Pilot Project Proposal

Name of Project:
Genome-scale design representation with SBOL

Proposer and Contact Information:
Dr. Jacob Beal
Raytheon BBN Technologies
Email: jakebeal@ieee.org   Phone: 617-873-7676
Representing: Synthetic Biology Open Language (SBOL) Steering Committee
(sbol-steering@googlegroups.com)

Background:
The Synthetic Biology Open Language (SBOL) is a key technology for supporting emerging biological engineering workflows. SBOL is a free and open community standard for the description and exchange of biological designs, supported by a diverse international community of researchers. Its development was motivated by the shortcomings of prior standards, such as FASTA and GenBank, with respect to the engineering of nucleic acid designs. These prior standards focus on the recording and annotation of natural sequence data, which has different challenges and requirements than the engineering and construction of sequences: the switch from "read" to "write" adds requirements such as reasoning about abstraction and (de)composition of components, integration of variant designs, and the tracking of assembly and validation processes. To address these challenges, the SBOL standard provides a "common core" set of relatively abstract representations of biological structure, function, and sequence, with a focus on abstraction and composition, and is broadly applicable across a wide range of workflow elements. Critically, SBOL also supports machine-interpretable links between this shared core and more specialized representations, such as numerical models, protocol automation scripts, LIMS tracking, and measurement data, allowing SBOL to serve as a "hub" for linking together a wide range of more specialized tools and processes without loss of information.

Technical Idea:
The SBOL community has already developed many of the representational and software capabilities necessary to support a large-scale construction effort such as GP-write. This pilot will study the application of these capabilities to key representation, coordination, and integration problems inherent in the GP-write vision. The products of this study will be:

1. A collection of representational examples, workflows, and best practice recommendations addressing key representation, coordination, and integration problems in GP-write, and
2. Extension of the existing standards as needed in support.

Utility:
Integration of software tools and data is likely to be one of the key challenges and risks in supporting GP-write: problems in data integration and curation frequently bedevil large-scale genomic efforts, and GP-write will be forging into new territory. Early investment in representations and software tools to support the GP-write effort will significantly enhance the likelihood of its success. Expected return on investment is further enhanced by leveraging an existing standard and community targeted specifically at the design and construction of complex genetic sequences.

"Fit" For GP-write:
The proposed pilot effort aims to address a set of critical issues that any GP-write effort will be forced to deal with. The existing SBOL community has already developed both tools and expertise well suited for addressing these issues. Taking advantage of this complementarity will avoid duplicating years of prior work, and will greatly reduce software- and data-related risk for GP-write, which is often severely and tragically underestimated in biological ventures.

Submit to: info@engineeringbiologycenter.org