Genome-specific models of bacterial cells and reverse engineering of prokaryote biology

Evgeni V. Nikolaev[1], Sergej V. Aksenov[2], Jordan C. Atlas[1], Michael L. Shuler[1] and Bruce W. Church[2]
[1]Cornell University, Ithaca NY and [2]Gene Network Sciences, Cambridge MA

Project Goals: Genome-specific models of bacterial cells and reverse engineering of prokaryote biology.

A deep and broad understanding of diverse prokaryote biologies is essential to the success of DOE biofuels and bioremediation missions. The optimization of a prokaryote phenotype to achieve DOE objectives requires models that connect environmental factors through organism state variables (genome, mRNA, proteins and metabolites) to phenotype. While there has been rapid progress in the systematic characterization and measurement of organism state variables via high throughput experimental techniques, the processes that connect these variables from environmental factors though to phenotypes are not understood on the scale of the entire organism or community of organisms. We present approaches based on 1) a system for rapid development of dynamical and modular coarse-grained models of bacterial cells and 2) reverse engineering of probabilistic functional models directly from high throughput omic-level data and systematic high-throughput interrogation of these models via forward simulation. The initial step towards the rapid development the dynamical models is the development of a coarse-grained model, describing pseudo-chemical interactions between lumped species. A hybrid model of interest can then be constructed by embedding genome-specific detail for a particular cellular subsystem (e.g. nucleotide biosynthesis), called here a *module*, into the coarse-grained model. Specifically, a new strategy for sensitivity analysis of the cell division cycle in a coarse-grained model is introduced to identify which pseudo-molecular processes should be delumped to implement a particular biological function in a growing cell (e.g. ethanol overproduction, pathogen viability). We are currently developing periodic boundary-value problem numerical methods to study the stability and robustness of coarse-grained models when parameters are allowed to vary. To illustrate the modeling principles and highlight computational challenges, the Cornell coarse-grained model of *Escherichia coli* B/r-A is used to benchmark the proposed framework. To reverse engineer coarse-grain probabilistic functional models, Gene Network Sciences has developed, NI Engine, a data-driven software platform capable of uncovering system wide models that connect phenotypes to environmental factors. This automated model discovery yields ensembles of models that span the diversity of processes consistent with the data when the quantity and/or quality of the data is low. We demonstrate that the performance of the reverse engineering algorithm as a function of data quantity or quality depends on the nature of the "query" applied to the ensemble of models. We are planning to apply the discussed modular dynamical model and reverse engineering approaches to construct large-scale models for *E.coli* and genomically related gram-negative organism *Shewanella oneidensis*.

mRNA-targeted In Situ Hybridization Without Template Amplification or Catalyzed Reporter Deposition

J Coleman, D. Culley, and **F Brockman***
Pacific Northwest National Laboratory, Richland WA 99352
(fred.brockman@pnl.gov)

Genomic technologies are needed to interrogate the spatial organization of gene expression and trafficking of material and information in microbial communities. In situ hybridization (ISH) supplies high-resolution spatial information on gene expression for many dozens to many hundreds of cells simultaneously; in contrast, this information is lost when conducting microarray analysis on RNA extracted from cultures.

While mRNA-targeted ISH has been popular for studying gene expression in eukaryotic cells, very little success has been achieved in applying mRNA-targeted ISH to prokaryotes. At present, detection of specific mRNA's in prokaryotes requires in situ-PCR and/or immunocytochemical amplification of hybridized molecules via catalyzed reporter deposition (CARD). However, these approaches are not amenable to more high-throughput analysis and the reagents are difficult or impossible to apply to many environmental samples. The objective of this project is to demonstrate the utility of near infrared dyes for direct (i.e., no PCR or post-hybridization signal amplification) ISH-based detection of specific mRNA's in prokaryotic cells.

To control transcript production a moderate copy number plasmid, containing the mut3 *gfp* gene encoding a short-lived protein controlled by the inducible pBAD promoter, was introduced into *Shewanella* and *E. coli*. Relative abundance qPCR was performed on induced and non-induced cultures to establish induction levels; induced cultures showed a 90x and 35x fold increase in *gfp* mRNA for *Shewanella* and *E. coli*, respectively. Singly labeled oligonucleotide probes targeting *gfp* mRNA (and as a control, the complementary DNA sense strand) were labeled with either Alexa-647 (near-IR) or Alexa-488. A 3-hour hybridization was found to produce maximum signal. The signal to noise ratio was 4 to 5 times greater for the 647 probe as compared to the 488 probe using bandpass filters and imaging conditions optimized for each probe. This was primarily due to low levels of cell autoflourescence in no-probe 'hybridizations' under the optimized imaging conditions used for the 488 probe; autoflourescence was not due to the presence of the Gfp protein as shown by comparison of cells possessing and lacking a Shine-Delgarno sequence upstream of the *gfp* gene. Antisense probes showed no hybridization to RNase treated cells. However, even in the best hybridizations – and even after inducer, permeabilization, fixation, and hybridization parameters were varied in an effort to increase the fraction of probe-positive cells – less than 50% of the cells were probe positive. This was consistent with the frequency of detection of Gfp protein-containing cells by flow cytometry, using the identical batch of cells.

ISH was performed targeting transcripts from the chromosomal, single copy gene *rps*H using singly-labeled Alexa-647 oligonucleotides. After optimization of conditions, the antisense probe produced signal in up to 60% of the cells, and signal was either not detected or clearly weaker in hybridizations on the identical batch of cells with the following controls: sense probe, antisense probe hybridized against RNase treated cells, and no-probe 'hybridizations'. Statistical significance (t-test, $p<0.05$ to $p<0.005$) was shown in each of the three cases, using datasets

comprised of randomly selected cells (i.e., inclusion of randomly selected cells that showed no hybridization).

In conclusion, we have successfully conducted ISH-based detection of specific mRNA's in prokaryotic cells using singly-labeled near-infrared (Alexa-647) oligonucleotides, without PCR or post-hybridization signal amplification. We are now targeting transcripts of single copy genes in a co-culture containing 4 microbes.

Metabolic Model for *M. genitalium* and Synthetic Circuit Design
Award No. DE-FG02-05ER25684, Costas D. Maranas, Penn State, PI

Madhukar S. Dasika[a] (presenter) and Costas D. Maranas[a,*]

[a]*Department of Chemical Engineering, The Pennsylvania State University, University Park, PA 16802*

Project Goals: In this project, we will highlight our progress towards the development of two complementary computational frameworks aimed at analysis and construction of biological networks. First, we will describe computational approaches that start from genome-scale metabolic reconstructions to identify viable growth environments. Subsequently, we develop computational frameworks that start from simple well-defined genetic elements to construct functional biological networks.

In this poster, we will highlight our progress towards the development of two complementary computational frameworks aimed at analysis and construction of biological networks. First, we will describe computational approaches that start from genome-scale metabolic reconstructions to identify viable growth environments. Subsequently, we develop computational frameworks that start from simple well-defined genetic elements to construct functional biological networks.

1. Generating Stoichiometric Model for the Smaller Free-Standing organisms: Mycoplasma Genitalium: A major hindrance to research and laboratory diagnosis of infection of small free standing organisms such as *Mycoplasma Genitalium* has been their cultivation *in vitro*. Researchers have reported the need for extremely fastidious growth requirements for *M. genitalium* and have typically used complex media to cultivate them. The use of complex undefined growth media has interfered with genetic analyses, estimation of growth requirements, characterization of auxotrophic mutants and examining the nutritional control of bacterial pathogenesis. In this work we develop a Flux Balance Analysis (FBA) based approach for guiding the design of growth medium for *M. genitalium*. To this end, the annotated genome sequence of *M. genitalium* is deployed within the SimPheny^{TM} platform to construct the genome-scale metabolic model for *M. genitalium*. Subsequently the genome-scale metabolic reconstruction is validated by comparing FBA simulation predictions with experimental data regarding gene essentiality studies. Finally the validated model is employed to predict viable uptake environments for *M. genitalium.* Our results indicate that FBA simulation predictions agree with experimental observations in 2/3 cases for “non-essential” genes and 29/41 cases for “essential” genes. Examination of uptake conditions reveals that *M. genitalium* can metabolize either glucose or fructose as substrates. Further, uptake of all amino acids, Riboflavin, Nicotinic acid, Spermidine, Putrescine, Acetate, co enzyme A, Adenine, Cytidine and 6 phospho D gluconate is found to be necessary for growth.

2. Computational design of synthetic biological circuits: Recent years has witnessed an increasing number of studies on constructing simple synthetic genetic circuits that exhibit desired properties such as oscillatory behavior, inducer specific activation/repression, etc. To meet these emerging challenges, in this work we introduce OptCircuit, an optimization based framework that automatically identifies the circuit components from a list and connectivity that brings about the desired functionality. The dynamics that govern the interactions between the elements of the

genetic circuit (promoters, repressors, etc.) are modeled using deterministic ordinary differential equations. The desired circuit response is abstracted as the maximization/minimization of an appropriately constructed objective function. The optimization framework has been applied on a variety of applications ranging from the design of circuits that exhibit a specific time course response and circuits that discriminate between the presence, absence and level of external stimuli. Overall, the results demonstrate the ability of the framework to (i) generate the complete list of circuit designs of varying complexity that exhibit the desired response; (ii) rectify a non-functional biological circuit and restore functionality by modifying an existing component and/or identifying additional components to append to the circuit; (iii) pinpoint what promoter's strength, interaction parameter or protein degradation constant to modify to meet the desired response.

New Tools to Probe Bacterial Chromosome Folding

Mark A. Umbarger*[1] (umbarger@fas.harvard.edu), Matthew A. Wright[1], Job Dekker[2], and **George M. Church**[1] (http://arep.med.harvard.edu/gmc/email.html)

1 Department of Genetics, Harvard Medical School and 2 Program in Gene Function and Expression and Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School

Project Goals: We are seeking to develop novel tools to generate new insights into the structure of bacterial chromosomes.

Emerging evidence suggests that the three-dimensional (3D) structure of the bacterial chromosome is dynamic and ordered with the sub-cellular positions of genetic loci mirroring their genomic position(*1-4*). Despite these advances, many questions remain. To derive new insights into the folding of the bacterial chromosome we have adapted two new technologies to the gram negative bacterium, *Caulobacter crescentus*. First, we have optimized the chromosome conformation capture (3C) assay in *Caulobacter* in order to interrogate the structure of small regions of the chromosome. Second, we have coupled multiplex polony sequencing(*5*) to the new large-scale 3C assay, carbon copy 3C (5C)(*6*), to enable the simultaneous assessment of the structure of the entire *Caulobacter* chromosome.

3C assesses the relative proximity of genomic loci by measuring the frequency at which these loci cross-link. In our optimized protocol synchronized swarmer cells are fixed with formaldehyde to cross-link DNA to protein and protein to protein. Chromatin is then digested with a restriction enzyme and is subsequently ligated at a low concentration such that most intermolecular ligation occurs between cross-linked restriction fragments. Since the frequency of cross-linking correlates with 3D distance, the proximity of restriction fragments within the structure of the chromosome can be measured by quantitating the frequency at which these fragments ligate (cross-link). Using PCR to quantitate ligation frequencies we have recapitulated the finding that in *Caulobacter* the origin and terminus of replication distantly localize(*3*) and have begun to address whether the arms of the chromosome physically interact.

A single *Caulobacter* 3C experiment generates a library of ligation products that is representative of the structure of the entire chromosome. To take full advantage of this library one must assess the frequency of thousands of different ligation products, a number not amenable to PCR. To facilitate the parallel quantitation of thousands of ligation products we have developed a modified version of the carbon-copy 3C (5C) technology recently reported by Dekker and colleagues(*6*). This technique, 5C-Polony Sequencing (5C-S), couples multiplex ligation-mediated amplification with polony-based sequencing to simultaneously determine the frequency of thousands of 3C ligation products. Using 5C-S we have probed the structure of an 800 kb region of the *Caulobacter* chromosome and are currently in the process of querying the structure of the entire chromosome. We are also beginning to use this novel technology to determine the effects of cellular and metabolic perturbations. Our preliminary results suggest that 5C-S is a powerful new tool for interrogating bacterial chromosome structure.

Works Cited:

1. A. A. Teleman, P. L. Graumann, D. C. Lin, A. D. Grossman, R. Losick, *Curr Biol* **8**, 1102-9 (Oct 8, 1998).
2. D. Bates, N. Kleckner, *Cell* **121**, 899-911 (Jun 17, 2005).
3. P. H. Viollier *et al.*, *Proc Natl Acad Sci U S A* **101**, 9257-62 (Jun 22, 2004).
4. H. J. Nielsen, Y. Li, B. Youngren, F. G. Hansen, S. Austin, *Mol Microbiol* **61**, 383-93 (Jul, 2006).
5. J. Shendure *et al.*, *Science* **309**, 1728-32 (Sep 9, 2005).
6. J. Dostie *et al.*, *Genome Res* **16**, 1299-309 (Oct, 2006).

Metabolic Flux Elucidation for Genome-Scale Models Using ^{13}C Labeled Isotopes

Award No. DE-FG02-05ER25684, Costas D. Maranas, Penn State, PI
Patrick F. Suthers[a] (presenter), Anthony P. Burgard[b], Madhukar S. Dasika[a], Farnaz Nowroozi[c], Stephen Van Dien[b], Jay D. Keasling[c], Costas D. Maranas[a,*]

[a]*Department of Chemical Engineering, The Pennsylvania State University, University Park, PA 16802*

[b]*Genomatica, Inc, 5405 Morehouse Drive, Suite 210, San Diego, CA 92121*

[c]*Department of Chemical Engineering, University of California - Berkeley, Gilman Hall, Berkeley, CA 94720-1462*

Project_Goals: A goal of this project is to develop a computational framework that combines a constraint-based modeling framework with isotopic label tracing for flux elucidation of large-scale metabolic networks.

A key consideration in metabolic engineering is the determination of fluxes of the metabolites within the cell. This determination provides an unambiguous description of metabolism before and/or after engineering interventions. Here, we present a computational framework that combines a constraint-based modeling framework with isotopic label tracing on the genome-scale. When cells are fed a growth substrate with certain carbon positions labeled with ^{13}C, the distribution of this label in the intracellular metabolites can be calculated based on the known biochemistry of the participating pathways. Most labeling studies focus on skeletal representations of central metabolism and ignore many flux routes that could contribute to the observed isotopic labeling patterns. In contrast, our approach investigates the importance of carrying out isotopic labeling studies using a more comprehensive reaction network consisting of 347 fluxes and 183 metabolites in *Escherichia coli* including global metabolite balances on cofactors such as ATP, NADH, and NADPH. The proposed procedure is demonstrated on an *E. coli* strain engineered to produce amorphadiene, a precursor to the anti-malarial drug artemisinin. The cells were grown in continuous culture on glucose containing 20% [U-^{13}C]glucose; the measurements are made using GC-MS performed on 13 amino acids extracted from the cells. We identify flux distributions for which the calculated labeling patterns agree within a few percent of the measurements alluding to the accuracy of the network reconstruction. Furthermore, we explore the robustness of the flux calculations to variability in the experimental MS measurements, as well as highlight the key experimental measurements necessary for flux determination. Finally, we discuss the effect of reducing the model, as well as shed light onto the customization of the developed computational framework to other systems.

Examining the molecular basis for utilization of alternative redox systems and hydrogen production by RubisCO-compromised mutants of nonsulfur purple bacteria

Rick A Laguna*****(laguna.2@osu.edu) and **F. Robert Tabita**

Department of Microbiology, The Ohio State University, 484 W. 12th Avenue. Columbus, OH 43210 USA

Project Goals: Understand the molecular basis for the utilization of alternative redox systems and maximize hydrogen production from nonsulfur purple bacteria.

The Calvin-Benson-Bassham (CBB) reductive pentose phosphate pathway is responsible for the incorporation of CO_2 into cellular carbon under autotrophic growth conditions. Under photoheterotrophic conditions, CO_2 is primarily used as an electron acceptor via the CBB cycle in order to maintain redox poise within the cell. The key enzyme of the CBB pathway, RubisCO, catalyzes the actual CO_2 reduction step. It was shown that gain-of-function adaptive mutant strains of *Rhodobacter sphaeroides, Rhodobacter capsulatus, Rhodospirillum rubrum*, and *Rhodopseudomomas palustris,* all of which have inactivated RubisCO genes (form I and II) could be selected. Such strains were capable of growth under photoheterotrophic conditions, unlike un-adapted strains. However, it was shown that adapted strains must utilize alternative systems to maintain redox balance since the CBB cycle is nonfunctional. The strains to be described accomplish this by de-repressing the synthesis of the nitrogenase enzyme complex, and thus utilize the inherent hydrogenase activity of this enzyme system to reduce protons, releasing copious amounts of hydrogen gas. *Rhodobacter sphaeroides* (strain 16PHC) and *Rhodopseudomonas palustris* (strain 2044) are being used as models to understand the molecular basis for utilization of the alternative redox system and for maximizing hydrogen production in these organisms. We further show that CbbR, the transcriptional regulator of the CBB system, is up-regulated in strain 16PHC. Moreover, when *cbbR* was inactivated, utilization of the nitrogenase complex and hydrogen evolution was negated in 16PHC. Regulation of the CBB system in strain 2044 differs from strain 16PHC but up-regulation of GlnK occurs in both organisms. Many additional proteins are up-regulated in strain 2044, as indicated by collaborative studies with the Oak Ridge National Laboratory proteomics group. Finally, hydrogen production by these various mutant strains was compared using typical growth substrates.

Optimization based automated curation of metabolic reconstructions

Award No. DE-FG02-05ER25684, Costas D. Maranas, Penn State, PI

Vinay Satish Kumar[1] (presenter), Madhukar S. Dasika[2], Costas D. Maranas[2]
[1]Department of Industrial and Manufacturing Engineering, [2]Department of Chemical Engineering, The Pennsylvania State University, University Park, PA 16802, USA

Project Goals: Currently, there exists tens of different microbial and eukaryotic metabolic reconstructions (e.g., *Escherichia coli*, *Saccharomyces cerevisiae*, *Bacillus subtilis*) with many more under development. All of these reconstructions are inherently incomplete with some functionalities missing due to the lack of experimental and/or homology information. A key challenge in the automated generation of genome-scale reconstructions is the elucidation of these gaps and the subsequent generation of hypotheses to bridge them.

Currently, there exists tens of different microbial and eukaryotic metabolic reconstructions (e.g., Escherichia coli, Saccharomyces cerevisiae, Bacillus subtilis) with many more under development. All of these reconstructions are inherently incomplete with some functionalities missing due to the lack of experimental and/or homology information. A key challenge in the automated generation of genome-scale reconstructions is the elucidation of these gaps and the subsequent generation of hypotheses to bridge them.

In this work, an optimization based procedure is proposed to identify and eliminate network gaps in these reconstructions. First we identify the metabolites in the metabolic network reconstruction which cannot be produced or consumed under any uptake conditions and subsequently we identify the reactions from a customized multi-organism database that restores the connectivity of these metabolites to the parent network using four mechanisms. This connectivity restoration is hypothesized to take place through four mechanisms: a) reversing the directionality of one or more reactions in the existing model, b) adding reaction from another organism to provide functionality absent in the existing model c) adding external transport mechanisms to allow for importation of metabolites in the existing model, and d) restoring flow by adding intracellular transport reactions in multi compartment models. We demonstrate this procedure for the genome scale reconstruction of *Escherichia coli* and also *Saccharomyces cerevisiae* wherein compartmentalization of intra-cellular reactions results in a more complex topology of the metabolic network. We determine that about 10% of metabolites in *E. coli* and 30% of metabolites in *S. cerevisiae* have no flux going through them. Interestingly, the dominant flow restoration mechanism is directionality reversals of existing reactions in the respective models.

In this study, we have proposed systematic methods to identify and fill gaps in genome-scale metabolic reconstructions. The identified gaps can be filled both by making modifications in the existing model, and by adding missing reactions by reconciling multi-organism databases of reactions with existing genome-scale models. Computational results provide a list of hypotheses to be queried further and tested experimentally.

A transcriptional regulatory network discovery system for microbes

P. Ortoleva, L. Ensman, K. Qu, F. Stanley, J. Sun, M. Trelinski, K. Tuncay

Center for Cell and Virus Theory, Indiana University

ABSTRACT

Discovering the network of biochemical processes underlying the behavior of Geobacteria and other microbes is obtained by creating a suite of interoperable systems biology modules. The workflow takes multiplex bioanalytical data as input, discovers the transcriptional regulatory network (TRN) and other process networks, and then uses cell simulation to derive microbial behavior, notably the biotechnical characteristics in the context of environmental remediation and energy production. To attain this goal we integrate a number of bioinformatics, cell modeling, and multiplex data/model integration tools. We have started this project with a TRN discovery system (a preliminary version is at *http://systemsbiology.indiana.edu*). Input to this system is microarray data on gene expression profiles generated by the bacterium in response to thermal, chemical, or gene insertion/deletion perturbations. A database provides a preliminary TRN which provides serves as a training set for the systems biology modules. Network inference using a similarity measure assumes that the activity of a transcription factor (TF) is represented by the expression of the gene that makes it. Failure to observe high correlation between mRNA level and TF activity in *E.coli* shows that this assumption does not hold. Therefore, in order to use expression data, we estimate the TF activities independent of expression level of the mRNA that translates into the TF. To accomplish this, we developed a novel algorithm to predict TF activities from expression levels of all genes that the TF regulates. This module is integrated with gene ontology and phylogenic similarity modules using a Bayesian framework. A second microarray data analyzer extracts transcription and RNA degradation kinetic rate constants and TF/gene binding constants at sites of gene regulation. The resulting TRN and information on the genes translated into the TF are fed into a final module which derives the general behavior and critical condition for dramatic changes in cell performance/characteristics. The workflow is demonstrated on several cellular systems.

Identifying distinct types of functionally associated proteins in the *Caulobacter crescentus* Pathway/Genome Database

Michelle L Green* and Peter D Karp, Bioinformatics Research Group, SRI International, 333 Ravenswood Ave, Menlo Park, CA 94025. green@ai.sri.com

The PathoLogic program constructs a Pathway/Genome database (PGDB) using a genome's annotation to predict the set of metabolic pathways present in an organism. PathoLogic determines the set of reactions composing those pathways from the list of enzymes in the organism, and computationally predicts operons for the organism. Pathologic includes predictors for protein complexes and transporters that require manual curator review.

Previously, we extended our pathway hole filler (PHFiller) to include genome context data (e.g., co-occurrence profiles, conserved gene neighbors, gene fusions) in the search for missing enzymes. Adding genome context data improved the coverage of PHFiller by eliminating the need for known enzyme analog sequences from other organisms. PHFiller-GC works by identifying functionally associated proteins for each known enzyme in a pathway and then predicting the probability that each functionally associated protein catalyzes the missing reaction in the pathway based on genome context data.

We have further extended the capability of the PHFiller-GC algorithm to predict additional types of functional associations beyond proteins that appear in the same pathway. These additional functional associations include:

1. Proteins that appear in the same complex.
2. Protein pairs where protein A transports a compound that is acted on by the pathway in which protein B operates as an enzyme.
3. Protein pairs whose genes appear in the same operon.
4. Protein pairs where protein A regulates transcription of the gene encoding protein B.

Our predictor integrates co-occurrence profiles, conserved gene neighbors, gene fusions, gene clusters, and co-expression profiles to identify candidate pairs and evaluate the probability that two genes are functionally associated by one or more of the above criteria. The predictor was trained using proteins from EcoCyc known to be associated by one or more of these functional association criteria. We performed cross-validation studies in EcoCyc to determine the predictive value of the algorithm for identifying known functionally associated protein pairs.

We also applied the full predictor (i.e., identifying any functional association) and the individual predictors (i.e., identifying pairs in the same pathway, same complex, etc.) to the *Caulobacter crescentus* PGDB, CauloCyc, and identified functional associations previously available only through manual curation.

Global Analyses of Two-Component Signal Transduction Pathways in *Caulobacter crescentus*

Michael T. Laub (laub@mit.edu), Emanuele G. Biondi, Jeffrey M. Skerker, Barrett S. Perchuk

Department of Biology, Massachusetts Institute of Technology

Two-component signal transduction systems, comprised of histidine kinases and their response regulator substrates, are the predominant means by which bacteria sense and respond to signals. These systems allow cells to adapt to prevailing conditions by modifying cellular physiology, including initiating programs of gene expression, catalyzing reactions, or modifying protein-protein interactions. These signaling pathways have also been demonstrated to play a role in coordinating bacterial cell cycle progression and development. We have initiated a system-level investigation of two-component pathways in the tractable model organism *Caulobacter crescentus*, which encodes 62 histidine kinases and 44 response regulators. Comprehensive deletion and overexpression screens have identified more than 40 of these 106 two-component genes as required for growth, viability, or proper cell cycle progression. We have also developed a systematic biochemical approach, called phosphotransfer profiling, to map the connectivity of histidine kinases and response regulators.

By combining these genetic and biochemical approaches, we have begun mapping pathways critical to growth and cell cycle progression. This includes a complex genetic circuit that controls the activity of CtrA, the master regulator of the *Caulobacter* cell cycle. At the heart of this circuit are two phosphorelays, one of which culminates in phosphorylation of CtrA and another which leads to proteolytic stabilization of CtrA. Both phosphorelays are initiated by the essential histidine kinase CckA. Once activated and stabilized by these two phosphorelays, CtrA triggers expression of target genes including the essential regulator *divK*. DivK then feeds back to down-regulate CckA, and consequently, CtrA. Our results thus define a negative feedback loop that drives cell cycle oscillations in *C. crescentus*.

We have also used our systematic phosphotransfer profiling technique to probe the molecular basis for specificity in two-component signaling pathways. We have found that histidine kinases are endowed with a global, kinetic preference *in vitro* for their *in vivo* cognate response regulators. This system-wide selectivity insulates two-component pathways from one another, preventing unwanted cross-talk. Moreover, it suggests that the specificity of two-component signaling pathways is determined almost exclusively at the biochemical level. By analyzing patterns of co-evolution between cognate histidine kinases and response regulators we have mapped the amino acids which dictate specificity in these pathways. Site-directed mutagenesis of these amino acids has been used to “rewire” signaling pathways. This serves as both a proof of specificity and may enable (i) the prediction of HK-RR pairs in other organisms and (ii) the rational design of novel signaling pathways for construction of biosensors or synthetic genetic circuits.

Genomics of Cellulosic Ethanol-Producing Bacteria

Christopher L. Hemme[1,4], Matthew W. Fields[2], Qiang He[3], Zhiguo Fang[1], Alla Lapidus[5], Cliff S. Han[6], David Bruce[6], Paul Richardson[5], Eddy Rubin[5] and **Jizhong Zhou**[1]

[1]Institute for Environmental Genomics, Department of Botany and Microbiology, University of Oklahoma, OK
[2]Department of Microbiology, Miami University, Oxford, OH
[3]Temple University, Philadelphia, PA
[1]Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN
[5]DOE Joint Genome Institute, Walnut Creek, CA
[6]Los Alamos National Laboratory, Los Alamos, NM

Recent global fluctuations in the supply and demand of petroleum have prompted a call for increased research into biologically-derived fuels (biofuels). Of particular interest are those processes involving production of biofuels from cellulosic biomass. To this end, the genomes of three strains of ethanol-producing bacteria (*Thermoanaerobacter ethanolicus* 39E, *T. ethanolicus* X514 and *Clostridium cellulolyticum*) have been sequenced. Strain 39E was isolated from a Yellowstone hot spring and is relatively well-characterized. Strain X514 is a metal-reducing bacteria isolated from the deep subsurface and is predicted to have been geographically isolated from 39E for ~200 MY. Metabolic reconstruction reveals insights into the carbon metabolism and niche adaptation of the two strains. Both strains are capable of metabolizing glucose and xylan to ethanol with a novel bifunctional secondary alcohol dehydrogenase serving as the terminal enzyme in the pathway. Slight differences are noted in the carbon metabolism of the two strains, including a complete KDPG metabolism pathway in 39E and the lack of a complete methylglyoxal shunt in X514. A survey of unique genes between the strains reveals lineage-specific gene expansions in the two strains including individual unique sugar transporter profiles and an increased number of P-type metal translocating ATPase genes in X514. In contrast to *Thermoanaerobacter*, *C. cellulolyticum* is capable of degrading a variety of cellulosic materials including cellulose, xylan, pectin, mannan and chitin. *C. cellulolyticum* employs a large extracellular cellulosome complex to degrade these materials and comparisons with a previously sequenced genome of *C. thermocellum* suggest a significant diversity in cellulosome composition. To complement this research, a request for sequencing the genomes of an additional 20 ethanol-producing Clostridia strains has been approved by JGI. Strains were chosen from among the genera *Clostridium*, *Thermoanaerobacter*, *Thermoanaerobacterium* and *Acetivibrio* based on prior knowledge, phylogeny, unique physiology and industrial applications. The expansion of the genomic database of industrially-important Clostridia is expected to provide substantial benefits in the understanding of this class of organisms.

M3 - MICROSCOPIES OF MOLECULAR MACHINES:

The structural basis for regulated assembly and function of the transcriptional activator NtrC

De Carlo[1], B. Chen[2], T. R. Hoover[3], E. Kondrashkina[4], B. T. Nixon[5] and E. Nogales[1]

[1] Life Sciences and Physical Biosciences Divisions, Lawrence Berkeley National Laboratory; and Howard Hughes Medical Institute, Department of Molecular and Cellular Biology, University of California at Berkeley
[2] Integrative Biosciences Graduate Degree Program - Chemical Biology, The Pennsylvania State University, University Park
[3] Department of Microbiology, University of Georgia
[4] BioCAT at APS/Argonne National Lab, Illinois Institute of Technology, 9700 South Cass Ave, Argonne
[5] Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park

In two-component signal transduction, an input triggers phosphorylation of receiver domains that regulate the status of output modules. One such module is the AAA+ ATPase domain in bacterial enhancer-binding proteins that remodel the σ54 form of RNA polymerase. Electron microscopy and X-ray scattering structures of the activated, full-length nitrogen-regulatory protein C (NtrC) reveal a novel mechanism for regulation of AAA+-ATPase assembly via the juxtaposition of the receiver domains and ATPase ring. Electron microscopy studies show that accompanying the hydrolysis cycle that is required for transcriptional activation, a major order–disorder transition occurs in the GAFTGA loops involved in σ54 binding, as well as in the DNA-binding domains.

CONFORMATIONAL VARIABILITY IN EUKARYOTIC TRANSCRIPTION COMPLEXES REVEALED BY CRYO-ELECTRON MICROSCOPY STUDIES

Grob, P.[1], Kostek, S. [1], DeCarlo, S.[1,3], Tjian, R. [1,3], Penczek, P.A.[4], Nogales, E. [1,2, 3]

[1] Molecular and Cell Biology Department, University of California, Berkeley, CA 94720, USA.
[2] Life Sciences and Physical Biosciences Divisions, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.
[3] Howard Hughes Medical Institute, Molecular and Cell Biology Department, University of California, Berkeley, CA 94720, USA.
[4] The University of Texas – Houston Medical School, Department of Biochemistry and Molecular Biology, 6431 Fannin, MSB 6.218, Houston, TX 77030, USA.

The multi-subunit transcription factor TFIID and RNA Polymerase II are essential elements of the transcription machinery in eukaryotes. They interact dynamically with other general transcription factors, activators, inhibitors, DNA and RNA to result in a fine-tuning of transcription. We were particularly interested in the structure of the endogenous human complexes in solution. We adopted the cryo-electron microscopy and single particle approach to obtain 3D reconstructions representing their “average” conformation. Additionally we performed an extensive statistical analysis of the data and obtained the associated 3D variance maps as well as covariance information. The localization and the amplitude of the variations have given us a novel insight into the dynamic of these complex molecular machines. For human TFIID this strategy has shown that the complex "breathes", moving several domains in a concerted manner that reshapes the putative DNA-binding cavities. In human RNA Polymerase II we have identified a more complex, although discreet, set of variable regions that can be interpreted in the context of the yeast crystallographic data.