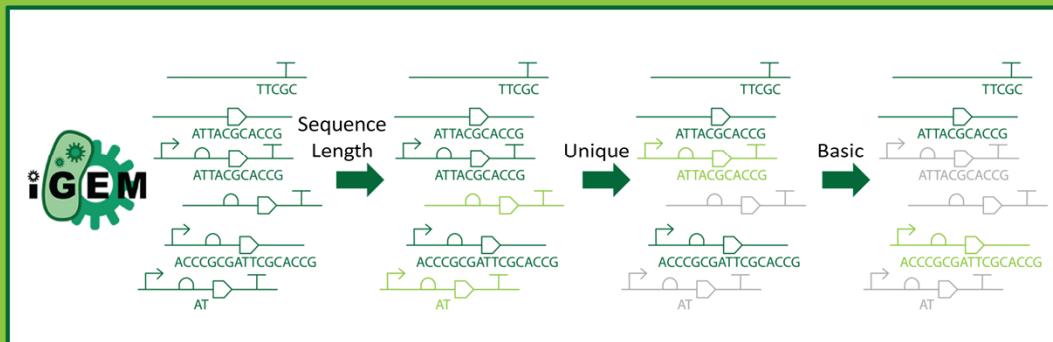
- Among ethics publications, collaboration groups tend to be isolated. Individuals might co-publish very frequently with the same individuals, but these collaboration groups tend not to collaborate with others.
- At the institutional level, collaborations tend to revolve around a handful of core institutions.
- The included image presents the largest connected component in the collaboration graph. Even within this community, one can see that subgroups are pretty distinct and tend to be connect by one to few individuals.
- Topic modeling uncovers a wide range of ethical concerns. Some topics are very concrete, e.g., CRISPR; other topics are policy related, i.e., regulation of these technologies and the allocation of new resources; still other topics are very abstract and philosophical, e.g., the definition of life.
- Linking back to non-ethics results
 - The next set of tasks will involve using many of the same NLP methods other project components have used so far (e.g., NER and custom embeddings) to connect these ethical concerns among frequently distinct communities to SBKS queries.

Data Mining Pipeline (Overview)



SBOL iGEM library was created, now being curated and validated.
Coarse and manual filtering, machine learning, and auto-annotation applied for poorly annotated sequences.
iGEM curation method will be applied to additional part libraries.
Developing methodology to add other part libraries to SynBioHub via Excel Spreadsheets.

iGEM Dataset Curation

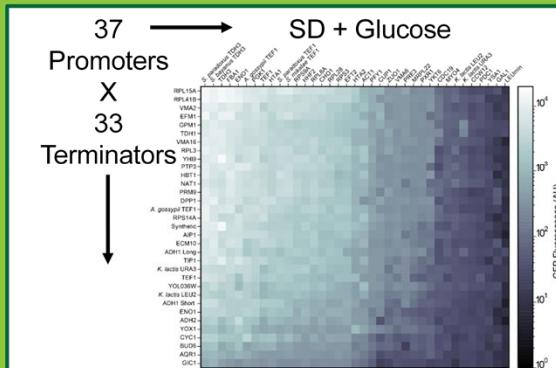


Further refinement performed using Snapgene auto annotation.



Other Libraries

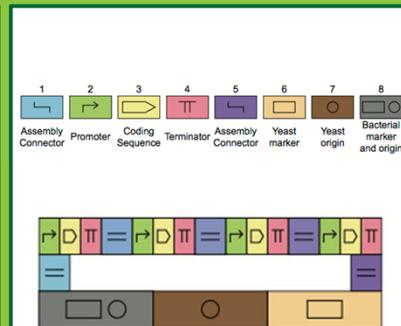
Foundry Yeast Parts



Young, et al. *Metabolic Engineering* 2018

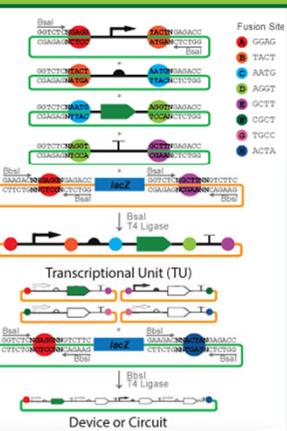
Casini, et al. JACS 2018

YTK



Lee, et al. ACS Synthetic Biology 2015

CIDAR



We would like to include libraries that meet several criteria:

1. Published with characterization data
 2. Available on Addgene for easy access and distribution
 3. Modular, particularly TypeIIIS assembly

The foundry parts library is a set of yeast parts developed by Eric Young at the MIT-Broad Foundry for massively parallel combinatorial design in yeast. This was applied to itaconic acid (met eng 2018) and several other molecules (JACS 2018)

The YTK parts library is a set of yeast parts developed by John Dueber that is on Addgene and used by the community.

The CIDAR MOclo kti is a set of *E. coli* parts developed by Doug Densmore that is on Addgene and used by the community, and easily adapted to other bacteria.

SynBioHub Sequence Search

- The ability to search becomes increasingly important as parts repositories grow.
- SBOLExplorer uses VSEARCH, a global alignment algorithm which is more effective at comparing similarities over entire sequences.

| Option | Value |
|----------------------------------|---------|
| Search Method | Global |
| Number of Results | 50 |
| Minimum Sequence Length | 20 |
| Maximum Sequence Length | 5000 |
| # of Failed Hits Before Stopping | 0 |
| Percent Match (0 to 1) | 0.8 |
| Pairwise Identity Definition | Default |



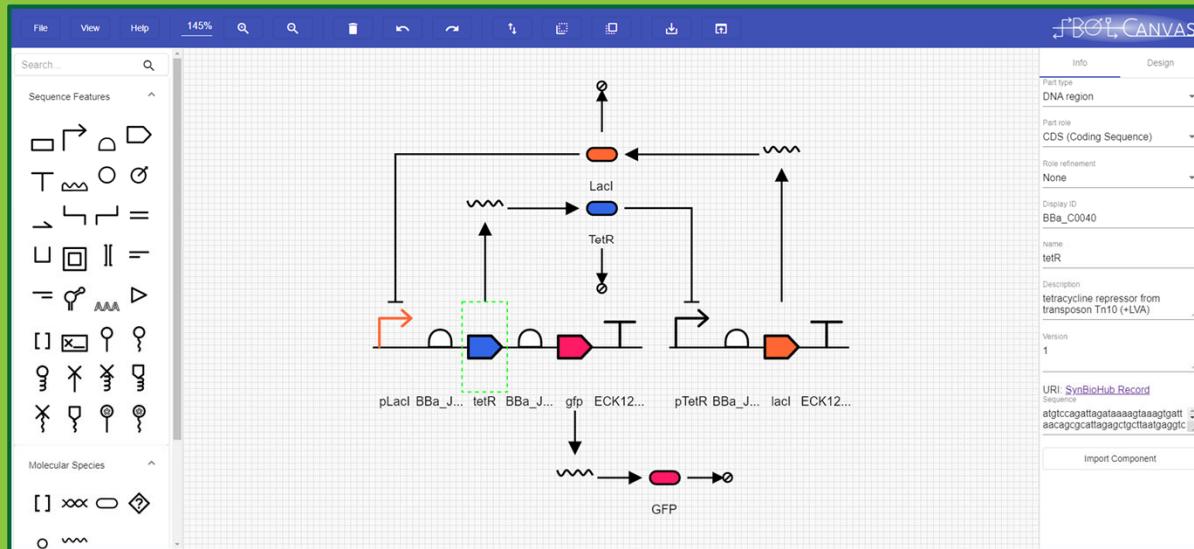
Searching by sequence will become a critical tool as repositories such as SynBioHub grow in size. Prior to this project, users could only search by keyword, or filters such as creator, or date.

The addition of a new sequence search tool allows users to search by a text sequence or FASTA/FASTQ file.

This was implemented using VSEARCH, an open-source tool using a global-alignment algorithm which is faster than BLAST for short sequences.

A command line API is also available for those who want to search via a GET request. More information can be found at the later poster session.

SBOLCanvas



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.



More Information (Poster Session II)

- Room 1
 - *Discovering Content through Text Mining for a Synthetic Biology Knowledge System*, M. Nguyen, et. al.
 - *The Social and Conceptual Organization of Synthetic Biology Ethics*, B. Sepulvado, et. al.
- Room 2
 - *Analysis of the SBOL iGEM Data Set*, J. Mante, et. al.
 - *Sequence-based Searching For SynBioHub Using VSEARCH*, E. Yu, et. al.
 - *VisBOL 2.0 - Improved Synthetic Biology Design Visualization*, B. Hatch, et. al.





Acknowledgements



This work was funded by the National Science Foundation under Grants No. 1939892, 1939929, 1939885, 193988, 1939951, and 1939860.



Affiliations



University
of Colorado
Boulder



VCU

UC San Diego

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN



WPI

NORC

at the UNIVERSITY of CHICAGO

THE
UNIVERSITY
OF UTAH®

SBKE