

# A Web-Based Validator and Validation API for the Synthetic Biology Open Language

Zach Zundel

University of Utah

IWBDA

August 19, 2016

## Essential information for synthetic DNA sequences

### To the Editor:

Following a discussion by the workgroup for Data Standards in Synthetic Biology, which met in June 2010 during the Second Workshop on Bidesign Automation in Anaheim, California, we wish to highlight a problem relating to the reproducibility of the synthetic biology literature. In particular, we have noted the very small number of articles reporting synthetic gene networks that disclose the complete sequence of all the constructs they describe.

To our knowledge, there are only a few examples where full sequences have been released. In 2005, a patent application<sup>1</sup> disclosed the sequences of the toggle switches published four years earlier in a paper by Gardner *et al.*<sup>2</sup>. The same year, Basu *et al.*<sup>3</sup> deposited their construct sequences for programmed pattern formation into GenBank<sup>3</sup>. Examples of synthetic DNA sequences derived from standardized parts that have been made available in GenBank include the refactored genome of the bacteriophage

gaps between key components are almost never reported, presumably because they are not considered crucial to the report. Yet, synthetic biology relies on the premise that synthetic DNA can be engineered with base-level precision.

Missing sequence information in papers hurts reproducibility, limits reuse of past work and incorrectly assumes that we know fully which sequence segments are important. For example, many synthetic biologists are currently realizing that translation initiation rates are dependent on more than the Shine-Dalgarno sequence<sup>4</sup>. Sequences upstream of the

start codon are crucial for translation rates, yet are underreported. Similarly, it has been demonstrated that intron length can affect the dynamics of genetic oscillators<sup>5</sup>. Many more such examples are likely to emerge.

Because full sequence disclosure is critical, we wonder why the common requirement by many journals to provide GenBank entries

for genomes and natural sequences has

and welcome contributions from the greater community.

### COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Jean Peccoud<sup>1</sup>, J Christopher Anderson<sup>2</sup>, Deepak Chandran<sup>3</sup>, Douglas Densmore<sup>4</sup>, Michal Galdzicki<sup>5</sup>, Matthew W Lux<sup>1</sup>, Cesar A Rodriguez<sup>6</sup>, Guy-Bart Stan<sup>7</sup> & Herbert M Sauro<sup>3</sup>

<sup>1</sup>Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, Virginia, USA. <sup>2</sup>Department of Bioengineering, QB3: California Institute for Quantitative Biological Research, University of California, Berkeley, California, USA.

<sup>3</sup>Department of Bioengineering, University of Washington, Seattle, Washington, USA.

<sup>4</sup>Department of Electrical and Computer Engineering, Boston University, Boston, Massachusetts, USA. <sup>5</sup>Biomedical and Health Informatics, University of Washington, Seattle, Washington, USA. <sup>6</sup>BIOFAB, Emeryville, California, USA. <sup>7</sup>Department of Bioengineering and Centre for Synthetic Biology and Innovation, Imperial College London, London, UK.  
e-mail: peccoud@vt.edu

1. Gardner, T.S. & Collins, J.J. US patent 6,841,376 (2005).
2. Gardner, T.S., Cantor, C.R. & Collins, J.J. *Nature* **403**, 339–342 (2000).
3. Basu, S., Gerchman, Y., Collins, C.H., Arnold, F.H. & Woicik, B. *Nature* **434**, 112–114 (2006).






















# Synthetic Biology Open Language



- A standard for storing, exchanging, and visualizing genetic data
- Goal is to replace formats such as FASTA and GenBank
- Several libraries and software tools support SBOL
- Adopted by ACS Synthetic Biology as preferred genetic data format

# SBOL Visual (Version 1.0)

 promoter	 origin of replication
 cds	 primer binding site
 ribosome entry site	 blunt restriction site
 terminator	 sticky restriction site
 operator	 5' overhang
 insulator	 3' overhang
 ribonuclease site	 assembly scar
 rna stability element	 signature
 protease site	 user defined
 protein stability element	

New symbols  
added on  
community  
consensus.

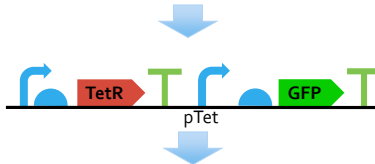
Quinn et al., PLoS Biology (2015)

# SBOL Data Model (Version 2.0)

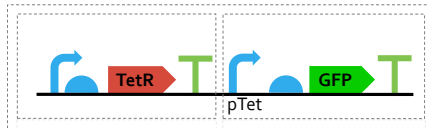
FASTA

ACTGTGCCGTTAAACGTGATTAAATCCGTACTGATAT...

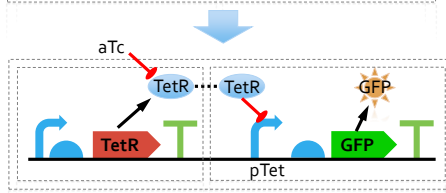
GenBank



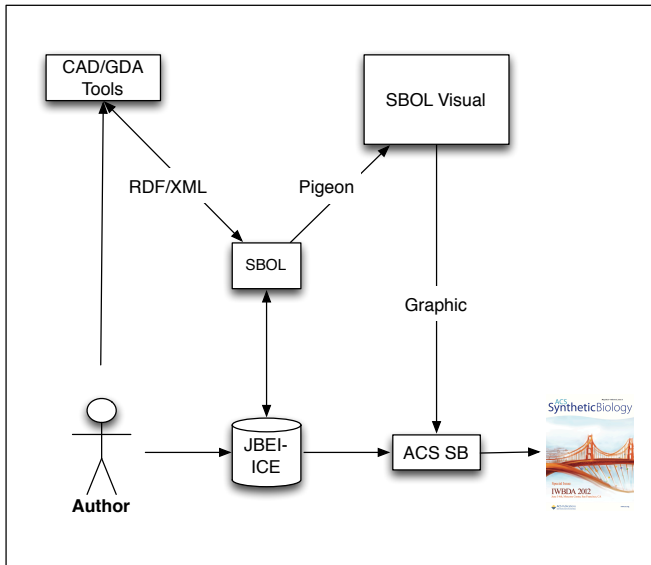
SBOL 1.1



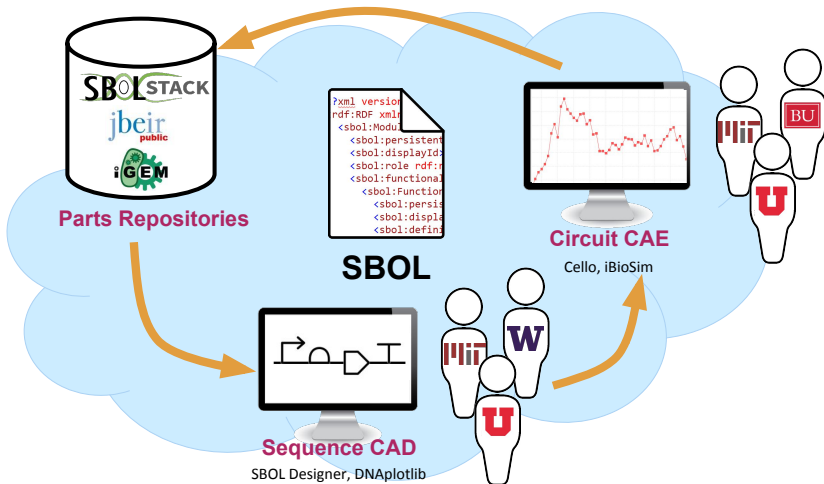
SBOL 2.0



# SBOL



# Synthetic Biology Workflow



# Library Support for SBOL 2.0

- Crucial to the success of a standard is software infrastructure to support developers' integration of the standard within their tools.
- There are several library implementations of the SBOL data structure, which provide an *application programmers interface* (API) for tool developers to interact with SBOL data objects.
  - libSBOLj - native Java library
  - libSBOL - C/C++ library
  - pySBOL - Python library
  - sboljs - Javascript library
- Library distributions include detailed documentation for the class definitions and the methods provided by the API.
- An online validator/converter powered by libSBOLj is available from the SBOL website.



- SBOL has many rules that delineate between valid and invalid data
- Two major classes of validation rules
  - Required: precise relationships and requirements for valid SBOL
  - Best practices: promote sensible and meaningful SBOL
- Validation rules should be both *unambiguous* and *machine-checkable*

# SBOL Validator

- A universally-accepted tool to check validity is needed
- As the standard evolves, it is useful to have a single verification point
- Based on libSBOLj verification methods
- Useful to many different members of the SBOL community
  - Developers: verify correct implementation of SBOL data standard
  - Authors: ensure data is properly represented before publication
  - Editors: verify SBOL compliance of software tools and libraries
- Converts between SBOL, FASTA, and GenBank to promote adoption
- Web-based - allows for central maintenance of validation methods

# Web Validator Demonstration

# SBOL Validation API

- RESTful JSON API allows for computational validation
- Removes need for end-user to include entire libSBOLj for validation
- Promotes single definition of valid SBOL

# SBOL Validation API

- Single endpoint
  - <http://apps.nonasoftware.org/SBOL-Validator/endpoint.php>
- Three-part JSON request
  - Validation parameters
  - File return
  - File(s) to validate
- Will return a JSON with validation result
- Please post feedback/bugs/feature requests to the issue tracker
  - <https://www.github.com/SynBioDex/SBOL-Validator/issues>

# Example SBOL Validation API Usage

```
import requests

url = "http://localhost/sbol-validator/endpoint.php"

request = {"validationOptions": {"output" : "FASTA",
                                "diff": False,
                                "noncompliantUrisAllowed": False,
                                "incompleteDocumentsAllowed": False,
                                "bestPracticesCheck": False,
                                "failOnFirstError": False,
                                "displayFullErrorStackTrace": False,
                                "topLevelToConvert": "",
                                "uriPrefix": "",
                                "version": "",
                                },
           "wantFileBack": True,
           "mainFile": open("sequence1.xml").read() }

resp = requests.post(url, json=request)
```



- 100+ people from all around the world.
- 30 universities, 14 companies, 8 other types of organizations.

# Organizations Supporting SBOL



**EPSRC**

Engineering and Physical Sciences  
Research Council



Office of  
Science



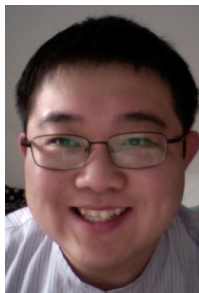
Current support for the development of SBOL provided by National Science Foundation Grants DBI-1356041 and DBI-1355909, and the Engineering and Physical Sciences Research Council under Grant Number EP/J02175X/1.



# Acknowledgments



Meher Samineni (Utah)



Zhen Zhang



Chris Myers



Supported by  
National Science Foundation Grants  
ECCS-0331270, CCF-07377655, CCF-0916042,  
CCF-1218095, and DBI-1356041.