DOE/SC-0140

Sections:
Systems Biology for DOE Energy Missions: Bioenergy
Computing for Bioenergy
Small Business Innovation Research (SBIR) and Small Business Technology Transfer (STTR)

U.S. DEPARTMENT OF ENERGY
Office of Science

Joint Meeting 2011

Genomic Science Awardee Meeting IX

and

USDA-DOE Plant Feedstock Genomics for Bioenergy Awardee Meeting

[Revised: April 14, 2011]

Crystal City, Virginia
April 10-13, 2011

Prepared for the
U.S. Department of Energy
Office of Science
Office of Biological and Environmental Research
Germantown, MD 20874-1290

http://genomicscience.energy.gov

Prepared by
Biological and Environmental Research Information System
Oak Ridge National Laboratory
Oak Ridge, TN 37830
Managed by UT-Battelle, LLC
For the U.S. Department of Energy
Under contract DE-AC05-00OR22725

of Molecular and Cellular Biology, Harvard University, Cambridge, Mass.; and [6]New York University Abu Dhabi, Abu Dhabi, UAE, and Center for Genomics and Systems Biology, Department of Biology, New York University, New York

Goals: The release of the complete genome sequence of *Chlamydomonas reinhardtii* has made this unicellular alga an ideal model for metabolic engineering; however, the annotation of the relevant genes has not been verified yet and the much-needed metabolic network model is currently unavailable. Using the integrated annotation and metabolic network modeling that we recently established (*Manichaikul et al., Nature Methods 2009*), we are engaged in efforts to: 1) assign enzymatic functions to the annotated proteome of *C. reinhardtii*, 2) experimentally verify or refine the structure of the annotated open reading frames (ORFs), and 3) build a genome-wide metabolic network model for the organism based on the assigned metabolic functions.

Results: We used the new *JGI* "filtered transcript models" (Chlre4_best_transcripts and Chlre4_best_proteins), and the *Augustus 5* models released through the *JGI* portal (http://genome.jgi-psf.org/Chlre4/Chlre4.home.html) for both functional assignments and structural annotation verifications. Enzymatic functional assignments were made by associating Enzyme Classification (EC) numbers through reciprocal blast searches against UniProt (and AraCyc) enzyme database (with over 100,000 protein entries). The best match for each translated ORF was identified (with an e-value threshold of 10^{-3}) and the EC number from the UniProt best match was transferred on to the ORF. We extended the EC assignments to the respective paralogs of the ORFs by clustering ORFs using BLASTCLUST (sequence identity cut-off of 35% and sequence length cut-off of 70%) within each annotation group (i.e., *Augustus* 5 and *JGI filtered models*). Altogether, we were able to assign over 900 enzyme annotations to 1,427 *JGI* and to 1,877 *Augustus* models. Over 93% of the EC terms were assigned to both *JGI* and *Augustus* models. We then carried out all possible pairwise alignments between the *JGI* and *Augustus* transcripts that had been assigned the same EC numbers by the above-mentioned procedure. In contrast to the high overlap between the two models in terms of EC assignments, less than half of each set were found to be 100% identical in sequence, indicating that the structural annotation of many of the two sets differ from one another.

To experimentally verify the structure of both *JGI* and *Augustus* ORF models, we carried out open reading frame (ORF) verification by RT-PCR on all ORFs that we had assigned EC numbers to (as well as a set of positive control ORFs). Following optimization of the RT-PCR procedure for high GC content of the *C. reinhardtii* transcriptome, we tested the structure of the metabolic-related ORF models by reverse transcription-PCR of the functionally annotated ORFs. Following cloning, we carried out 454FLX sequencing of the ORFs. Based on alignment of the 454FLX reads to the ORF predicted sequences, we obtained more than 90% coverage for 80% of the metabolic ORFs. Only 99 ORFs were not verified using this experimental pipeline. We obtained expression evidence for 93% of the metabolic ORFs in the algal cells grown under constant light and in the presence of acetate.

Using our in-house generated functional annotation (*described above*), combined with literature and publicly available database resources, we have reconstructed the first genome-scale reconstruction of *C. reinhardtii* metabolic network, accounting for all pathways and metabolic functions indicated. The reconstruction accounts for 1,080 genes, associated with 2,190 reactions and 1,068 unique metabolites. Our reconstruction accounts for multiple wavelengths of light and includes considerable expansion of fatty acid metabolism over previous reconstructions. Further, the metabolic network reconstruction provides a greater level of compartmentalization than existing reconstructions of *C. reinhardtii*, with the inclusion of the lumen as a distinct component of the chloroplast for photosynthetic functionality, and the eyespot used to guide the flagella in phototaxis.

Conclusion: Our validated and comprehensive genome-scale reconstruction of *C. reinhardtii* metabolism provides a valuable quantitative and predictive resource for metabolic engineering toward improved production of biofuels and other commercial targets. The verified metabolic ORF clones will provide the experimental resource needed for downstream experiments and will be made available to the research community.

The DOE support for this research was generously made through a grant from the Office of Science (Biological and Environmental Research), U.S. Department of Energy, grant No. DE-FG02-07ER64496 to J.P. and K.S.-A.

Computing for Bioenergy

100

Multiscale Coarse-Grain Simulation Studies of Cellulosic Biomass

Goundla Srinivas[1]* (srinivasg@ornl.gov), Dennis Glass,[1] Xiaolin Cheng,[1] Jeremy Smith,[1,2] Mark S. Gordon,[3] Yana Kholod,[3] Monica H. Lamm,[3] Sergiy Markutsya,[3] John Baluyut,[3] and Theresa L. Windus

[1]Oak Ridge National Laboratory, Oak Ridge, Tenn.; and [2]Biochemistry and Cellular and Molecular Biology, University of Tennessee, Knoxville; and [3]Applied Mathematics and Computational Science, Ames Laboratory, Department of Chemistry, Iowa State University, Ames

Project Goals:

1. **To develop computationally affordable large scale coarse-grain force field for cellulosic biomass.**
2. **Study structure and dynamics of cellulosic biomass from a micro and macroscopic viewpoint in order to understand biomass recalcitrance.**

Understanding cellulose structure and dynamics is important for improving the process of conversion of biomass to ethanol. Here, we explore cellulose fibril structure and dynamics and catalytic degradable pathways by developing multiscale methods for extending the time and length scales accessible to biomolecular simulation on massively parallel supercomputers. For this purpose, high level quantum mechanics calculations are performed using the fragment molecular orbital (FMO) method in GAMESS (General Atomic and Molecular Electronic Structure System). The FMO simulations are used to study the relevant catalyzed reaction paths, with the solvent represented by the effective fragment potential (EFP) method. The combined FMO/EFP methods are used with molecular dynamic simulations to provide coarse grained (CG) potentials using the force matching method. Further, accurate parameterization of glucose monomers derived from quantum mechanical calculations is used as input for classical molecular dynamics which in turn is utilized to develop a large-scale force field for the cellulose fibril. Without the use of constraints the CG crystalline fibril is found to remain stable (>1μs). Analyzing the static coherent structure factors reveals the ability of the present CG model to differentiate between intra-plane and inter-plane interactions in the fibril. The model is successfully extended to represent various amorphous cellulose fibrils as well. Further, we have carried out a REACH CG analysis of the cellulose fibril using information on correlated motions. Together these simulation results contribute to our understanding of biomass recalcitrance to hydrolysis.

101

Bayesian Computational Predictions of Gene Regulation in the α-Proteobacteria

Charles E. Lawrence[1] (charles_lawrence@brown.edu), **Lee Newberg**[2,3] (lee.newberg@wadsworth.org), William Thompson[1] (william_thompson_1@brown.edu), and **Lee Ann McCue**[4]* (leeann.mccue@pnl.gov)

[1]Center for Computational Molecular Biology and the Division of Applied Mathematics, Brown University, Providence, R.I.; [2]The Wadsworth Center, New York State Department of Health, Albany; [3]Department of Computer Science, Rensselaer Polytechnic Institute, Troy, N.Y.; and [4]Pacific Northwest National Laboratory, Richland, Wash.

http://www.brown.edu/Research/CCMB/

Project Goals: see below

Decreasing America's dependence on foreign energy sources and reducing the emission of greenhouse gases through the development of biofuels are important national priorities. These priorities have catalyzed research on cellulosic ethanol as a clean, renewable energy source to replace fossil fuels, and biohydrogen as a carbon-free energy carrier. Turning these biofuels into viable alternative energy sources requires further research into the degradation of cellulose and fermentation of the resulting sugars, and the metabolic and regulatory networks of biohydrogen production. The genomes of many of the microbial species capable of these processes have been sequenced by the GSP and other programs, and many more are expected soon. These sequence data provide a wealth of information to explore nature's solutions for the production of biofuels. In particular, among the α-proteobacterial species with genome sequence data available are several species with metabolic capabilities of interest, including efficient fermentation of sugars to ethanol and the ability to produce hydrogen. Understanding the regulatory mechanisms and complex interplay of metabolic processes in these species is key to realizing the promise of biofuels. Thus, our research goal is to identify the ensemble of solutions that have been explored by the α-proteobacteria to regulate the metabolic processes key to biofuel production. We have predicted putative regulatory sites and motifs in clades from 63 α-proteobacterial species, to identify regulatory mechanisms and reconstruct the ancestral states of the regulatory networks for the efficient fermentation of sugars to ethanol and the production of biohydrogen.

102

Vertically Integrated Metabolic Engineering Process: Computational and Experimental Optimization of *Escherichia coli* Toward Fatty Acid Production

George Church (guido@genetics.med.harvard.edu), Nicholas Guido,[1]* and Graham Rockwell[1,2]

[1]Dept. of Genetics, Harvard Medical School; and [2]Bioinformatics Program, Boston University

http://arep.med.harvard.edu/

Project Goals: The goal for this work is to integrate computational, mutagenesis and selection tools toward a generalized method for the optimization of production strains synthesizing high-value products. We are modeling and mutagenizing strains of *Escherichia coli* to increase fatty acid production as an initial test case.

Advancements in both computational and metabolic selection are required to complement new in vivo mutagenesis technology enabling metabolic engineering from design to strain production. Here we build new in silico strategies, based on flux balance analysis (FBA), to provide gene level engineering targets for the optimization of complex metabolic processes at the genome scale, previously unobtainable by other methods. Fatty acid production is the test case. These engineering instructions are brought to realization via multiplexed recombination (MAGE), optimizing transcription and translation levels, demonstrating that the model successfully predicts metabolic targets to increase fatty acids. We also develop a unique in vivo selection method to select for cells that have increased fatty acid production as a result of our genetic manipulations. This work shows that computational instructions can be used in conjunction with

mutagenesis and specific selections to create a process of metabolic engineering which represents a general method for the optimization of biological metabolite production.

103

Computational Study of Gating Elements for Proton Pumping in Cytochrome c Oxidase

S. Yang* (syang@chem.wisc.edu), J**. Mitchell**, and Q**. Cui**

BACTER Graduate Program, University of Wisconsin, Madison

Cytochrome c Oxidase (CcO) is the terminal component of the electron transfer pathway in the cellular respiration. It reduces oxygen to water and utilizes the released chemical free energy to pump protons across the membrane in a stoichiometric and efficient fashion. After decades of structural, kinetic, mutagenesis and computational analysis, the identity of the gating element(s) in CcO remains hotly debated. Two studies have been published recently, which focused on the conformational preference of a key side-chain and water molecules. In one of them, the side-chain of a key branching residue, Glu 286 (*R. sphaeroides* number) was found to strongly prefer the downward orientation when deprotonated, which presumably shut off the back flow of protons. Therefore, it was proposed that Glu 286 is the essential valve that minimizes leakage of the pump. In the other study, the orientation of the water molecules in the active site switched depending on the redox state of heme a and the binuclear center. This led to the proposal that water wire reorientation is another major element that controls the appropriate branching of proton transfers, which was shown by elegant kinetic analysis as crucial to the thermodynamic efficiency of proton pumping.

In this report, we systematically analyze the relevant energetics using elaborate molecular models for CcO. The results show that neither Glu 286 rotation nor water wire reorientations acts as the robust gating factor, and the observations from the previous studies are most likely due to the use of simplified models of CcO, which highlights the importance of carefully benchmarking the molecular model for an analysis of complex proton pumping systems.

104

Uncovering Mechanisms of Cellobiose Hydration

Madeleine Pincu[1]* (mpincu@uci.edu), John P. Simons,[2] and **Robert B. Gerber**[3]

[1]University of California-Irvine, Department of Chemistry; [2]University of Oxford, Chemistry Department, Physical and Theoretical Chemistry Laboratory, Oxford, United Kingdom; and [3]The Hebrew University, The Fritz Haber Research Center, Jerusalem, Israel, and University of California-Irvine, Department of Chemistry

Project Goals: We aim to study with theoretical tools saccharides—the most abundant bio-molecules on earth. Our studies are designed to shed light on the fundamental structure, energy configuration and temperature-dependent dynamic behaviour of saccharides in isolation and in the hydrated state. We also seek to understand the effect of ions and salts on these compounds as these have important implications to the understanding of biological systems. To validate our theoretical findings, we collaborate with several experimentalists, in particular with Professors J. P. Simons (of Oxford) and Ilana Bar (of Beer Sheva University), who provide us with spectroscopy data for some of our systems. We expect this work to illuminate our understanding of the chemistry of polymers of sugar, in particular of cellulose—a source of bio-energy. In the process, we will develop and validate new computational tools designed to simulate moderate to large size bio-molecular polymers within reasonable time frames.

We explore conformational and structural dynamics of cellobiose and their micro-hydrated complexes isolated in the gas phase, with a combination of theoretical and experimental tools. Their structures at low temperature have been determined through double resonance, IR-UV vibrational spectroscopy conducted under molecular beam conditions, substituting D_2O for H_2O to separate isotopically, the carbohydrate (OH) bands from the hydration (OD) bands. Car-Parrinello (CP2K) simulations, employing dispersion corrected DFT potentials and conducted 'on-the–fly' from ~20K to ~300K, were used to explore the consequences of raising the temperature on the infra-red vibration spectrum.

Our findings are: (1) Good agreement between experiment and theory, which gives us confidence in our theoretical methods. (2) Increasing hydration from 1 to 10 H_2O molecules reveals a persistent motif developing, in which hydration proceeds with water forming a bridging network across the trans glycosidic bond, (from OH6' to OH4'). This "bridge" remains stable for the 5+ps duration of the (CP2K) dynamic simulation at ~300K, despite individual fluctuations in the intra- and inter-molecular hydrogen bonding. (3) Although in nature the trans-cellobiose conformer is more stable than cis, in solution at room temperature, and although our calculations predict that their micro-hydrated energies should equalize for a water cluster containing ~12 molecules, we find that a cluster of 10 water molecules organizes with either cis or trans cellobiose at surfactant position relative to that of the water cluster.

The connection between hydration in small complexes and in the extended cluster is discussed and important implications of the results for properties of saccharides are suggested.

105

The Transcriptional Architecture of *Thermotoga maritima*

H. Latif[1]* (halatif@ucsd.edu), V.A. Portnoy,[1] Y. Tarasova,[1] A.C. Schrimpe-Rutledge,[2] H. Nagarajan,[1] R.D. Smith,[2] J.N. Adkins,[2] D.H. Lee,[1] Y. Qiu,[1] **B.O. Palsson**,[1] and K**. Zengler**[1]

[1]Department of Bioengineering, University of California San Diego, La Jolla; and [2]Pacific Northwest National Laboratory, Richland, Wash.

Project Goals: Validate and improve the existing genome-scale metabolic reconstruction of *T. maritima*. Reconstruct the transcriptional regulatory network (TRN) of *T. maritima*. Generate and use the integrated in silico model of *T. maritima* to assess growth conditions for optimal production of hydrogen.

The hydrogen producing, hyperthermophilic microorganism *Thermotoga maritima* is the focus of this work. Previously, the *T. maritima* metabolic model has been constructed containing 479 metabolic genes, 565 metabolites and 646 reactions. The metabolic model was then updated to include 478 protein structures yielding the first three dimensional metabolic reconstruction. In addition, we have extended the metabolic model by including all reactions for transcription and translation (see abstract by Lerman et al.), laying the ground work for *in silico* gene expression predictions. Fundamental for exploring the flow of information from genes, to transcripts, to proteins is the structural and operational genome annotation of *T. maritima*. This multi-level annotation is being accomplished through the development and integration of various genome-wide, experimentally derived data. Customized high density, whole genome microarrays together with RNAseq data were utilized for elucidation of the *T. maritima* transcriptome. Growth conditions included exponential phase, hydrogen induced stationary phase, heat shock, acid shock and carbon limiting stationary phase. Combining these data types resulted in expression of >97% of the genome. Furthermore, proteomic data was generated from exponential phase and stationary phase cultures using LC-MS/MS and mapped against a stop-to-stop database. Over 41,000 unique peptides were mapped to ~1,400 coding regions, covering over 73% of the entire proteome. The integration of proteome and transcriptome data provided a highly improved structural annotation (ORFs) of the *T. maritima* genome. The operational annotation, consisting of operons and transcription units, was subsequently resolved by including genome-wide transcription start site (TSS) data, determined at single base pair resolution. Integrating transcriptomics, proteomics, and TSS data yielded the transcriptional architecture of *T. maritima.* This data now provides the basis upon which the transcriptional regulatory network for this hyperthermophilic bacterium will be built.

106

Discovery of Phenotype-Related Biochemical Processes with Application to Biological Hydrogen Production

Andrea M. Rocha[1]* (amrocha@mail.usf.edu), Matthew C. Schmidt,[2,3] William Hendrix,[2,3] Kanchana Padmanabhan,[2,3] Yekaterina Shpanskaya,[2] James R. Mihelcic,[1] and **Nagiza F. Samatova**[2,3] (samatovan@ornl.gov)

[1]University of South Florida, Tampa; [2]North Carolina State University, Raleigh; and [3]Oak Ridge National Laboratory, Oak Ridge, Tenn.

Project Goals: The purpose of this project is to develop computational approaches to identify phenotype-related cellular subsystems and demonstrate predictability of these techniques to hydrogen production by dark fermentative and acid-tolerant bacteria.

Many microbial communities in natural environments exhibit phenotypes that directly cause particular diseases, convert biomass or wastewater to energy, or degrade various environmental contaminants. Understanding of how interacting biochemical pathways in these communities realize specific phenotypic traits (e.g., carbon fixation, hydrogen production) is critical for addressing health, bioremediation, or bioenergy problems that, arguably, cannot be solved by experiments alone.

To complement experimental approaches, in this study, we develop graph-theoretical and statistical methods for *in silico* prediction of the cellular subsystems that are related to the expression of a target phenotype. First, our Network Instance-Based Biased Subgraph Search (**NIBBS**) is capable of comparing hundreds of genome-scale metabolic networks to identify *metabolic subsystems* that are statistically biased toward phenotype-expressing organisms. NIBBS accurately approximates the set of all *biased* network motifs in a set of metabolic networks. From the results obtained, for example, we were able to predict a number of metabolic subsystems that are likely related to biohydrogen production, such as acetate and butyrate fermentation, fatty acid biosynthesis, amino acid metabolism, and nitrogen metabolism. Also, the genome-scale comparative network analysis enabled us to predict pathway cross-talks, including those involved in production of Acetyl-CoA.

Second, our α,β*-motifs* approach allows for identification of *functional modules* that, in addition to metabolic subsystems, could include their regulators, sensors, transporters, and even uncharacterized proteins that are predicted to be related to the target phenotype. By comparing hundreds of genome-scale networks of functionally associated proteins, our method identifies those functional modules that are enriched in at least α networks of phenotype-expressing organisms but may still appear in no more than β networks of organisms that do not exhibit the target phenotype. Using α,β*-motifs* approach, for instance, we were able to identify

clusters of genes responsible for synthesis, metal insertion, or regulation of hydrogenase and nitrogenase enzymes complexes. Within hydrogen producers, these two complexes play important roles in production of hydrogen.

Third, our Dense ENriched Subgraph Enumeration (DENSE) algorithm allows for incorporating partial *prior* knowledge about the proteins involved in a phenotype-related process and enriches that knowledge with newly identified sets of functionally associated proteins present in individual phenotype expressing organisms. When applied to a network of functionally associated proteins in the dark fermentative, hydrogen producing bacterium, *Clostridium acetobutylicum*, we were able to predict known and novel relationships including those with regulatory, signaling, and uncharacterized proteins.

Finally, we integrate the information obtained from application of these complementary approaches not only to allow for further understanding of networks related to hydrogen production, but also to provide insights into metabolic networks and system controls involved in expression of microbial traits. Such information is necessary for advancing engineering approaches that result in more efficient biohydrogen production.

This research is supported by both the Office of Biological and Environmental Research and by the Office of Advanced Scientific Computing Research of the U.S. Department of Energy. The work by A.M.R. was supported by the Delores Auzenne Fellowship and the Alfred P. Sloan Minority PhD Scholarship Program.

107

Ultrascale Computational Modeling of Phenotype-Specific Metabolic Processes in Microbial Communities

Nagiza F. Samatova[1,2]* (samatovan@ornl.gov), **Chongle Pan**[2] (pcl@ornl.gov), **Robert (Bob) L. Hettich**[2] (hettichrl@ornl.gov), and **Jill Banfield**[3] (jbanfield@berkeley.edu)

[1]North Carolina State University, Raleigh; [2]Oak Ridge National Laboratory, Oak Ridge, Tenn.; and [3]University of California, Berkeley

Project Goals: Computational modeling methods are aimed to be developed for revealing phenotype-specific metabolic processes and their crosstalks and applied to the critical DOE problem of acid mine drainage (AMD). The apex of the project is a procedure for: (1) identification and expression-level characterization of phenotype-related genes; (2) reconstruction of phenotype-specific metabolic pathways; and (3) elucidation of symbiotic/competing interplays between these pathways.

Many microbial communities in natural environments exhibit phenotypes of interest to DOE, such as oxidization of pyrite ore that leads to acid mine drainage, breaking down the lignocellulosic barrier of biomass, and the biodegradation of various environmental contaminants. Addressing bioremediation and bioenergy problems will require an understanding of how interacting biochemical pathways in these communities realize specific phenotypic traits (nitrogen and carbon fixation, resistance to heavy metals, tolerance to pH perturbations, etc.). This problem cannot be solved by experiments alone. There is a need for computational modeling methods that will reveal phenotype-related "signals'" and their combinatorial interplay by comparing potentially hundreds of microorganisms with millions of genes organized into thousands of metabolic pathways, which are uncertainly defined. These methods are being applied to the acid mine drainage (AMD) community to help answer long-standing questions regarding the role of fine-scale variation in adaptation to dynamic environmental conditions and community composition.

We performed and published quantitative proteomics comparison of field AMD biofilm to laboratory AMD biofilm. To enable laboratory studies of growth, production, and ecology of AMD microbial communities, a culturing system was designed to reproduce natural biofilms, including organisms that are recalcitrant to cultivation. A comprehensive metabolic labeling-based quantitative proteomic analysis was utilized to verify that natural and laboratory communities were comparable at the functional level. Results confirmed that the composition and core metabolic activities of laboratory-grown communities were similar to a natural community, including the presence of active, low abundance bacteria and archaea that have not yet been isolated. However, laboratory growth rates were slow compared to natural communities, and this correlated with increased abundance of stress response proteins for the dominant bacteria in laboratory communities. Modification of cultivation conditions reduced the abundance of stress response proteins and increased laboratory community growth rates. This was the first application of a metabolic labeling-based quantitative proteomic analysis at the community level and resulted in a model microbial community system ideal for testing physiological and ecological hypotheses

We also performed and published quantitative proteomics study of pH perturbation to functionally characterize laboratory-cultivated acidophilic communities sustained in pH 1.45 or 0.85 conditions. The distributions of all proteins identified for individual organisms indicated biases for either high or low pH, and suggests pH-specific niche partitioning for low abundance bacteria and archaea. Although the proteome of the dominant bacterium, *Leptospirillum* group II, was largely unaffected by pH treatments, analysis of functional categories indicated proteins involved in amino acid and nucleotide metabolism, as well as cell membrane/envelope biogenesis were generally more abundant at high pH. Results indicate solution pH may play an important role in shaping community membership and biofilm structure. Proteomic analysis of communities also revealed differences in the number of phage proteins detected across biological replicates. Stochastic spatial heterogeneity of viral outbreaks may also play a role in shaping community

structure. Quantitative proteomic comparisons showed distinct differences in community composition and metabolic function of individual organisms during different pH treatments, and confirms the importance of specific geochemical parameters that 'fine-tune' acidophilic microbial community structure and function at the species and strain level.

In support of these proteomics studies, we developed, published, and released a *de novo* sequencing algorithm, **Vonode**, to exploit the potential of high-resolution MS/MS data by using a unique tag scoring function and a novel type of spectrum graphs. When compared to an established *de novo* sequence algorithm, PepNovo v2.0, the Vonode algorithm inferred sequence tags for 11,422 (vs. 2,573) spectra at an average length of 5.5 (vs. 6.0) residues with 84% (vs. 65%) accuracy of inferred consensus sequence tags.

We also developed and released a module for **ProRata**, a data analysis algorithm for quantitative proteomics, to address the following two critical needs: (1) to combine multiple replicates and to assess the reproducibility of measurements to obtain reliable quantification information and (2) to compare two unlabeled field samples of interest to a labeled reference sample grown in the laboratory because we cannot label metabolic labeling to a field sample.

We developed graph-theoretical and statistical methods for *in silico* prediction of the cellular subsystems that are related to the expression of a target phenotype: (1) the Network Instance-Based Biased Subgraph Search (NIBBS) is capable of comparing hundreds of genome-scale metabolic networks to identify *metabolic subsystems* that are statistically biased toward phenotype-expressing organisms; (2) ther α,β-motifs approach allows for identification of *functional modules* that, in addition to metabolic subsystems, could include their regulators, sensors, transporters, and even uncharacterized proteins that are predicted to be related to the target phenotype; and (3) the Dense ENriched Subgraph Enumeration (DENSE) algorithm allows for incorporating partial *prior* knowledge about the proteins involved in a phenotype-related process and enriches that knowledge with newly identified sets of functionally associated proteins present in individual phenotype expressing organisms.

From the results obtained, for example, we were able to predict cellular subsystems that are likely related to various phenotypes such as acid tolerance, biohydrogen production, aerobic and anaerobic respiration, etc. Also, the genome-scale comparative network analysis enabled us to predict pathway crosstalks and to perform a systematic study on various mechanisms underlying crosstalks.

This research is supported by both the Office of Biological and Environmental Research and by the Office of Advanced Scientific Computing Research of the U.S. Department of Energy.

108

Functional Annotation of Hierarchical Modularity

Kanchana Padmanabhan[1,2]* (kpadman @ncsu.edu), Kuangyu Wang,[1] and **Nagiza F. Samatova**[1,2] (samatovan@ornl.gov)

[1]North Carolina State University, Raleigh; and [2]Oak Ridge National Laboratory, Oak Ridge, Tenn.

Project Goals: The goals of this project are (1) To develop a method that would assess functional coherence and provide annotation of hierarchically structured modules; (2) To design a biologically relevant functional coherence scoring metric; (3) To reconstruct a hierarchical modularity of cellular organization from a "bag of genes" allowing multi-functional genes or proteins to be part of multiple functional modules in the hierarchy.

Network motifs are recurring, statistically significant patterns of node interactions that act as building blocks of complex networks. In biological networks of molecular interactions in a cell, such as protein-protein interaction networks or gene transcriptional regulatory networks, network motifs that are biologically relevant are also functionally coherent, or form functional modules, such as a ribosomal module synthesizing proteins or a signal transduction system governing bacterial chemotaxis. These functionally homogenous modules combine in a hierarchical manner into larger, less cohesive subsystems, thus revealing one of the essential design principles of system-level cellular organization and function—*hierarchical modularity*.

Arguably, hierarchical modularity has not been explicitly taken into consideration by most, if not all, functional annotation systems. Instead, a functional module is traditionally viewed as a "bag of genes," and methods that assess its functional coherence, or provide functional annotation, analyze this bag in its entirety. As a result, the existing methods would often fail to assign a statistically significant functional coherence score to biologically relevant molecular machines (see Table 1).

To address this gap, we developed a methodology for hierarchical functional annotation of biological cal network motifs. Given the hierarchical taxonomy of functional concepts (e.g., Gene Ontology) and the association of individual genes or proteins with these concepts (e.g., GO terms), our method will assign a *Hierarchical Modularity Score* (**HMS**) to each node in the hierarchy of functional modules; the HMS score and its *p*-value measure functional coherence of each module in the hierarchy. While existing methods annotate each module with a set of "enriched" functional terms in a bag of genes, our complementary method provides hierarchical functional annotation of the modules and their components that are hierarchically organized.

A hierarchical organization of functional modules often comes as a bi-product of cluster analysis of gene expression

Table 1: Statistical significance of protein pairs' functional coherence in *Saccharomyces cerevisiae*.

Protein Pair				*p*-value (pair/module/module size)					Ref
ID	Description	ID	Description	HMS	[1]	[2]	[3]	[4]	
SRB2	Subunit of the RNA polymerase II mediator complex	RPB9	RNA polymerase II subunit B12.6	0.022/ 2.2 e^{-16}// 57	0.48	0.12/ 1.0/ 57	0.333/ 1.0/ 57	0.98/ 1.0/ 57	[5]
SNU13	RNA binding protein	DIB1	17-kDa component of the U4/ U6aU5 tri-snRNP	0.003/ 0.003/ 2	0.23	0.01/ 0.01/ 2	0.1/ 0.1/ 2	0.42/ 0.42/ 2	[6]
HAP1	Zinc finger transcription factor involved in the complex regulation of gene expression in response to levels of heme and oxygen	RPM2	Protein subunit of mitochondrial RNase P	0.006/ 3.207e^{-8} / 3	0.214	0.1/ 1.0/ 3	0.02/ 1.0/ 3	0.51/ 1.0/ 3	[8]
NSR1	Nucleolar protein that binds nuclear localization sequences	DBP2	Essential ATP-dependent RNA helicase of the DEAD-box protein family	0.012/ 0.012/ 2	0.44	0.1/ 0.1/ 2	0.13/ 0.13/ 2	0.74/ 0.74/ 2	[7]

Table 2: Skill metrics for *Saccharomyces cerevisiae* KEGG experiments.

	KEGG	Heidke Score	Pierce Score	Gerrity Score
HMS	Level 1	1	1	1
[1]		1	1	1
HMS	Level 2	0.917	0.923	0.887
[1]		0.61	0.79	0.65
HMS	Level 3	0.916	0.923	0.931
[1]		0.71	0.73	0.73

data or protein interaction data. Otherwise, our method will automatically build such a hierarchy by directly incorporating the functional taxonomy information into the hierarchy search process and by allowing multi-functional genes to be part of more than one component in the hierarchy. In addition, its underlying HMS scoring metric ensures that functional specificity of the terms across different levels of the hierarchical taxonomy is properly treated.

We have evaluated our method using *Saccharomyces cerevisiae* data from KEGG and MIPS using GO ontology as the underlying hierarchical taxonomy of functional concepts. Table 2 illustrates biological relevance of the hierarchical modularity built by our method from a set of genes in various KEGG pathways: at various levels of the hierarchy, the corresponding modules match quite well with the manually-curated hierarchy of pathways in KEGG. We obtained similar results for the protein complexes in the MIPS database. We provide literature evidence for several functional modules that have been identified by HMS as significant both at the protein pairs and at the module levels but have been missed by some existing methods (see examples in Table 1).

This research is supported by both the Office of Biological and Environmental Research and by the Office of Advanced Scientific Computing Research of the U.S. Department of Energy.

References

1. Pandey J., *et al.* Functional characterization and topological modularity of molecular interaction networks. BMC *Bioinformatics*, 2010, vol: 11, pp: S35.
2. Bauer S., *et al.* Ontologizer 2.0-a multifunctional tool for go term enrichment analysis and data exploration. *Bioinformatics*, 2008, vol: 24(14), pp: 1650.
3. Boyle E. I., *et al.* GO::TermFinder–open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics*, 2004, vol: 20(18), pp: 3710.
4. Huang D. W., *et al.* Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 2009, vol: 4(1), pp: 44.
5. Myer V. E., *et al.* RNA polymerase II holoenzymes and subcomplexes. *The Journal of Biological Chemistry*, 1998, vol: 273, pp: 27757.
6. Scott W. S., *et al.* Purification of the yeast U4/U6U5 small nuclear ribonucleoprotein particle and identification of its proteins. *Proceedings of the National Academy of Sciences*, 1999, vol: 96(13), pp: 7226.
7. Tai S.L., *et al.* Acclimation of *Saccharomyces cerevisiae* to low temperature: A chemostat-based transcriptome analysis. *Molecular Biology of the Cell*, 2007, vol: 18, pp: 5100.
8. Hach A., *et al.* A new class of repression modules is critical for heme regulation of the yeast transcriptional activator Hap1. *Molecular and Cellular Biology*, 1999, vol: 19, pp: 4324.

Student Oral Presentation–Tuesday

109

Elucidation of Symbiotic and Competing Metabolic Pathway Crosstalks

Kuangyu Wang[1]* (kwang2@ncsu.edu), Kanchana Padmanabhan,[1,2] and **Nagiza F. Samatova**[1,2] (samatovan@ornl.gov)

[1]North Carolina State University, Raleigh; and [2]Oak Ridge National Laboratory, Oak Ridge, Tenn.

Project Goals: This project is focused on a systems-level understanding of biological nature of metabolic pathway crosstalks in bacterial organisms. The overall goal is 1) to provide a systematic characterization of metabolic pathway crosstalk mechanisms, such as those due to physical and genetic interactions, 2) to construct pathway crosstalk networks based on various types of functional genomics data and provide biological interpretations for discovered patterns in such networks, and 3) to confirm and validate the predicted crosstalks using literature searches and gene expression data across different conditions using *Escherichia coli* as our model organism.

Given the complex nature of a phenotype in a microbial community, it is likely that not a single metabolic pathway but a group of phenotype-related pathways jointly function to accomplish a particular phenotype. For example, in a community, interplays between pathways may exist not only in a cell, but also across cells of different species. From a metabolite-centric perspective, pathways can interact cooperatively by exchanging intermediates, competitively by inputting or outputting common intermediates, or incompatibly by requiring different conditions to function. From a gene-centric perspective, genes in pathways can be regulated in a correlated manner or an anti-correlated manner. In this study, we aim to systematically characterize the mechanisms underlying various metabolic pathway cross-talks.

Traditionally, metabolic pathways are viewed as a linear sequence of metabolic reactions, where the product metabolites of one reaction are used as the substrate metabolites of the next reaction. Arguably, such an abstraction de-emphasizes the prevalence of crosstalks (non-additive interactions) between individual pathways. A better understanding of underlying mechanisms of pathway crosstalks will help predict processing of various environmental signals and conduct metabolic engineering for the purpose of bioenergy production. Using *Escherichia coli* as our model organism, here we present a systematic study that fuses genomics and proteomics information in order to predict crosstalks between metabolic pathways such as those extracted from KEGG.

A large number of non-additive interactions or crosstalks exist globally; however, the mechanisms underlying pathway crosstalks are limited. Although there is no universal categorization of biological mechanisms that underlie crosstalks, in this study, we have made attempts towards a possible classification that takes into consideration various types of evidences such as 1) physical via direct binding, 2) biochemical via phosphorylation, and 3) functional via transcriptional regulation.

By analyzing various types of biological networks, we infer the clues about putative metabolic pathway crosstalks. We further analyze those clues for possible positive and negative correlations and identify those that are hypothesized to be orthogonal. We devise methodology for understanding the higher-level organization of pathway crosstalk networks derived from these clues and provide biological interpretation of and literature evidence for the discovered patterns in those networks.

This research is supported by both the Office of Biological and Environmental Research and by the Office of Advanced Scientific Computing Research of the U.S. Department of Energy.

110

Numerical Optimization Algorithms and Software for Systems Biology: von Bertalanffy 1.0 : A COBRA Toolbox Extension to Thermodynamically Constrain Metabolic Models

Ronan M.T. Fleming[1]* (ronan.mt.fleming@gmail.com), Ines Thiele,[2] and **Michael A. Saunders**[3]

[1]Science Institute and Center for Systems Biology, University of Iceland, Reykjavik; [2]Center for Systems Biology and Faculty of Industrial Engineering, Mechanical Engineering and Computer Science, University of Iceland, Reykjavik; and [3]Dept of Management Science and Engineering, Stanford University

http://opencobra.sourceforge.net

Project Goals: This project aims to reconstruct genome-scale models of metabolism and macromolecular synthesis and to develop algorithms capable of solving the resulting large, stiff and ill-scaled matrices. We aim to combine state of the art reconstruction and constraint-based modeling and 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*.

In flux balance analysis of genome scale stoichiometric models of metabolism, the principal constraints are uptake or secretion rates, the steady state mass conservation assumption and reaction directionality. Here, we introduce an algorithmic pipeline for quantitative assignment of reaction directionality in multicompartmental genome scale models based on an application of the second law of thermodynamics to each reaction. Given experimental or computationally estimated standard metabolite species Gibbs energy and metabolite concentrations, the algorithms bounds reaction Gibbs energy, which is transformed to *in vivo* pH, temperature, ionic strength and electrical potential. This toolbox may be used to distinguish between thermodynamically feasible and thermodynamically infeasible metabolic pathways. This is a critical first step in the rational design of novel strategies for production of biofuels. This cross platform MATLAB extension to the COnstraint-Based Reconsruction and Analysis (COBRA) toolbox, is computationally efficient, extensively documented and open source.

Student Oral Presentation–Tuesday

111

Numerical Optimization Algorithms and Software for Systems Biology: Existence of Positive Equilibria for Mass Conserving and Mass-Action Biochemical Reaction Networks with a Single-Terminal-Linkage Class

Santiago Akle[1]* (akle@stanford.edu), Onkar Dalal[1] (onkar@stanford.edu), Ronan Fleming[2] (ronan.mt.fleming@gmail.com), **Michael Saunders**[3] (saunders@stanford.edu), Nicole Taheri[1]* (ntaheri@stanford.edu), and **Yinyu Ye**[3] (yinyu-ye@stanford.edu)

[1]Institute for Computational and Mathematical Engineering, Stanford University, Stanford, Calif.; [2]Science Institute and Center for Systems Biology, University of Iceland, Reykjavik; and [3]Department of Management Science and Engineering, Stanford University, Stanford, Calif.

Project Goals: Develop a convex optimization algorithm for computing thermodynamically feasible reaction fluxes in a general instance of a genome-scale integrated metabolic and macromolecular biosynthetic network.

A steady state of a chemical reaction network is a set of chemical concentrations that remain constant for the induced reaction rates. In this work we assume the *law of mass-action* governs the rate of the reactions, i.e. the rate of a reaction is proportional to the concentrations of the participating species. More specifically, if Y_{ij} is the stoichiometric coefficient for species i in reaction j, k_j is the thermodynamically feasible rate constant for reaction j, and c_i is the concentration of species i, then the rate of reaction j is

$$v_j = k_j \prod_i c_i^{Y_{ij}}.$$

Assuming the reactions are *mass conserving*, and that the directed graph corresponding to the set of reactions forms a strongly connected component, we show that, regardless of the rate constants, there exists at least one steady state where all concentrations are positive.

We also establish the parallel between steady states and a fixed point of a mapping that arises from solving a strictly convex optimization problem, which allows us to find such steady states in randomly constructed networks.

Research supported in part by DOE Grant DE-SC0002009
Research supported in part by NSF Grant GOALI 0800151and DOE Grant DE-SC0002009

112

*Zea mays i*RS1563: A Comprehensive Genome Scale Model of Maize Metabolism

Rajib Saha* (rus184@psu.edu), Patrick F. Suthers, and **Costas D. Maranas**

Department of Chemical Engineering, The Pennsylvania State University, University Park

Project Goals: Develop a genome-scale model for maize that meets rigorous standards on gene-protein-reaction (GPR) associations, elementally and charged balanced reactions and a biomass reaction abstracting the relative contribution of all biomass constituents.

The scope and breadth of genome-scale metabolic reconstructions has continued to expand over the last decade. Herein, we introduce a genome-scale model for a plant with direct applications to food and bioenergy production (i.e., maize). Maize annotation is still underway which introduces significant challenges in the association of metabolic functions to genes. The developed model *Zea mays i*RS1563 (see Figure 1) is designed to meet rigorous standards on gene-protein-reaction (GPR) associations, elementally and charged balanced reactions and a biomass reaction abstracting the relative contribution of all biomass constituents. *Zea mays i*RS1563 can be viewed as a mathematically structured database of maize metabolism. The metabolic network contains 1,563 genes and 1,825 metabolites involved in 1,985 reactions from primary and secondary maize metabolism. For approximately 42% of the reactions direct literature evidence for the participation of the reaction in maize was found. As many as 445 reactions and 369 metabolites are unique to the maize model compared to the AraGEM model for *A. thaliana*. 674 metabolites and 893 reactions are present in *Zea mays i*RS1563 that are not accounted for in maize C4GEM. All reactions are elementally and charged balanced and localized into six different compartments (i.e., cytoplasm, mitochondrion, plastid, peroxisome, vacuole and extracellular). *Zea mays i*RS1563 accounts for the fact that photosynthesis in maize (i.e., a C4 plant) occurs in two separate cell types (i.e., mesophyll cell and bundle sheath cell). A biomass equation is established that quantifies the relative abundance of different constituents of dry plant cell biomass. GPR associations are also established based on the functional annotation information and homology prediction accounting for monofunctional, multifunctional and multimeric proteins, isozymes and protein complexes. We describe results from performing flux balance analysis under different physiological conditions, (i.e., photosynthesis, photorespiration and respiration) of a C4 plant and also explore model predictions against experimental observations for two naturally occurring mutants (i.e., *bm1* and *bm3*). The developed model corresponds to the largest and more complete to-date effort at cataloguing metabolism for a plant species.

By making use of high throughput enzymatic assays, proteomic and transcriptomic data across different parts of the maize plant, *Zea mays i*RS1563 could serve as the starting

point for the development of tissue-specific maize models as well as for other important C_4 plants such as Sorghum and switch grass. By accounting for both primary and some secondary metabolism pathways of maize, *Zea mays* *i*RS1563 can be used to explore *in silico* the effect of genetic modifications aimed at plant cell wall modification and/or starch storage on the overall metabolic state of the plant (e.g., biomass precursor availability, cofactor balancing, redox state, etc.). By taking full inventory of plant metabolism optimal gene modifications could be pursued for a variety of targets in coordination with experimental techniques. These may include (i) increase cellulose and hemicellulose production, (ii) starch yield, (iii) disruption of recalcitrant lignin subunits, and (v) nitrogen efficiency.

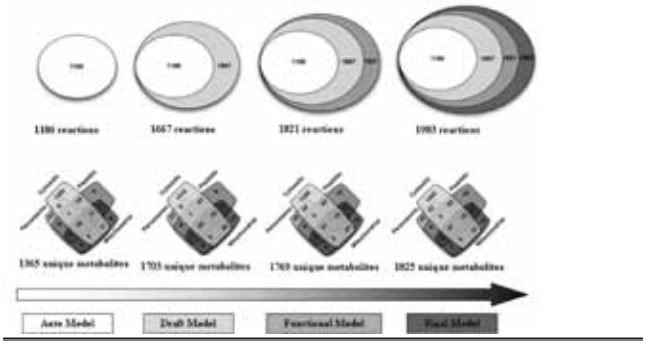

Figure 1. Successive development of *Zea mays* *i*RS1563: evolution of total number of reactions and metabolites moving from Auto, Draft, Functional and Final models

113

OptForce Directed *Escherichia coli* Genetic Modifications for Maximizing Malonyl-CoA Availability

Sridhar Ranganathan[1]* (sur152@psu.edu), Peng Xu,[2] Hila Dvora,[2] Mattheos A.G. Koffas,[2] and **Costas D. Maranas**[3]

[1]Huck Institutes of Life Sciences, [2]Chemical and Biological Engineering, and [3]Department of Chemical Engineering, The Pennsylvania State University, University Park

Project Goals: Develop an integrated computational and experimental study aimed at improving the availability of malonyl-CoA in *Escherichia coli* by deploying our OptForce methodology to predict the minimal set of genetic manipulations.

Malonyl-CoA is an important precursor metabolite for the biosynthesis polyketides, fatty acids, biofuels (microdiesel) and plant-specific secondary metabolites. However, malonyl-CoA is directly consumed for the production of amino acids, phospholipids and biomass leaving behind only a residual amount available for overproduction targets. Metabolic engineering of microorganisms for products derived from malonyl-CoA requires achieving a fine balance between malonyl-CoA available for cellular growth and product synthesis. Efficiently harnessing malonyl-CoA continues to be one of the key barriers in the biosynthesis of long-chain biofuels ($<C_6$) and pharmaceutical compounds. Metabolic engineering efforts aimed at improving intracellular malonyl-CoA availability have so far focused on reaction steps adjacent to malonyl-CoA. Examples include overexpression of acetyl-CoA carboxylase that directly produces malonyl-CoA and elimination of competing bioconversions catalyzed by acetate kinase and alcohol dehydrogenase. Even though existing genetic manipulations have managed to significantly improve malonyl-CoA availability, numerous engineering possibilities are yet to be explored. Our group recently introduced the computational strain design procedure OptForce[1] that can hierarchically pinpoint genetic interventions that lead to yield improvements.

In this work, we present milestones achieved from an integrated computational and experimental study aimed at improving the availability of malonyl-CoA in *Escherichia coli*. We deployed our OptForce methodology to predict the minimal set of genetic manipulations in *E. coli* wild-type strain BL21 Star™ that overproduces malonyl-CoA. Using OptForce we identified a hierarchy of metabolic modifications that cooperatively redirect carbon flux towards malonyl-CoA while ensuring biomass production at pre-specified rates. Interventions predicted by OptForce can be ranked based on their quantitative impact towards achieving the overproduction target thus providing a way to prioritize the implementation of genetic interventions. In this work, up-regulations for glycolytic reactions, glyceraldehyde-3-phosphate dehydrogenase (GAPD) and phosphoglycerate kinase (PGK), pyruvate dehydrogenase (PDH), and acetyl-CoA carboxylase (ACCOAC) were predicted as the most important. All of these interventions directly contribute towards precursors of the malonyl-CoA pathway. OptForce suggested reducing the activity of TCA reactions (malate dehydrogenase (MDH), fumarase (FUM) and aconitase (ACONTa/b)) instead of eliminating them (to ensure production of all biomass components). In addition, knockouts for succinyl-CoA synthetase (SUCOAS) and propionyl-CoA:succinyl-CoA transferase (PPCSCT) were predicted that reduce the drain of malonyl-CoA towards by-products. The complete set of engineering interventions and alternatives suggested by OptForce can be represented in the form of a logic decision tree (see Figure 1).

By successively implementing the hierarchy of OptForce suggestions, we have successfully constructed a recombinant strain of *E. coli* that exhibits improved malonyl-CoA levels. We demonstrate the efficacy of this strain by incorporating a set of heterologous pathways that use three moles of malonyl-CoA for the production of naringenin. Naringenin is a flavanone that is a low-molecular weight plant-specific polyphenolic compound. Upon translating the OptForce predictions to the gene level using the gene-protein-reaction (GPR) associations for these reactions, we implemented genetic modifications (see Figure 1) for overproducing naringenin in *E. coli*. The evolution of naringenin yield as more interventions are accumulated highlights the synergistic effect of combining beneficial mutants (Δ*fumC* and Δ*sucC*) and overexpression targets (*acc*, *pgk*, *gapA* and *pdh*) predicted

by OptForce. Specifically, a titer of 474 mg/L of naringenin production was observed which is up-to-date the highest yield achieved in a lab-scale fermentation process.

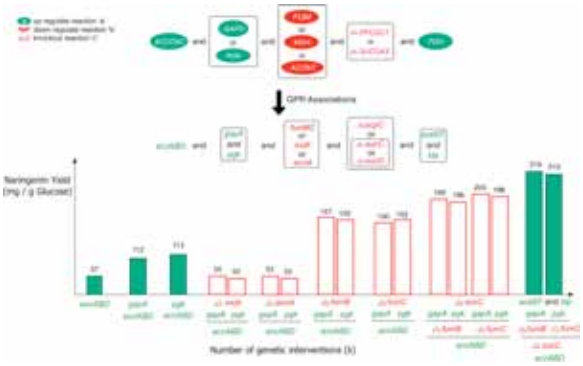

Figure 1. Reaction interventions predicted by OptForce procedure and the corresponding gene associations for the overproduction of malonyl-CoA and naringenin synthesis.

Reference

1. Ranganathan, S., Suthers, P. F., Maranas, C. D., OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions. *PLoS Comput Biol* 2010, 6, e1000744.

Student Oral Presentation–Monday

114

MetRxn: Reaction and Metabolite Standardization and Congruency across Databases and Genome-Scale Metabolic Models

Akhil Kumar,[2]* Patrick Suthers,[1] and **Costas D. Maranas**[1]

[1]Department of Chemical Engineering, and [2]Department of Computer Science and Engineering, The Pennsylvania State University, University Park

Project Goals: Create a knowledgebase for biochemical transformations with standardized metabolite and reaction entries that encompasses existing databases and all publicly available genome-scale metabolic models.

The ever-accelerating pace of DNA sequencing and annotation information generation is spearheading the global inventorying of metabolic functions across all kingdoms of life. Increasingly, metabolite and reaction information is organized in the form of community, organism, or even tissue-specific genome-scale metabolic reconstructions. These reconstructions account for reaction stoichiometry and directionality, gene to protein to reaction associations, organelle reaction localization, transporter information, transcriptional regulation and biomass composition. Already over 35 genome-scale models are available for eukaryotic, prokaryotic and archaeal species and are becoming indispensable for computationally driving engineering interventions in microbial strains for targeted overproductions, elucidating the organizing principles of metabolism and even pinpointing drug targets. A key barrier to the pace of extraction of metabolic knowledge from data is our inability to directly make use of metabolite/reaction information from databases (e.g., BRENDA, KEGG, BioCyc, UM-BBD, PubChem, ChEBI, Reactome.org, Rhea, etc.) or other metabolic models due to incompatibilities of representation, duplications and errors. Therefore, the inadvertent inclusion of multiple replicates of the same metabolite, stoichiometrically inconsistent and/or elementally/charge unbalanced reactions can lead to erroneous model predictions and missed opportunities to reveal (synthetic) lethal gene deletions, repair network gaps and quantify metabolic flows. There have already been a number of efforts aimed at addressing some of these limitations. The Rhea database aggregates reaction data primarily from IntEnz and ENZYME whereas Reactome.org is a collection of reactions primarily focused on human metabolism. Research towards integrating genome-scale metabolic models with large databases has so far been even more limited. An important step forward is Model SEED which is a web resource that generates draft genome-scale metabolic models drawing from an internal database that integrates KEGG with 13 genome scale models (including six of the models in the BiGG database). Motivated by this challenge we recently carried out an initial construction of the web-based resource MetRxn that integrates, using internally consistent descriptions, metabolite and reaction information from 6 databases and 34 metabolic models. The MetRxn content generation follows the general steps outlined in Figure 1. Metabolite and reaction data was first downloaded from BRENDA, KEGG and BioCyc using a variety of methods based on protocols such as SOAP, FTP and HTTP. We subsequently pre-processed the data into flat files that were imported into MetRxn. All original information pertaining to metabolite name, abbreviations, metabolite geometry, related reactions, catalyzing enzyme and organism name, gene-protein-reaction associations, and compartmentalization was retained. For all 34 genome-scale models ancillary information culled form the corresponding publications was also imported. The "raw data" from both databases and models was unified using standard SQL scripts on a MySQL server. We used Marvin (Chemaxon) to analyze all 231,085 raw metabolite entries containing structural information (out of a total of 322,936 entries). Metabolite atom bond connectivity was calculated at a fixed pH of 7.2 and converted into standard Isomeric SMILES format. Metabolites were also annotated with Canonical SMILES using the OpenBabel Interface from Chemspider. Metabolites with missing structural information were revisited during the reaction reconciliation step. After generating the initial metabolite associations, we identified reaction overlaps using the reaction synonyms and reaction strings along with the metabolite SMILES representations. During this step, reactions were flagged as single-compartment or two-compartment (i.e., transport reactions). Using the corrected metabolite elemental composition and protonation states, reactions are evaluated for charge and elementally balance. We used a linear optimization program to charge and elementally balance all reactions. Currently, the MetRxn knowledgebase is stored on the Open Source MySQL database system using the MyISAM storage engine.

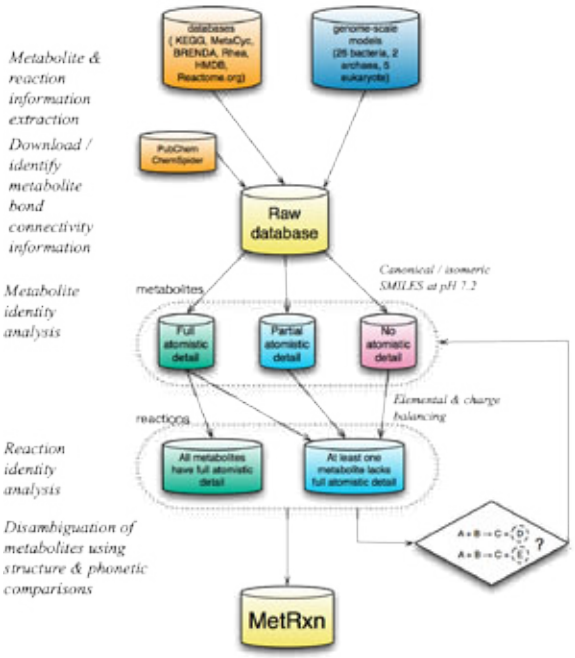

Figure 1. Flowchart abstracting the iterative workflow implemented in constructing the first-phase of the MetRxn knowledgebase. MetRxn contains, so far, over 62,345 distinct metabolites and 56,142 reactions that are charge and elementally balanced.

115

Improving the *i*MM904 *S. Cerevisiae* Metabolic Model Using Essentiality and Synthetic Lethality Data

Ali R. Zomorrodi[1]* (zomorrodi@engr.psu.edu) and **Costas D. Maranas**[1]

[1]Department of Chemical Engineering, the Pennsylvania State University, University Park

Project Goals: Use gene essentiality and synthetic lethality predictions to improve genome-scale models of metabolism.

Saccharomyces cerevisiae is the first eukaryotic organism for which a multi-compartment genome-scale metabolic model was constructed. Since then a sequence of improved metabolic reconstructions for yeast has been introduced. These metabolic models have been extensively used to elucidate the organizational principles of yeast metabolism and drive yeast strain engineering strategies for targeted overproductions. They have also served as a starting point and a benchmark for the reconstruction of genome-scale metabolic models for other eukaryotic organisms. Despite the successive improvements in the details of the described metabolic processes, even the recent yeast models (e.g., *i*MM904 and Yeast 4.0) remain significantly less predictive than the latest *E. coli* model (i.e., *i*AF1260). This is manifested by its significantly lower specificity in predicting the outcome of grow/no grow experiments in comparison to the microbial models. Contrasting the predicted growth phenotype of single mutant strains with the available experimental data under various growth conditions is the established standard for testing the accuracy of genome-scale metabolic models. These comparisons result in four different outcomes: GG or NGNG when both model and experimental data either imply growth (G) or no growth (NG) for the mutant strain, NGG when the model predicts that the gene deletion is lethal but the experiment shows that it is viable, and finally GNG when the model predicts that the mutant strain would be viable but *in vivo* observations show a lethal effect. Our group recently introduced a mathematical procedure termed GrowMatch for reconciling both NGG and GNG growth prediction inconsistencies across different substrates.

Here, we demonstrate that additional layers of correction and improvement can be gleaned by making use of synthetic gene lethal information. As shown in Figure 1, comparisons of model predicted synthetic lethal interactions with available experimental data reveal a number of additional ways that model and experiment may disagree. Notably, the "no growth" phenotype in this case could be due to either essentiality (ES) or synthetic lethality (SL) of the gene deletions. For example, GES and GSL inconsistencies refer to cases where the *in silico* deletion of a gene pair is not lethal (i.e., *G*rowth) but *in vivo* they are lethal due to gene essentiality or synthetic lethality (i.e., *ES*sential or *S*ynthetic *L*ethal). Similarly, ESG and ESSL represent mismatches where the single deletion of one of the genes *in silico* is lethal (i.e., *ES*sential), however, their simultaneous deletion *in vivo* results in either a viable strain (i.e., *G*rowth) or a lethal phenotype (i.e., *S*ynthetic *L*ethal), respectively. Finally, SLG and SLES denote inconsistencies where the model implies that only the double gene mutation is lethal (i.e., *S*ynthetic *L*ethal) but experimental observations support either growth (G) or lethality of any of the two single gene deletions (i.e., *ES*sential), respectively.

In this project we make use of the automated GrowMatch procedure for restoring consistency with single gene deletion experiments in yeast and extend the procedure to make use of synthetic lethality data using the genome-scale model *i*MM904 as a basis. In addition to essentiality and synthetic lethality we also explored disagreements in auxotrophy complementation, where model predicted supplementation rescue (i.e., auxotrophy) scenarios are inconsistent with experimental data. Overall, we identified and vetted using literature sources 90 distinct model modifications along with 30 regulatory constraints for minimal and YP media. The incorporation of the suggested modifications led to a substantial increase in the fraction of correctly predicted lethal knockouts (i.e., specificity) from 38.84% (87 out of 224) to 53.57% (120 out of 224) for the minimal medium and from 24.73% (45 out of 182) to 40.11% (73 out of 182) for the YP medium. Synthetic lethality predictions improved from 12.03% (16 out of 133) to 23.31% (31 out of 133) for the minimal medium and from 6.96% (8 out of 115) to 13.04% (15 out of 115) for the YP medium. Given that these improvements in the model were achieved using only

the partial list of synthetic lethality data currently available in literature, a far larger contribution of synthetic lethals in providing model refinement strategies is expected as more synthetic lethality data are becoming available. Overall, this study provides a roadmap for the computationally driven correction of multi-compartment genome-scale metabolic models and demonstrates the value of synthetic lethals as curation agents.

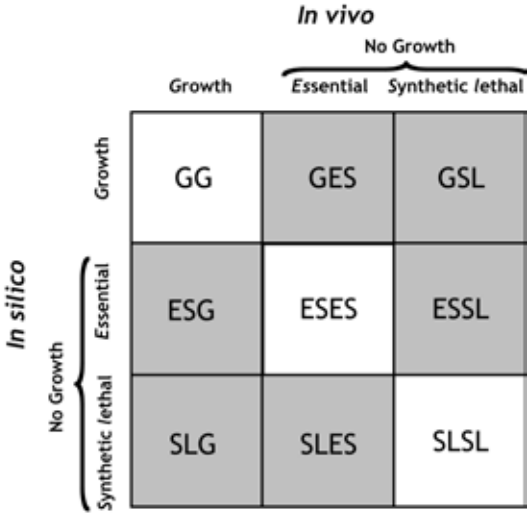

Figure 1. Different types of mismatches between *in silico* predictions and *in vivo* observations for double gene perturbations. The abbreviations G, ES and SL in this figure refer to ***Growth***, ***Essential*** and ***Synthetic Lethal***, respectively. Here, 'No Growth' can be due to either essentiality or synthetic lethality of single or double gene deletions.

115A‡

submitted post-press

Construction of an *E. coli* Genome-Scale Atom Mapping Model for MFA

Prabhasa Ravikirthi[1]* (pxr173@psu.edu), Patrick F. Suthers,[2] and **Costas D. Maranas**[2]

[1]Department of Cell and Developmental Biology and [2]Department of Chemical Engineering, The Pennsylvania State University, University Park

Project Goals: To generate a genome-scale atom mapping model of *E. coli* for metabolic flux analysis.

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 metabolic networks. Here we describe a procedure that uses a compound matching algorithm based on the graph theoretical concept of pattern recognition along with relevant reaction information to automatically generate genome-scale atom mappings which trace the path of atoms from reactants to products for every reaction. The procedure is applied to the *i*AF1260 metabolic reconstruction of *Escherichia coli* yielding the genome-scale *i*sotope *m*apping model imPR90068. This model maps 90,068 non-hydrogen atoms that span all 2,077 reactions present in *i*AF1260 and contains a total of 1.37 x 10^{157} isotopomers (with 8.34 x 10^{93} ^{13}C isotopomers).

The isotope mapping model imPR90068 contains mappings for reactions that were previously lumped or completely absent from earlier isotope mapping models (even in imPS1485 (Suthers et al. 2007)). These new additions include 68 reactions involved in the metabolism of amino acids (see Figure 1), 65 reactions involved in central metabolism, 153 reactions in nucleotide biosynthesis and salvage pathways, 225 reactions in glycerophospholipid metabolism, 160 reactions in cofactor and prosthetic group biosynthesis and 181 reactions in alternate carbon metabolism. The incorporation of more than 1,100 new reactions involved in various parts of *E. coli* metabolism together with the inclusion of more than 800 metabolites compared to the previous largest imPR1485 model implies that alternate metabolic routes could now fully be taken into account during flux elucidation using MFA. Additionally, it allows for the possible labeling of substrates other than glucose.

This paper also introduced the computational infrastructure for tracing all atoms present in every reaction in the *i*AF1260 metabolic reconstruction of *E. coli* from reactants to products to create a genome-scale mapping database. This automated procedure can be efficiently leveraged for genome-scale models of other organisms to create isotope mapping databases. Common reactions already present in *i*AF1260 can be directly culled from the imPR90068 reaction-mappings database thus significantly reducing the effort needed to construct other organism-specific mapping models. The potential to improve our understanding of flux allocation in different organisms is alluded by the gap in the size of genome scale vs. isotope mapping models. For example, there exists a 50-fold difference in the size of the genome-scale reconstruction of *Bacilus subtilis* that spans 1,020 reactions (Oh et al. 2007) and its current isotope mapping model (Dauner et al. 2001) that accounts for only 25 reactions (all from central metabolism). Approximately 70% of reactions in its genome-scale model have an exact match to reactions in *i*AF1260. It is expected that incorporating reactions into the mapping model already present in the genome-scale model could shed light onto metabolic pathway usage patterns. We can do the same for organisms with multiple compartments such as *Saccharomyces cerevisiae*.

However, the ability to elucidate fluxes using the full complement of reactions and metabolites present in genome-scale level reconstructions comes at the expense of requiring additional labeling data. While lumped isotope models (Antoniewicz et al. 2007b; Kim et al. 2008; Suthers et al. 2007) typically require the analysis of spectra (i.e., NMR or GC/MS) for only about 20-50 fragments, using the totality of mapped isotopomers in imPR90068 will require significantly higher numbers of carefully chosen labeled fragments. This makes even more pertinent the use of methods such as OptMeas (Chang et al. 2008; Suthers et

al. 2010), EMU representations (Antoniewicz et al. 2007a) and systematic reaction step aggregation techniques (e.g., SLIPs (Quek 2009)), as well as advances in metabolomics.

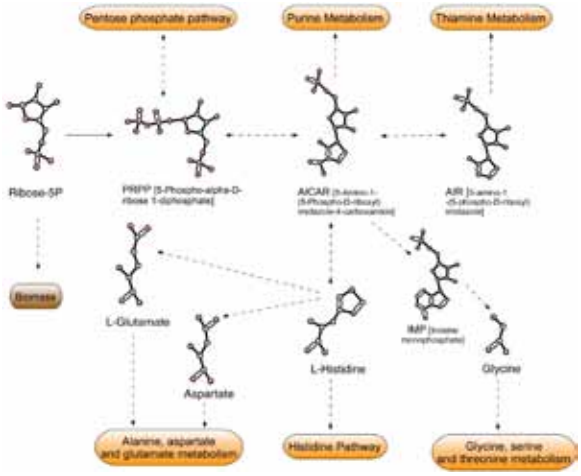

Figure 1. An example of the expanded scope of the genome-scale isotope mapping model imPR90068. In imPS1485 Ribose-5P production was directly routed to biomass as a stand-in substitute for histidine. In imPR90068 R5P downstream conversion is linked to other amino acid synthesis pathways.

Small Business Innovation Research (SBIR) and Small Business Technology Transfer (STTR)

116

Microbioreactor Technology for Obligate Anaerobes

Harry Lee (harrylee@pharyx.com) and Paolo Boccazzi* (boccazzi@pharyx.com)

Pharyx, Inc., Boston, Mass.

http://www.pharyx.com

Project Goals: see below

Anaerobic microorganisms have evolved biochemical pathways that can be exploited for industrial applications. These include the ability to breakdown environmental pollutants for bioremediation, the breakdown of cellulose into simple sugars for biofuels, and the production of specialty chemicals. However, there remains a tremendous challenge to the scale-up of bioenzymatic activities to industrial processes. While systems biology approaches and metabolic engineering promise to contribute to our understanding of these systems, a key bottleneck is in conducting controlled experiments to ground these approaches with high quality data. Thus far, experiments are frustrated by the laborious set-up and operation of stirred tank bioreactor systems, which for anaerobic microbiology is further encumbered by the requirement of an anaerobic environment. The absence of easy to use systems also holds back more traditional microbiology approaches such as mutagenesis and screening and directed evolution.

We are developing a parallel bioreactor system, based on microfluidic integration technology and disposable microbioreactor modules, with application specific customizations for anaerobic fermentation. These customizations are aimed to enable up to 32 simultaneous anaerobic fermentations under controlled conditions, with online monitoring of growth kinetics and other phenotypes such as enzyme activity. A unique feature of this system is the ability to operate it in ambient air through careful inoculation port and reactor and control module design, or to operate it within an anaerobic bag, taking advantage of its compact size.

Project goals are 1) determine microbioreactor designs that will support anaerobic inoculation and fermentation, 2) identify optimal materials for fabricating anaerobic bioreactors, 3) determine the range of process parameters where microbioreactor data corresponds to serum tubes and stirred tank fermentors, 4) monitor enzyme activity on-line.

In our initial work, we demonstrated passive anaerobic fermentations in simple microbioreactor devices. These experiments highlighted problems including gas bubble generation by anaerobes, which spoiled optical density measurements, and the significance of absorbed oxygen in plastic materials. This motivated a new anaerobic device design with anaerobic mixing, bubble removal, and fluid injection to enable pH controlled anerobic fermentations. Figure 1 shows bubble-free online measured optical density in the microbioreactor and at line measured samples from a stirred tank bioreactor for uncontrolled *C. acetobutylicum* fermentations. Figure 2 shows a comparison between controlled and uncontrolled fermentations of *B. fibrisolvens* in both a stirred tank and the microbioreactor.

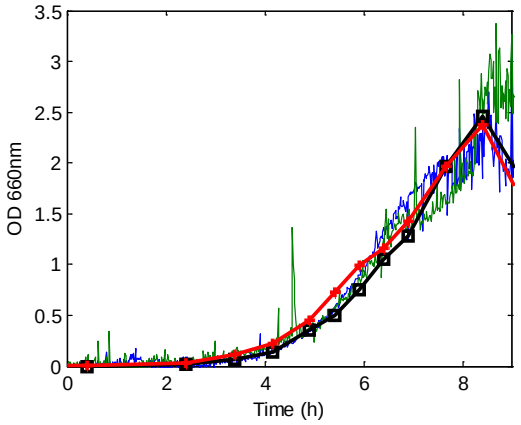

Figure 1. Microbioreactor (solid lines) and bench top stirred tank (+ and square markers) fermentations of *C. acetobutylicum*. With time alignment to account for varying lag phase, and calibration between microbioreactor and spectrophotometer optical density, there is excellent correspondence between both fermentation systems.

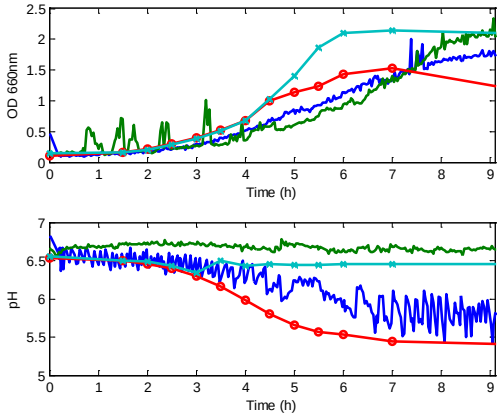

Figure 2. Optical density and pH curves for Microbioreactor (solid lines) and bench top stirred tank bioreactor (lines with x and o markers) of *B. fibrisolvens*. In the stirred tank, there was a clear effect of pH control, which maintained a steady growth rate to the end of the fermentation while the uncontrolled case showed a slower growth rate after 4.5 hours. Similar behavior was observed in the microbioreactor.

117

Semantic Indexing of the Green Technology Patent Literature

George M. Garrity* (garrity@msu.edu), Charles T. Parker, and Catherine Lyons

NamesforLife, LLC and Michigan State University

Project Goals: NamesforLife, LLC has developed a novel technology that resolves uncertainty about the meaning of biological names or other dynamic terminologies. It uses those terms to create persistent links to related information, goods, and services available on the Internet, even if the terms have changed.

Under a Phase I/II STTR, NamesforLife, LLC created a suite of software tools and techniques to manage dynamic terminologies and an underlying term set (an up-to-date list of over 14,000 validly published names of bacteria and archaea, including all of the synonyms and homonyms, links to appropriate taxonomic literature, key genetic and genomic data). The company's N4L tools can automatically detect and tag bacterial names in HTML and XML documents with a high degree of precision. An interactive browser-based application (N4LGuide) provides end users direct access to correct nomenclatural information along with links to key data (16S sequence and genome sequence data) while reading the literature. It uses ISO standard Digital Object Identifier (DOI) technology to create links at each occurrence of a validly published name in HTML documents. The company has also developed batch tools (the N4L Semantic Tagger) that can embed N4L-DOIs into XML versions of scientific articles that are created as part of the contemporary publishing workflow and used to create human readable content in various forms (e.g., HTML, PDF, ink-on-paper). The company has also developed a unique way of tracking the occurrence of biological names in the literature, based on the usage of our tools (the N4L Contextual Index).

While initially intended as a tool for readers, authors, and publishers of scientific literature, N4L tools can also be applied to other documents where bacterial names appear. As proof of principle, the company processed approximately 250,000 U.S. patents and patent applications with the Semantic Tagger and then indexed the tagged documents using Apache Lucene to provide end users with additional search and retrieval capabilities. Simple graphical tools were added to support limited on-demand analyses of search results. These tools are designed to support data mining by non-commercial organizations, highlighting trends in commercialization of biodiversity research. This work also led to the discovery of "terminological fingerprints" that could be used to classify patents and other documents using externally managed term sets.

To validate the concept of "terminological fingerprinting", the company processed the EPO Green Technology collection of patents, which consists of approximately 362,000 documents. In addition to detecting bacterial names, the N4L Semantic Tagger was modified to recover patent classifications (IPC and ECLA), applicants, assignees, inventors, patent references, patent metadata, and patent titles, as is common in patent landscaping.

A total of 3,845 patents were found that made reference to 3,385 distinct bacterial and archaeal names held in the NamesforLife database. Of these, 626 names were unique to non-U.S. patents. The number of names per patent (name vectors) ranged from 1 – 1,290, with an average of 13 names and a median of 5 names. In addition to name occurrence,

frequency data for each name occurrence per patent was tabulated. The resulting name vectors were then used to further examine the associations among the patents based on the IPC and ECLA classifications. Simple associations could be derived directly from the captured data. However, more complex patterns involving multiple many-to-many relationships could only be ascertained from the cross-products of underlying contingency and frequency data.

The results were then examined using routine approaches for exploratory data analysis and visualization (e.g., principal components analysis, robust clustering, 2D scatter plots, 3D spin plots and heatmaps). Each of these methods revealed strong evidence of terminological fingerprints in the patents. However, those methods did not scale well or suffered from other limitations. Hexagonal binning was, however, found to be suitable for visualizing the complex relationships inherent in the patent data. The company is currently developing interactive hexagonal bin plots as a means of selecting subsets of patents that involve related technologies and microorganisms.

As DOE research on biofuels, bioremediation and carbon sequestration moves from the laboratory into production or commercial environments, a number of important policy and business decisions must be made that demand correct information. These include establishing the patentability of a given technology, freedom to operate, and potential infringement of patents held by competitors, both in the U.S. and abroad. Failure to pay careful attention to these issues can have serious consequences beyond the payment of stiff penalties for infringement. These include lost opportunities arising for technology licensing, failure to detect and understand regional disparities, rapid growth in patent coverage of technologies by competitors and migration of technology across international borders. The scientific and technical literature provides an incomplete view of any field having commercial potential because the underlying technologies are typically not revealed in public until absolutely necessary, and then only after patent applications have been filed. While patents with corresponding papers are not uncommon as a means of announcing important new developments, they are not obligatory. Therefore, an awareness of developments in the field requires a thorough review of both bodies of literature. NamesforLife is building tools to simplify such searches, using its proven approach to indexing through the creation of persistent links to externally managed terminologies that common to both bodies of literature. This approach integrates well with existing commercial, academic and USPTO data mining capabilities.

This research is supported by the Office of Biological and Environmental Research of the U.S. Department of Energy under Phase I and II STTR Awards DE-FG02-07ER86321 A001 - A005.

118

Accelerating Metagenomics Using Graphics Processing Units

Matthew Hudson* (mhudson@illinois.edu)

University of Illinois

Project Goals: Development of faster database searching algorithms for high throughput sequencing.

The goal of this project is to develop software for DNA sequence processing from DOE projects that runs on the graphics processing units available in modern computers. In particular, the aim is to accelerate the matching of sequence reads from experiments using high-throughput next-generation sequencing platforms (such as metagenomics projects) to large databases such as those maintained at GenBank. The most critical need is for a way to compare hundreds of millions of DNA reads to a protein database using the translated BLAST algorithm BLASTX. Using the existing versions of the BLAST software, searching one read against a complete protein database such as GenBank non-redundant protein (NR) takes many seconds to a few minutes on one CPU. One sequence run from the latest sequencing platforms (e.g. Illumina HiSeq) can produce hundreds of millions of sequence reads. Searching this data against NR using conventional BLAST thus has an unacceptably high computational cost of millions of CPU hours, the cost of analysis vastly exceeding the cost of generating the data. In this project we have developed MulticoreWare BLASTX (MBLASTX). This new-generation software package employs algorithm-level acceleration and the use of the GPU (the powerful parallel graphics processors found on most modern computers) to accelerate sequence searches more than 1,000 times compared to the performance of the latest version of NCBI BLAST on the latest computers. The accuracy of the search results is over 99% compared to NCBI BLAST. This software has accelerated the processing of DNA sequence enough to replace expensive and energy-hungry supercomputers used for this purpose with ordinary, much cheaper computers, lowering the cost and energy consumption of an essential analysis step by more than 1,000 fold.