The DOE Systems Biology Knowledgebase (KBase): Progress towards a system for collaborative and reproducible inference and modeling of biological function

Adam P. Arkin[*1] (aparkin@lbl.gov), Jason Baumohl[1], Aaron Best[2], Jared Bischof[2], Ben Bowen[1], Tom Brettin[2], Tom Brown[2], Shane Canon[1], Stephen Chan[1], John-Marc Chandonia[1], Dylan Chivian[1], Ric Colasanti[2], Neal Conrad[2],, Brian Davison[3], Matt DeJongh[6], Paramvir Dehal[1], Narayan Desai[2], Scott Devoid[2], Terry Disz[2], Meghan Drake[3], Janaka Edirisinghe[2], Gang Fang[7], José Pedro Lopes Faria[2], Mark Gerstein[7], Elizabeth M. Glass[2], Annette Greiner[1], Dan Gunter[1], James Gurtowski[5], Nomi Harris[1], Travis Harrison[2], Fei He[4], Matt Henderson[1], Chris Henry[2], Adina Howe[2], Marcin Joachimiak[1], Kevin Keegan[2], Keith Keller[1], Guruprasad Kora[3], Sunita Kumari[5], Miriam Land[3], Folker Meyer[2], Steve Moulton[3], Pavel Novichkov[1], Taeyun Oh[8], Gary Olsen[9], Bob Olson[2], Dan Olson[2], Ross Overbeek[2], Tobias Paczian[2], Bruce Parrello[2], Shiran Pasternak[5], Sarah Poon[1], Gavin Price[1], Srivdya Ramakrishnan[5], Priya Ranjan[3], Bill Riehl1, Pamela Ronald[8], Michael Schatz[5], Lynn Schriml[10], Sam Seaver[2], Michael W. Sneddon[1], Roman Sutormin[1], Mustafa Syed[3], James Thomason[5], Nathan Tintle[6], Will Trimble[2], Daifeng Wang[7], Doreen Ware5, David Weston[3], Andreas Wilke[2], Fangfang Xia[2], Shinjae Yoo[4], Dantong Yu[4], Bob Cottingham[3] , Sergei Maslov[4] , Rick Stevens[2]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA, [2]Argonne National Laboratory, Argonne, IL, [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 computational 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 including (meta)genomic assembly, annotation, network inference and modeling, thereby 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.

Systems biology is driven by the ever-increasing wealth of data resulting from new generations of genomics-based technologies. With the success of genome sequencing, biology began to generate and accumulate data at an exponential rate. In addition to the massive stream of sequencing data, each type of technology that researchers use to analyze a sequenced organism adds another layer of complexity to the challenge of understanding how different biological components work together to form a functional living system. Achieving this systemslevel understanding of biology will enable researchers to predict and ultimately design how biological systems will function under certain conditions. A collaborative computational environment is needed to bring researchers together so they can share and integrate large, heterogeneous datasets and readily use this information to develop predictive models that drive scientific discovery.

The advancement of systems biology relies not only on sharing the results of projects through traditional