164. Collaborative Workspaces in the DOE Systems Biology Knowledgebase

Christopher S. Henry*[1] (chenry@mcs.anl.gov), Jason Baumohl[2], Aaron Best[1], Jared Bischof[1], Ben Bowen[2], Tom Brettin[1], Tom Brown[1], Shane Canon[2], Stephen Chan[2], John- Marc Chandonia[2], Dylan Chivian[2], Ric Colasanti[1], Neal Conrad[1], Brian Davison[3], Matt DeJongh[6], Paramvir Dehal[2], Narayan Desai[1], Scott Devoid[1], Terry Disz[1], Meghan Drake[3], Janaka Edirisinghe[1], Gang Fang[7], José Pedro Lopes Faria[1], Mark Gerstein[7], Elizabeth M. Glass[1], Annette Greiner[2], Dan Gunter[2], James Gurtowski[5], Nomi Harris[2], Travis Harrison[1], Fei He[4], Matt Henderson[2], Adina Howe[1], Marcin Joachimiak[2], Kevin Keegan[1], Keith Keller[2], Guruprasad Kora[3], Sunita Kumari[5], Miriam Land[3], Folker Meyer[1], Steve Moulton[3], Pavel Novichkov[2], Taeyun Oh[8], Gary Olsen[9], Bob Olson[1], Dan Olson[1], Ross Overbeek[1], Tobias Paczian[1], Bruce Parrello[1], Shiran Pasternak[5], Sarah Poon[2], Gavin Price[2], Srivdya Ramakrishnan[5], Priya Ranjan[3], Bill Riehl[2], Pamela Ronald[8], Michael Schatz[5], Lynn Schriml[10], Sam Seaver[1], Michael W. Sneddon[2], Roman Sutormin[2], Mustafa Syed[3], James Thomason[5], Nathan Tintle[6], Will Trimble[1], Daifeng Wang[7], Doreen Ware[5], David Weston[3], Andreas Wilke[1], Fangfang Xia[1], Shinjae Yoo[4], Dantong Yu[4], **Robert Cottingham[3], Sergei Maslov[4], Rick Stevens[1], Adam P. Arkin[2]**

[1]Argonne National Laboratory, Argonne, IL, [2]Lawrence Berkeley National Laboratory, Berkeley, CA, [3]Oak Ridge National Laboratory, Oak Ridge, TN, [4]Brookhaven National Laboratory, Upton, NY, [5]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, [6]Hope College, Holland, MI, [7]Yale University, New Haven, CT, [8]University of California, Davis, CA, [9]University of Illinois at Champaign-Urbana, Champaign, IL, [10]University of Maryland, College Park, MD

http://kbase.us

Project Goals: The KBase project aims to provide the capabilities needed to address the grand challenge of systems biology: to predict and ultimately design biological function. KBase enables users to collaboratively integrate the array of heterogeneous datasets, analysis tools and workflows needed to achieve a predictive understanding of biological systems. It incorporates functional genomic and metagenomic data for thousands of organisms, and diverse tools for (meta)genomic assembly, annotation, network inference and modeling, allowing researchers to combine diverse lines of evidence to create increasingly accurate models of the physiology and community dynamics of microbes and plants. KBase will soon allow models to be compared to observations and dynamically revised. A new prototype Narrative interface lets users create a reproducible record of the data, computational steps and thought process leading from hypothesis to result in the form of interactive publications.

Collaboration, provenance, data consolidation, data standards, and data validation are all major objectives of the KBase project, and the Workspace is the engine being used to drive towards these goals. The Workspace is an online data store in KBase, where the wide variety of data entities (e.g., metabolic models, genome annotations, phenotype data) being produced and consumed by KBase analysis pipelines are stored.

User-generated data including genomes, contigs, phenotype data, and expression data may be loaded directly into the Workspace for analysis by KBase tools. The Workspace stores the many derived data products (e.g., genome annotations, metabolic models) produced by KBase analyses, with all objects generated from an analysis pipeline being interconnected all the way back to the raw data. For example, a

metabolic model is linked to the genome it was constructed from, the genome is linked to the contigs with the genome sequence, and the contigs are linked to the reads they were assembled from. The Workspace also facilitates collaboration by enabling project data, which could encompass thousands of objects, to be rapidly shared with a few select collaborators or—if desired— with the entire research community (for example, during the publication phase of a research project).

The Workspace supports full versioning of all data objects, enabling a researcher to rapidly access, compare, and restore prior versions of all objects stored in the Workspace. Provenance is also provided for data in the Workspace, providing researchers with detailed information on the analysis pipelines and parameters that produced each data object. We anticipate that these features will greatly facilitate the evaluation, validation, and replication of even complex systems biology studies.

All objects stored in the Workspace are typed, with each data type specified in detail (e.g., required fields, field types, field indexing). For example, a Genome object in the workspace must include a scientific name, a domain, and a list of features with functional annotations and locations on the chromosome. Data objects are validated against these specifications, enabling the Workspace to enforce data standards, as well as metadata requirements. The Workspace also provides a rich API for users to specify their own new data types, supporting the rapid integration of new tools, data, and analysis pipelines by the KBase user community. In this poster, we will highlight the functionality encompassed in the Workspace, we will explore the typed objects already available within the Workspace, and we will demonstrate how these typed objects connect to the many analysis tools and pipelines already available in KBase.

KBase is funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research.