

logic for predicting protein location became evident when we discovered that representatives of all six specialized protein translocation systems (T1SS-T6SS) known to occur in gram negative bacteria were present in at least one sequenced *Shewanella*. We developed a series of rules to identify substrates of specialized secretion systems as either bioinformatics tools were not available to identify their substrates or their predictions were not particularly robust. For example, combining domain information and proteomics data for the NiFe hydrogenase orthologs allowed us to identify these proteins as substrates of the twin arginine translocation (TAT) system. Substrates of the TAT secretion system are expected to include proteins that possess metallic redox active centers and therefore all proteins having such domains, including the NiFe hydrogenases, were carefully evaluated for the presence of N-terminal targeting peptide recognized by this secretion pathway. In *Shewanella*, the NiFe hydrogenases have an unusually long targeting peptide that was validated by proteome analysis (68 amino acids) but routinely missed by both TatP and Tatfind algorithms. The identification of outer membrane proteins was also not very accurate using a single computational tool. The Bomp beta barrel prediction tool, for example, inconsistently detected outer membrane proteins within ortholog groups even after gene model adjustment. Therefore, we supplemented these analyses by searching for a C-terminal outer membrane targeting consensus motif. Since it is known that some outer membrane proteins do not encode this domain at the C-terminus (e.g. OmpA family proteins, secretins) we also used location-informative domains to assist in identification of outer membrane proteins. Other systems, such as the type II secretion system (T2SS) that translocate periplasmic proteins across the outer membrane have no universally recognized targeting motif, but are instead believed to be recognize targeting signals that are species-specific. In *Shewanella* it is known that at least three lipoproteins are substrates of this system. A comparative analysis of these lipoproteins with other proteins deduced from the genome sequence revealed a putative targeting motif similar to those described for extracellular proteins in other bacteria, providing us a means to expand the number of predicted T2SS substrates in this Genus.

We estimate that approximately 40% of the predicted proteome for each strain of *Shewanella* is translocated out of the cytoplasm. These extracytoplasmic proteins play a central role in modulating the interactions of members of this genus with their external environments and in generating the energy and accessing the nutrient necessary to support growth and metabolism. As part of PNNL's new Foundational Science Focus area on Biological Systems Interactions we intend to employ this general strategy to identify secreted proteins in new model organisms and microbial communities to facilitate future studies directed at developing a broader understanding of microbial interactions.

Computing for Systems Biology

230

Standards in Genomic Sciences: Launch of a Standards Compliant Open-Access Journal for the 'Omics Community

G. M. Garrity^{1,5*} (garrity@msu.edu), N. Kyrpides,^{2,5} D. Field,^{3,5} P. Sterk,^{3,5} H.-P. Klenk,^{4,5} and the Editorial and Advisory Boards of Standards in Genomic Sciences

¹Michigan State University, East Lansing; ²DOE Joint Genome Institute, Walnut Creek, Calif.; ³NERC Center for Ecology and Hydrology, Oxford, United Kingdom; ⁴DSMZ – German Collection of Microorganisms and Cell Cultures GmbH, Braunschweig, Germany; ⁵Genomic Standards Consortium, Seattle, Wash.

Project Goals: The goal of DOE funding was to underwrite a pre-launch meeting of the Editorial and Advisory Boards of Standards in Genomic Sciences. The meeting was held at Michigan State University on March 12–14, 2009.

Standards in Genomic Sciences (SIGS) is an open-access e-journal that was created to promote the data standardization efforts of the Genomic Standards Consortium (GSC). The GSC was founded in 2005 by an international community of like-minded scientists to work toward improving the descriptions of our rapidly growing collection of genomes and metagenomes [1,2]. In the absence of metadata standards, the difficulty of exchanging and integrating genomic data into analytical models and public knowledgebases increases while the overall value of each subsequent sequence diminishes. This is problematic because the ease and cost of producing sequence data have dropped sharply while the cost of annotation and documentation has increased.

Membership in the GSC consists of biologists, bioinformaticians, and computer scientists, with representatives from the National Center for Biological Information (NCBI), European Molecular Biology Laboratory (EMBL), National Institute of Genetics Japan (NIG), J. Craig Venter Institute (JCVI), DOE Joint Genome Institute (JGI), European Bioinformatics Institute (EBI), Sanger Institute, and a number of other international research organizations involved in cross-cutting research. As a first step toward accomplishing organizational goals, the GSC published the "Minimum Information about a Genome Sequence" (MIGS) specification, which describes the core information that should be reported with each new genome or metagenome [3]. The GSC has led the development of a richer set of descriptors within GCDML (Genomic Contextual Markup Language), an XML variant for mark-up and transport of genomic and metagenomic data and the Genomic Rosetta Stone, a proposed resolver for mapping genome identifiers across databases [4,5]. The GSC also participates in initiatives on data

standardization, including Biosharing and the sequencing finishing standards recently described by Chain et al. [6,7]

The rationale for SIGS is to provide a venue for publication of highly structured, standardized publications of genome and metagenome sequences in accordance with MIGS and to report on other efforts that promote data standardization and data sharing [8]. Whereas peer-reviewed publications of genomes were once commonplace in a number of journals, the current trend is for many general and discipline-specific publications to eschew such papers, leading to a loss of contextual information that is critical for analyzing and interpreting genome sequence data [9]. SIGS aims to counter this trend and to provide concise, standardized descriptions of the sequencing and annotation methods along with biological information about the source organism, which we refer to as short genome reports. To that end, and with the generous support from the Michigan State University Foundation to fund the editorial office and the U.S. Department of Energy Office of Science, Biological and Environmental Research Program to convene the first meeting of the editorial and advisory boards, an open-access publication was launched to help meet those needs.

Publication of SIGS began in July 2009. At the end of October, the journal had published 28 short genome reports on bacterial and archaeal species, many of which were derived from the *Genomic Encyclopedia of Bacteria and Archaea* collaboration between the DOE Joint Genome Institute (JGI) and the German Collection of Microorganisms and Cell Cultures (DSMZ). We anticipate publishing at least another 14 short genome reports before the close of 2009, along with approximately five to six additional papers, bringing the total to approximately 60 published articles in the first volume.

Growth of the journal, to date, has been largely organic, through the journal website, search engines and forward linking of SIGS DOIs on the websites of other journals that have been cited in SIGS articles. Published articles have been downloaded > 4,750 times. Readership of SIGS is worldwide, with visitors to the site coming from over 60 countries during November, with the preponderance of visitors coming from North America and western Europe. Approximately half the daily visitors are new to the site, and the bulk of that traffic appears to be directed to the site either by search engines or by direct linking from other sites. We anticipate that traffic will continue to grow as additional content is published and SIGS becomes accepted in the major literature indices (PubMed Central, PubMed, ISI Web of Science). Our goal is to engage with the GTL community to solicit feedback and discuss additional unmet publishing needs.

References

1. Field D., Hughes J. Cataloguing our Current Genome Collection. *Microbiology* 2005; 151: 1016-1019. PubMed doi:10.1099/mic.0.27914-0.
2. Field D., Garrity G., Morrison N., Sterk P., Selengut J., Thomson N., Tatusova T. Meeting Report: eGenomics: Cataloguing Our Complete Genome Collection I. *Comp Funct Genomics* 2006; 6: 357-362. doi:10.1002/cfg.493.

3. Field D., Garrity G., Gray T., Morrison N., Selengut J., Sterk P., Tatusova T., Thomson N., Allen M.J., Angiuoli S.V., et al. The Minimum Information about a Genome Sequence (MIGS) Specification. *Nat Biotechnol* 2008; 26: 541-547. PubMed doi:10.1038/nbt1360.
4. Kottmann R., Gray T., Murphy S., Kagan L., Kravitz S., Lombardot T., Field D., Glockner F.O. A Standard MIGS/MIMS Compliant XML Schema: Toward the Development of the Genomic Contextual Data Markup Language (GCDML). *OMICS* 2008; 12: 115-121. PubMed doi:10.1089/omi.2008.0A10.
5. Van Brabant B., Gray T., Verslyppe B., Kyrpides N., Dietrich K., Glockner F. O., Cole J., Farris R., Schriml L. M., De Vos P., et al. Laying the Foundation for a Genomic Rosetta Stone: Creating Information Hubs Through the Use of Consensus Identifiers. *OMICS* 2008; 12: 123-127. PubMed doi:10.1089/omi.2008.0020.
6. Chain P. S., Grafham D.V., Fulton R. S., Fitzgerald M. G., Hostetler J., Muzny D., Ali J., Birren B., Bruce D. C., Buhay C., et al. Genomics. Genome Project Standards in a New Era of Sequencing. *Science* 2009; 326: 236-237. PubMed doi:10.1126/science.1180614.
7. Field D., Sansone S. A., Collis A., Booth T., Dukes P., Gregurick S. K., Kennedy K., Kolar P., Kolker E., Maxon M., et al. Megascience. 'Omics Data Sharing. *Science* 2009; 326: 234-236. PubMed doi:10.1126/science.1180598.
8. Garrity, G. M., Field D., Kyrpides N., Hirschman L., Sansone S. A., Angiuoli S., Cole J. R., Glockner F.O., Kolker E., Kowalchuk G., et al. Toward a Standards-Compliant Genomic and Metagenomic Publication Record. *OMICS* 2008; 12: 157-160. PubMed doi:10.1089/omi.2008.A2B2.
9. Liolios K., Mavromatis K., Tavernarakis N., Kyrpides N. The Genomes On Line Database (GOLD) in 2007: Status of Genomic and Metagenomic Projects and Their Associated Metadata. *Nucleic Acids Res* 2008; 36: D475-D479. PubMed doi:10.1093/nar/gkm884.

231

NamesforLife Semantic Resolution Services for the Life Sciences

Charles T. Parker,¹ Dorothea Rohlf's,¹ Sarah Wigley,¹ Nicole Osier,¹ Catherine Lyons,¹ and George M. Garrity^{1,2,*} (garrity@namesforlife.com)

¹NamesforLife, LLC, East Lansing, Michigan and
²Michigan State University, East Lansing

Project Goals: The overall objective of the Phase II study is to develop and deploy a set of convenient, easy to use semantic services that provide end-users with on-demand access to key information. This ensures that they can accurately interpret the meaning of any bacterial or archeal name when encountered in digital content.

Within the Genomes-to-Life Roadmap, the DOE recognizes that a significant barrier to effective communication in the life sciences is a lack of standardized semantics that accurately describe data objects and persistently express knowledge change over time. As research methods and biological concepts evolve, certainty about correct interpre-

tation of older data and published results decreases because both become overloaded with synonymous (multiple terms for a single concept) and polysemous terms (single terms with multiple meanings). Ambiguity arising from rapidly evolving terminology is a common and chronic problem in science and technology. N4L services are being developed to address this problem. The core of N4L consists of a data model, an XML schema, and an expertly managed vocabulary that is interlinked with Digital Object Identifiers (DOIs) to form a transparent semantic resolution service that disambiguates terminologies, makes them actionable, and provides direct links back to key literature and data resources.

Objectives - The overall objective of the current Phase II study is to deploy a set of convenient, easy to use semantic services that allow end-users to accurately interpret the meaning of a biological name or other dynamic term encountered in digital content, on demand and without having to query external resources or to leave the material they are reading or searching. The service can be used by database owners, publishers, or other information providers, to semantically enable their offerings; making them readily discoverable by their clients, even when the definition of a name or term has changed.

Curatorial Efforts - We significantly extended the scope of our data curation and built a framework for distributing and enhancing N4L information services to different categories of users. The target vocabulary consists of the validly published names of Bacteria and Archaea, which provides a rich and complex set of interrelated terms and interlinked resources that have a high value to the GTL community. At the end of 4Q 2009, there were 13043 validly published names (of which 3022 are synonyms of varying complexity) corresponding to 12630 taxonomic concepts and 8976 biological entities. Currently, new validly published names appear in the literature at a rate of 3.9 names/day. This number is however significantly lower than the number of names that have no standing in the literature (14.9 names/day) that appear in INSDC records and the GenBank taxonomy. Trivial names appearing on INSDC records add further confusion to the process and occur at a rate more than five-fold higher than validly published names.

The NamesforLife data have undergone further refinement to improve their accuracy. All names, taxon and exemplar records have been asserted by literature, corresponding to 9474 references, including 277 involving judicial opinions that affect the legitimacy and valid publication of 433 names. This has significance to the GTL program as some genomes are currently posted under rejected names (e.g. *Sinorhizobium medicae*). We have also verified all of the strains, culture collection deposits, and 16S rRNA sequences used in taxonomic assertions based on a review of the published record. This addresses a growing problem that has arisen from more than a decade of automated data harvesting, coupled with transitive data closure, leading to numerous systematic errors that are being routinely re-propagated.

N4L BrowserTool - The N4L BrowserTool provides a means of wide-spread distribution of our semantic

resolution services to end-users of scientific and technical literature, published in digital form and distributed via the web. The tool is currently distributed as a Firefox extension and provides on-the-fly client-side mark-up of bacterial and archaeal names with links to NamesforLife information objects that can be actuated on demand. Alpha testing of the BrowserTool ran from May - December 2009. Large-scale beta-testing is scheduled for January-February 2010 with a product release in March 2010.

N4L Autotagger - The N4L Autotagger provides publishers and other content providers with a method for enabling and enhancing content during composition. This results in articles that contain persistent links to N4L information objects and allow readers to view such content in any browser. Collaborative work is underway with the Society for General Microbiology to enhance and enable content appearing in the International Journal of Systematic and Evolutionary Microbiology.

N4L Contextual Index - The BrowserTool and AutoTagger are designed to recognize bona-fide nomenclatural events in pre-composition XML and HTML, thus allowing for high-fidelity harvesting of new/modified names and associated references from the taxonomic literature automatically. These tools can also capture the metadata for each source, thus allowing us to track all such events. This information is being used to create a contextual index that enhances the value of N4L tools as each successive use can be placed into a variety of larger contexts and used for a variety of purposes, ranging from resource planning to plotting research trends at both a fine-grained (taxon specific) and global level. In addition to the scientific literature, we are building the necessary infrastructure to permit the use of our tools to uncover prior art in areas of interest to the DOE (e.g. bio-energy/biobased feedstocks/genomics) in the U.S. and EPO patent literature.

This research is supported by the Office of Biological and Environmental Research of the U.S. Department of Energy under Phase II STTR Award DE-FG02-07ER86321 A001

232

Numerical Optimization Algorithms and Software for Systems Biology: An Integrated Model of Macromolecular Synthesis and Metabolism of *Escherichia coli*

Ines Thiele^{1*} (ithiele@hi.is), R.M.T. Fleming,¹ A. Bordbar,² R. Que,² and B.O. Palsson²

¹Center for Systems Biology, University of Iceland, Reykjavik, Iceland and ²Bioengineering Dept., University of California, San Diego, La Jolla

Project Goals: This project aims to reconstruct a genome-scale model of metabolism and macromolecular synthesis and to develop algorithms capable of solving the resulting large, stiff and ill-scaled matrices. This project combines

state of the art reconstruction and constraint-based modeling analysis tools with high-end linear optimization solvers and convex flux balance analysis. The incorporation of thermodynamic information in addition to environmental constraints will allow an accurate assessment of feasible steady states. While we will prototype the reconstruction and algorithm developments with *Escherichia coli*, we will employ the resulting networks to determine thermodynamically favorable pathways for hydrogen production by *Thermotoga maritima*.

Systems biology is a rapidly growing discipline. It is widely believed to have a broad transformative potential on both basic and applied studies in the life sciences. In particular, biochemical network reconstructions are playing a key role as they provide a framework for investigation of the mechanisms underlying the genotype-phenotype relationship. The constraint-based reconstruction and analysis approach was applied to reconstruct the transcriptional and translational (tr/tr) machinery of *Escherichia coli*. This reconstruction, denoted 'Expression-matrix' (E-matrix), represents stoichiometrically all known proteins and RNA species involved in the macromolecular synthesis machinery. It accounts for all biochemical transformations to produce active, functional proteins, tRNAs, and rRNAs known to be involved in *E. coli*'s tr/tr machinery. An initial study investigated basic properties of the E-matrix, including its capability to produce ribosomes, which was found to be in good agreement with experimental data from literature. Furthermore, quantitative gene expression data could be integrated with, and analyzed in the context of, the resulting constraint-based model. Adding mathematically derived constraints to couple certain reactions in the model allowed the quantitative representation of the size of steady state protein and RNA pools.

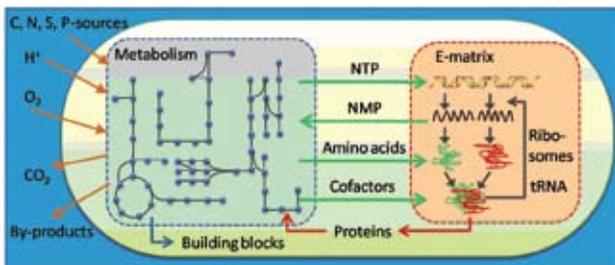


Figure 1: Functional synergy between the metabolic network and the macro-molecular synthesis network in *E. coli*.

The E-matrix was integrated with the genome-scale *E. coli* metabolic model and extended the transcriptional and translational reactions to encompass genes encoding all the respective metabolic enzymes. The resulting Metabolite-Expression-matrix (ME-matrix), exceeds the predictive capacity of the metabolic model and it can, for example, be used to predict the biomass yield since it represents the production of almost 2,000 proteins. *E. coli*'s ME-matrix is the first of its kind and represents a milestone in systems biology as it demonstrates how to quantitatively integrate 'omics'-datasets into a network context, and thus, to study the mechanistic principles underlying the genotype-phenotype

relationship. We will show some possible applications which include protein engineering, interpretation of adaptive evolution, and minimal genome design. An integration of the ME-matrix with remaining cellular processes, such as regulation, signaling, and replication, will be a next step to complete the first whole-cell model.

Building on this reconstruction effort we now started to construct the ME-matrix for *Thermotoga maritima* based on published data. Furthermore, significant advances have been made in incorporating thermodynamic constraints with metabolic networks, as shown in the accompanying poster "Numerical Optimization Algorithms and Software for Systems Biology". This work sets the stage for the goal of thermodynamically favorable pathways for hydrogen production by *Thermotoga maritima*.

233

The Ribosomal Database Project: Tools and Sequences for rRNA Analysis

J.R. Cole* (colej@msu.edu), Q. Wang, B. Chai, J. Fish, E. Cardenas, R.J. Farris, D.M. McGarrell, G.M. Garrity, and J.M. Tiedje

Michigan State University, East Lansing

Project Goals: The Ribosomal Database Project (RDP; <http://rdp.cme.msu.edu>) offers aligned and annotated rRNA sequence data and analysis services to the research community. These services help researchers with the discovery and characterization of microbes important to bioenergy production, biogeochemical cycles, including those affecting carbon and climate, and bioremediation.

Updated monthly, the RDP maintained 1,281,097 aligned and annotated quality-controlled rRNA sequences as of December 2009 (Release 10.17). The *myRDP* features have grown to support a total of over 2,500 active researchers using their *myRDP* accounts to analyze over 4,000,000 pre-publication sequences in 38,708 sequence groups, while the RDP Pyrosequencing Pipeline has been used by over 700 researchers to analyze next-generation sequences.

New NCBI/EMBL Short Read Archive Submission Tool:

Because it is very difficult for researchers to submit their next-generation rRNA sequence data to the three INSDC databases (GenBank, EMBL, and DDBJ), RDP developed a combination of web and downloadable programs, the *myRDP* SRA PrepKit, to allow users to prepare and edit their submissions. This package provides an effective solution to the difficult and confusing process involved in preparing metadata documents that are required for submission to the GenBank Short Read Archive (SRA) or EMBL European Read Archive (ERA), the two databases for reads generated from ultra-high-throughput sequencing technologies. It can be applied on sequence data generated from 454 (GS 20, FLX and Titanium), Illumina/Solexa, ABI SOLiD and Helicos platforms. It transforms the

preparation of six separate XML document types required for each submission into a clear flow of tasks implemented in easy-to-understand forms for collecting metadata about the study, samples, experiments, analyses, and sequencing runs. A set of suggested attributes in the data forms assist researchers in providing metadata conforming to the MIMS Minimal Information about a Metagenome Sequence specification, and the upcoming MEINS Minimal Information about an Environmental Sequence specifications (Field et al., 2008, *Nat. Biotechnol.* 26:541; http://gensc.org/gc_wiki/index.php/MIENS). The user can save unfinished work for later sessions and copy individual components to new submissions to avoid repetitive entry of shared data. In addition, a provided Java Web Start program creates a Fastq file from sequence reads. The *myRDP* Submission Web Start program makes it easy to perform the tasks needed to finalize your submission. A help page outlining the workflow is also provided. (The USDA provided additional funding for the *myRDP* SRA PrepKit.)

RDP Pyrosequencing Pipeline: This toolkit has been used by 777 researchers (unique e-mail addresses) to analyze their next-generation sequence data. This pipeline offers a collection of tools that automate the data processing and simplify the computationally intensive analysis of large sequencing libraries. A number of new functions have been developed for the pipeline, including a new distance matrix tool that generates distance matrices in two popular formats used by third-party tools such as Mothur (Schloss, 2009, *Appl. Environ. Microbiol.* 75:7537). All the tools now accept compressed files to reduce the upload time of large amounts of sequence data. The Initial Process, Aligner, and Clustering tools have been enhanced to return graphical summary files that provide a visual representation of sequence quality and diversity. (The USDA and NIEHS provided additional support for the RDP Pyrosequencing Pipeline.)

Other RDP tools have been used, on average, in **18,633 analysis sessions per month** by an average of **5,634 researchers** (unique IPs). These include the **RDP Classifier**, which is also available as an open-source package through SourceForge and has been **downloaded 729 times**, the online Infernal secondary-structure based aligner (Nawrocki, 2009, *Bioinformatics* 25:1335) trained by RDP on representative bacterial and archaeal alignments, the **RDP Sequence Match** program for finding nearest neighbors, the **RDP Library Compare** program for determining differentially represented taxa between two environmental libraries, the **RDP Probe Match** program for determining taxonomic coverage of primers and probes, the **RDP Tree Builder** for rapid phylogenetic tree construction, and the **RDP Hierarchy Browsers** that provide entry to the RDP sequences in taxonomic order, by publication, or by completed genome (many genomes contain multiple rRNA operons). A **new RDP Multi-Classifer** is being provided as a command-line tool to accommodate the growing need for taxonomy-based analyses of large numbers of sequences in multiple samples. This tool combines the functions of both RDP Classifier and Library Compare, and thus provides a convenient solution for researchers to use as standalone tools or to be integrated into their own analysis workflow.

RDP Web Services have been expanded to provide interfaces for the RDP Classifier, Sequence Match, Probe Match and *myRDP* tools. There are, on average, **198,632 SOAP requests** received per month. Usage examples are provided in Java and Ruby. Researchers can incorporate these web services into their own analysis pipelines to make use of these popular RDP tools.

This research is supported by the Office of Science (BER), U.S. Department of Energy under Grant No. DE-FG02-99ER62848.

References

1. Cole, J. R., Q. Wang, E. Cardenas, J. Fish, B. Chai, R.J. Farris, A.S. Kulam-Syed-Mohideen, D.M. McGarrell, T. Marsh, G.M. Garrity, and J.M. Tiedje. 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.* 37 (Database issue): D141-D145; doi:10.1093/nar/gkn879.
2. Wang, Q, G.M. Garrity, J.M. Tiedje, and J.R. Cole. 2007. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Appl. Environ. Microbiol.* 73:5261-7; doi:10.1128/AEM.00062-07.

234 Identifying Proteins from Microbial Communities

William R. Cannon^{1*} (William.Cannon@pnl.gov), Mitchell Rawlins,¹ Gaurav Kulkarni,² Andy Wu,² Ananth Kalyanaraman,² Douglas Baxter,¹ Mary Lipton,¹ and Steven Callister¹

¹Pacific Northwest National Laboratory, Richland, Wash. and ²Washington State University, Pullman, Wash.

Project Goals: See below.

The lack of reliable genome sequences currently limits the effectiveness of proteomics studies of microbial communities because of the difficulty in identifying peptides. Characterizing the proteomics of microbial communities requires (1) the computational interpretation and integration of high-throughput experimental data, (2) the leveraging of existing sources of knowledge from multiple domains, and (3) searching for solutions that meet criteria on multiple levels in a large search space. Our goal is to develop novel methods needed to describe the proteins and metagenomic functional processes occurring within unsequenced microbial communities being investigated as part of DOE's missions in carbon sequestration, bioremediation and bioenergy research.

235

Student Presentation

Identifying the Mediators of Environmental Changes Through Integration of Steady State and Time-Course Gene Expression Profiles in *Shewanella oneidensis* MR-1

Qasim K. Beg,¹ Mattia Zampieri,^{2,5} Sara Baldwin^{2*} (baldwin2@bu.edu), Niels Klitgord,² Margrethe H. Serres,⁴ Claudio Altafini,⁵ and Daniel Segre^{1,2,3}

¹Dept. of Biomedical Engineering, ²Bioinformatics Program, and ³Dept. of Biology, Boston University, Boston, Mass.; ⁴Josephine Bay Paul Center, Marine Biological Laboratory, Woods Hole, Mass.; and ⁵International School for Advanced Studies, Trieste, Italy

Project Goals: In this study, we combine mRNA microarray and metabolite measurements with statistical inference and dynamic flux balance analysis to study the transcriptional response of *S. oneidensis* as it passes through exponential, stationary, and transition phases. By measuring time-dependent mRNA expression levels during batch growth of *S. oneidensis* MR-1 under two radically different nutrient compositions, we obtain detailed snapshots of the regulatory strategies used by this bacterium to cope with gradually decreasing nutrient availability.

The dynamics of transcriptional regulation in microbial growth is an environment-dependent process. This dynamics is strongly controlled by two main factors: the wiring of the underlying regulatory network, and the time-dependent array of environmental stimuli. Understanding the interplay between these two factors is a fundamental challenge in systems biology, particularly relevant for the study of microbial systems, often adapted to rapidly changing environments. Certain genes may be activated as a response to the lack of a specific nutrient, and therefore display a strong dependence on environmental conditions; others may be more generally associated with growth rate, or growth phase requirements, and could therefore show similar behavior across different media. We address these questions in the environmental microbe *Shewanella oneidensis* MR-1, whose versatile respiratory functions make it a key player in environmental and bioenergy applications.

In this study, we combine mRNA microarray and metabolite measurements with statistical inference and dynamic flux balance analysis to study the transcriptional response of *S. oneidensis* as it passes through exponential, stationary, and transition phases. By measuring time-dependent mRNA expression levels during batch growth of *S. oneidensis* MR-1 under two radically different nutrient compositions (minimal lactate medium and LB medium), we obtain detailed snapshots of the regulatory strategies used by this bacterium to cope with gradually decreasing nutrient availability. In addition to traditional clustering, which provides a first indication of major regulatory trends and transcription factors activities, we implement a new approach for Dynamic Detection of Transcriptional Triggers (D2T2). This new

method allows us to infer a putative topology of transcriptional dependencies, with special emphasis on the nodes at which external stimuli are expected to affect the internal dynamics. In parallel, we address the question of how to compare transcriptional profiles across different time-course experiments. Our growth derivative mapping (GDM) method makes it possible to relate with each other points that correspond to the same relative growth rate in different media. This mapping allows us to discriminate between genes that display an environment-independent behavior, and genes whose transcription seems to be tuned by specific environmental factors.

Several observed transcript time-courses raise interesting biological questions. For example, we observe a coupling between nitrogen-related genes and the glycogen biosynthesis/degradation pathway. To help rationalize the observed patterns, we measure extracellular metabolites and show how transcription and metabolism can be interpreted in the context of a dynamic flux balance analysis model.

236

Computational Design of Microbial Cross-Feeding Induced by Synthetic Growth Media

Niels Klitgord^{1*} (niels@bu.edu) and Daniel Segre^{1,2} (dsegre@bu.edu)

¹Program in Bioinformatics and ²Dept. of Biology and Dept. of Biomedical Engineering, Boston University, Boston, Mass.

<http://prelude.bu.edu>

Project Goals: We seek to develop algorithm for engineering novel microbe-microbe interactions. Our method, based on stoichiometric genome-scale models of metabolism, is aimed at identifying environment that induce cross-feeding interactions. We envisage that such a “synthetic ecology” approach will be relevant for environmental and bioenergy applications.

Microbial ecosystems are ubiquitous on our planet, and play a major role in the global balance of the biosphere, as well as in the ongoing efforts for establishing renewable bioenergy sources. Since most microbe-microbe and microbe-environment interactions are likely mediated by metabolic intermediates, understanding the flow of metabolism between microbes constitutes a fundamental unsolved challenge. Here, towards addressing this challenge, we show how stoichiometric genome-scale models of metabolism can be extended to the ecosystem level, helping identify, understand and engineer interactions between pairs of microbial species. Specifically, we propose a novel suite of algorithms that can identify artificial environments predicted to induce mutualistic interactions between two given microbial species, by efficiently searching for growth media that sustain growth of two species only when simultaneously present. Our strategy is based on two major steps: *First*, we implement a procedure for automatically joining together the

stoichiometric models for two species, embedding them into a common environment. *Second*, we search the space of possible nutrient combinations for media that could not sustain growth of each species alone, but allow growth of both species simultaneously.

We validated our approach using three organism pairs of increasing complexity. The first is a simple toy model, in which one can arbitrarily pre-define expected mutualistic interactions. The second is a special case of the naturally occurring interactions between methanogenic archaea and hydrogen-producing microorganisms, which was recently analyzed in detail using flux balance models. The third is an experimentally engineered synthetic biological system of two yeast strains that can grow only in the presence of each other, because each of them is unable to synthesize a specific essential metabolite. In addition to recapitulating these known interactions, we will use our approach to generate new experimentally testable predictions of environments that induce interactions between pairs of environmentally relevant microorganisms, including *Shewanella oneidensis*. Selected predictions will be tested experimentally. We envisage that these algorithms will make it possible to engineer novel metabolism-based interactions between pairs of microbial species, helping develop a new computationally-driven synthetic ecology discipline

237

Multi-Scale Spatially Distributed Simulations of Microbial Ecosystems

William J. Riehl^{*1} (briehl@bu.edu), Niels Klitgord,¹ Christopher J. Marx,² Nathaniel C. Cady,³ and Daniel Segre^{1,4} (dsegre@bu.edu)

¹Graduate Program in Bioinformatics, Boston University, Boston, Mass.; ²Dept. of Organismic and Evolutionary Biology, Harvard University, Cambridge, Mass.; ³College of Nanoscale Science and Engineering, University at Albany, N.Y.; and ⁴Depts. of Biology and Biomedical Engineering, Boston University, Boston, Mass.

<http://prelude.bu.edu>

Project Goals: The goals of this project are: (1) to extend current genome-scale models to include spatio-temporal dynamics, (2) to allow more realistic simulations of microbial growth for individual species and ecosystems; and (3) to enable open source development of models for the study of renewable bioenergy sources, bioremediation challenges and ecosystem balance.

Genome-scale models of microbial metabolism represent the most advanced synthesis of genomic information, biochemical knowledge, and computational efficiency relevant for developing a predictive, quantitative understanding of microbial ecosystems. These models are becoming increasingly relevant for use in a number of endeavors, such as bioenergy production, bioremediation, and carbon and nitrogen cycling in the biosphere. As automated annotation pipelines,

network gap-filling algorithms, and high throughput experimental methods improve, we will gradually approach the capacity to model virtually any sequenced microbe using this approach. Yet, some of the most fundamental properties of natural microbial ecosystems crucially depend on aspects that are well beyond the stoichiometries of individual biochemical species. These include contact- or metabolite-mediated interactions between different microbes, dynamical changes of the environment, spatial structure of the underlying geography and evolutionary competition between distinct subpopulations.

We present the early stage development of a broadly applicable and user-friendly platform for modeling these interactions by performing spatially distributed time-dependent flux balance based simulations of microbial ecosystems. We use a modified version of dynamic flux balance analysis (dFBA) to implement the dynamics of the system. By taking advantage of the computational efficiency involved in flux balance model calculations, we implement a spatially structured lattice of interacting metabolic subsystems. These subsystems represent a level of detail that is intermediate between a fine-grained single-cell modeling approach, and a broad global population modeling approach, and performs akin to a cellular automaton.

This platform has been developed with the capacity to bridge multiple spatial and temporal scales, making it possible to observe long term dynamics of microbial populations growing in a given environmental setting, based on constant updates of local nutrient availabilities and exchanges, and ultimately determined by the activity of individual metabolic reactions present in each microbial species. Thus, it can be used as a platform for modeling the spatial and temporal growth of a single bacterial species in a Petri dish, biofilm formation on complex substrate morphologies, seasonality of microbial communities in a specific geographical setting, or the growth and diffusion of a microbe that has been genetically engineered toward bioremediation in a contaminated body of water.

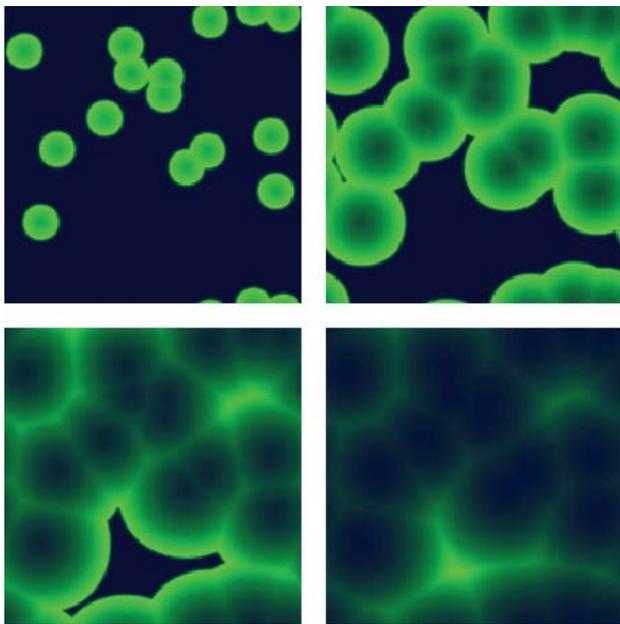


Figure 1. A sample run of the modeling platform, showing the effects of a model of *E. coli* colonies growing and merging together on a 2D surface with limited nutrient.

We present a prototype of our platform, which uses the open-source GNU Linear Programming Kit (GLPK) for performing the dFBA calculations, and a Java-based language (Processing) for coordinating the simulations and rapid visualization. We have applied this prototype to the analysis of several different examples, including the growth of a single species in a 2-Dimensional environment (see Figure 1) and syntrophic growth of microbial species. In future work, computer simulations will be integrated with experiments, allowing us to (i) calibrate the simulation parameters towards faithful representation of microbial growth patterns, and (ii) perform pilot studies on microbial ecosystem dynamics.

238

Ground and Transition State Binding Calculations to Improve Cytochrome P450_{BM3} Reactivity and Specificity

Costas D. Maranas,¹ George A. Khoury^{1*} (khoury@psu.edu), Michael J. Janik,¹ Patrick C. Cirino,¹ and Ping Lin²

¹Dept. of Chemical Engineering and ²Materials Simulation Center, Pennsylvania State University, University Park

<http://maranas.che.psu.edu>

Project Goals: The main goal of this work is to assess the impact of molecular interactions between an enzyme and its substrate at the ground and transition states on reactivity and substrate specificity. The identified trends

are currently used to inform a method for systematically re-designing Cytochrome P450_{BM3} to hydroxylate ethane.

In this work, we introduce the combined use of ground state and transition state calculations to understand how specific mutations present in engineered variants of cytochrome P450_{BM3} confer improved reactivity. The cytochrome P450_{BM3} monooxygenase has been the target of extensive directed evolution by other groups. The fatty acid hydroxylase is functionally expressed at high levels in *E. coli* and has been engineered to convert small alkanes to their corresponding alcohols, with an emphasis in biofuel production. We first identified and calculated the ground and transition state structures for the rate-limiting step using quantum mechanical methods. Next, we computationally assessed the effects of 14 different experimentally isolated mutations in P450 mutant 535-h (3 mutations lie in the active site) on interactions with the ground and transition state structures with a newly developed computational saturation mutagenesis procedure. The general trend found was that some mutations are important for improving substrate binding, while other mutations in different positions are important for improving transition state stabilization. We find that calculations at both ground and transition state appear to be important for rational enzymatic design. In the design phase, we systematically chose design positions based on sequence, structure, and energetic factors, and customized the Iterative Protein Redesign and Optimization (IPRO) framework to identify the energetically optimal mutations with the ground and transition states. We report on the general trends from the optimal designs predicted by IPRO.

239

Student Presentation

Improving Metabolic Models Using Synthetic Lethality Data and Generating Genome-Scale Isotope Mapping Models for Flux Elucidation

Patrick F. Suthers^{1*} (suthers@enr.psu.edu), Alireza Zomorodi,¹ Prabhasa Ravikirthi,^{2*} and **Costas D. Maranas**¹

¹Dept. of Chemical Engineering, and ²Dept. of Cell and Developmental Biology, Pennsylvania State University, University Park

<http://maranas.che.psu.edu>

Project Goals: The project goal of the research described here is twofold: First to improve the quality of genome-scale metabolic models by making use of gene essentiality and synthetic lethality experimental data. The second goal is to combine existing metabolic reconstructions with information from atom transitions to generate genome-scale isotope mapping models.

A pair of non-essential genes is referred to as synthetic lethal if the simultaneous deletion of both genes is lethal but the single gene deletions are not. One can generalize the concept of synthetic lethality to reactions or extend it further

by considering gene/reaction groups of increasing size where only the simultaneous elimination of all genes/reactions is lethal. Previous studies have demonstrated the utility of synthetic lethal predictions for the curation of genome-scale metabolic models. We recently used synthetic lethality information to identify twenty-one model improvements for the genome-scale model of *Escherichia coli*, *iAF1260*. In this talk, we discuss the systematic identification of synthetic lethal gene combinations for the most recent genome-scale metabolic model of yeast, (i.e., *iMM904*) for a variety of different growth medium conditions. By contrasting the *in silico* lethality predictions with *in vivo* observations we identified/corrected many missing regulatory mechanisms in yeast. The incorporation of the altered regulatory mechanisms into the genome-scale metabolic model led to a substantial increase in the accuracy of the *in silico* gene essentiality predictions. Overall, this study demonstrates the utility of synthetic lethality information for correcting genome-scale metabolic models.

Metabolic flux analysis (MFA) has so far been restricted to lumped networks lacking many important pathways, partly due to the difficulty in automatically generating isotope mapping matrices for genome-scale networks. Here we describe a procedure for the largely automated generation of atom mappings for genome-scale metabolic reconstructions. The developed procedure uses a compound matching algorithm based on the graph theoretical concept of pattern recognition along with relevant reaction heuristics to automatically generate genome-scale atom mappings which trace the path of atoms from reactants to products for every reaction in any given reconstruction. When applied to the *iAF1260* metabolic reconstruction of *Escherichia coli*, the genome-scale isotope mapping model *imPR90068* is obtained. The model maps 90,068 non-hydrogen atoms, contains 1.37×10^{157} distinct isotope forms and accounts for all 2,077 reactions present in *iAF1260* (the previous largest mapping model included 238 reactions). The expanded scope of *imPR90068* allows for tracking of labeled atoms through pathways such as cofactor and prosthetic group biosynthesis and histidine metabolism. We also discuss how using an elementary metabolite unit (EMU) representation of *imPR90068* significantly reduces the number of variables during MFA.

240 Computational Pathway Identification and Strain Optimization for Chemical and Biofuel Production

Sridhar Ranganathan^{1*} (sur152@psu.edu), Patrick F. Suthers,² and Costas D. Maranas²

¹Huck Institutes of the Life Sciences and ²Dept. of Chemical Engineering, Pennsylvania State University, University Park

<http://maranas.che.psu.edu>

Project Goals: The main goal of this work is to develop new methods to discern novel pathways for chemical and biofuel production and to elucidate strain engineering strategies that will ensure production at desired target levels.

We present an integrated computational base to support pathway identification and strain optimization with an emphasis on biofuel production. An efficient graph-based algorithm is presented for the exhaustive identification of all pathways enabling the production of a targeted biofuel molecule. The algorithm is based on a min-path formulation. It searches over a database of biotransformations that spans reactions from KEGG, Metacyc, BRENDA and other resources with an emphasis on C4+ alcohols. The identified pathways are then integrated into the genome-scale model of the production host (e.g., *Escherichia coli*). We describe the application of the OptForce computational framework to pinpoint engineering modifications (knock-outs/up/down) that are required for the targeted biofuel overproduction. This is accomplished by classifying reactions (and combinations thereof) in the metabolic model depending upon whether their flux values must increase, decrease or become equal to zero to meet the pre-specified overproduction target. A “force set” can then be extracted that contains a sufficient and non-redundant set of reactions that need to be directly changed to meet the production requirements. We apply the integrated framework for the production of 1-butanol, isobutanol, and other alcohols in *E. coli* using the most recent *in silico E. coli* model, *iAF1260*. We also examine the production of succinate in *E. coli*. The proposed computational workflow not only recapitulates existing pathways and engineering strategies but also reveals novel and non-intuitive ones that boost production by using and performing coordinated changes on sometimes distant pathways.

241 COBRA Toolbox 2.0: In Silico Systems Biology Suite

Jan Schellenberger^{1*} (jschelle@ucsd.edu), Richard Que,² Andrei Osterman,³ Bernhard O. Palsson,² and Karsten Zengler²

¹Bioinformatics Program, and ²Dept. of Bioengineering, University of California, San Diego; and ³Burnham Institute for Medical Research, La Jolla, Calif.

http://systemsbiology.ucsd.edu/Downloads/Cobra_Toolbox

Project Goals: This project is focused on a systems-level understanding of biological hydrogen production using *Thermotoga maritima* as a model organism. The project will address the basic science required to improve our understanding of hydrogen production from various carbon sources including glucose, cellulose, starch and xylan by this thermophilic microorganism. The overall goal is 1) to reconstruct the regulatory and metabolic network in

T. maritima using various sets of “omics” data, 2) to integrate regulatory and metabolic networks into one “integrated” genome-scale model, 3) to confirm and validate the ability of the integrated model to predict processing of various environmental signals.

With the advent of whole genome sequencing in the late 1990s, it became possible to build genome scale metabolic models. Since then, this field has undergone a renaissance in terms of 1) size and scope of reconstructions, 2) number of reconstructions and 3) number of analysis tools. The first version of the COBRA (Constraint Based Reconstruction and Analysis) toolbox was published in 2007 to combine many of these emerging methods into one easy to use package. We present version 2.0 here.

The COBRA toolbox is a set of Matlab scripts. Constraint Based models are loaded from various sources into a COBRA specific data structure. The user can then manipulate these models by using the command line or simple scripts. Methods can be chained to create simple data pipelines. The scope of COBRA falls under 8 categories as shown in Figure 1. New to version 2.0 are methods for gap filling, C13 analysis, visualization and thermodynamics. Also new in version 2.0 is a test case suite which gives examples of use of the different methods and expected results.

The objective of the COBRA toolbox is to abstract away details of implementation of constraint based methods. For the end user this reduces development time, cuts down on bugs and makes the code easier to share with other research groups. For the *Thermotoga* project, the COBRA toolbox is used to refine the model, analyze high throughput data, and visualize the results.

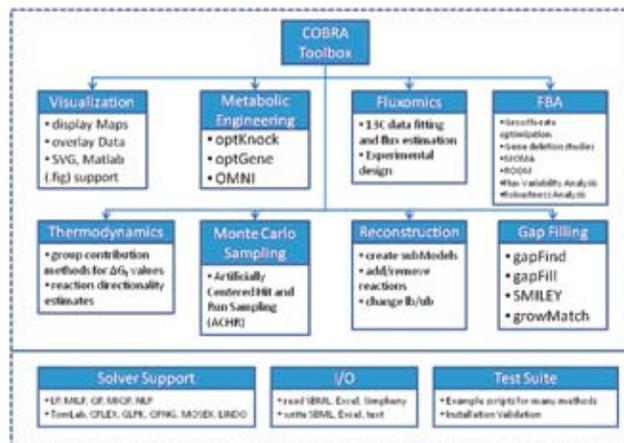


Figure 1: Features of the COBRA toolbox: Top) Scripts are available for methods in eight areas of metabolic systems biology. Bottom) Linear and Quadratic solvers are implemented through a simple yet flexible API in a vendor independent fashion. A set of test scripts are present to validate proper installation as well as demonstrate examples of use.

Reference

1. Becker, S.A., Feist, A.M., Mo, M. L., Hannum, G., Palsson, B.Ø., Herrgard, M.J. Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox., *Nat. Protocols*, 2, 727-738 (2007).

242

Numerical Optimization Algorithms and Software for Systems Biology: Optimality Principles in Nonequilibrium Biochemical Networks

Ronan M.T. Fleming^{1*} (ronan.mt.fleming@gmail.com), Chris Maes,² Ines Thiele,³ Bernhard Ø. Palsson,⁴ Yinyu Ye,⁵ and Michael A. Saunders⁵

¹Science Institute and Center for Systems Biology and ²Center for Systems Biology, University of Iceland, Iceland; ³Institute for Computational and Mathematical Engineering, Stanford University, Stanford, Calif.; ⁴Dept. of Bioengineering, University of California, San Diego; and ⁵Dept. of Management Science and Engineering, Stanford University, Stanford, Calif.

<http://www.hi.is/~rfleming>

Project Goals: Simultaneous prediction of metabolic fluxes and concentrations in *Escherichia coli*.

We derive a new optimization problem on a steady-state non-equilibrium network of biochemical reactions, with the property that mass conservation, energy conservation, the second law of thermodynamics and the proportionality of reaction rate to reactant concentration, all hold at the problem solution. These nonlinear, non-convex constraints are enforced without recourse to linearization or any other form of approximation. This method provides the first computationally tractable method for enforcing thermodynamic, energy, and mass-conservation constraints, at genome scale. Moreover, the formalism has a clear thermodynamic interpretation and suggests a new optimality principle for non-equilibrium biochemical networks. This method may be used for simultaneously predicting reaction rate (flux) and metabolite concentrations in genome-scale biochemical networks. In particular, we demonstrate its utility for simultaneous integration of metabolomic and fluxomic data in *Escherichia coli*, in order to predict unmeasured concentrations and fluxes.