Scientific Annual Report 2007

Virginia Bioinformatics Institute
Washington St. (0477)
Blacksburg, VA 24061
T: 540.231.2100; **F**: 540.231.2606
e-mail: info@vbi.vt.edu | web: www.vbi.vt.edu

The following scientific reports are not intended as publications and should not be cited without specific permission by the primary author. These reports are only an overview of each research group's activities. For more specific details about the groups' work, please refer to the refereed publications at the end of each report.

Table of Contents

Faculty Reports (cont.)

Faculty Fellow Report

Dear friends,

The publication of this annual report coincides with seven years since the inception of the Virginia Bioinformatics Institute (VBI). From day one, VBI was conceived with the goal of tapping into the power of computing and the life sciences and the innovative potential of a networked, collaborative body of researchers to solve key challenges in the life science arena. We set out to build an institute focused on creating a culture of transdisciplinary collaborative science, one that could make significant contributions to the development of life science knowledge, technologies and useful products aimed at society’s problems. As readers of this report will note, we have traveled a considerable way along the path to building a collaborative research culture, an interactive environment that is gaining momentum in terms of scientific productivity and other levels of output.

Our goal in producing this report is to describe this year’s exciting developments that have taken place under the guidance of faculty members at VBI, and which involve our many collaborators both national and international. For me personally, it is always an exciting time to reflect on research findings and developments and project these tangible achievements into the future of the institute and, indeed, of biology itself. One thing is clear, VBI is a scientific enterprise that has only just begun and many exciting challenges lay ahead. We need to remain responsive to what is a rapidly changing scientific environment and rigorously pursue new opportunities as they arise.

Our research achievements are gaining traction in the wider scientific community, which I believe reflects progress in our peer-reviewed scientific output as well as the outreach activities of faculty and staff at the institute. Work in bioinformatics is leading to new models, tools and a comprehensive cyberinfrastructure that are being adopted by the scientific community. We are eagerly looking at ways to integrate diverse data on cellular components over many scales of organization and in multiple domains of knowledge. Strengths in the study of cellular components are serving as a solid foundation for studies of cellular networks as we pursue our research in systems biology. In complexity science, we are also building an exciting portfolio of projects that presently spans diverse areas, for example, infectious disease epidemiology, public health informatics, social networks, decision-making, and computational behavioral economics.

As we move ahead, we will build further on the many assets that we have put in place through our transdisciplinary, collaborative approach to science. We have every reason to be optimistic about the potential of VBI. I would like to thank all employees of the institute, past and present, for the effort and commitment they have shown in making VBI the success story that it is today.

Sincerely,

Bruno Sobral
Executive and Scientific Director
Virginia Bioinformatics Institute

2007 Research Reports

from the Faculty at the Virginia Bioinformatics Institute

Chris Barrett

cbarrett@vt.edu

Professor, Virginia Bioinformatics Institute

Professor, Dept. Computer Science, Virginia Tech

Group contributors: Kofi Adasi, V. S. Anil Kumar, Karla Atkins, Richard Beckman, Sanket Bedare, Keith Bisset, Deepti Chafekar, Jiangzhuo Chen, Abhijit Deodhar, Stephen Eubank, Ajit Kulkarni, Bryan Lewis, Yi Ma, Achla Marathe, Madhav V. Marathe, Farid Merchant, Henning Mortveit, Joyce Randall, Paula Stretz, Jyotirmaya Tripathi, Xiaoyu Zhou.

Network Dynamics and Simulation Science Laboratory

Abstract. The Network Dynamics and Simulation Science Laboratory has established funded programs of at least $15 million over three years in the area of complex systems, including programs with the National Institutes of Health, the Centers for Disease Control and Prevention, the Defense Threat Reduction Agency and the National Science Foundation. We have established a presence in the National Capital Region, which is located in Virginia Tech's National Capital Region offices, and play a leading role in the institutional initiative in Policy Informatics for Complex Systems. The group is pursuing new programs in communication networks, commodity markets and high performance computing, and continues to develop diverse tools for reasoning about complex systems. Simulation performance is also being studied on TeraGrid architectures. We have used these tools in several stakeholder-designed studies supporting policy planning for pandemics. The United States Department of Health and Human Services has used these studies to create the Community Strategy for Pandemic Influenza Mitigation. The environments we are creating will also be used in a new Centers for Disease Control and Prevention Public Health Informatics program. The group has also developed a new understanding of the dynamics of reaction-diffusion systems over large, irregular, but structured, networks. Both the tools and their theoretical underpinnings will be leveraged as part of a large new multi-year research program devoted to building a national comprehensive incident management system.

Keywords: interaction-based computing; complex systems; infectious diseases; public health; grid computing; service-oriented architectures; population dynamics; social networks; communication networks; commodity markets; discrete dynamical systems; computational complexity; algorithmic semantics; algorithmic theory.

Scientific Progress

Significant progress has been made in achieving important programmatic goals during the last year. Here we highlight some of the achievements. In the reporting period, we continued the development of Simfrastructure – a service and grid computing oriented modeling tool for socio-technical, biological, and information systems. New tools provide capabilities for activity modification, data management for storing, organizing and retrieving simulation output, and visualization of geographical distributions.

We have also continued the development of Simdemics – a scalable high performance computing-based service environment for general reaction diffusion systems. New capabilities include handling complex intervention strategies, representing dynamically co-evolving social networks and improved scaling performance. Other milestones include the design and implementation of EpiFast, a highly scalable algorithm for simulating epidemics and other reaction diffusion systems. For static social networks, EpiFast is three orders of magnitude faster than traditional methods of simulating analogous diffusion processes.

We have refined the synthetic population of 250 million individuals endowed with daily activity patterns performed at specific, real locations. Portions of the population were used in stakeholder-defined studies. New methods were developed for refining activities for subpopulations such as military populations and bases.

We have also participated in three large and three small stakeholder-defined studies related to policy planning for pandemic influenza. This includes studies done for the United States Department of Health and Human Services as part of the National Institutes of Health Modeling Infectious Disease Agent Study (MIDAS) project and for the Defense Threat Reduction Agency to support military planning for force readiness.

We have also developed a new tool for finding optimal sequestration group sizes in the event of pandemics. The tool is based on a service-oriented architecture and has a number of useful features, including personalization, a graphic user interface, and efficient high performance computing-based algorithmic methods.

The group has developed a prototype network analysis toolkit, GALIB, which is designed to handle large complex networks. Several new algorithmic methods were developed and integrated within the toolkit.

In the reporting period, we obtained a number of new theoretical results pertaining to the study of large complex systems:

- New provable algorithmic results for assessing the fundamental performance limits of wireless networks
- New algorithms based on discrete dynamical systems to analyze cascading failures in inter-connected complex networks
- Comprehensive characterization of an important inverse problem, called the pre-image existence problem for finite graph based discrete dynamical systems
- New results on the stability of finite threshold function-based discrete dynamical systems used to study gene annotations

Programmatic progress

The work of the Network Dynamics and Simulation Science Laboratory has resulted in progress in several key program areas. We have established a Defense Threat Reduction Agency project on modeling the spread of infectious diseases related to civil and military planning for situation assessment and response analysis. We are also a partner institution and co-principal investigator in the Centers for Disease Control and Prevention Center of Excellence in Public Health Informatics, a project which is led by the University of Utah School of Medicine. The Network Dynamics and Simulation Science Laboratory has established a Department of Defense project to develop a comprehensive national incident characterization and management system. We

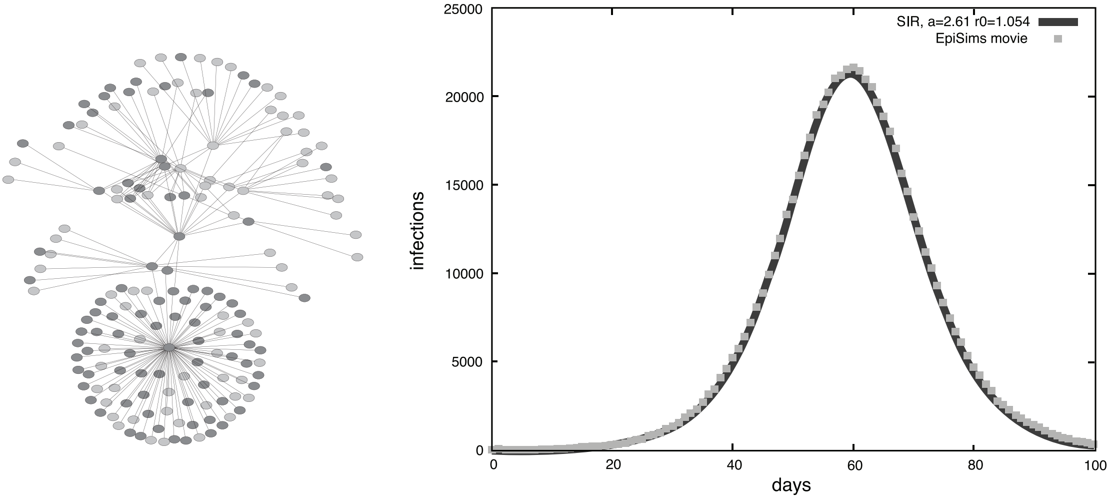


Figure. Contagious disease simulation and epidemiology.

continue to contribute as a principal institution in the National Institutes of Health Modeling Infectious Disease Agent Study (MIDAS) project.

National and international leadership

We have been advising a National Institutes of Health steering committee for developing modeling methods for contagious diseases. Advice is also being given to the United States Department of Health and Human Services and White House on pandemic influenza planning. Members of our group have participated in a European Union program review and planning committee on complexity science. We have also served on an external advisory committee for the Institute of Scientific Exchange at Torino, Italy and on the External Advisory Board of the National Center for Advanced Secure Systems at the National Center for Supercomputing Applications. In addition, members of the group have served on several National Science Foundation panels and as program committee members for the International Parallel and Distributed Processing Symposium (IPDPS), the Intergovernmental Panel on Climate Change (IPDPS), the Workshop on Algorithm Engineering and Experiments (ALENEX) and Ajantasaisen Liikenneinformaation T&K-Ohjelma (AINO; the Research and Development Program on Real-Time Transport Information). We also served on the Lagrange Foundation Selection Committee for the Lagrange Prize.

Outreach and education

A new senior undergraduate Virginia Tech course entitled Algorithms for Real World Problems was introduced and was well received by the undergraduate students. This course will be offered again in Fall 2007. The Science of Complex Networks seminar series has also been very well received and continues to attract good external as well as internal speakers. The group has been involved in the graduation of three Master's students, serves on the committee of four Master's students and five PhD students. In addition, it jointly supervises two PhD students.

We co-organized a mini-symposium within CanaDAM'07 (1st Canadian Discrete and Algorithmic Mathematics Conference, May 28-31, 2007, Banff Conference Center, Alberta, Canada) on the topic of Discrete Dynamical Systems over Graphs. Members of the group also developed an exhibit on the impact of measles vaccination, which was shown at the Marian Koshland (National Academy of Sciences) Museum, Washington, DC. In the reporting period, the Network Dynamics and Simulation Science Laboratory hosted a Modeling Infectious Disease Agent Study (MIDAS) site visit at Blacksburg, VA. The group also represented the Virginia Bioinformatics Institute at Supercomputing 06 (participation at the Virginia Tech booth). As part of outreach efforts, we also participated in VBI's High School program to foster interest in science in high school students in the New River Valley area.

Conferences and workshops

Barrett C (2006) *Transdisciplinary Science and Socio-Technical Problems*, Invited Presentation at the 26th Annual Conference of Center for Nonlinear Studies, Socio-Technical Systems: Bridging the Scales 2006, Oppenheimer Center, Los Alamos National Laboratory, Los Alamos, NM, USA.

Barrett C (2006) *Agency and Detail in the Simulation of Very Large Networks*, Invited Presentation to Salishan Conference Department of Energy Annual High Speed Computing Conference, Gleneden Beach, Oregon, August, 2006.

Barrett C (2006) *High Performance Computing Based Modeling for Weapons of Mass Destruction Problems: Invited Presentation to Department of Defense, Office of the Secretary of Defense, Defense Threat Reduction Agency*. Invited Talk organized by Defense Threat Reduction Agency/Weapons of Mass Destruction, September 2006.

Barrett C (2006) *Large Scale Detailed Simulation of Epidemics*, Invited Presentation to Canadian National Research Council, Canadian Congress on Computing, Social Sciences and Humanities, July 2006.

Barrett C (2006) *Modeling for Decisions in Epidemic Crisis*, Invited Presentation to EU Symposium at Brussells, December 2006.

Eubank S (2006) *Cyberinfrastructure for Epidemiological Simulations*, Invited Presentation to NCSA Infectious Disease Informatics at the University of Illinois at Urbana-Champaign, September 7, 2006.

Eubank S (2006) *Workshop on Modeling Community Containment*, Institute of Medicine of the National Academies, Washington DC, October 25-27, 2006.

Eubank S (2006) *Infectious Disease Informatics: Surveillance, Modeling and Response*, NCSA at the University of Illinois at Urbana-Champaign, September 7-8, 2006.

Eubank S (2006) *Collective and Individual Perspectives on Complex Systems: Seeing the forest and the trees?* University of Manchester Summer School on Complex Systems, Ambleside, England, August 4, 2006.

Eubank S (2006) *New Models, New Perspectives*, Congress of Epidemiology, Seattle, WA, June 24, 2006.

Eubank S (2006) *Social Network Structure and Epidemiology*, SUNY Computer Science Department Seminar, Albany, NY, April 17, 2006.

Eubank S (2006) *Social Network Structure and Epidemiology*, MIT Networks and Dynamics Seminar, Cambridge, MA, April 13, 2006.

Marathe MV (2006) *Computational Considerations in the Study of Real Very Large Complex Networks*, NCSA at University of Illinois at Urbana-Champaign, October 27, 2006.

Marathe MV (2006) *SIAM Spatial Data Mining: Consolidation and Renewed Bearing*, Plenary talk at SIAM Spatial Data Mining Workshop, Hyatt Regency Bethesda, Bethesda, Maryland, October 27, 2006.

Marathe A (2006) *Modeling Tools to Study Energy Markets*, Deans' Forum on Energy Security and Sustainability, Virginia Tech, Blacksburg, VA, October 16.

Marathe A (2006) *Tools to Design Energy Markets & Trading Strategies*, Energy Showcase Symposium, Virginia Tech, Blacksburg, VA, November 29-30, 2006

Mortveit H (2006) *Mathematical Models in Population Dynamics*, Pan American Advanced Studies Institute, University of El Salvador, San Salvador, El Salvador, July 17-28, 2006.

Anil Kumar VS (2007) *An Algorithm for Spread of Disease on a Network Model*, Annual Infectious Disease Epidemiology Spring Seminar Series, Harvard School of Public Health, Boston, MA, May 14, 2007.

Barrett (2007) *High Performance Computing Based Epidemiological Modeling*, Invited Presentation to the Defense Threat Reduction Agency R&D Enterprise and Weapons of Mass Destruction, February 2007.

Barrett (2007) *High Performance Computing Based Simulation,* Invited Presentation to Defense Threat Reduction Agency Director and Senior Management, February 2007.

Bissett K (2007) *Urgent Computing for Interaction-based Socio-Technical Simulations,* Invited Presentation to Argonne National Laboratory, Argonne, IL 2007, April 25-26, 2007.

Eubank S (2007) *Exhibit on impact of measles vaccination*, Invited Presentation to Marian Koshland (National Academy of Sciences) Museum, Washington DC, March 1, 2007.

Eubank S (2007) *Comparing Individual and Collective Models of Infectious Disease Spread*, Invited Presentation to 1st Atlantic Coast Conference on Mathematics in the Life and Biological Sciences, Blacksburg, VA, May, 2007.

Marathe MV (2007) Workshop on Algorithm Engineering and Experiments, Society for Industrial and Applied Mathematics, New Orleans, LA, January 6, 2007.

Peer-reviewed publications in reporting period

Anil Kumar VS, Marathe MV, Thakur M, Sundaram R, Thulasidasan S (2006) Scaling laws for the internet over urban regions, CAIDA (Cooperative Association for Internet Data Analysis). In *ISMA 2006 WIT: Workshop on the Internet Topology*. Proceedings are in electronic form.

Anil Kumar VS, Parthasarathy S, Marathe MV, Srinivasan A, Zust S (2006) Provable Algorithms for Parallel Generalized Sweep Scheduling for Unstructured Meshes. *Journal of Parallel and Distributed Computing* **66:** 807-821.

Atkins K, Barrett CL, Beckman RE, Bissett K, Chen J, Eubank S, Anil Kumar VS, Lewis B, Macauley M, Marathe A, Marathe MV, Mortveit HS, Stretz P (2006) Simulated pandemic influenza outbreaks in Chicago: NIH DHHS Study Final report. *NDSSL Internal Report No. 06-023,* 2006.

Atkins K, Chen J, Anil Kumar VS, Macauley M, Marathe A (2006) Locational market power in power markets. In *29th IAEE International Conference,* 7-10 June, Potsdam, Germany.

Atkins K, Chen J, Kumar VSA, Marathe A (2006) Structural Analysis of Electrical Power Networks, *Proc. 3rd International Conference on Critical Infrastructures,* September 24-27, 2006.

Barrett CL, Beckman RE, Chen J, Eubank S, Anil Kumar VS, Marathe MV (2006) Fast diffuse: Percolation based methods for studying epidemics on directed networks. *NDSSL Technical Report No. 06-022*.

Barrett CL, Eubank S, Marathe MV (2006) Modeling and Simulation of Large Biological, Information and Socio-Technical Systems: An Interaction Based Approach. *Interactive Computing: A new Paradigm*, pp. 353-394.

Barrett CL, Hunt III HB, Marathe MV, Ravi SS, Rosenkrantz DJ, Stearns RE, Thakur M. (2006) Computational complexity of analyzing the dynamic reliability of interdependent infrastructures *Proceedings of the Third International Conference on Critical Infrastructures*.

Barrett CL, Hunt III HB, Marathe MV, Ravi SS, Rosenkrantz DJ, Stearns RE, Thakur M (2006) Dichotomy Theorems for Reachability Problems in Sequential Dynamical Systems. *Journal of Computer and System Sciences* **72**(8): 1317-1345.

Bisset K, Atkins K, Barrett CL, Beckman R, Eubank S, Marathe A, Marathe MV, Mortveit H, Stretz P, Vullikanti AK (2006) Synthetic data products for societal infrastructures and proto populations: Data set 1.0. *NDSSL Technical Report 06-006.*

Chen J, Kumar VSA, Marathe A, Atkins K (2006) Model Based Spatial Data Mining for Power Markets, *SIAM-DM 06, Spatial Data Mining conference*, April 22, 2006.

Chen J, Marathe M, Rajaraman R, Sundaram R (2006) The confluent capacity of the internet: congestion vs. dilation. In *Proceedings of the 26th International Conference on Distributed Computing Systems (ICDCS)*, pp 5-15.

Duncan C, Kobourov S, Anil Kumar VS (2006) Optimal Constrained Graph Exploration. *ACM Transactions of Algorithms* **2:** 380-402.

Eubank S, Anil Kumar VS, Marathe MV, Srinivasan A, Wang N (2006) Structure of social contact networks and their impact on epidemics. In *AMS-DIMACS Special Volume on Epidemiology* Vol. 70, pp. 181-213.

Istrate G, Hansson A, Marathe MV, Thulasidasan S, Barrett CL (2006) Semantic compression of TCP traces. In *Proceedings of the IFIP Conference on Networking*, Vol. 3976, pp 123-135.

Anil Kumar VS, Marathe M, Parthasarathy S, Srinivasan A. (2007) Minimum Weighted Completion Time. In Kao M (ed.), *Encyclopedia of Algorithms*. Springer Verlag. In press September 2007.

Atkins K, Chen J, Anil Kumar VS, Marathe A (2007) Structural properties of electrical networks. *International Journal of Critical Infrastructure*. In press.

Atkins K, Marathe A, Barrett C (2007) A Computational Approach to Modeling Commodity Markets. *Computational Economics*. In press.

Anil Kumar VS, Marathe MV, Parthasarathy S, Srinivasan A. (2007) Scheduling on unrelated machines under tree-like precedence constraints. *Algorithmica*. In press.

Barrett CL, Bisset K, Eubank S, Anil Kumar VS, Marathe MV, Mortveit HS (2007) Modeling and Simulation of Large Biological, Information and Socio-Technical Systems: An Interaction-Based Approach. *Proceedings of the Short Course on Modeling and Simulation of Biological Networks, AMS Lecture Notes Series: PSAPM*, revised and accepted, January 2007.

Barrett CL, Hunt III HB, Marathe MV, Ravi SS, Rosenkrantz DJ, Stearns RE, Thakur M (2007) Predecessor existence problems for finite discrete dynamical systems. *Theoretical Computer Science*. In press.

Barrett CL, Hunt HB III , Marathe MV, Ravi SS, Rosenkrantz DJ, Stearns R, Thakur M (2007) Computational Aspects of Analyzing Social Network Dynamics. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI 07)* pp. 2268-2273. Hyderabad, India

Hunt HB III, Marathe MV, Stearns RE (2007) The Complexities of Unquantified, Quantified, and Stochastic Constrained Satisfaction Problems, *Discrete Applied Mathematics.* In press.

Mozumder P, Marathe A (2007) Causality relationship between electricity consumption and GDP in Bangladesh. *Energy Policy* **35:** 395-402.

Allan Dickerman

dickerman@vt.edu

Assistant Professor, Virginia Bioinformatics Institute

Group contributors: Johanna Craig, Eric Nordberg, Elena Shulaeva, Yuying Tian.

Phylogenomics and Systems Biology of Plants and Plant Pathogens

Abstract. Analyzing and describing the diversity of phylogenetic patterns that characterize the different components of the genomes of organisms allow us to focus on both the dominant vertical component of inheritance and the alternative modes of descent that are shaped by duplication, deletion, and lateral gene transfer. Pulling out the dominant vertical component is useful for fully defining a supported phylogenetic tree model for the group. Making comparisons to this dominant tree model can reveal genes that exhibit alternative patterns of ancestry. The website, genetrees.vbi.vt.edu, which was made public in December, 2006, exports this kind of information and analysis methodology to scientists and the public. The complex genomic patterns generated by evolution reflect population dynamics of organisms in their environments, which can impact human interests in many ways. We are involved in different kinds of environmental surveys of microbes, largely using ribosomal RNA, but with different detection methods. Dr. Marc Fisher, a recent Ph.D. from Virginia Tech's Department of Entomology, analyzed diversity of bacterial endosymbiont in termite guts, doing much of his molecular work in our lab with follow-on work continuing. We are also sampling for microbial diversity in the atmosphere, working with Dr. David Schmale of Virginia Tech's Department of Plant Pathology, Physiology and Weed Science who samples with remote-controlled aircraft. A major effort is a United States Department of Agriculture-funded project to develop a microarray strategy for identifying plant pathogens from infected tissue, which involves using Affymetrix microarray design for rRNA. Our group is also providing bioinformatics support to outside collaborators on projects focusing on functional genomics and systems biology in the plants *Arabidopsis* and *Populus*.

Keywords: phylogenomics; gene phylogeny; essential genes in *Arabidopsis*; plant pathogen; microarray; environmental sampling.

Scientific Progress

The GeneTrees research effort, including a database and website (genetrees.vbi.vt.edu) devoted to phylogenetic analysis of all homologies to be found among the completely sequenced genomes, has yielded fruit in the 2006-2007 period. A paper on the database was published in the database issue of *Nucleic Acids Research* (Tian and Dickerman, 2007). Another branch of this work was a study of the phylogeny of the α-proteobacteria using all the then-available sequenced genomes conducted by Drs. Kelly Williams, Bruno Sobral and Allan Dickerman (Williams et al, 2007). This work produced a species tree for 72 α-proteobacteria from a concatenation of alignments for 104 well-behaved protein families. Most of the tree structure supported currently accepted relationships; however, the *Hyphomonadaceae* family was split from the other *Rhodobacterales*, forming an expanded group with *Caulobacterales* that also included *Parvularcula*. The three earliest-branching α -proteobacterial orders were the *Rickettiales*, followed by the *Rhodospirillales*, and finally the *Sphingomonadales*. The main uncertainty is whether the expanded *Caulobacterales* group is more closely associated with the *Rhodobacterales* or the *Rhizobiales*. The mitochondrial branch was placed within the *Rickettsiales* as sister to the combined *Anaplasmataceae* and *Rickettsiaceae*, all subtended by the *Pelagibacter* branch. *Pelagibacter* genes will serve as a useful addition to the bacterial outgroup in future evolutionary studies of mitochondrial genes, including those that have transferred to the eukaryotic nucleus.

Plant pathochip microarray

Our United States Department of Agriculture (USDA) project entitled "Highly Parallel Pathogen Microarrays for Plant Biosecurity" with VBI's Dr. Chris Lawrence as co-principal investigator has yielded interesting initial results. We have obtained a minimum order of 45 Affymetrix "Phylochips" designed by Dr. Gary Andersen of Lawrence Berkeley National Laboratory in Berkeley, Calif. This chip with 500 000 features has probes to 16S RNA sequences of 9263 'taxa' of prokaryotes. Although there is an emphasis on organisms of concern to human biosecurity, such as *Bacillus anthracis* that has over 2000 probe pairs, there are many direct plant pathogens represented, such as *Pseudomonas syringae*. Although we do not have the specific probe sequences, we can infer which probes represent each of a large range of taxonomic groups of varying diversity, given information in the support data files for the chip provided by Affymetrix and the Andersen lab. Work with Dr. Chris Lawrence has yielded successful demonstrations of the system on laboratory infections of *Arabidopsis* leaves with four pathogens. We anticipate receiving samples of naturally infected crop samples from collaborator Dr. Stephen Goodwin of Purdue University. Over the course of the coming year, we will be working with Affymetrix to design and build a new chip focusing on plant pathogens, both eukaryotic and prokaryotic.

Poplar xylem systems biology

As proteome studies have grown in importance in the biological sciences, we were fortunate to have the opportunity to participate in a grant submission with Drs. Eric Beers in Virginia Tech's Department of Horticulture and Amy Brunner in Forestry to pursue studies of the proteins involved in wood formation in *Populus*, the poplar tree. This builds on prior work we have done with Dr. Beers on the expression profile of *Arabidopsis* vascular tissue, including xylem (Zhao et al, 2005), and more recent work by Dr. Brunner in *Populus* using a *Populus* expression microarray. The proposed work will focus on tracking the interacting partners of proteins shown by transcriptomics to be uniquely involved in the biology of xylem differentiation. Techniques used will pair yeast two-hybrid screens with affinity purification mass spectrometry.

SeedGenes

The long-standing SeedGenes project led by Dr. David Meinke at Oklahoma State University was renewed in August, 2006 for three years. VBI's Yuying Tian worked on this project to develop a BioMoby web service

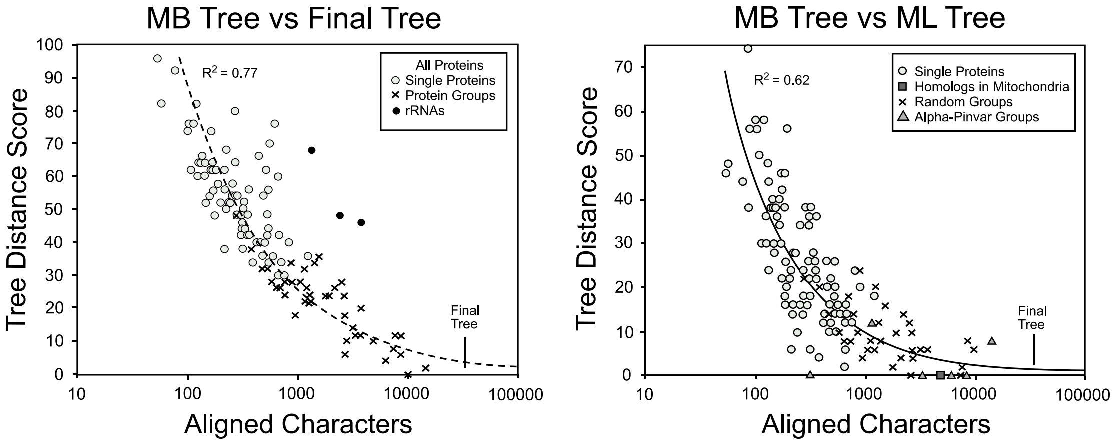


Figure. The need for large datasets for robust phylogeny inference is shown by these plots of two measures of the reduction in topology disparity among trees inferred from increasingly large datasets representing 72 genomes of α-proteobacteria. Approximately 10 000 aligned amino-acid positions are needed to approach convergence to the level obtained by the 33 000 aligned positions (the full dataset included). One implication is that a single protein homology group of 100-300 amino acids will regularly fail to reveal the historical pattern. ML, maximum likelihood; MB, maximum likelihood bounded tree.

to serve mutant phenotype images from the SeedGenes database. This effort contributed to the Arabidopsis Web Services Demonstration Project and is listed among deployed web services at the official web site (http://bioinfo.mpiz-koeln.mpg.de/araws/web-services/public-ws/). This demonstration effort helps integrate our group into the evolving bioinformatics community.

Environmental microbiology

One effort in the reporting period was the preparation and submission of a proposal to the National Science Foundation's Microbial Observatory program. Dr. David Schmale of Virginia Tech's Department of Plant Pathology, Physiology and Weed Science Department led this work with Dickerman and Dr. Craig Woolsey as co-principal investigators. The thrust of the proposed project is to survey the airborne load of microbes, including bacteria, fungal spores and other propagules over agricultural fields using the technology of remote and autonomously controlled small aircraft. The proposal was declined, but we believe we have laid a strong basis for a re-submission supported by more developed preliminary data. In particular, we are working on an application of VBI's Roche GS-FLX™ high-throughput sequencing system to survey microbial diversity via PCR of small-subunit ribosomal RNA from aerial samples collected over a nine-month period.

Marc Fisher, a graduate student in Virginia Tech's Department of Entomology who did much of his DNA sequencing work in our lab, held his Ph.D. defense on April 25, 2006, entitled "Comparison of Subterranean Termite (Rhinotermitidae: Reticulitermes) Gut Bacterial Diversity Within and Between Colonies and to Other Termite Species Using Molecular Techniques (ARDRA and 16S rRNA Gene Sequencing)". Marc accepted a position as Product Technology Specialist for Dow AgroSciences for the East Coast. Two manuscripts have been submitted from Marc Fisher's work on termite endosymbionts and are scheduled for publication soon. Work continues on this project, deepening Marc's diversity surveys using VBI's GS-FLX™ sequencer.

Conferences and workshops

Dickerman A (2006) *Highly Parallel Pathogen Microarrays for Plant Biosecurity*. Invited talk, American Phytopathological Society Meeting, Project Directors Meeting for the United States Department of Agriculture-Cooperative State Research, Education, and Extension Service (USDA-CSREES) Plant Biosecurity Competitive Program, Quebec, Canada, 31 July 2006.

Publications in reporting period

Snyder EE, Kampanya N, Lu J, Nordberg EK, Karur HR, Shukla M, Soneja J, Tian Y, Xue T, Yoo H, Zhang F, Dharmanolla C, Dongre NV, Gillespie JJ, Hamelius J, Hance M, Huntington KI, Jukneliene D, Koziski J, Mackasmiel L, Mane SP, Nguyen V, Purkayastha A, Shallom J, Yu G, Guo Y, Gabbard J, Hix D, Azad AF, Baker SC, Boyle SM, Khudyakov Y, Meng XJ, Rupprecht C, Vinje J, Crasta OR, Czar MJ, Dickerman A, Eckart JD, Kenyon R, Will R, Setubal JC, Sobral BWS (2007) PATRIC: The VBI PathoSystems Resource Integration Center. *Nucleic Acids Research* **35**: D401-D406.

Tian YY, Dickerman AW (2007) GeneTrees: a phylogenomics resource for prokaryotes. *Nucleic Acids Research* **35**: D328-D331.

Williams KP, Sobral BW, Dickerman AW (2007) A Robust Species Tree for the Alphaproteobacteria. *J Bacteriol* **189**: 4578-4586.

Tyler BM, Tripathy S, Zhang XM, Dehal P, Jiang RHY, Aerts A, Arredondo FD, Baxter L, Bensasson D, Beynon JL, Chapman J, Damasceno CMB, Dorrance AE, Dou DL, Dickerman AW, Dubchak IL, Garbelotto M, Gijzen M, Gordon SG, Govers F, Grunwald NJ, Huang W, Ivors KL, Jones RW, Kamoun S, Krampis K, Lamour KH, Lee MK, McDonald WH, Medina M, Meijer HJG, Nordberg EK, Maclean DJ, Ospina-Giraldo MD, Morris PF, Phuntumart V, Putnam NH, Rash S, Rose JKC, Sakihama Y, Salamov AA, Savidor A, Scheuring CF, Smith BM, Sobral BWS, Terry A, Torto-Alalibo TA, Win J, Xu ZY, Zhang HB, Grigoriev IV, Rokhsar DS, Boore JL (2006) Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science* **313:** 1261-1266.

Other cited references

Zhao CS, Craig JC, Petzold HE, Dickerman AW, Beers EP (2005) The xylem and phloem transcriptomes from secondary tissues of the Arabidopsis root-hypocotyl. *Plant Physiology* **138**: 803-818.

Ina Hoeschele

inah@vbi.vt.edu

Professor, Virginia Bioinformatics Institute

Professor, Dept. of Statistics, Virginia Tech

Group contributors: Bing Liu, Alberto de la Fuente, Yongcai Mao, Lei Bao, Lucia Gan.

Genetic Architecture of Complex Diseases

Abstract. The common theme of our research is the use of data from systems genetics experiments to further our understanding of how the joint action and interaction of multiple genes determines complex diseases or phenotypes at the organism level. We have evaluated and continue to evaluate different methods for high-dimensional mapping of expression quantitative trait loci. We have implemented a structural equation modeling analysis based on maximum likelihood inference and a genetic algorithm that can reconstruct networks consisting of several hundred gene and expression Quantitative Trait Loci nodes. This method is based on genetical genomics or systems genetics experiments that provide causal inference and strong constraints on network topology. We have analyzed several microarray expression and proteomic profiling experiments comparing different models and methods for differential expression analysis. Our research group has performed some efficiency enhancing modifications to our deterministic method for haplotype reconstruction in complex human pedigrees and we are finalizing a review paper on this topic.

Keywords: systems genetics; statistical genetics; gene network inference; design and analysis of microarray experiments; Quantitative Trait Loci (QTL) mapping; haplotype inference in complex pedigrees.

Scientific Progress

High-dimensional mapping of expression quantitative trait loci and construction of an encompassing (directed) gene network

Several expression Quantitative Trait Loci (eQTL) mapping approaches were evaluated: Single Profile Analysis (SPA), Principal Component (PC) mapping, and *cis*-, *cis-trans-* and *trans*-eQTL mapping (Liu, de la Fuente and Hoeschele, unpublished results). Results on a genetical genomics dataset from yeast showed that *cis*- and *trans*-eQTL mapping and PC mapping greatly increased power for eQTL detection as compared to SPA. Both *trans*- and PC-eQTL mapping incorporated information from correlated expression profiles showing that SPA needs to be replaced with a method that utilizes the correlation structure of the (high-dimensional) expression profiles. Several related methods are currently being evaluated, including alternative dimension-reduction, an improved version of *cis*, *cis-trans-* and *trans*-eQTL mapping, and Bayesian analysis across all expression profiles. With the identified eQTL regions, using DNA sequence information, genes physically located in an eQTL region are identified as candidate causal regulators of target genes affected by the eQTL. For regulator-target pair identification within each eQTL region, we use local structural models that performed well on simulated data. For the yeast data set, the identified regulator-target pairs from all eQTL regions were assembled into an encompassing network using the eQTL results from PC- and *cis*-, *cis-trans-* and *trans*-mapping. The network included 28 609 regulator-target pairs.

Structural equation modeling for causal network inference in genetical genomics and systems genetics experiments

We implemented Structural Equation Modeling (SEM) to sparsify the encompassing network by producing a set of sub-models containing fewer edges and being well supported by the data (Liu, de la Fuente and Hoeschele, unpublished results). Typically, SEM has been implemented for, at most, tens of variables. Based on a factorization of the likelihood and a strongly constrained search space, our algorithm can construct networks of several hundred genes and eQTL. Parameters are estimated by maximum likelihood and structure inference is based on a penalized likelihood ratio and an adaptation of Occam's Window model selection. The likelihood function is factorized into a product of conditional likelihoods of single genes (not contained in a cycle) and of subsets of genes that constitute cyclic components. The likelihood of a cyclic component is maximized using a genetic algorithm. The SEM algorithm was evaluated on artificial data generated with random network topologies and nonlinear ordinary differential equations. For the simulated networks, the SEM analysis had an average power of 90% and a false discovery rate (FDR) of 10%, so the linear SEM performed well on this type of non-linear data. We are now implementing and comparing deterministic and stochastic Bayesian SEM analyses incorporating prior information on the network topology. In collaboration with Dr. Brett Tyler at the Virginia Bioinformatics Institute and co-workers, we plan to apply this methodology to a soybean genetical genomics experiment for quantitative resistance to the pathogen *Phytophthora sojae*. This should allow us to construct a causal network based on 300 recombinant inbred lines that are profiled for disease resistance, expression of 37 420 genes and thousands of DNA markers.

Haplotyping in complex human pedigrees

Haplotyping in pedigrees is an essential component of genetic studies on pedigrees (e.g., linkage analyses and association studies). To identify a set of haplotype configurations with the highest likelihoods for a large pedigree with a large number of linked markers, we have previously proposed a conditional enumeration haplotyping method by setting a threshold for the conditional probabilities of the ordered genotypes at unordered person-markers to delete ordered genotypes with low conditional probabilities. In our current work (Gao and Hoeschele, in press), we present a rapid algorithm based on a modification of our previous

Table. Comparison among the original and modified conditional enumeration haplotyping methods (denoted by "Original" and "Modified", respectively) and SimWalk2 (2.83) based on analyses of a 163- and a 198-member pedigree. For smaller marker distances, the enumeration method substantially outperforms SimWalk2 (times in boldface) in Central Processing Unit (CPU) time while finding multiple haplotypes with higher likelihood. The times shown in the table are for a 2.00 GHz Intel (R) Xeon(TM) CPU (1,047,546 KB RAM, MS Window 2000). Thresholds λ and α are used to eliminate haplotype configurations with low likelihood (Gao and Hoeschele, 2007).

Number of individuals in the pedigree (*N*)	Distance between adjacent markers (cM)	Number of loci in the (single) linkage group	Method	Thresholds		Number of haplotype configurations with the estimated highest log-likelyhood	Time on the CPU (h:min:sec)
				λ	α		
163	10	10	Original	0.835	-	-266.223 (17)	4:15:20
			Modified	0.98	-2.2	-265.221 (18)	0:58:57
			SimWalk2	-	-	-271.001 (1)	1:09:11
198	2.0	15	Original	0.97	-	-281.575 (16)	**0:07:41**
			Modified	0.995	-3.0	-281.575 (33)	**0:10:35**
			SimWalk2	-	-	-369.891 (1)	**160:42:34**

method, which finds and sets an additional threshold for the conditional probabilities of (partial) haplotype configurations to early eliminate configurations with low conditional probabilities. The new algorithm is much more efficient than our former method and the widely used software SimWalk2 (see table).

Analysis of expression and proteomics profiling experiments

We have analyzed microarray experiments within the soybean-*Phytophthora sojae* project. The researchers were interested in evaluating the effects of multiple factors and their interactions on the expression of 37 420 soybean genes in a split-plot model with multiple random factors using Proc Mixed of SAS and control of the FDR. An optimal procedure for FDR control in a complex, multi-factorial experiment remains to be investigated. We have analyzed a proteomics experiment with 26 444 molecular components quantified in samples of 15 patients with renal clear-cell carcinoma before and after surgical removal of the tumor as a first step in the identification of biomarkers. To identify components with differential expression, we compared paired *t*-test with normality based and permutation *p*-values, Wilcoxon test, and a regularized *t*-test.

Conferences and workshops

Hoeschele I (2006) Gene network inference from genetical genomics experiments using structural equation modeling. *6th Annual Meeting of the Complex Trait Consortium*, Chapel Hill, NC, May 6-10, 2006.

Stock KF, Distl O, Hoeschele I (2006) Estimation of genetic parameters for categorical, continuous and molecular genetic data in multivariate animal threshold models using Gibbs sampling. *2006 Meeting of the European Association of Animal Production*.

Stock KF, Distl O, Hoeschele I (2006) Multivariate estimation of genetic parameters for categorical, continuous and molecular genetic data in threshold models using Gibbs sampling. *CME conference on genome-wide association studies: design and analysis*. Yale School of Public Health, October 26-27.

Dorrance A, Mideros S, St. Martin S, Saghai Maroof M, Zhou L, Tripathy S,
Mao Y, Hoeschele I, Tyler B (2006) Expression from eight soybean genotypes with different levels of partial resistance following inoculation with *Phytophthora sojae*. *North Central American Phytopathological Meetings*. Fargo, North Dakota.

Saghai Maroof MA , Tucker D, Skoneszka J, Dorrance AE, Mideros M, St. Martin SK, Zhou L, Tripathy S, Mao Y, Hoeschele I, Tyler BM (2006) Genomics of disease resistance in soybean: QTL mapping and expression profiling. *11th Biennial Conference on the Molecular and Cellular Biology of the Soybean*. August 5-8, Lincoln, Nebraska.

Hoeschele I, Liu B, de la Fuente A, Mao Y, Zhou L, Hanlon R, Dorrance AE, St. Martin S, Saghai Maroof MA, Tyler BM (2007) Inference of gene networks controlling quantitative resistance in soybean against *Phytophthora sojae*. *International Plant and Animal Genome Conference XV*, San Diego, January 2007.

Zhou L, Mideros SX, Tripathy S, Mao Y, Torto-Alalibo G, Li H, Constanzo S, Liu B, St. Martin S.K., Saghai Maroof MA, Hoeschele I, Dorrance AE, Tyler BM (2007) Whole genome transcriptional profiling reveals diverse mechanisms of quantitative resistance in soybean to *Phytophthora sojae*. *International Plant and Animal Genome Conference XV*, San Diego, January 2007.

Peer-reviewed publications in reporting period

Betthauser JM, Pfister-Genskow M, Xu H, Gouleke PJ, Lacson JC, Koopang RW, Liu B, Hoeschele I, Eilertsen KJ, Leno GH (2006) Nucleoplasmin facilitates reprogramming and in vivo development of bovine nuclear transfer embryos. *Molecular Reproduction and Development* **73**:977-986.

Stock KF, Distl O, Hoeschele I (2007) Influence of priors in Bayesian estimation of genetic parameters for multivariate threshold models using Gibbs sampling. *Genetics, Selection, Evolution* **39**:123-137.

Gao G, Hoeschele, I (2007) A note on a haplotyping method in pedigrees. *Genetics, Selection, Evolution* (in press).

Hoeschele I (2007) Mapping quantitative trait loci in outbred pedigrees. In *Handbook of Statistical Genetics*, Balding DJ, Bishop M, Cannings C (eds) pp 477-525. Wiley. (in press).

Reinhard Laubenbacher

reinhard@vbi.vt.edu

Professor, Virginia Bioinformatics Institute

Professor, Dept. of Mathematics, Virginia Tech

Group contributors: Abdul Salam Jarrah, Reinhard Laubenbacher, John McGee, Alan Veliz Cuba, Paola Vera-Licona.

Computational Systems Biology

Abstract. One of the central problems in systems biology is to infer biochemical networks from system-wide experimental measurements, including gene regulatory, metabolic, and signaling networks, so-called "top-down" modeling. This report details the theoretical and applied advances our group has made toward the development of mathematical tools to solve this problem. We use techniques from discrete mathematics and symbolic computation, implemented in open source symbolic computation software. The techniques are being tested on both simulated and published data sets, as well as a data set generated as part of a collaborative yeast systems biology project, in collaboration with the Mendes and Shulaev groups at the Virginia Bioinformatics Institute.

Keywords: modeling and simulation of biological networks; dynamical systems theory; reverse-engineering of biological networks; yeast systems biology.

Scientific Progress

Introduction

The major focus of the Applied Discrete Mathematics Group remains the development and application of algorithms and software to reverse-engineer biochemical networks, in particular gene regulatory networks, from large-scale system measurements such as DNA microarray data. We have chosen the modeling paradigm of time-discrete dynamical systems on finite state sets. This class of models includes in particular Boolean networks which have a long history as models of gene regulatory and other biochemical networks. The medium-term goal of the group is to release software that allows computationally unsophisticated users to reverse-engineer networks based on their experimental data. The paradigm for this software is Copasi, a biochemical network simulator that makes simulation capabilities accessible to a broad audience of users without a modeling background.

In the past year, we have improved our existing algorithms, advanced the development of new ones that address shortcomings in the existing software, and validated the algorithms and software on simulated as well as real data sets. Furthermore, we have carried out a theoretical mathematics program that focuses on the study of polynomial dynamical systems over finite state sets, in support of the reverse-engineering algorithm development. This report details progress in these three areas.

Development of reverse-engineering algorithms

Our basic approach to reverse-engineering was published in 2004 (Laubenbacher and Stigler, 2004). The algorithm had several shortcomings as published, even though it performed well on appropriately chosen data sets. What distinguishes the algorithm from many others is that it provides a nonlinear dynamical system model of the network rather than a linear approximation or only a wiring diagram. The drawback is that it requires more and special types of data in order to perform well. In the reporting period, we have developed three other, related algorithms that exploit different types of mathematical structure that is inherent in the problem. The first one is an algorithm that provides as output only a wiring diagram and hence has less stringent data requirements than the original algorithm (published in Jarrah et al, 2007). The basic modeling framework employed by the group is the class of polynomial dynamical systems

$$f = (f_1, \dots, f_n): X^n \rightarrow X^n$$

in variables $x_1, \dots, x_n$, representing genes or gene products. The f_i are polynomials in the x_i, with coefficients in a finite set of states X, denoted by $X[x_1, \dots, x_n]$.

Suppose we have observed a collection of state transitions of one network node in response to system perturbations as follows:

$$\{(\boldsymbol{s}_i, t_i) \mid i=1, \dots , r\}.$$

Here, each $\boldsymbol{s}_i$ represents a vector in n variables with entries assumed to be from the set X of possible values. (That is, we assume that the experimental measurements have been discretized into finitely many states, e.g., "up-regulated, down-regulated, unchanged.") And each t_i represents a new state in X. The problem is to find all *minimal* (with respect to inclusion) sets of variables $y_1, \dots , y_m \in \{x_1, \dots , x_n\}$ such that $M \cap k[y_1, \dots , y_m] \neq \emptyset$, where M is the space of all polynomial functions that interpolate the observed data. Since the wiring diagram of a given model for a given variable x is derived from the variables that appear in the transition function f for x, the problem translates into finding all minimal sets of variables for which one can find a function f in those variables that explains the given data set. Each such minimal set corresponds to a minimal wiring diagram for x. We obtain a collection of minimal wiring diagrams for the network by carrying out this search for each variable.

We have developed an algorithm that solves this problem, have implemented it and applied it to several different data sets, with positive results. The implementation is in the computer algebra systems *Macaulay2*, and forms part of a larger reverse-engineering software package we are developing together with *Macaulay2* co-developer Michael Stillman at Cornell University.

A second algorithm addresses a problem with model selection in the original algorithm that depends on the fact that one needs to choose a particular total ordering of the variables and the resulting most likely model may depend on this choice. Using a geometric object called the *Gröbner fan* of the set of all functions that evaluate to 0 on the given data points (Dimitrova et al, 2007b) one can systematically explore all possible choices. It still requires more thorough testing and validation.

Finally, the third algorithm is motivated by the fact that the original algorithm relies on exact interpolation of the given data and is therefore susceptible to overfitting of noisy data. It takes a very different approach to the problem of model selection, employing an evolutionary algorithm to search the space of all possible models, optimizing a cost function that measures data fit and model complexity.

Method validation and comparison

We have done extensive work on the validation of these algorithms, as well as a comparison of the performance of other reverse-engineering algorithms on data generated from

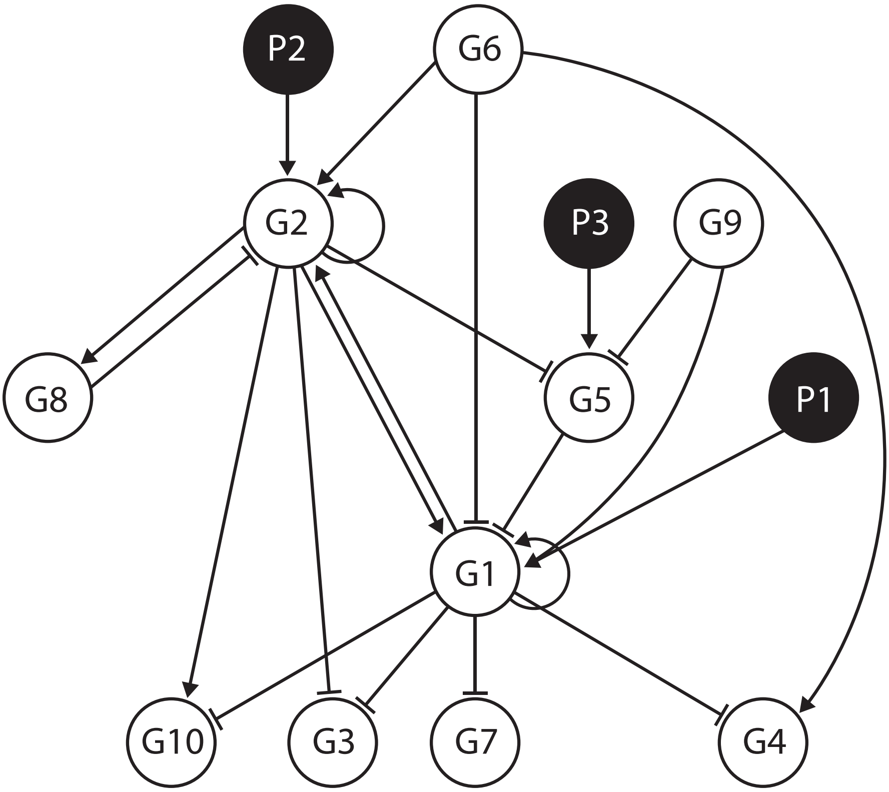


Figure. Simulated gene regulatory network.

one of the networks we have been using for this purpose. In order to test the algorithms without the confounding effect of using particular data discretization algorithms we have been using data generated from a published Boolean model of the segment polarity network in *Drosophila melanogaster* (Albert and Othner, 2003) containing 21 variables. In Stigler et al (2007), data from this model were used successfully to reverse-engineer the underlying dynamic network. The data have also been used for model validation (Vera-Licona et al, unpublished results).

In order to test the methods using real-valued data, we have used a ten-node simulated gene network, with data generated by Copasi. The wiring diagram of the network is depicted in the figure. The network includes three perturbations as well as different time scales to simulate data with realistic characteristics. Data from this network were used in Camacho et al (2007) as well as in Dimitrova et al, 2007a.

Development of a theoretical foundation

In the course of algorithm development several mathematical problems have arisen that are best solved in the long term by developing a rigorous mathematical foundation for the class of finite dynamical systems that forms the basis for our approach. Since the same class of systems also plays a central role elsewhere, for instance in agent-based simulation, such a foundation is more broadly applicable, and we are working on related problems in collaboration with the Barrett group.

Our efforts are focused on two fundamental problems:

1. *Using the structure of the polynomial functions describing a finite dynamical system, derive features of the resulting dynamics.* Such a capability is crucial in improving the model selection feature of our algorithms, since it would allow the preferential choice of models with specified dynamic characteristics. For instance, in modeling a system that only shows steady state behavior, it is preferable to select models that do not have any periodic attractors. Colon-Reyes et al, 2006 contains results pertinent to this question. Also, in a submitted paper (Jarrah et al, 2007), we have designed and implemented an algorithm that solves this problem completely for linear systems.
2. *Characterize computationally special classes of models that are constructed from biologically meaningful functions.* Several classes of such functions have been proposed for the construction of Boolean network models of biological systems. Examples include canalyzing and nested canalyzing Boolean networks. A successful solution of this problem would allow the preferential selection of models constructed from such functions. In Jarrah et al (2007), we have solved part of this problem for the case of nested canalyzing Boolean functions. We have parameterized this class by the points in a toric algebraic variety of the field with two elements. In the process we discovered that nested canalyzing Boolean functions, introduced by S. Kauffman in 2003, are identical to unate cascade functions, studied in computer science and engineering since the 1970s. It was shown recently that this class consists precisely of those Boolean functions that lead to binary decision diagrams with shortest average path length (Butler et al, 2005).

Conferences and workshops

Laubenbacher R (2006) *Discrete Models in Epidemiology,* Summer School on Mathematical Modeling in Epidemiology, San Salvador, El Salvador, July 23-24, 2006.

Laubenbacher R (2006) *Mathematical Methods in Computational Biology,* Annual Meeting of the Society for the Advancement of Native Americans and Chicanos in Science (SACNAS), Tampa, FL, October 26-29, 2006.

Laubenbacher R (2006) *A Computational Algebra Approach to Yeast Systems Biology,* Center for Genomics, Cuernavaca, Mexico, September 25, 2006.

Laubenbacher R (2007) *Parametric Inference of Biochemical Network Models,* SIAM Minisymposium on Mathematical Modeling of Complex Systems in Biology, Annual Joint Meetings of the AMS-MAA-SIAM, New Orleans, LA, January 6, 2007.

Laubenbacher R (2007) *Parametric Inference of Biochemical Network Models,* Keynote lecture, 2nd Bioinformatics Research Symposium, Clemson University, January 18, 2007.

Laubenbacher R (2007) *Polynomial dynamical systems over finite fields, with applications to modeling and simulation of biological networks,* IMA Workshop on Applications of Algebraic Geometry to Biology, Dynamics, and Statistics, University of Minnesota, March 7, 2007.

Laubenbacher R (2007) *Discrete models in systems biology*. Gulbenkian Institute, Lisbon, Portugal. April 27, 2007.

Peer-reviewed publications in reporting period

Choi V, Huang Y, Lam V, Potter D, Laubenbacher R, Duca K (2006) Using formal concept analysis for microarray data comparison. *5th Asia-Pacific Biocomputing Conference*.

Colon-Reyes O, Jarrah A, Laubenbacher R, Sturmfels B (2006) Monomial dynamical systems over finite fields. *Complex Systems* **4**(16): 333-342.

Camacho D, Vera-Licona P, Mendes P, Laubenbacher R (2007) A comparison of reverse-engineering methods using an in silico network. *Annals of New York Academy of Sciences*. In press.

Castiglione F, Duca KA, Jarrah A, Laubenbacher R, Hochberg D, Thorley-Lawson DA (2007) Simulating Epstein-Barr Virus Infection with C-ImmSim. *Bioinformatics* **23:** 1371-1377.

Choi V, Huang Y, Lam V, Potter D, Laubenbacher R, Duca K (2007) Using formal concept analysis for microarray data comparison. *Journal of Bioinformatics and Computational Biology* **6**(1): 57-66.

Dimitrova E, Jarrah A, Stigler B, Laubenbacher R (2007) A Groebner fan-based method for biochemical network. *Proceedings of The International Symposium on Symbolic and Algebraic Computation 2007 (ISSAC 2007)*. In press.

Jarrah A, Laubenbacher R (2007) Discrete models of biochemical networks: The toric variety of nested canalyzing functions. *Proceedings of the Second International Conference on Algebraic Biology*. In press.

Jarrah A, Laubenbacher R, Stigler B, Stillman M (2007) Reverse-engineering of polynomial dynamical systems. *Advances in Applied Mathematics*.In press.

Jarrah A, Raposa B, Laubenbacher R (2007) Nested canalyzing, polynomial, and unate cascade functions. *Physica D.* In press.

Knoebel A, Laubenbacher R**,** Lodder J, Pengelley D (2007) Mathematical Masterpieces. In *Undergraduate Texts in Mathematics. Readings in Mathematics.* Springer-Verlag, New York. In press.

Laubenbacher R (2007) Mathematical Methods in Computational Biology. In *Symposia in Pure and Applied Mathematics*, Laubenbacher R (ed), American Mathematical Society. In press.

Stigler B, Jarrah A, Stillman M, Laubenbacher R (2007) Reverse-engineering dynamic networks. *Annals of New York Academy of Sciences.* In press.

Other cited references

Albert R, Othmer H (2003) The topology of the regulatory interaction network predicts the expression pattern of the segment polarity genes in *Drosophila melanogaster. J. Theor. Biol.* **223:** 1-18.

Laubenbacher R, Stigler B (2004) A computational algebra approach to the reverse-engineering of gene regulatory networks. *J. Theor. Biol.* **229:** 523-537.

Butler JT, Sasao T, Matsuura M (2005) Average path length of binary decision diagrams. *IEEE Trans. on Computers* **54:** 1041-1053.

Jarrah A, Laubenbacher R, Stillman M, Vera-Licona P (2006) An efficient algorithm for finding the phase space structure of linear finite dynamical systems. *J. Symbolic Comp.* Submitted.

Dimitrova E, Vera-Licona P, McGee J, Laubenbacher R (2007) Comparison of data discretization methods for the inference of biochemical networks. *BMC Bioinformatics.* Under review.

Vera-Licona M, Jarrah A, McGee J, Garcia L, Laubenbacher R (2007) An optimization algorithm for the inference of biochemical networks. In preparation.

Chris Lawrence

lawrence@vbi.vt.edu

Associate Professor, Virginia Bioinformatics Institute

Associate Professor, Dept. of Biological Sciences, Virginia Tech

Group contributors: Mihaela Babiceanu, Megan Calliham, Yangrae Cho, Amanda Cronin, Kwang-Hyung, Kim, Mauricio LaRota, Graciela Santopietro.

Fungal Pathogenomics

Abstract. Many pathological relationships have developed between ubiquitous airborne fungi (molds) and prospective hosts across kingdoms. Fungi cause devastating diseases of plants, animals, humans and even other fungi. In some cases, co-evolution of host and potential pathogen has resulted in the selective tailoring of the innate immune system of plants, animals and other eukaryotes to recognize, respond to and persevere against attempted invasive infection. Most ubiquitous fungi do not typically cause invasive infections in plants and humans except in certain species or in an immune-compromised host, respectively. Even though the innate human immune system is very effective at detecting and thwarting attempted colonization and infection by fungi, many chronic airway diseases such as allergy, asthma and various forms of sinusitis are the result of both deleterious alterations in airway functionality and overzealous innate and adaptive immune responses leading to chronic inflammation. In extreme cases, these immune system defects can result in morbidity in humans. More often, however, they significantly affect an individual's daily quality of life and productivity. In our laboratory, the majority of research thus far has been focused on pathogenomics of the ubiquitous fungus *Alternaria*. Besides being an important genus of necrotrophic fungi causing economically important plant diseases, *Alternaria* has long been clinically associated with chronic respiratory diseases such as IgE-dependent mold allergy, life-threatening forms of asthma, chronic sinusitis and other allergic airway disorders. We continue to develop not only a new model pathosystem for plant-fungal genomics research (The *Alternaria brassicicola*-Brassicaceae interaction), but also a new model fungal genomics system for chronic airway disorder research centered upon the more clinically relevant species, *Alternaria alternata*.

Keywords: *Alternaria*-Brassicaceae pathosystem; necrotrophic fungal-plant interactions; genome sequence; *Alternaria brassicicola*; chronic respiratory disorders; fungal biotechnology.

Scientific Progress

Introduction

In this post-genome, technologically savvy era, we are in the unique position to further elucidate the underlying etiologies and molecular signaling events associated with plant and human chronic diseases caused by various types of sensitivities to fungi. It is quite clear that molecular recognition and subsequent signaling play critical roles in these pathosystems and determine the overall outcome of the interaction and severity of the disease in question. By understanding the delicate interplay of host- and pathogen-derived molecules in these diseases, novel therapeutic strategies, new diagnostics, drug targets and, eventually, medicines may be able to be developed over the next decade to reduce the deleterious impacts of ubiquitous fungi of several kingdoms on the planet.

The Alternaria brassicicola genome sequencing project

In 2004, the NSF-USDA Interagency Microbial Genome Sequencing Program funded the project entitled "The Genome Sequence of the Model Necrotrohpic Fungus, *Alternaria brassicicola*". The project team was assembled at the Virginia Bioinformatics Institute (VBI) and included collaborators at the Washington University Genome Sequencing Center (WUGSC) in St. Louis, Mo., who played a major role in the Human Genome Project, as well as scientists at other institutions and a corporate partner (Solexa, Inc., Hayward, Calif.). Collectively this team had expertise in areas such as whole genome shotgun sequencing (WGS), BAC library construction, Expressed Sequence Tag (EST) sequencing, database development, genome annotation, and unique transcript profiling technologies.

During this project, WUGSC has produced a whole genome shotgun sequence of *A. brassicicola* to 6.3x coverage. A BAC library has been used to generate a BAC-based fingerprint physical map of the genome. The BAC library has been assembled into contiguous sequences and a minimum tile of BACs has been end-sequenced to provide a framework to facilitate assembly. A 0.3x coverage fosmid library has also been constructed and end sequenced as an additional resource to benefit assembly. Massively parallel signature sequencing (MPSS) of fungal mRNA by Solexa Inc. is ongoing and is being used as an additional annotation feature primarily for gene prediction. One MPSS library has been created from heavily infected plant tissue and will provide a quantitative glimpse at the fungal genes expressed during pathogenesis.

A bioinformatics platform called Alternaria Community Annotation Database (ACAD), which allows for a web-based display of the annotated draft genome and community finishing and annotation, is near completion. Machine annotation will be followed soon by manual curation of select genes of interest. For machine annotation of the *A. brassicicola* genome we adapted an EnsEMBL genome annotation pipeline (Alternaria Genome Annotation Pipeline) this year to make use of Timelogic's FPGA accelerating cards and their bioinformatics algorithms (Decypher) by writing modules that extend the current software-only capacity of the pipeline.

In summary, this project is nearing completion and the resulting annotated genome sequence information has already proven invaluable for data mining including the identification of putative fungal pathogenicity factors of plants (proteins and metabolites). In addition, by using a combination of proteomics and immunological approaches, the *A. brassicicola* genome resources have allowed for the identification of a subset of immunoreactive/antigenic proteins with a potential role in human respiratory diseases. In a consortium project with the Allergic Airway Diseases Laboratory at the Mayo Clinic, Rochester Minn., we have identified a subset of *Alternaria* proteins that potently cause eosinophil degranulation, dendritic cell activation and T-cell maturation using mammalian immune cells and a mouse model of *Alternaria*-induced airway inflammation.

Alternaria functional genomics and related technologies

While sequencing the *A. brassicicola* genome, we concomitantly began working towards the development of an improved high throughput functional genomics platform for molecular manipulation of *Alternaria* species in a general sense. In this regard, we developed an extremely efficient homologous recombination-based gene knockout and overexpression system that was easily adapted from *A. brassicicola* to *A. alternata* (Cho et al, 2006). This system has already proven useful for identifying and functionally characterizing plant virulence factors (Cho et al, 2007; Kim et al, 2007). As described in these and our unpublished studies, we have knocked out well over 100 genes in *A. brassicicola* in two years and have identified several novel virulence factors of plants among secondary metabolite biosynthesis, signal transduction and transcription factor genes.

The efficiency of homologous recombination-based gene knock-out (KO) and molecular manipulation approaches far surpass what is available for the other major asthma-associated fungus, *Aspergillus fumigatus*. It is noteworthy to mention that the *A. fumigatus* genome sequence has been completed and some progress has been made at identifying fungal antigens potentially contributing to allergy and allergic airway disorders. However, when a genome sequence for *A. alternata* becomes available, more rapid progress than in virtually any other host-fungal system can be made. This involves the identification of immunogenic fungal products and characterization of host responses/signaling events associated with *Alternaria*-specific and fungal-associated respiratory disorders in a general sense. We have already begun amassing genome-scale information for *A. alternata* through the execution of several EST sequencing projects in collaboration with the Allergic Airway Disease Laboratory at the Mayo Clinic. Our long-term vision is the further development of *A. alternata* as the next and most efficient model organism for studying life-threatening, fungal-associated respiratory disorders and, in particular, asthma.

In addition to identification of fungal genes important for plant and human health, we have also been studying the response of cultivated Brassicas and the model plant *Arabidopsis* to *A. brassicicola* infection using ESTs, whole genome microarray and MPSS platforms (Cramer et al, 2006). In the future we will begin functional characterization of host genes differentially expressed during fungal infection through the use of *Arabidopsis* T-DNA KO mutants and the use of silencing and overexpression approaches.

Acknowledgements

These projects were supported by the National Research Initiative of the USDA Cooperative State Research, Education and Extension Service, grant number # 2004-35600-15030, by the National Science Foundation Award # DBI-0443991, by the National Institute for Allergy and Infectious Diseases Award #AI49235, and Mayo Research Foundation.

Conferences and workshops

La Rota CM, Tyler BM, Mitchell TK, Brown S, Knudson D, Clifton S, Haudenschild C, Meyers B, Lawrence CB (2006) *The genome of the necrotrophic fungus,* Alternaria brassicicola. 24th Fungal Genetics Meeting, Asilomar, CA, March, 2007.

La Rota CM, Weeks C, Tyler BM, Clifton S, Lawrence CB (2006) *Modules to extend the capabilities of the EnsEMBL automated annotation pipeline for its connection to Timelogic's decypher hardware accelerator engine and Xgrid computer clusters: Application to the Alternaria genome sequencing project.* 9th Computational Genomics Conference. Baltimore, MD, October 28-31, 2006.

Cho Y, La Rota CM, Kim K, Scott DC, Santopietro G, Calliham M, Cronin A, Lawrence CB (2007) *High throughput discovery of virulence factors among transcription factors of* Alternaria brassicicola. 24th Fungal Genetics Meeting, Asilomar, CA, March 20-25, 2007.

Lawrence CB (2007) *Secrets of the mold: the use of a fungal genome for biomedical and biotechnological applications*. Invited talk, Astronaut High School, Titusville, FL, April 24-25, 2007.

Lawrence CB (2007) *The* Alternaria brassicicola *genome project and functional genomics*. Invited talk, 24th Fungal Genetics Conference, Pacific Grove, CA, March 20-25, 2007.

Lawrence CB (2007) *Identification of virulence factors among secondary metabolite-related genes in* A. brassicicola. Invited talk, 24th Fungal Genetics Conference, Pacific Grove, CA, March 20-25, 2007.

Lawrence CB (2007) *Current status of the Alternaria genome project and applications for airway disorder research*. Invited talk, Mayo Medical School, Rochester, MN, March 13, 2007.

Lawrence, CB (2007) *Alternaria Pathogenomics*. Invited talk, University at Buffalo School of Medicine and Biomedical Sciences, Buffalo, NY, July 12, 2006.

Lawrence CB (2007) *The* Alternaria brassicicola *genome project: current status of annotation and applications for health-related research across host kingdoms*. Invited talk, International Plant and Animal Genome Conference XV, San Diego, CA, January 13-17, 2007.

Peer-reviewed publications in the reporting period

Cho Y, Davis JW, Kim K, Wang J, Sun Q, Cramer RA, Lawrence CB (2006) A high throughput targeted gene disruption method for *Alternaria brassicicola* functional genomics using Linear Minimal Element (LME) constructs. *Molecular Plant-Microbe Interact.* **19**: 7-15.

Cramer RA Thon M, Cho Y, Craven KD, Knudson DL, Mitchell TK, Lawrence CB (2006) Bioinformatic analysis of expressed sequence tags derived from a compatible *Alternaria brassicocola—Brassica oleracea* interaction. *Molecular Plant Pathology* **7**: 113-124.

Cho Y, Cramer RC, Kim K, Pryor BM, Lawrence CB (2007) The Amk1 Map Kinase Regulates Virulence Factors in *Alternaria brassicicola. Fungal Genetics and Biology* **44**: 543-553.

Kim K, Cho Y, Cramer RC, Lawrence CB (2007) Functional analysis of the *Alternaria brassicicola* non-ribosomal peptide synthetase gene *AbNPS2* reveals a role in conidial cell wall construction. *Molecular Plant Pathology* **8**: 23-29.

Intellectual property

Lawrence CB, Cho Y (03/15/07) Targeted and non-targeted gene insertions using a linear minimal element construct. US Utility Patent Application Filed, Virginia Tech Intellectual Properties (VTIP).

Kita H, Ponikau J, Lawrence CB (10/13/06) Fungus induced inflammation and eosinophil degranulation. US Utility Patent Application Filed, VTIP and Mayo Medical Ventures.

Iuliana Lazar
lazar@vt.edu
Assistant Professor, Virginia Bioinformatics Institute
Assistant Professor, Dept. of Biological Sciences, Virginia Tech
Adjunct Assistant Professor, Dept. of Biomedical Engineering, Virginia Tech

Group contributors: Jenny M. Armenta, Abdulilah A Dawoud, Yang Xu.

Proteomic Studies for Biomarker Discovery in MCF7 Breast Cancer Cells Using Conventional and Microfluidic Mass Spectrometry Instrumentation

Abstract. The development of novel technologies and bioanalytical strategies for fast proteomic profiling of cancerous cells and tissues is essential for speeding up the discovery process of early disease biomarkers. Moreover, the perfection of such technologies can lead to their ultimate implementation for fast, large-scale population screening applications. To date, we have developed bioanalytical protocols that have enabled the confident identification of ~2600 proteins ($P<0.001$) from only ~40 µg of cellular extracts. Over 200 of these proteins were correlated with cellular processes relevant to cancer, and more than 25 were previously reported as putative cancer biomarkers. Our long-term objective is focused on transferring this technology onto high-throughput microfluidic platforms for the detection of protein co-expression patterns. The reliable quantitation of all protein components and the identification of post-translationally modified proteins, in particular phosphoproteins, are the major topics that are pursued.

Keywords: oncoproteomics, biomarkers, microfluidics, mass spectrometry.

Scientific Progress

Comparative analysis of non-labeling and iTRAQ-based methods for protein quantification in the MCF7 breast cancer cell line

Breast cancer is one of the most frequent forms of cancer affecting women worldwide and in the United States. Although the intricate pathways that lead to breast cancer on-set and development remain to be completely elucidated, certain hormones such as estradiol are known to play a major role in its development. Innovative treatments are constantly being implemented to fight this disease. For example tamoxifen has proven effective in improving survival rates. To gain a better insight into the global molecular mechanisms that intervene in cancer development, and the specific effects that estradiol and tamoxifen have on the progression/regression of breast cancer, a quantitative proteomic study was conducted using the MCF7 breast cancer cell line cultured in the presence of estradiol and/or tamoxifen.

Two-dimensional liquid chromatography followed by mass spectrometry detection (2DLC-MS/MS) has emerged as an attractive technology for quantitative proteomics. Both labeling- and non-labeling-based approaches have been explored. Among the labeling techniques, iTRAQ (Isobaric tags for relative and absolute quantification) has received much attention. In this method, peptides are labeled with isobaric tags at the N-terminus and the lysine side chains. MS/MS fragmentation produces signature ions that can be used to obtain quantitative information. Non-labeling techniques rely on the use of internal standards that are spiked in the sample. In our study, four MCF7 cell extracts were digested with trypsin and treated separately with iTRAQ labeling reagents (*m*/*z* = 114–117). The four fractions were combined in a 1:1:1:1 ratio and separated by strong cation-exchange reverse-phase liquid chromatography (SCX/RPLC)-MS/MS. For the non-labeling-based strategy, the MCF7 cell extracts were spiked with standard proteins, digested with trypsin, and prefractionated by SCX. The sample sub-fractions were then subjected to RPLC coupled to an ion trap LTQ mass spectrometer.

The qualitative studies of MCF7 resulted in the identification of ~2600 proteins with P<0.001 and a rate of false positive identification of ~0.4–2.5%. The limits of detection were in the low fmol range (1–40 fmol). Proteins were consistently quantified at various concentration levels (0.005–0.5 µM), with much less variability for proteins present at higher concentrations and quantified by several unique peptides. The variability of the iTRAQ ratios for all quantified proteins in the MCF7 cell extract is shown in Figure 1. Two sets of labeled extracts were mixed in a 1:1 ratio and analyzed by RPLC-MS/MS. These studies will be further used for the identification of proteins

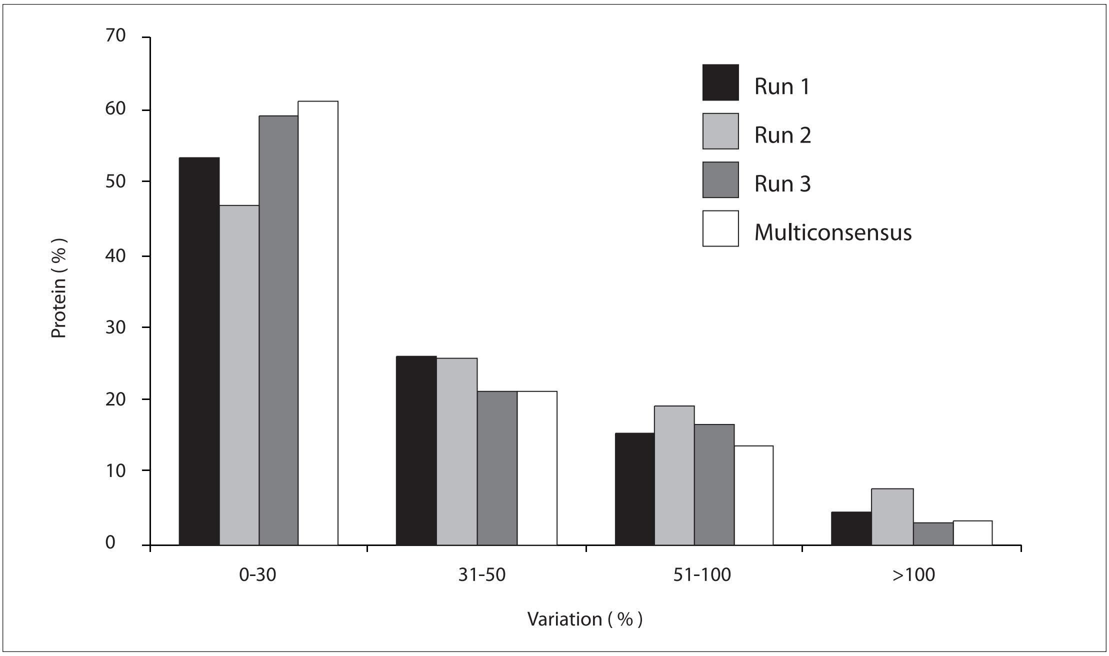


Figure 1. Variation (%) of iTRAQ ratios for quantified proteins in the MCF7 cell extract (calculated from triplicate runs and a multiconsensus analysis).

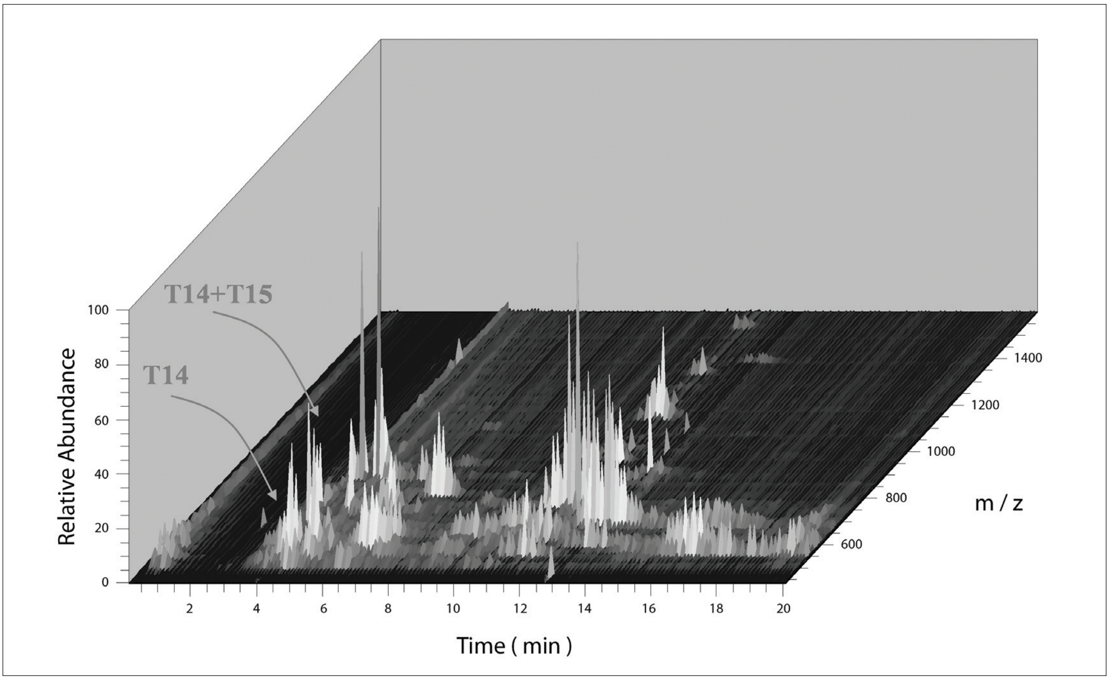


Figure 2. Two-dimensional-view chromatogram of a microfluidic LC-data-dependent triple-play MS analysis of a mixture of 10 proteins (tryptic digest). Phosphorylated peptides were detected from low intensity signals in the presence of multiple, intense, non-phosphorylated peptides.

(putative biomarkers) that are differentially expressed between cells cultured with estradiol alone, or estradiol and tamoxifen.

Microfluidic liquid chromatography/mass spectrometry platform for phosphoproteomics analysis

Protein phosphorylation is involved in many regulatory biological processes and is considered to be one of the most important post-translational modifications of proteins. Protein phosphorylation plays a central role in critical cellular events such as regulation of gene expression and protein synthesis, events that ultimately lead to the control of cell division, growth, and differentiation. Furthermore, phosphorylation is considered to be a key event in regulating cellular signal transduction and amplification. Hence, the development of novel and reliable technologies for the analysis of proteins and their post-translational modifications, in particular, has recently received much attention and interest. The implementation of a fully integrated microfluidic platform interfaced with mass spectrometry detection for the analysis of phosphoproteins has been demonstrated in this project. The microfluidic platform (dimensions 3" x 1") comprises two individual sample-processing systems: one for performing direct sample infusion, and one for performing microfluidic liquid chromatography (LC) separations. Various MS detection strategies, specific for the study of post-translational modifications, were conducted using α-casein as a model protein (Figure 2). Neutral loss ion mapping, data-dependent triple-play and neutral loss analysis, and *in-situ* dephosphorylation followed by LC separation and MS detection, were performed. Consistent results in identifying phosphopeptides with conventional and microfluidic instrumentation have been obtained. Unlike with conventional instrumentation, however, the microfluidic device enabled the completion of each analysis from only a few microliters of sample, in less than 10 min, and on a disposable platform that facilitates multiplexing and contamination-free analysis. The microfluidic platform will be further utilized for the analysis of targeted phosphorylated biomarker components from the MCF7 breast cancer cell line.

Acknowledgment

This work was supported by National Science Foundation (NSF) Career grant BES-0448840.

Intellectual property

Lazar IM (08/24/06) Microfluidic device and method for on-chip mass spectrometry detection and complex sample processing, US Utility Patent Application.

Conferences and workshops

Sarvaiya HA, Lazar IM (2006) *Microfluidic chip for biomarker screening in cancer cell extracts*, 54th Conference on Mass Spectrometry and Allied Topics, Seattle, WA, USA, May 28-June 1, 2006.

Lazar IM, Armenta JM, Dawoud AA, Yang X (2007) *A novel microfluidics approach for biomarker screening applications*, AACR Special Conference on Advances in Proteomics in Cancer Research, Amelia Island, FL, USA, February 27-March 2, 2007.

Peer-reviewed publications in reporting period

Lazar IM, Trisiripisal P, Sarvaiya HA (2006) Microfluidic liquid chromatography system for proteomic applications and biomarker screening. *Anal. Chem.* **78**(15): 5513-5524.

Sarvaiya HA, Yoon JH, Lazar IM (2006) Proteome profile of the MCF7 cancer cell line: a mass spectrometric evaluation. *Rapid Commun. Mass Spectrom*. **20**: 3039-3055.

Bissel P, Geherin S, Igarashi K, Gandour RD, Lazar IM, Castagnoli N Jr (2006) Mass spectrometric studies on 4-aryl-1-cyclopropyl-1,2-dihydropyridinyl derivatives: an examination of a novel fragmentation pathway. *J. Mass Spectrom*. **41**(12): 1643-1653.

Pedro Mendes

mendes@vbi.vt.edu

Associate Professor, Virginia Bioinformatics Institute

Adjunct Associate Professor, Dept. of Biochemistry, Virginia Tech

Adjunct Associate Professor, Dept. of Cancer Biology, Wake Forest University

Professor, School of Computer Science, University of Manchester

Group contributors: Diogo Camacho, Autumn Clapp, Hui Cheng, Stefan Hoops, Adaoha Ihekwaba, Aejaaz Kamal, Xing Jing Li, Ana M. Martins, Bharat Mehrotra, Pedro Mendes, Saroj Mohapatra, Revonda Pokrzywa.

Modeling Biochemical Networks

Abstract. Biochemical networks are central to biological function and computer modeling provides a particularly useful way to understand their workings. Biochemical models are the ideal means to design and predict the effect of interventions such as curing diseases, improving crop yields, and designing biotechnology. This report summarizes the activities carried out in our group in several aspects of biochemical modeling: construction of models, simulation software, and methods for analyzing omics data. We also summarize our activities in community-driven efforts such as the Yeast Systems Biology Network and the Systems Biology Markup Language (SBML) forum.

Keywords: systems biology; modeling and simulation of biochemical networks; reverse-engineering of biochemical networks; metabolomics; data analysis.

Scientific Progress

The Biochemical Networks Modeling Group carries out research in the broad area of biochemical networks; our main goal is to develop computational methods for studying these systems using data from experimental observations. Biochemical networks are central to biological function and computer modeling provides a particularly useful way to understand their workings. We have been pursuing two separate strategies of building biochemical models: a) from molecular information obtained *in vitro* with purified molecules, or b) from genome-wide measurements of the effect of perturbations *in vivo*. The first approach is well established in principle and is currently the most popular. It can be referred to as a "bottom-up" strategy because its principle is to combine a number of model entities which are well characterized, and derive the system's behavior from the interactions of these elements. The second approach is less well developed and is based on observing the intact system and how it responds to perturbations to infer how the system is organized internally (a "top-down" strategy). The first approach requires by necessity that we have previously identified the components of the system and how they are organized, which in terms of biochemistry means that we must have already identified all of the relevant reactions, enzymes/genes, and metabolites; this is usually conveyed in the form of a biochemical network. In the top-down approach, the objective is to discover the underlying network from the cell's behavior, also known as "reverse-engineering". The group's research in the reporting period was in both of these areas and included development of software infrastructure for both types of activity, as well as actual modeling of specific networks.

A common component of the two modeling strategies described above is that both require the simulation of the behavior of the system through a mathematical model. We have for a long time been involved in this activity mainly through the development of the simulator COPASI (Complex Pathway Simulator). This year, the COPASI project achieved two important milestones: the publication of an article describing the program and its algorithms (Hoops et al, 2006), and the establishment of a commercial licensing scheme that allows for-profit corporations to use the software for a small fee (non-profit use is free of charge). In the meantime, we have released two stable versions of COPASI and the total number of downloads is above 5000. We have also carried out several outreach activities in the form of demonstrations and tutorials. COPASI is the product of a long-term collaboration with the group of Ursula Kummer, which has recently moved to the University of Heidelberg.

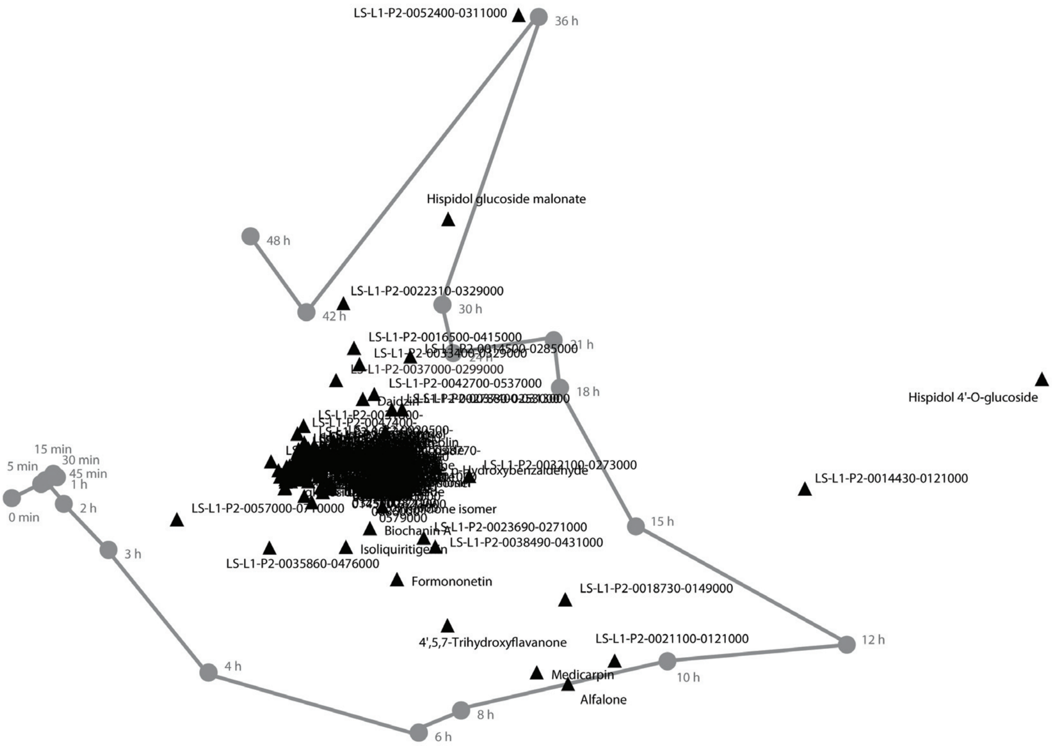


Figure. Biplot display of the response of *M. truncatula* cells to a fungal elicitor. Depicted are metabolites of the phenylpropanoid, flavonoid, and isoflavonoid pathways (triangles) and the data points of the time course after addition of elicitor (circles).

Another important piece of software infrastructure that we have worked on are databases, both of biochemical networks as well as of data from large-scale omics experiments. Biochemical network databases are equally important for both the bottom-up and top-down approaches, while the latter also needs databases to organize the omic experimental data that are used for the inferences. The group has been developing two such databases: B-Net is a biochemical network reference database and DOME is an integrative functional genomics database. B-Net was developed first as a means of annotating data in DOME, but we have also made it independent as it can be used as a reference for many other purposes. B-Net is a relational schema that includes highly detailed information about metabolism, signaling, and genetic networks, allowing for tracing and classifying the evidence that is behind the information stored. B-Net was used in the Yeast 2D Annotation Jamboree that took place in Manchester, England, under the auspices of the Yeast Systems Biology Network, and which unified two independent reconstructions of yeast metabolism. The final unified model will be made available publicly using B-Net.

In the last year, we have become part of a new collaboration between several metabolomics researchers, which is funded by the National Institute of General Medical Sciences (NIGMS). The Metabolomics Network for Drug Response Phenotype, led by Duke University, is carrying out studies to integrate metabolomics and pharmacogenomics to achieve a deeper understanding of the drug-response phenotype. The network is carrying out clinical studies on two important drug classes, an 3-hydroxy-3-methyl-glutaryl coenzyme A (HMG CoA) reductase inhibitor (simvastatin) and a specific serotonin reuptake inhibitor (escitalopram). Our DOME database is being used to store and integrate data obtained from various partners of the project.

Further work on metabolomics data analysis has continued in the group. In particular, we continued studying the application of machine learning algorithms to the analysis of metabolomics data. In collaboration with the group of Nancy Love in the Department of Civil and Environmental Engineering at Virginia Tech, we employed the GA-DFA algorithm to identify biomarkers of community toxicity in active sludge cultures (Henriques et al, 2007). The same algorithm has been used to identify molecules of importance in the response of *Medicago truncatula* cells to biotic stresses, and cancer cells to oxidative stress. Another data analysis method that we have been developing for metabolomics, proteomics and gene expression data is the biplot. The biplot is a visualization technique that is the multidimensional equivalent of scatter plots. It has been used to visualize multidimensional data sets and relate the variables observed with the experiments. We have noted that when the experiments are part of a time course, the resulting biplot is equivalent in some ways to a phase space diagram. We have been developing this concept using data from the *M. truncatula* and *Vitis vinifera* projects of previous years.

The data analysis methods described above are important for top-down modeling as they provide unbiased analyses of the data that lead to clues of how the network is organized. However, it is expected that much more information may be obtained if the experiments are designed to fulfill specific methods for network reverse-engineering. For this, we continued the development of a method to reverse-engineer biochemical networks using time course observations that also makes use of specific null mutants (genetic perturbations). This method is based on a continuous model of the dynamics of a generic biochemical network, which is then calibrated through nonlinear regression using the experimental data. A particularly appealing aspect of this method is that in its iterations, the method suggests the next experiment to carry out, closing the experiment-analysis loop. This research project is in parallel with a similar method being developed by the Laubenbacher group at the Virginia Bioinformatics Institute using discrete mathematics and using data obtained from the Shulaev group. As part of this effort, we have created artificial (simulated) gene networks to produce simulated experiments that are then used to attempt to infer the original network. These artificial data were used to benchmark other reverse engineering methods (Camacho et al, 2007).

Finally, we have also carried out research in the context of bottom-up modeling, where we continued collecting enzyme kinetic data to calibrate a yeast pentose-phosphate pathway model. The pentose phosphate pathway is extremely important as it is the source of building blocks for DNA and proteins, and it is also responsible to keep the redox balance of the cell, particularly under oxidative stress conditions. The first three enzymes of the pathway, which constitute the oxidative branch of this pathway, were purified and assayed resulting in kinetic parameters that are being used in the construction of a model of this pathway. Other experimental work of the group included comparing techniques for extraction of metabolites, proteins and mRNA from yeast (Martins et al, 2007).

Conferences and workshops

Pokrzywa R, Broeckling CD, Sumner LW, Mendes P (2006) *Identifying major determinants in temporal metabolomics data with machine learning*, 3rd Metabolomics Society Conference, Boston, MA, June 24-29, 2006.

Cheng H, Mehrotra B, Smith JT, Sumner L, Dixon RA, Mendes P (2006) *Biplot display for metabolomics data*, 3rd Metabolomics Society Conference, Boston, MA, June 24-29, 2006.

Mendes P, Cheng H, Mehrotra B, Kamal A, Li XJ (2006) *Metabolomics as a component of systems biology: issues in data fusion*, 3rd Metabolomics Society Conference, Boston, MA, June 24-29, 2006.

Mendes P (2006) *Building dynamic models from functional genomics data: a systems biology challenge*, Gordon Research Conference on Macromolecular Organization and Cell Function, South Hadley, MA, August 7-11, 2006.

Mendes P (2006) *Top-down modeling of biochemical networks – a grand challenge in systems biology*, International Workshop on Integrative Bioinformatics, Harpenden, UK, September 4-6, 2006.

Camacho D, Vera-Licona P, Laubenbacher R, Mendes P (2006) *Comparing reverse engineering methods using an* in silico *network*, DREAM Workshop, New York, NY, September 7-8, 2006.

Mendes P (2006) In silico *models for reverse engineering: complexity and realism versus well-defined metrics, DREAM Workshop*, New York, NY, September 7-8, 2006.

Mendes P, Sahle S (2006) *Advanced model analysis with COPASI*, International Conference on Systems Biology, Yokohama, Japan, October 8, 2006.

Hoops S (2006) *COPASI – Software for modeling and simulation of biochemical networks*, University of Alabama at Birmingham, Birmingham, AL, October 20, 2006.

Mendes P (2006) *Adventures in reverse engineering of biochemical networks*, Pittsburgh University, Pittsburgh, PA, December 5, 2006.

Cheng H, Mehrotra B, Mendes P (2007) *To gain insight into metabolomics time course data using biplot display*, Pacific Symposium of Biocomputing, Maui, Hawaii, January 3-7, 2007.

Mendes P (2007) *Software development for computational systems biology: where we should be going and why*. Computational Cell Biology Workshop, Cold Spring Harbor Laboratory, NY, March 6-9, 2007.

Mendes P, Camacho d, Vera-Licona P, Laubenbacher R (2007) *Benchmarking parameter estimation and reverse engineering strategies*, Parameter Estimation in Systems Biology, Manchester, UK, March 28-29, 2007.

Mendes P (2007) *Integrated data management for kinetic modelling and experimental results*, Keynote presentation at Workshop on Storage and Annotation of Reaction Kinetics Data, Heidelberg, Germany, May 21-23, 2007

Peer-reviewed publications in reporting period

Camacho D, Vera-Licona P, Mendes P, Laubenbacher R (2007) Comparison of reverse-engineering methods using an *in silico* network. *Annals of the New York Academy of Sciences.* In press.

Henriques ID, Aga DS, Mendes P, O'Connor SK, Love NG (2007) Metabolic footprinting: a new approach to identify physiological changes in complex microbial communities upon exposure to toxic chemicals. *Environ. Sci. Technol. 41*, 3945-3951.

Hoops S, Sahle S, Gauges R, Lee C, Pahle J, Simus N, Singhal M, Xu L, Mendes P, Kummer U (2006) COPASI - a COmplex PAthway SImulator. *Bioinformatics 22*, 3067-3074.

Martins AM, Sha W, Evans C, Martino-Catt S, Mendes P, Shulaev V (2007) Comparison of sampling techniques for parallel analysis of transcript and metabolite levels in *Saccharomyces cerevisiae*. *Yeast. 24*, 181-188.

Mendes P (2006) Metabolomics and the challenges ahead. *Brief. Bioinform. 7*, 127.

Sansone SA, Fan T, Goodacre R, Griffin JL, Hardy NW, Kaddurah-Daouk R, Kristal BS, Lindon J, Mendes P, Morrison N, Nikolau B, Robertson D, Sumner LW, Taylor C, van der Werf M, van Ommen B, Fiehn O (2007) The metabolomics standards initiative, *Nature Biotechnol. 25*, 846-848.

Biswarup Mukhopadhyay

biswarup@vbi.vt.edu

Assistant Professor, Virginia Bioinformatics Institute

Adjunct Professor, Depts. of Biochemistry and Biology, Virginia Tech

Group contributors: Christopher Case, Deanna Colton, Lakshmi Dharmarajan, Eric Johnson, Karla Piedl, Endang Purwantini, Jason Rodriguez, Jennifer Stieber, Dwi Susanti, Lindsay Von Herbulis, Ban Wang.

Extreme Condition Microbiology and Enzymology

Abstract. We work on extreme condition microbiology and enzymology, with the goal of developing therapeutics for tuberculosis (TB) and diabetes, new energy production processes, and a better understanding of how metabolic reactions and pathways arose and evolved from inorganic reactions and how microorganisms survive or thrive in extreme environments. The approaches we use include microbiology, bioinformatics, proteomics, protein structure modeling, protein and small molecule biochemistry, enzymology and enzyme kinetics. We have found that the activity of phosphoenolpyruvate carboxykinase, an enzyme playing a major role in the development of type 2 diabetes and attainment and maintenance of dormancy by *Mycobacetrium tuberculosis* (causative agent of TB) in the human body, can be controlled by interfering with residues that are away from the active site, but connected to it via long-range interactions. This observation can be used to develop therapeutics that will help maintain blood glucose level in type 2 diabetics at an optimal level and will allow for the development of drugs for treating tuberculosis. In our work on the mycobacteria we have identified a protein that is involved in cell envelop synthesis. We are exploring the possibility of this protein being a target for the development of TB drugs. We are characterizing clinical strains of *M. tuberculosis* from Indonesia to understand the mechanism for the development of more virulent and drug resistant tuberculosis. Work on *Methanocaldococcus jannaschii*, a hyperthermophilic archaeon that lives in deep-sea hydrothermal vents, has allowed us to develop a hypothesis on the evolution of methanogenesis, sulfate reduction and anaerobic methane oxidation. In our work on the conversion of coal and paraffin to methane to increase natural gas production, we have isolated microorganisms with interesting and useful metabolic properties.

Keywords: evolution; metabolism; methanogenesis; sulfate reduction; methane oxidation; mine explosion mitigation; tuberculosis; diabetes.

Scientific Progress

Structure-function analysis of phosphoenolpyruvate carboxykinase (PEPCK)

PEPCK catalyzes nucleotide-dependent reversible synthesis of phosphoenolpyruvate (PEP) from oxaloacetate (Case et al, 2006; Mukhopadhyay et al, 2001). Two types of PEPCKs are known: GTP-dependent and ATP-dependent. PEP synthesis is the first committed and one of the rate controlling steps of gluconeogenesis and glycerogenesis in certain archaea and bacteria, as well as several eukaryotes, including human. Consequently, it contributes to the development of type 2 diabetes in humans and is perhaps required for attainment and maintenance of dormancy or latency of *Mycobacterium tuberculosis* within a human host; *M. tuberculosis* causes tuberculosis (TB). These two implications have been the motivation for our work on PEPCK. We work on GTP-PEPCKs from human liver cytosol and *Mycobacetrium smegmatis*, a close non-pathogenic relative of *M. tuberculosis*. In both cases we use recombinant forms that are structurally manipulated, expressed and purified from *E. coli*. In 2001 we had reported that the mycobacterial PEPCKs are similar to the human liver enzyme both in primary structure and kinetic properties (Mukhopadhyay et al, 2001). The scientific focus of our work has been to identify the catalytically essential and influential residues of GTP-PEPCKs and elucidate the key differences between the mycobacterial and human enzymes. Our approaches used are structural modeling, evolutionary deductions, site-directed mutagenesis, and kinetic and biophysical studies. We collaborate with Pete Dunten from the Stanford Synchrotron Radiation Laboratory to help determine the crystal structures of variant enzymes. Information derived from these studies will allow us to develop therapeutics that will lower but not eliminate the activity of human PEPCK and thereby achieve normal blood glucose levels in a type 2 diabetic person. A complete inhibition of PEPCK is not an option because this enzyme nourishes our body between meals and a lack of its activity would lead to hypoglycemia. Another goal is to develop drugs that will selectively target *M. tuberculosis* PEPCK and thereby prevent the attainment and maintenance of the organism's dormant state. Mycobacterial dormancy is the reason why a lengthy drug treatment is necessary for treating TB. Once the active disease starts, it would be preferable to let all mycobacterial cells come out of dormancy, for only the actively growing cells of *M. tuberculosis* are sensitive to most anti-TB drugs. We have recently reported the identification of two catalytically critical residues and a catalytically influential residue of *M. smegmatis* GTP-PEPCKs (Case et al, 2006). We have also studied the kinetic properties of human liver cytosolic GTP-PEPCK and found that it is inhibited by high levels of bicarbonate or CO_2 (Case and Mukhopadhyay, 2007). Generation of variants of the human enzyme and studies on their kinetic and biophysical properties are in progress (Case, Dharmarajan and Mukhopadhyay, unpublished data). Our results indicate that it is possible to influence activity of a PEPCK by interfering with a site that is away from the active site but connected to it via long-range interactions. Such sites can be exploited to develop therapeutic agents that do not interact directly with the active site. This mode has an advantage because an active site directed reagent might interact with other enzymes with similar active site characteristics and thereby have side effects. Also, these off sites might not be well conserved between the human and the mycobacterial enzymes and thereby will allow for the development of drugs that target mycobacterial PEPCK but do not interact with the human liver enzyme.

Mycobacterial physiology and genomics of clinical isolates of Mycobacterium tuberculosis

Coenzyme F_{420} metabolism of the mycobacteria

We have been investigating the role of F_{420} in basic biology and pathogenicity of the mycobacteria. F_{420} is a deazaflavin derivative. It is structurally similar to flavins but functionally acts as a hydride transfer coenzyme similar to the nicotinamides. F_{420} is present in all known methanogenic archaea, but it is rare in the bacterial domain, where it is primarily found in the actinobacteria such as mycobacteria. All mycobacteria examined thus far contain F_{420}. In 1996-1998 it was discovered that these bacteria express an F_{420}-specific glucose-6-phosphate

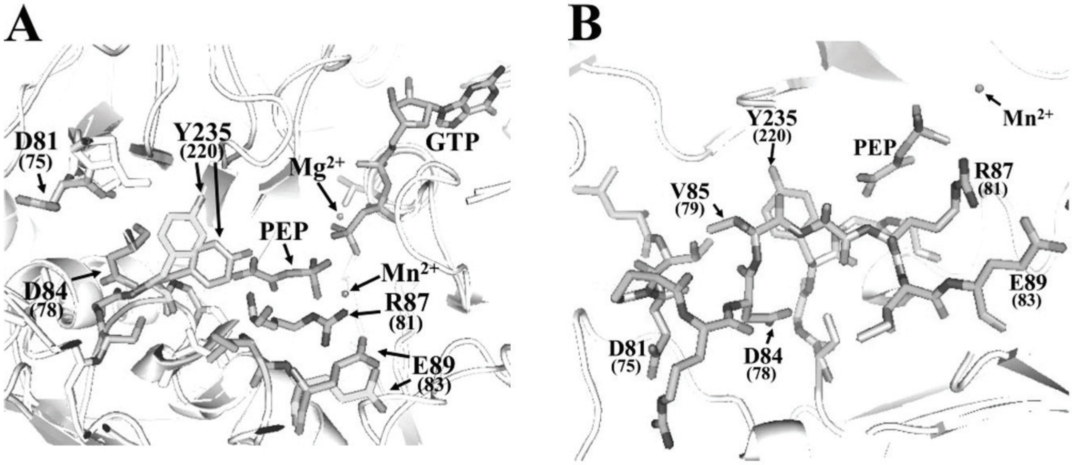


Figure. Phosphoenolpyruvate (PEP)-binding site of human cytosolic phosphoenolpyruvate carboxykinase (PEPCK). The numbers in parentheses correspond to the equivalent residues in *M. smegmatis* PEPCK. (A) Superposition of PEPCK-Mn^{2+}-GTP analog complex with and without bound PEP. Both forms have the same structure except for Glu89 and Tyr235, which undergo a shift because of PEP binding; the carbons at the shifted position are highlighted. (B) Sequence element being investigated (79LTDPRDVARIE89) and the adjacent element containing Tyr235 (233SGYGG237). The structure shown is for the PEP-bound form. The crystal structure of human liver cytosolic PEPCK was viewed and manipulated by using PyMOL version 0.97. The Protein Data Bank codes are 1KHF and 1KHB.

dehydrogenase (Fgd) (Purwantini and Daniels, 1998; Purwantini and Daniels, 1996; Purwantini et al, 1997). The use of Fgd-generated reduced F_{420} (H_2F_{420}) in the mycobacteria is unknown. Our ongoing investigation indicates that H_2F_{420} is utilized in the defense against oxidative damage, a role relevant to mycobacterial defense against the bactericidal effect of macrophages (Purwantini and Mukhopadhyay, 2007, under review). We have identified mycobacterial cell envelope synthesis as another point of use for F_{420} and H_2F_{420} (Purwantini, Stieber and Mukhopadhyay, 2007, unpublished results). It has been shown that both F_{420} and Fgd are required for the activation of PA824, a new anti-tuberculosis drug containing nitroimidazopyran nucleus (Stover et al, 2000). Since the activation of a drug is not a normal cellular activity, we are investigating which cellular machineries are exploited for converting this pro-drug to an active form. The goal of these investigations is to identify of new cellular targets for the development of new TB drugs. The emergence of MDR- and XDR-TB as a major threat to world population call for rapid development of these drugs. The last effective drug specifically developed for treating TB, was introduced in 1966 (Maggi et al, 1966; Willcox, 2000). Therefore, it is imperative that one looks into targets that are fundamental to the survival of a mycobacterial cell within the human host. Our works fits within this area.

Mechanism for the development of more virulent and MDR/XDR strains of *Mycobacterium tuberculosis*

This is a collaborative project involving Institut Teknologi Bandung and Rotinsulu Pulmonary Hospital, both in Bandung, Indonesia, and our laboratory. Our goal is to determine the genetic and biochemical basis for the development of more virulent and multi-drug or extensively drug-resistant strains of *M. tuberculosis*. This project is based on the hypothesis that population life styles (economic status, movement, environment) and treatment methodologies, which determine the immune system of a patient and chemical environment within the infected immune cells, promote changes in the genome of the pathogen which, in turn, leads to increased virulence and drug resistance genotypes. It is possible that the movement of the *IS6110* element in the chromosome, which could promote or block the expression of certain genes, increases virulence or imparts drug resistance phenotype. Our current investigation stems from the following observation from one of our projects that was funded by the Government of Indonesia and the Virginia Bioinformatics Institute in 2004-2006. From an *IS6110* RFLP-based study we found that almost all of the *M. tuberculosis* strains isolated from patients from one area of Bandung were unique. Some of these strains are in the MDR and XDR categories. The patients represented a rather impoverished population with a great amount of mobility in search of jobs. We are currently studying an area with a comparatively stable population to examine whether it also has the above-described variability. We are also investigating whether the strains isolated from a particular patient at various stages of disease and treatment are of different genotypes. Our immediate future work will involve the sequencing of the whole genomes of the collected strains, employing one of the cutting edge and low cost technologies, such as that offered by the Roche GS-FLX^{TM} system.

Biology of hydrothermal vent microbes and evolution of metabolism

The life in and around deep-sea hydrothermal vents is sustained by the thermal energy from the Earth's mantle. This energy allows the generation of high-energy chemical species such as hydrogen, sulfide and reduced iron. Autotrophic microorganisms that inhabit the vent environment utilize these compounds as sources of energy and fix CO_2 to generate cells and energy-rich compounds that nourish other organisms that are heterotrophs. The chemical and physical conditions of the vent to a great extent mimic the conditions that existed on early Earth. Therefore, a study of the vent autotrophs might shed light on the early earth-life processes. With this promise we work on the physiology and biochemistry of *Methanocaldococcus jannaschii,* a hyperthermophilic chemolithoautotrohic, strict hydrogenotrophic, methanogenogenic archaeon that lives in the deep-sea hydrothermal vents. Our techniques used have included bioinformatics, genomics, proteomics, and enzymology. We have discovered interesting regulatory systems and new enzymes in this organism (Johnson and Mukhopadhyay, 2005; Mukhopadhyay et al, 2000). Our current bioinformatics- and proteomics-based studies have provided an indication that the organism might harbor an ancient form of a membrane-based energy transduction system that is commonplace in many extant organisms (Case, Rodriguez and Mukhopadhyay, 2006-2007, unpublished data). A similar investigation by Robert E. Blankenship and coworkers at Washington University in St. Louis in collaboration with our laboratory has shown that the ancestors of two key proteins for nitrogen fixation (NifD and NifH) and photosynthesis (BchL and BchN/B) exist in *M. jannaschii*, which is neither a phototroph or nitrogen fixing organism (Staples et al, 2007). The roles of these ancestral proteins, called Nif-like proteins (NflD and NflH), in *M. jannaschii* are currently under investigation. We elaborate below how one of our earlier discoveries (Johnson and Mukhopadhyay, 2005) has led to the following current investigation.

Sulfite Reduction in Methanogenic Archaea - New Enzymes and New Hypotheses on Two Ancient Metabolisms

We are investigating whether at one time hydrogenotrophic methanogenesis and dissimilatory sulfate reduction, two of the oldest energy conserving respiratory systems on Earth, existed in one organism. If that was the case, the same organism might have carried out anaerobic methane oxidation, which is expected to be an ancient metabolism as well. This investigation follows our discovery of a new type of sulfite reductase (called Fsr) that allows *Methanocaldococcus jannaschii*, a hyperthermophilic methanogenogenic archaea, to utilize sulfite as sole sulfur source (Johnson and Mukhopadhyay, 2005). *M. jannaschii* is an inhabitant of the deep-sea hydrothermal vents (Jones et al, 1983). Fsr uses coenzyme F_{420} as an electron carrier, whereas all previously reported sulfite reductases use nicotinamides, cytochromes or membrane resident complexes as electron carriers (Johnson and Mukhopadhyay, 2005; Johnson and Mukhopadhyay, 2007). Fsr is a structural and functional chimera of two enzymes, H_2F_{420} dehydrogenase (FpoF or FqoF) and siroheme containing dissimilatory sulfite reductase (Dsr) (Johnson and Mukhopadhyay, 2005). FpoF/FqoF, an electron funneling unit of a membrane-bound energy transduction system, is present in certain late evolving methylotrophic methanogens and *Archaeoglobus fulgidus*, a sulfate reducing archaeon (Johnson and Mukhopadhyay, 2005). Dsr, composed of DsrA and DsrB subunits, are found in a group of anaerobic bacteria and certain archaea, such as *A. fulgidus,* that utilize sulfate as a terminal electron acceptor for the oxidation of complex materials such as carbohydrates and hydrocarbons, as well as simple compounds such as acetate and hydrogen (Widdel, 1988). The discovery of Fsr and use of sulfite by *M. jannaschii* was a surprise, because methanogens have been generally considered sensitive to sulfite. This oxyanion inhibits methanogenesis (Baldertson and Payne, 1976), which is the only source of energy for the methanogens (Wolfe, 1992). From these observations we conjectured that sulfite reductase was present in ancient methanogens and the enzyme helped these organisms to cope with the toxicity of sulfite that was probably generated from limited oxidation of sulfide during the early and gradual oxygenation of earth (Johnson and Mukhopadhyay, 2005). One problem with this hypothesis has been that the distribution of Fsr in methanogens is rather limited. In our paper in 2005 we reported that only three methanogens carried Fsr and all of these organisms are thermophiles (Johnson and Mukhopadhyay, 2005). Since then, by analyzing newly available genome sequences, we have found Fsr homologs in a psychrophilic methanogen and an uncultured archaeon, which is a member of a consortium performing sulfate reduction driven anaerobic methane oxidation (Johnson and Mukhopadhyay, 2007). Therefore, it is possible that more Fsr homologs will be found if investigations focus on extreme environments. However, another observation brought more certainty for sulfite metabolism existing in methanogens. We found that every methanogen for which a whole genome sequence is available carries a homolog of the C-terminal half of Fsr (or the DsrA domain of Fsr) (Johnson and Mukhopadhyay, 2007; Susanti and Mukhopadhyay, 2007, unpublished data). We call these ORFs small sulfite reductase homologs. Even the methanogens that possess Fsr homologs carry the small size sulfite reductase. In fact, a small size sulfite reductase has been purified from *Methanosarcina barkeri*, a late evolving methylotrophic mesophilic methanogen (Moura et al, 1986). In all, it seems that sulfite reduction is fundamental to the methanogens and this observation brings up the possibility of some of these organisms carrying either active or remnants of sulfate reduction pathway genes. The development of a complete sulfate reduction pathway must have followed the appearance of sulfite reductase. Sulfite is toxic to cells of all types and an obligatory intermediate in the reduction of sulfate. Therefore, sulfite reductase must be expressed in an organism that reduces sulfate. Our unpublished data show the presence of sulfate reduction genes in *M. jannaschii* and other methanogens (Dharmarajan and Mukhopadhyay, 2007, unpublished data), which make it more logical to look for an extant organism or signatures of an extinct organism with abilities to perform methanogenesis, sulfate reduction and anaerobic methane oxidation.

Enhancement of natural gas production: bioconversion of coal and paraffin to methane

According to a report from the Energy Information Administration, "Total natural gas consumption in the United States is projected to increase from 22.0 trillion cubic feet in 2005 to 26.1 trillion cubic feet in 2030" (Energy Information Administration, 2007). Within the same period the US natural gas production is projected to increase to only 21.2 trillion cubic feet. As it stands today, the only way to fill this gap between the projected consumption and production would be to import liquefied natural gas (LNG). Coalbed methane (CBM) accounts for about 9% of the total natural gas production in the USA. We are investigating the possibility of increasing CBM production by converting coal in the underground coalbeds to methane through microbial degradation. Research on a similar process for generating methane from paraffin that clogs natural gas wells and makes mining and transportation of petroleum cumbersome is also in progress in our laboratory. These investigations involve the development of coal or paraffin bioconversion enrichments by using natural gas well water as the source of microorganisms. Microbes from other sources and existing culture collections are also being investigated for bioconversion abilities. From these promising enrichments we have isolated microorganisms and characterized them at the physiological and phylogenetic levels and we are beginning genomics studies with these isolates. Studies on their metabolic abilities and their relations to the ultimate goal of converting coal and paraffin to methane are currently in progress.

Mitigating mine explosion: microbial conversion of methane at low oxygen levels

With goal of developing a microorganism-based technology for continuous removal of methane from sealed coalmines, we have been isolating microorganisms that can carry out methane oxidation at low methane and oxygen concentrations. The environments that we have been studying for finding microorganisms with such capabilities are plant surfaces, methane seeps and coalmines that emit methane. This search has led to the isolation of methane consuming bacteria that are currently under detailed characterization (Purwantini, Piedl and Mukhopadhyay, 2007, unpublished data). We have also isolated bacteria that may play roles in maintaining structure of soil and the bio-mineralization of complex carbohydrates (Stieber, Colton, Wang, Johnson and Mukhopadhyay, 2007, unpublished data). We are beginning genomics investigations on these organisms.

Conferences and workshops

Johnson EF, Mukhopadhyay B (2007) A novel coenzyme F_{420}-dependent sulfite reductase and a small size sulfite reductase in methanogenic archaea. In *Proceedings of the International Symposium on Microbial Sulfur Metabolism*, Dahl C and Friedrich CG (eds), Springer, New York, N.Y. (in press).

Mukhopadhyay B (2007) *Extreme Condition Microbiology*. The Virginia Tech Network Dynamics and Simulation Science Laboratory Complexity Science Seminar, Blacksburg, Virginia, March 2007.

Mukhopadhyay B (2006) *A novel coenzyme F_{420}-dependent sulfite reductase from* Methanocaldococcus jannaschii*: a detoxification system and a clue to an ancient metabolism*. International Symposium on Microbial Sulfur Metabolism (ISMSM), Münster, Germany, June 2006.

Mukhopadhyay B (2006) *A new type of sulfite reductase - a novel coenzyme F_{420}-dependent enzyme from the methanearchaeon* Methanocaldococcus jannaschii. Gordon Conference on Molecular Basis of Microbial One-carbon Metabolism, Magdalen College, Oxford, UK, August 2006.

Purwantini E, Stieber JP, Mukhopadhyay B (2007) *Involvement of Rv0406c homolog of* Mycobacterium smegmatis *in Mycolic Acid Synthesis*. 2007 General Meeting of the American Society for Microbiology, Toronto, Canada, May 2007.

Peer-reviewed publications in reporting period

Case CL, Concar EM, Boswell KL, Mukhopadhyay B (2006) Roles of Asp^{75},

Asp^{78}, and Glu^{83} of GTP-dependent phosphoenolpyruvate carboxykinase from *Mycobacterium smegmatis*. *J. Biol. Chem.* **281:** 39262-39272.

Lai H, Kraszewski JL, Purwantini E, Mukhopadhyay B (2006) Identification of the pyruvate carboxylase genes in *Pseudomonas aeruginosa* PA01 and development of a *P. aeruginosa*-based over-expression system for a_4- and a_4b_4-type pyruvate carboxylases. *Appl. Environ. Microbiol.* **72:** 7785-7792.

Seleem MN, Ali M, Boyle SM, Mukhopadhyay B, Witonsky SG, Schurig GG, Sriranganathan N (2006) Establishment of gene expression system in *Ochrobactrum anthropi*. *Appl. Environ. Microbiol.* **72:** 6833-6836.

References

Balderston WL, Payne WJ (1976) Inhibition of methanogenesis in salt marsh sediments and whole-cell suspensions of methanogenic bacteria by nitrogen oxides. *Appl. Environ. Microbiol.* **32:** 264-9.

Case CL, Concar EM, Boswell KL, Mukhopadhyay B (2006) Roles of Asp^{75}, Asp^{78}, and Glu^{83} of GTP-dependent phosphoenolpyruvate carboxykinase from *Mycobacterium smegmatis*. *J. Biol. Chem.* **281:** 39262-72.

Case CL, Mukhopadhyay B (2007) Kinetic characterization of recombinant human cytosolic phosphoenolpyruvate carboxykinase with and without a His_{10-}tag. *Biochimica et Biophysica Acta* In press.

Dunten P, Belunis C, Crowther R, Hollfelder K, Kammlott U, Levin W, Michel H, Ramsey GB, Swain A, Weber D, Wertheimer SJ (2002) Crystal structure of human cytosolic phosphoenolpyruvate carboxykinase reveals a new GTP-binding site.
J. Mol. Biol. **316**:257-64.

Johnson EF, Mukhopadhyay B (2005) A new type of sulfite reductase, a novel coenzyme F_{420}-dependent enzyme, from the methanarchaeon *Methanocaldococcus jannaschii*. *J. Biol. Chem.* **280:** 38776-86.

Johnson EF, Mukhopadhyay B (2007) A novel coenzyme F_{420}-dependent sulfite reductase and a small size sulfite reductase in methanogenic archaea. In *Proceedings of the International Symposium on Microbial Sulfur Metabolism*, Dahl C and Friedrich CG (eds), Springer, New York, N.Y. In press.

Jones WJ, Leigh JA, Mayer F, Woese CR, Wolfe RS (1983) *Methanococcus jannaschii* sp. nov., an extreme thermophilic methanogen from a submarine hydrothermal vent. *Archiv. Microbiol.* **136:** 254-261.

Maggi N, Pasqualucci CR, Ballotta R, Sensi P (1966) Rifampicin: a new orally active rifamycin. *Chemotherapy* **11:** 285-92.

Moura I, Lino AR, Moura JJ, Xavier AV, Fauque G, Peck Jr. HD, LeGall J (1986) Low-spin sulfite reductases: a new homologous group of non-heme iron-siroheme proteins in anaerobic bacteria. *Biochem. Biophys. Res. Commun.* **141:** 1032-41.

Mukhopadhyay B, Concar EM, Wolfe RS (2001) A GTP-dependent vertebrate-type phosphoenolpyruvate carboxykinase from *Mycobacterium smegmatis*. *J. Biol. Chem.* **276:** 16137-45.

Mukhopadhyay B, Johnson EF, Wolfe RS (2000) A novel pH2 control on the expression of flagella in the hyperthermophilic strictly hydrogenotrophic methanarchaeaon *Methanococcus jannaschii*. *Proc. Natl. Acad. Sci. USA* **97:** 11522-7.

Purwantini E, Daniels L (1998) Molecular analysis of the gene encoding F_{420}-dependent glucose-6-phosphate dehydrogenase from *Mycobacterium smegmatis*. *J. Bacteriol.* **180:** 2212-9.

Purwantini E, Daniels L (1996) Purification of a novel coenzyme F420-dependent glucose-6-phosphate dehydrogenase from *Mycobacterium smegmatis*. *J. Bacteriol.* **178:** 2861-6.

Purwantini E, Gillis TP, Daniels L (1997) Presence of F_{420}-dependent glucose-6-phosphate dehydrogenase in *Mycobacterium* and *Nocardia* species, but absence from *Streptomyces* and *Corynebacterium* species and methanogenic Archaea. *FEMS Microbiol. Lett.* **146:** 129-34.

Staples CR, Lahiri S, Raymond J, Von Herbulis L, Mukhophadhyay B, Blankenship RE (2007) The expression and association of group IV nitrogenase NifD And NifH homologs in the non-nitrogen fixing archaeon *Methanocaldococcus jannaschii*. *J. Bacteriol*. In press.

Stover CK, Warrener P, VanDevanter DR, Sherman DR, Arain TM, Langhorne MH, Anderson SW, Towell JA, Yuan Y, McMurray DN, Kreiswirth BN, Barry CE, Baker WR (2000) A small-molecule nitroimidazopyran drug candidate for the treatment of tuberculosis. *Nature* **405:** 962-6.

Widdel F (1988) Microbiology and ecology of sulfate- and sulfur-reducing bacteria, p. 469-585. In Biology of anaerobic microorganisms, Zehnder A (ed), New York: John Wiley and Sons.

Willcox PA (2000) Drug-resistant tuberculosis. *Curr. Opin. Pulm. Med.* **6:** 198-202.

Jean Peccoud

peccoud@vt.edu

Associate Professor, Virginia Bioinformatics Institute

Adjunct Associate Professor, Dept. of Biological Sciences, Virginia Tech

Group contributors: David Ball, Bill Baumann, Yizhi Cai, Brian Hartnett, Rebecca Shelton.

Synthetic Biology: Design Of Artificial Gene Networks

Abstract. The Synthetic Biology Group at VBI streamlines the design and fabrication of artificial gene networks. Computer-assisted design of genetic systems could bring significant benefits to the biomedical community and the biotechnology industry. However, the lack of calibrated genetic parts remains a major limitation. It is extremely desirable to associate a quantitative effect with the use of specific DNA sequences in an expression cassette. Biologists have long recognized that some promoters are stronger than others. This indicates that it is reasonable to attempt to associate a number to a promoter sequence and to generalize the approach to any genetic part. We are aggressively developing methods to calibrate genetic parts by specifying calibration protocols and highly customized nonlinear optimization algorithms capable of analyzing the data they generate. We mostly rely on flow cytometry data to characterize genetic parts but a high-throughput live cell imaging system would give additional useful information. Since no commercial instrument meets the requirements of this approach, we have initiated the development of a customized microscope. Now that commercial contractors can assemble long DNA molecules readily, assisting potential users of gene synthesis services in the design of their constructs is important. We are therefore developing software tools that can guide users through the design of new DNA molecules or help them verify previously designed molecules.

Keywords: genetic parts; artificial gene networks; computer-assisted design; live cell imaging; aptamers; parameter estimation; stochastic modeling; nonlinear optimization.

Scientific Progress

In 2000, the description of Gardner's bistable toggle switch and Elowitz's repressilator is often regarded as the founding act of synthetic biology. Since then, many artificial gene networks have been reported. These studies demonstrate that a model-driven approach to the design of genetic constructs can lead to sophisticated phenotypes that could be useful in biomedicine. However, seven years after the first publications, this technology has yet to find its first application. The field seems to face challenges in moving beyond the proof-of-concept stage. Our hypothesis is that there is a significant "entry barrier" for the technology preventing potential users of the technology from accessing it. Researchers who could use a bistable genetic switch to control the expression of a transgene do not have the expertise or the resources to develop the construct. Even scientists who have demonstrated the expertise to develop such constructs seem to experience difficulties in customizing them for different organisms or applications. Our group develops experimental protocols, algorithms and mathematical models, software tools, and instruments that will enable people who could use synthetic genetic systems in basic biological research, pre-clinical applications, or industrial applications to gain access to this technology.

The design of genetic constructs

Today, gene synthesis technology enables molecular biologists to assemble long DNA molecules that may include multiple genes and their regulatory sequences. We refer to these molecules as "genetic constructs" or just "constructs". As the throughput of construct manufacturing increases, the design of complex genetic constructs becomes the bottleneck of the process. It becomes easier to assemble complex DNA molecules than to design them. A natural way of designing complex constructs consists of combining basic building blocks also known as "biological parts" or "genetic parts". These parts are small DNA fragments implementing specific biological functions. The mechanisms of gene expression require that certain structural constraints be met in order for a construct to be functional. Parts of different types need to be placed in a particular order and next to each other in order to ensure that coding sequences are properly transcribed and translated. Certain parts are functional in a specific context whereas other parts have proved functional in different organisms than the one from which they originate. For instance promoters are often functional in specific organisms or even cell types whereas genes coding for proteins can often be expressed in multiple species. The design of complex genetic constructs such as artificial gene networks therefore requires an intimate knowledge of gene expression mechanisms.

A

Step	Production	String
1	P01	**S**S
2	P03	HNB**S**
	P02	HNBA**M**G
3	P06	H**N**BAMMG
4	P10	HFDBA**M**MG
	P08	HFDBACE**M**G
	P08	**H**FDBACECEG
5	P15	HHFDBACECE**G**
	P14	HHFDB**A**CECEGG
6	P21	HHFDBOACECEGG
7		h02h01f01d04b02o01a08c08e15c05e14g01g02
		tataa..gcgtttata

B

```
[taataaacgcagaaaggcccacccgaaggtgagccagtgtga] [gagagcgttcaccgacaaacaacagataaaacgaaaggccc
agtctttcgactgagcctttcgttttatttgatgcctgg] [ttaagc...cacca] [catcgaaccggtttcct] [tcctttgcat
accctgctgatgtgctcattataaccgccagtggtatttatgtcaacaccgccagagataatttatcaccgcagatggttatctgt
gcatgc] [ttatcaaaaaccatggtttttgataa] [ccatcgaatggctgaaatgagctgttgacaattaatcatccggctcgtat
aatgtgtggaattgtgagcggataacaatttcacacagga] [aggaaaccggttatg] [atgagca...ttacaa] [aggaattta
aatg] [atgcgt...aaataa] [ccaggcatcaaataaaacgaaaggctcagtcgaaagactgggcctttcgttttatctgttgtt
tgtcggtgaacgctctc] [tcacactggctcaccttcgggtgggcctttctgcgtttata]
```

C

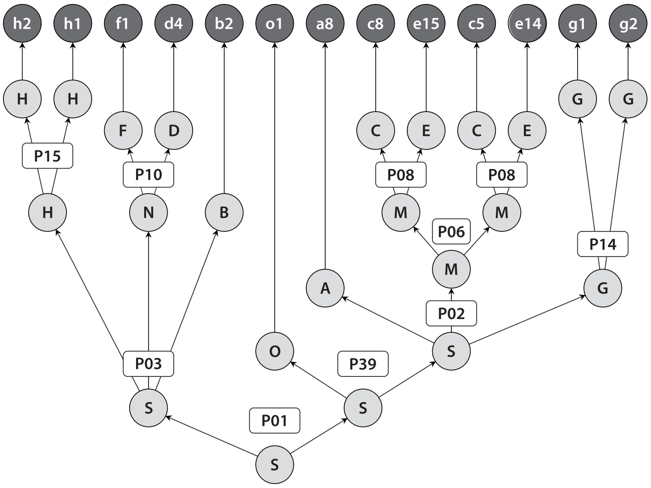


Figure. Design and verification of genetic constructs. (A) The successive applications of productions starting from S provide a framework to guide the design of genetic constructs. (B) The verification of an existing DNA sequence requires the use of a lexical analyzer to identify the parts from which the sequence is composed. (C) The symbolic description of the sequence provided by the lexical analyzer can be parsed using an LR algorithm.

Context-free grammars

We have shown that context-free grammars (CFG), a formalism used in computer science to specify programming languages, can formalize the design principles of genetic constructs. This approach provides a path to organizing libraries of genetic parts according to their biological functions that correspond to the syntactic categories of the CFG. It also provides a framework to the systematic design of new genetic constructs consistent with the design principles expressed in the CFG. This framework can be used to build software wizards that guide users in the design of their constructs by offering them design options consistent with structural constraints in the grammar and the availability of genetic parts. Using parsing algorithms, this syntactic model also enables the verification of existing constructs. We have illustrated these possibilities by describing a CFG that generates the most common architectures of genetic constructs in *Escherichia coli*. We have compiled a library containing close to 100 genetic parts according to the syntactic categories of this CFG. The architecture of previously published constructs was represented using the library part identifiers and verified by LR(0) parsing. The lexical analysis of the DNA sequences of 30 previously published artificial gene networks revealed a large number of discrepancies between the network theoretical sequences and their actual sequences. This uncontrolled genetic variability can explain discrepancies between the observed and predicted phenotypes of the network. It also illustrates the need for combining computer-assisted design with gene synthesis to capture the potential benefits of standardizing genetic parts.

Conferences and workshops

Peccoud J (2006) *Dynamic partitioning for hybrid simulation*. Systems Biology Markup Language Forum, The Eleventh Workshop on Software Platforms for Systems Biology, Tokyo, Japan, 12-13 October, 2006.

Intellectual property

Peccoud J, Cai Y (03/30/07) Software for design and verification of synthetic genetic constructs, US provisional.

Peer-reviewed publications in reporting period

Griffith M, Courtney T, Peccoud J, Sanders WH (2006) Dynamic partitioning for hybrid simulation of the bistable HIV-1 transactivation network. *Bioinformatics* **22**: 2782-2789.

Andy Pereira

pereiraa@vbi.vt.edu

Professor, Virginia Bioinformatics Institute

Group contributors: Utlwang Batlang, Amal Harb, Arjun Krishnan, Madan Reddy, Peter Wittich.

Plant-Environment Interaction Systems Biology

Abstract. The interaction and adaptation of plants to environmental signals and stresses are complex and need to be analyzed in a network model using a systems biology approach. We have initiated studies using transcription factor genes that we have identified from *Arabidopsis* and which provide abiotic (drought and salt) stress resistance when expressed in *Arabidopsis* and rice. Our group has analyzed the downstream-regulated network of genes that probably have a cellular role in abiotic stress resistance. When compared to the altered transcriptome in engineered stress-resistant genotypes the drought transcriptome reveals distinct subsets of genes represented in Gene Ontology (GO) functional categories that overlap and which are unique to specific drought-resistance mechanisms. By transforming the stress resistance genes into *Arabidopsis* and rice, we can compare the downstream transcriptional network and the associated *cis*-elements to develop a network model conserved between plant species. This network model can be linked to other gene/protein interaction information that is being generated. These studies will help identify the cellular network that determines a phenotype across plant species.

Keywords: abiotic stress; *Arabidopsis*; rice; switchgrass; water use efficiency; drought; microarray; transcription factors.

Scientific Progress

Stress systems biology

Adaptation and resistance to dehydration/drought have evolved in land plants through some common conserved mechanisms that can be addressed in the sequenced model plants *Arabidopsis* and rice. In response to drought stress, which is the decreased availability of water, there is an induction of genes involved in cell protection and in the regulation of drought responses.

We have identified the transcriptional factor genes *SHN* and *HRD* that confer drought resistance when overexpressed in Arabidopsis and rice by using a gain-of-function strategy in *Arabidopsis*. We compared three AP2/ERF transcription factor overexpressing *Arabidopsis* lines (*SHN*, *HRD*, *DREB1A*) that confer drought resistance for physiological parameters and transcriptome changes using oligonucleotide microarrays. In addition, wild-type *Arabidopsis* plants subjected to 'progressive drought' conditions, which simulates gradual field drought stresses, were analyzed for transcriptome changes and the results compared to published *Arabidopsis* seedling dehydration studies (AtGenExpress).

The drought resistance genotypes reveal a network of downstream-regulated genes or 'regulons' classified under specific Gene Ontology (GO) categories. A comparison of the three drought resistance gene regulons reveal a number of shared genes as well as some specific to each drought resistance gene. The drought transcriptome, namely the genes differentially regulated by progressive drought stress, shows significant overlap with subsets of the drought resistance regulons. We propose that the shared subset of the drought transcriptome and the drought resistance regulon are responsible for the drought resistance phenotype.

Comparative systems biology

Rice orthologs of the drought resistance genes are being validated by overexpression in *Arabidopsis* and rice. The orthologs will be used in microarray experiments to reveal their respective drought resistance regulons. The drought resistance genes will be transformed into switchgrass, a promising bioenergy crop, to examine the drought resistance phenotype as well as uncover the drought resistance regulons by transcriptome profiling on a Roche GS-FLX^{TM} genome sequencer.

Quantitative analysis of the phenotype, metabolome and transcriptome in drought resistant genotypes of diverse plant species will reveal conserved downstream pathways leading to drought stress resistance.

	8-day drought	6-h dehydration	*DREB1A*	*SHN*	*HRD*
Response to water deprivation	■	■	■		
Response to abscisic acid	■	■	■		
Response to reactive oxygen species (ROS)	■				
Response to salinity					■
Response to temperature	■	■	■		■
Response to biotic stimulus		■			
Protein biosynthesis	▨	▨			■
Ribosome biogenesis	▨	▨			
Photosynthesis	▨		▨		
Regulation of transcription				■	▨
DNA packaging	▨	▨			■
Cell wall formation		▨			
Carbohydrate metabolism			■	▨	
Lipid metabolism				■	▨
Sulfur metabolism				▨	
Hormone metabolism				▨	▨
Alcohol metabolism				▨	
Nucleotide metabolism					■
Amino acid metabolism		■			
Glycosinolate metabolism		■	▨	▨	▨
Transport			▨	■	■

Representation of different processes regulated during progressive drought (8-day drought), acute dehydration (6 h), and in overexpression lines *DREB1A*, *SHN* and *HRD*. The analysis suggests that the drought resistance conferred by overexpression of these transcription factors comes from partially independent mechanisms and pathways. Black: induced; Grey: repressed. The genes in each category have ≥1.5-fold change, and the categories were chosen based on statistical enrichment of Gene Ontology (GO) terms.

Functional genomics tools

We are developing activation tag strategies to conduct functional genomics in a number of plants after genome sequencing. In *Arabidopsis*, an *En-I* transposon-based activation tagging population is being generated in ecotype Col. This approach is revealing novel morphological mutants that are being catalogued. These mutant lines will be screened for abiotic stress parameters using high-throughput systems that are under development. The mutant lines will be available for collaborative research during the process of their generation. Additional activation tag populations are being initiated in tomato, rice and Brassica in collaborative work.

Conferences and workshops

Karaba A, Dixit S, Trijatmiko KR, Aharoni A, Nataraja KN, Udayakumar M, Pereira A (2006) Improvement of water use efficiency together with drought resistance in rice. *4th International Rice Functional Genomics Symposium*, October 9-11, 2006, Montpellier, France.

Karaba A, Dixit S, Trijatmiko KR, Aharoni A, Nataraja KN, Udayakumar M, Pereira A (2006) Improvement of water-use efficiency in crop plants. *5th Plant Genomics European Meetings*, October 11-14, 2006, Venice, Italy.

Peer-reviewed publications in reporting period

Marsch-Martinez N, Greco R, Becker JD, Dixit S, Bergervoet JHW, Karaba A, de Folter S, Pereira A (2006) BOLITA, an Arabidopsis AP2/ERF-like transcription factor that affects cell expansion and proliferation/differentiation pathways. *Plant Mol. Biol.* **62**: 825-843.

Salentijn EMJ, Pereira A, Angenent GC, van der Linden CG, Krens F, Smulders MJM, Vosman B (2007) Plant translational genomics: from model species to crops. *Mol Breeding* **20**: 1-13.

Dharmendar Rathore

rathore@vbi.vt.edu

Assistant Professor, Virginia Bioinformatics Institute

Group contributors: Rana Nagarkatti, Dewal Jani.

Malaria Pathogenesis and Drug Discovery

Abstract. The pathogenesis of malaria infection has its roots in the rapid degradation of host cell hemoglobin by the parasite. This process releases heme, which is toxic to the parasite, and a unique, albeit poorly understood parasite-specific mechanism converts it into a non-toxic and insoluble material termed hemozoin. We have recently identified a novel *Plasmodium* protein that can produce hemozoin. Based on its function, we have named this protein Heme Detoxification Protein. In the reporting period, we have shown that Heme Detoxification Protein is critical for the survival of the parasite, is highly conserved and follows an unusual trafficking route to the food vacuole. To date, this trafficking route has never been observed for any Plasmodial protein. Furthermore, we performed a high throughput screening of compound libraries for the identification of novel chemical entities that can inhibit Heme Detoxification Protein–heme interactions. This screening has led to the identification of 12 novel chemical compounds with potent antimalarial activity on *Plasmodium falciparum* parasites in culture.

Keywords: malaria, drug discovery, heme detoxification, hemozoin, protein trafficking, high throughput screening, target identification, Lipinski rules.

Scientific Progress

Malaria drug discovery - Developing HDP as an antimalarial drug target

As Heme Detoxification Protein (HDP) interacts with heme (for converting it into hemozoin), we optimized a spectrophotometric assay for high throughput screening where HDP–heme interactions were measured both in the absence (control) or presence of chemical compounds. In this protein-based assay, we screened approximately 80 000 novel chemical entities for their potential to inhibit HDP–heme interactions. After a medicinal chemistry triage to remove potentially toxic molecules, 330 compounds showing >40% inhibition of HDP–heme interactions were evaluated for their anti-parasitic activity on *Plasmodium falciparum* parasites in culture. This screen identified 12 compounds with potent antimalarial activity with median inhibitory concentration (IC_{50}) of <1 mM. All 12 compounds were found to follow Lipinski's rule of 5.

These compounds were evaluated on both chloroquine-sensitive and a multidrug-resistant K1 strain of *P. falciparum* parasites where, unlike chloroquine, these compounds showed comparable activity on both chloroquine-sensitive and multidrug-resistant strain of the parasites. Some of the compounds were found to have activities in the low nanomolar range (15–50 nM), making them as active as chloroquine although they have not undergone any lead optimization. Subsequently, we have shown that these compounds could inhibit hemozoin production *in vitro*. This suggests that their anti-parasitic activity is mediated by their potential to inhibit heme detoxification.

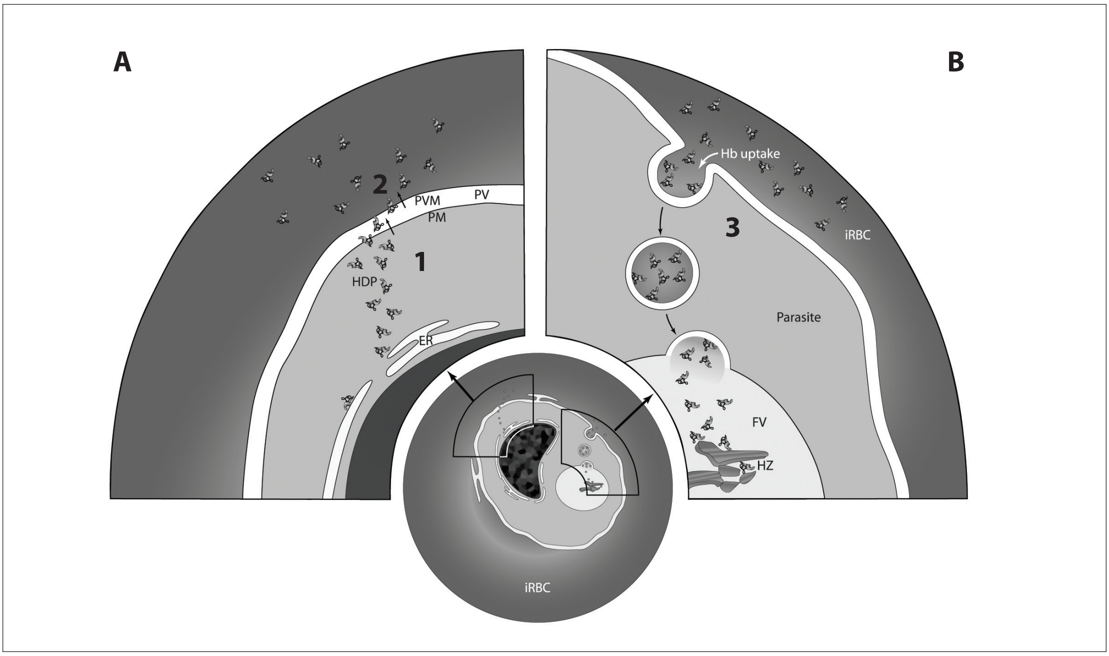
Next, we investigated 8 out of 12 compounds for their antimalarial activity in an *in vivo* rodent (*Plasmodium berghei*) malaria model and found that a single oral dose of the compound FP1 can cause up to 40% decrease in the levels of parasitemia in infected animals. Though FP1 had moderate antimalarial activity *in vitro* it was most effective under *in vivo* conditions. This suggests that validated hits would likely require structural optimization to improve their bioavailability. The animal studies were performed at the Swiss Tropical Institute with the financial support from the Medicines for Malaria Venture (MMV). Thus, our results suggest that HDP has the potential of being an antimalarial drug target and the validated hits could be developed into drug candidates with potent antimalarial activities.

Intracellular trafficking of HDP in malaria

Inside an infected cell, >75% of the hemoglobin is degraded and most of the heme released is converted into hemozoin. Although we have previously shown that both native and recombinant HDP can produce hemozoin *in vitro*, to be biologically relevant, HDP should be present in the food vacuole, the site of hemozoin production. We discovered that HDP could indeed be found in the food vacuole, in close vicinity of the hemozoin crystals. This observation strongly supports our assertion that HDP is responsible for hemozoin synthesis in the parasite *in vivo*. However, the protein lacks an N-terminal signal sequence or a transmembrane anchor by which it could be transported to the food vacuole.

We hypothesized that HDP has an unconventional trafficking mechanism whose discovery will be important for understanding pathogenesis as well as for developing HDP as a therapeutic target. Subsequently, by cryo-immunoelectron microscopy we showed that immediately after merozoite invasion, the ring stage parasite expresses and secretes HDP to the cytosol of infected red blood cells before any detectable amount of hemozoin can be found inside the parasite.

Export of HDP to the host cell cytosol and its presence in the food vacuole during the trophozoite stage led us to investigate the trafficking link between the two distinct intracellular sites within a parasite-infected red blood cell. In the malaria parasite, a continuous



HDP trafficking is circuitous and unique. (A) HDP crosses the plasma membrane of the parasite (1). Once in the parasitophorous vacuole, an unknown targeting signal in HDP leads to its trafficking across the parasitophorous membrane (2). (B) Finally, HDP, along with host hemoglobin, is endocytosed, transported and delivered to the food vacuole, the site of hemozoin formation. ER, endoplasmic reticulum; FV, food vacuole; Hb, hemoglobin; HDP, Heme Detoxification Protein; HZ, hemozoin; iRBC, infected red blood cell; PM, parasitophorous membrane; PV, parasitophorous vacuole; PVM, parasitophorous vacuole membrane.

outbound trafficking to the parasitophorous membrane of the host red blood cell is equally matched with a concomitant inbound trafficking of host hemoglobin. Studies have shown that hemoglobin intake starts with a cytostomal-mediated pinocytic uptake that subsequently pinches off creating a transport vesicle. The vesicle undertakes a retrograde trip and fuses with the food vacuole to deliver hemoglobin. Our experiments revealed that HDP exploits the hemoglobin transport machinery to access the food vacuole. We detected HDP uptake along with the endocytic import of host hemoglobin, its presence in the transport vesicles and, finally, its delivery to the food vacuole (See figure).

Our results are the first demonstration that the malaria parasite is capable of importing its own protein after secreting it into the host cytosol. This pathway could indeed be in use for other food vacuole-bound parasite proteins. We are currently working on the identification of unique targeting signals encoded in HDP that lead to its secretion and subsequent endocytic uptake by the parasite.

Conferences and workshops

Jani D, Nagarkatti R, Beatty W, Rathore D (2006) A highly conserved *Plasmodium* protein is responsible for hemozoin formation in the malaria parasite. *XVII Molecular Parasitology Meeting*, Woods Hole, MA.

Jani D, Nagarkatti R, et al (2006) Identification and characterization of a novel *Plasmodium* protein responsible for hemozoin formation—Implications for antimalarial drug development. *American Society of Tropical Medicine & Hygiene (ASTMH) Annual Meeting*. 12-16 November 2006, Atlanta, GA, USA.

Nagarkatti R, Jani D, et al (2007) Targeting the Newly Characterized *Plasmodium* Protein HDP that Synthesizes Hemozoin; Identification of Novel Inhibitors by High Throughput Screening. *Keystone Symposium – Drug Against Protozoal Parasites.* January 28-February 1 2007, Tahoe City, California.

Intellectual Property

Rathore D, Jani D, Nagarkatti N (2007) A novel therapeutic target for protozoal diseases, US Utility Non-provisional.

Peer-reviewed publications in reporting period

Rathore D (2007) American Society of Tropical Medicine and Hygiene – 55th Annual Meeting. 12–16 November 2006, Atlanta, GA, USA. *IDrugs* **10**(2): 93-95.

David Samuels

dsamuels@vbi.vt.edu

Assistant Professor, Virginia Bioinformatics Institute

Group contributors: Jonghoon Kang, Harsha Rajasimha, Zhou Song, Passorn Wonnapinij.

Modeling Mitochondrial Biology: Partnering With Clinical Researchers

Abstract. The focus of our research group is to work closely with clinical and wet-lab researchers to apply computational techniques to biomedical research involving mitochondrial dysfunction. Our work published in the past year has concentrated on two areas: the toxicity mechanisms of antiviral drugs and the dynamics of pathogenic mitochondrial DNA (mtDNA) mutations. In modeling the toxicity mechanisms of antiviral drugs such as AZT (azidothymidine, zidovudine), we have been developing new computational models of the metabolism of these drugs. We have developed a detailed computational model of the mitochondrial DNA polymerase and its interaction with the activated antiviral drugs. This DNA polymerase model may be generalized in the future to build a model of viral DNA polymerases (reverse transcriptases in most cases). Over the past year, our modeling of pathogenic mtDNA mutations has focused on the analysis of new clinical data that have been derived from experiments designed to test our mathematical models. The predictions of the models have been confirmed in the new data, and new unexpected behavior has also been found in these experiments. We are now in the process of developing new hypotheses based on these new clinical data, which we will incorporate into the next generation of models.

Keywords: azidothymidine (AZT); antiviral drugs; nucleoside analogs; nucleotide metabolism; mitochondria; DNA; muscle; hematopoietic stem cells.

Scientific Progress

The bulk of our research effort falls under the domain of "cellular networks", including biochemistry models and more general models of cell biology. Our biochemical modeling work in this period has focused on the biochemistry of the nucleoside analog drug azidothymidine (AZT), which is used to treat Human Immunodeficiency Virus/ Acquired Immunodeficiency Syndrome (HIV/AIDS). This prodrug is metabolized in the cell to its active form, AZT triphosphate, which then interferes with the production of viral DNA. However, the prodrug is also metabolized within mitochondria, where the active drug metabolites have a toxic effect. Several different mechanisms have been suggested for this toxicity and we are modeling this metabolism to assess the affects of the major candidates for the toxicity mechanism. We are taking two approaches to this problem. In one approach, we have developed a biochemical model of the metabolic pathways that phosphorylate both AZT and the four natural nucleosides (see Figure). We have developed both a deterministic differential equation model (using Mathematica) and a discrete stochastic model (using Copasi). We have been surprised to find no significant differences in the results from these two different models, even in cases where the copy numbers of some important chemical species decreased to less than 100 molecules. In a separate approach, in collaboration with Professor Cao of the Computer Science Department at Virginia Tech, we have used the Stochastic Simulation Algorithm (SSA) to develop a detailed model of the mitochondrial DNA polymerase (pol-γ) and its interaction with these drugs. In this research, the stochastic mutation mechanisms that we need to reproduce in the model require an SSA approach to the simulation. The most common hypothetical mechanism for nucleoside analog toxicity is focused on the effect of these drugs metabolites on pol-γ activity.

A related area of research for our group is the modeling of the long-time-scale dynamics of the population of mitochondrial DNA molecules (mtDNA) in single cells. Our focus in this area over the past year has been on applying the results of our simulations to the analysis of new clinical data. The experiments reported in Durham et al. (2006) were designed to test hypotheses from our modeling of the acquired mtDNA mutations that form as a part of the natural aging process. Based on those simulations, which were published in a series of papers from 1999 to 2003, we had made the hypothesis that clonal expansions of acquired neutral mtDNA mutations would be found in cell samples from elderly subjects. The prevailing view was that only clonal expansions of pathogenic mtDNA mutations would be found. In Durham et al. (2006), we analyzed skeletal muscle tissue from an 85-year-old subject and found clonal expansions of acquired neutral mtDNA mutations in 15% of the muscle fibers, which confirms the hypothesis based on our simulations.

The clinical experiments reported in Pyle et al. (2007) were designed specifically to test the results of a simulation of hematopoietic stem cells developed by Genetics, Bioinformatics, and Computational Biology (GBCB) graduate student Harsha Rajasimha, in our group. The simulation predicted a slow exponential decline in pathogenic mtDNA mutation in blood samples, a potential explanation for why blood samples have proven to be of little use in the diagnosis and monitoring of the progression of mitochondrial diseases, forcing clinicians to rely on invasive and painful muscle biopsies. Testing the hypothesis from the simulation required the analysis of blood samples taken 5 to 20 years apart. The Pyle et al. (2007) paper reports the experimental findings and focuses on an unexpected result, which is still not explained. The prediction of the blood stem cell simulations was confirmed in these new clinical data, and a manuscript describing the simulation and analyzing a broad range of clinical data in light of the simulation has now been submitted.

In these publications, and another that is currently in press, the predictions from the modeling were confirmed in the experiment. However other behaviors were also found in the experiments that were not expected based on the modeling. We are now exploring

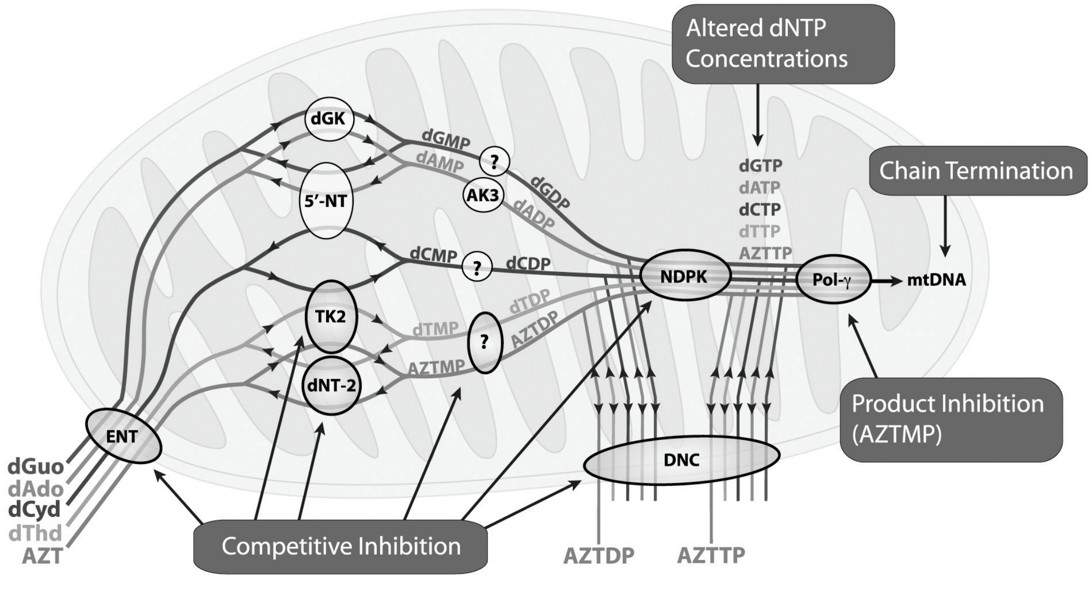


Figure. Metabolism of the antiviral drug AZT and the four natural nucleosides in mitochondria. Four possible mechanisms of the toxicity of AZT are included. Adapted from Samuels (2006).

hypotheses based on these new results, and are incorporating these hypotheses into our next generation of models. These papers illustrate our commitment to "turning the wheel" between experiment and theory. This process of building theoretical models based on clinical data, designing and carrying out our new experiments based on the model results, and then using this new information for further modeling, requires long term and committed collaborations.

Conferences and workshops

Samuels D (2006) Simulations of mtDNA precursor metabolism in mitochondria, Department of Pathology, Emory University School of Medicine, Atlanta, May 2006.

Samuels D (2007) DNA precursor metabolism in mitochondria: Simulations of drug toxicity and genetic diseases, Greater Philadelphia Bioinformatics Alliance, Thomas Jefferson University, February 2007.

Peer-reviewed publications in reporting period

Samuels DC, Carothers AD, Horton R, Chinnery PF (2006) The power to detect disease associations with mitochondrial DNA haplogroups. *The American Journal of Human Genetics* **78**(4): 713-720.

Durham SE, Samuels DC, Chinnery PF (2006) Is selection required for the accumulation of somatic mitochondrial DNA mutations in post-mitotic cells? *Neuromuscular Disorders* **16**(6): 381-386.

Samuels DC (2006) Mitochondrial AZT metabolism. *IUBMB Life* **58**(7): 403-408.

Samuels DC (2007) Computational models of antiviral toxicity. *Current Opinion in Drug Discovery & Development* **10**(1): 43-48.

Pyle A, Taylor RW, Durham SE, Deschauer M, Schaefer AM, Samuels DC, Chinnery PF (2007) Depletion of mitochondrial DNA in leucocytes harbouring the 3243A -> G mtDNA mutation. *Journal of Medical Genetics* **44**(1): 69-74.

João Setubal

setubal@vbi.vt.edu

Associate Professor, Deputy Director, Virginia Bioinformatics Institute

Associate Professor, Dept. of Computer Science, Virginia Tech

Group contributors: Chris Lasher, Jian Sun, Tsai-Tien Tsen, Andrew Warren, and PATRIC project team members.

Bacterial Genomics and Bioinformatics

Abstract. The Setubal research group works primarily on bioinformatics for bacterial genome annotation and sequence analysis. New bacterial genomes continue to become available at an exponential rate thanks to new sequencing technologies. Comparative genomics is one of the main beneficiaries of the surge in sequencing since it has become cheap enough to sequence several strains of the same species as well as species from phylogenetically under-represented groups. This presents exciting opportunities for our research group. In addition to specific genome analyses (which currently cover the genera *Agrobacterium, Brucella, Rickettsia* [α-proteobacteria], *Azotobacter, Coxiella, Pseudomonas,* and *Xanthomonas* [γ-proteobacteria]), current topics of interest include automated genome annotation, web-based infrastructure for genome annotation and analysis, and exploitation of ortholog groupings for comparative studies. During this reporting period significant advances were achieved for the PATRIC (PathoSystems Resource Integration Center) project, exemplified by the first published description of PATRIC in the 2007 database issue of *Nucleic Acids Research* and the release and publication of PATRIC's first genomics analysis tool, GenVar. In other projects, progress was also made, notably in the generation of hundreds of new Gene Ontology (GO) terms for plant-associated microbes in the Plant-Associated Microbe Gene Ontology (PAMGO) project.

Keywords: bacterial genomics; bioinformatics; *Agrobacterium*; *Azotobacter*; *Pseudomonas*; *Xanthomonas*; genome annotation; biodefense.

Scientific Progress

The description that follows gives brief updates on the various projects underway in the Setubal research group. For more information on these projects, please refer to the 2006 VBI Scientific Annual Report or to http://staff.vbi.vt.edu/setubal.

Agrobacterium biovars genome project

Annotation and analyses have proceeded and two manuscripts are in preparation: one describing the two new species sequenced (*Agrobacterium vitis* S4 and *Agrobacterium radiobacter* K84) and the other reporting re-annotation results for the *A. tumefaciens* C58 genome, which reconciles and integrates the two previously published sequences and annotations (Wood et al, 2001; Goodner et al, 2001). A companion to these two papers is a book chapter surveying new findings on the C58 genome that have been published since 2001 (Slater et al, 2007). Sequences and annotation are already publicly available at http://agro.vbi.vt.edu/public.

Azotobacter vinelandii genome project

Annotation and analyses are still ongoing for the *Azotobacter vinelandii* genome project.

Xanthomonas axonopodis pv aurantifolii genome project

A manuscript is in preparation describing the *Xanthomonas axonopodis* pv *aurantifolii* genome project.

Pseudomonas syringae genome project

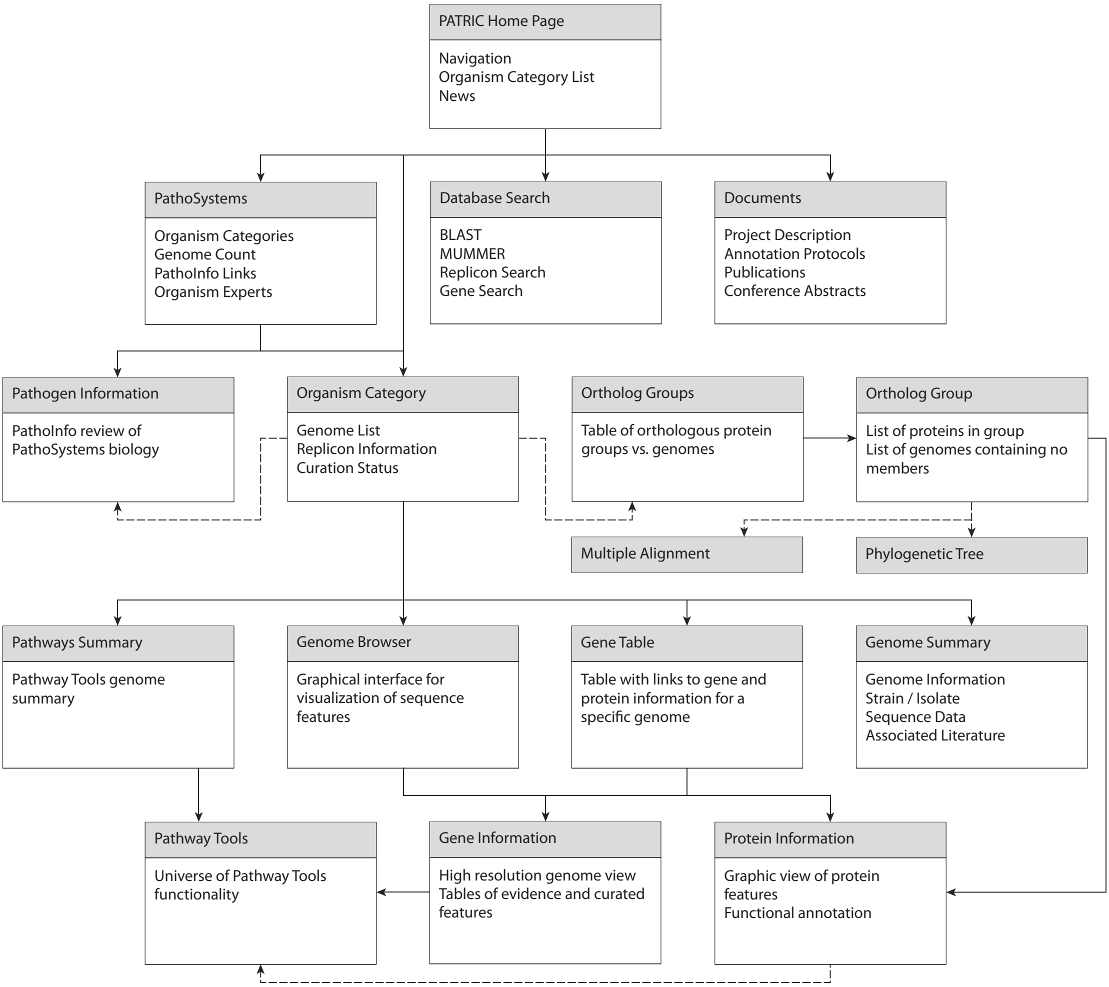
The genome of the PtoT1 strain was sequenced to 20X coverage using the newly acquired Roche GS-FLX™ sequencing machine at the Virginia Bioinformatics Institute. This is a joint project with Boris Vinatzer from the Department of Plant Pathology, Physiology, & Weed Science at Virginia Tech. Virginia Tech's Institute for Biomedical and Public Health Sciences funded the sequencing. Annotation and analyses are ongoing. The sequence is publicly available at http://staff.vbi.vt.edu/jcslab/pseudomonas.

Plant-Associated Microbe Gene Ontology project

More than 450 new terms have been created and the annotation of *A. tumefaciens* C58 is ongoing. More information is available at http://pamgo.vbi.vt.edu.

Genome Annotation Tool (GAT)

Many improvements have been made that include the following: a revised SRS-like search engine; Gene Ontology (GO) term assignment support for both automated and manual GO curation; an automated start codon change pipeline to check and assign correct gene start positions; and inclusion of ortholog and paralog information obtained from the OrthoMCL tool. The scientific community has used GAT heavily for genome annotation in the *Agrobacterium* biovars and *Azotobacter vinelandii* genome projects.



Conceptual map of PATRIC web site (http://patric.vbi.vt.edu). Arrows show the relationship between the principal data type on a page and related data on neighboring pages. Solid arrows represent 'drilling down' to more specific information (e.g. from genome to gene). Dashed arrows represent links between different views of conceptually similar data (e.g. between ortholog group and phylogenetic tree). This figure represents only a subset of the pages and links on the actual web site (from Snyder et al, 2007).

Genome Reverse Compiler (GRC)

New developments are as follows. GRC now uses the Gene Ontology (GO) to create consensus annotations based on the functions suggested by top BLAST results. The consensus annotation is created by utilizing the DAG structure of the ontology and optimizing for functional detail, support from multiple hits, and the amount of amino acid conservation. GRC also adjusts the start sites of maximum length open reading frames (ORFs) based on multiple alignments. In the spirit of open source, GRC enables statistics to be generated that detail its performance on gene finding and functional annotation using a reference annotation file. This allows the user to make modifications to GRC and then immediately observe the impact of these changes.

PathoSystems Resource Integration Center (PATRIC)

Over the past year a description of the project and database/website was published in the *Nucleic Acids Research* database issue (Snyder et al, 2007). Figure 1 gives an overview of some of the resources made available by PATRIC.

The first genomic analysis tool developed by PATRIC, GenVar, was released, and its description published (Yu et al, 2007). GenVar is designed to analyze an annotated genome and automatically identify missed gene calls and sequence variants such as genes with disrupted reading frames (split genes) and those with insertions and deletions (indels). For a given genome to be analyzed, GenVar relies on a database containing closely related genomes (such as other species or strains) as well as a few additional reference genomes. GenVar also helps identify gene disruptions that are probably caused by sequencing errors. In the paper published by our group (Yu et al, 2007), GenVar's capabilities were demonstrated on four *Brucella* genomes. *Brucella* is an important human pathogen and zoonotic agent. The analysis revealed hundreds of missed gene calls, new split genes and indels, several of which are species-specific and hence provide valuable clues to understanding how the genome of *Brucella* influences pathogenicity and host specificity. During the reporting period, a summary of results concerning PATRIC's support for coronavirus research was also published (Baker et al, 2006).

During the reporting period, I was chairman of the Ninth Annual Computational Genomics Conference, held in Baltimore, October 28-31, 2006. It received nearly 100 abstracts, out of which a program committee selected 23 oral presentations and 5 tutorials. One of the conference highlights was a talk by Barry J. Marshall, from the University of Western Australia and 2005 Nobel Laureate. The abstract booklet is available at http://www.

cpe.vt.edu/genomics. Bruno Sobral and Carol Bult (Jackson Laboratory) were co-chairpersons of the event. In November 2006, I gave a talk about the PAMGO project at the 27th Annual Crown Gall meeting, held on November 18 and 19 at Williams College, Williamstown, MA. Crown Gall is the meeting that brings together *Agrobacterium* researchers from around the world.

Conferences and workshops

Ninth Annual Computational Genomics Conference, Baltimore, October 28-31, 2006.

The Plant-Associated Microbe Gene Ontology (PAMGO) project, 27th Annual Crown Gall meeting, November 18 and 19, 2006, Williams College, Williamstown, MA.

Peer-reviewed publications in reporting period

Slater S, Goodner B, Setubal J, Goldman B, Wood D, Nester E (2007) The *Agrobacterium tumefaciens* C58 genome. In *Agrobacterium*, Tzfira T, Citovsky V (eds), Springer, New York. In press.

Snyder EE, Kampanya N, Lu J, Nordberg EK, Karur HR, Shukla M, Soneja J, Tian Y, Xue T, Yoo H, Zhang F, Dharmanolla C, Dongre NV, Gillespie JJ, Hamelius J, Hance M, Huntington KI, Jukneliene D, Koziski J, Mackasmiel L, Mane SP, Nguyen V, Purkayastha A, Shallom J, Yu G, Guo Y, Gabbard J, Hix D, Azad AF, Baker SC, Boyle SM, Khudyakov Y, Meng XJ, Rupprecht C, Vinje J, Crasta OR, Czar MJ, Dickerman A, Eckart JD, Kenyon R, Will R, Setubal JC, Sobral BWS (2007) PATRIC: The VBI PathoSystems Resource Integration Center. *Nucleic Acids Research* **35**: D401-D406.

Yu G, Snyder E, Boyle S, Crasta O, Czar M, Mane S, Purkayastha A, Sobral B, Setubal JC (2007) A Versatile Computational Pipeline for Bacterial Genome Annotation Improvement and Comparative Analysis, with Brucella as a Use Case. *Nucleic Acids Research.* In press.

Baker SC, Jukneliene D, Purkayastha A, Snyder EE, Crasta OR, Czar MJ, Setubal JC, Sobral BW (2006) Developing bioinformatic resources for coronaviruses. *Advances in Experimental Medicine and Biology*. **581**:395-398.

Other cited references

Wood DW, Setubal JC, Kaul R, Monks DE, Kitajima JP, Okura VK, Zhou Y, Chen L, Wood GE, Almeida NF, Woo L, Chen Y, Paulsen IT, Eisen JA, Karp PD, Bovee D, Chapman P, Clendenning J, Deatherage G, Gillet W, Grant C, Kutyavin T, Levy R, Li MJ, McClelland E, Palmieri A, Raymond C, Rouse G, Saenphimmachak C, Wu Z, Romero P, Gordon D, Zhang S, Yoo H, Tao Y, Biddle P, Jung M, Krespan W, Perry M, Gordon-Kamm B, Liao L, Kim S, Hendrick C, Zhao ZY, Dolan M, Chumley F, Tingey SV, Tomb JF, Gordon MP, Olson MV, Nester EW (2001) The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. *Science* **294**: 2317-2323.

Goodner B, Hinkle G, Gattung S, Miller N, Blanchard M, Qurollo B, Goldman BS, Cao Y, Askenazi M, Halling C, Mullin L, Houmiel K, Gordon J, Vaudin M, Iartchouk O, Epp A, Liu F, Wollam C, Allinger M, Doughty D, Scott C, Lappas C, Markelz B, Flanagan C, Crowell C, Gurson J, Lomo C, Sear C, Strub G, Cielo C, Slater S (2001) Genome sequence of the plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58. *Science* **294:** 2323-2328**.**

Vladimir Shulaev

vshulaev@vbi.vt.edu

Associate Professor, Virginia Bioinformatics Institute

Associate Professor, Dept. of Horticulture, Virginia Tech

Adjunct Associate Professor, Dept. of Cancer Biology, Wake Forest University School of Medicine, Winston-Salem, NC

Group contributors: Sunil Bajad, Diego Cortes, Sarah Holt, Deborah Reed, Wei Sha, Joel L. Shuman, Leepika Tuli.

A High-throughput Metabolomics Platform For Systems Biology

Abstract. The Biochemical Profiling Group at the Virginia Bioinformatics Institute is developing a high-throughput metabolomics platform for metabolic biomarker discovery, gene function elucidation and as a tool for systems biology. Our platform is based on a combination of untargeted metabolite profiling, targeted analysis, and metabolic fingerprinting. For sample analysis, we employ mass spectrometry-based analytical techniques due to their high sensitivity and wide range of covered metabolites. Our platform has been successfully used in collaborative systems biology projects to study oxidative stress response in the yeast *Saccharomyces cerevisiae*, to identify unique metabolic signatures associated with the progression of malignancy in human breast epithelium cells as well as in the response to various drugs by the malaria parasite *Plasmodium falciparum*, and to elucidate early metabolic responses to abiotic stress in plants.

Keywords: metabolomics; systems biology; metabolite profiling; mass spectrometry; cancer; malaria; oxidative stress; *Plasmodium falciparum*; *Saccharomyces cerevisiae*; *Arabidopsis*.

Scientific Progress

Metabolomics is aimed at measuring profiles of all, or a large number of, cellular metabolites. Progress in metabolomics depends on the development of more powerful analytical methods as well as robust bioinformatics tools for data analysis. Our group is developing a high-throughput metabolomics platform for metabolic biomarker discovery, and gene function elucidation as an integral component of systems biology studies. There are three major approaches used in metabolomics studies: targeted analysis, metabolite profiling, and metabolic fingerprinting. Targeted analysis is used to measure the concentration of limited numbers of known metabolites precisely. It is a quantitative approach and provides low limits of detection for known metabolites. Metabolite profiling usually is an untargeted high-throughput measurement of the levels of a large set of metabolites. The advantage of metabolite profiling, unlike targeted analysis, is that it does not require the compounds of interest to be known *a priori* and can therefore be used to identify novel metabolic biomarkers or survey the global metabolic changes in a system. Metabolic fingerprinting considers a total metabolite profile, or fingerprint, as a unique pattern characterizing a snapshot of the metabolism in a particular cell line or tissue.

Mass spectrometry-based analytical methods for simultaneous analyses of hundreds and thousands of analytes from a single cell or tissue sample have recently been developed. Successful development and application of such methods requires mass spectrometers capable of fast data acquisition rates combined with high sensitivity and specificity. Triple quadrupole instruments are increasingly used in metabolomics due to their high sensitivity and specificity. Typically, they are used to quantitate one or few analytes using selected reaction monitoring mode (SRM). SRM mode offers high specificity and has been successfully used for quantitation of hundreds of analytes simultaneously. Current triple quadrupole instruments allow simultaneous measurements of as many as 300 analytes in a sample using the SRM mode.

We have been working on the further development of metabolomics methodology based on simultaneous analysis of over 150 SRM transitions using hydrophilic interaction liquid chromatography (HILIC)–tandem mass spectrometry. We have optimized and validated the method with uninfected and *Plasmodium falciparum*-infected human red blood cell samples. Using this method, 81 of the compounds were detected in control red blood cells or red blood cells infected with a ring stage parasite (see Table). As some of the compounds were specifically detected in infected red blood cells, we expect that the number of detectable compounds would be much higher in trophozoite and other stages of the infection.

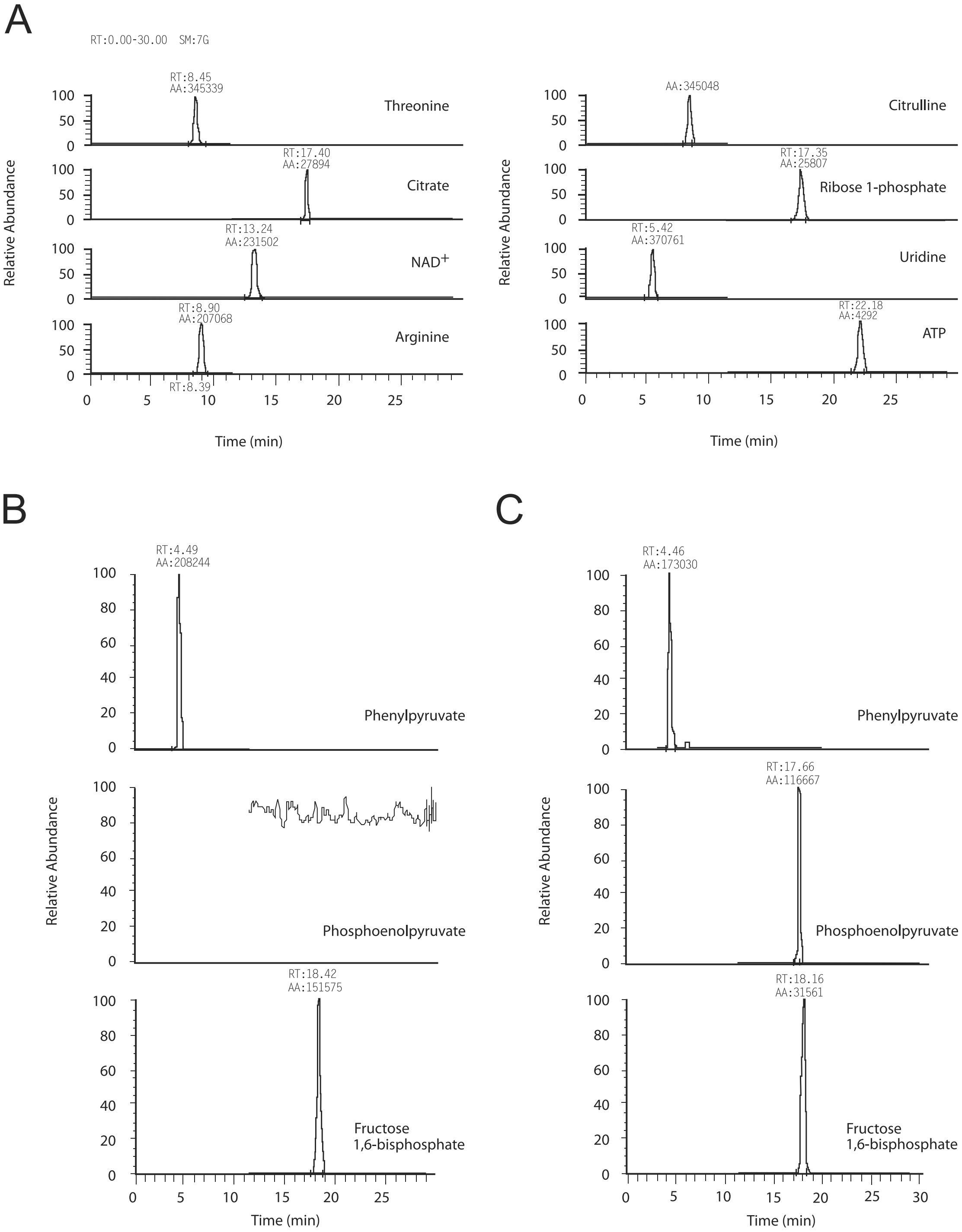


Figure. Liquid chromatography-mass spectrometry (LC-MS) analyses. (A) Representative chromatogram of the LC-MS separation of uninfected red blood cells using Selected Reaction Monitoring (SRM) method. More than 150 SRM transitions can be analyzed using this LC/MS/MS method. Using this method, we can identify changes in specific metabolites during development or following drug treatment. (B and C) An example of differential expression of phosphoenolpyruvate between non-infected red blood cells (B) and ring (C) samples.

We will continue to use this approach to study metabolic changes following *P. falciparum* treatment with antimalarial drugs.

For non-targeted liquid chromatography–mass spectrometry (LC-MS) analysis of plant and animal extracts, we have developed applications using a reversed phase monolithic capillary column. This method can be used to detect thousands of components in a tissue extract. We used this approach to identify early metabolic responses to oxidative stress in plants. Separation of the *Arabidopsis* tissue extract was performed on a 450 x 0.1 mm RP-18 capillary monolithic column using a ThermoFisher LTQ mass spectrometer operated in the Electrospray Ionization (ESI) mode with constant positive/negative switching. Data were processed with the ACD MS Manager software package using a CODA component detection algorithm.

Table. List of the compounds that can be detected using single liquid chromatography-mass spectrometry (LC/MS/MS) method. Compounds detected in uninfected and *P. falciparum*-infected human red blood cells are shown in bold.

Type	Compounds
Amino acids	**Alanine**, **Arginine**, **Asparagine**, Aspartate, Cysteine, **Glutamate**, **Glutamine**, **Glycine**, **Histidine**, **(Iso)leucine**, **Lysine**, **Methionine**, **Phenylalanine**, **Proline**, **Serine**, **Threonine**, **Tryptophan**, **Tyrosine**, Valine
Amino acid derivatives/ precursors	*N*-α-Acetylornithine, ***S*-Adenosyl-L-homocysteine**, ***S*-Adenosyl-L-methionine**, *p*-Aminobenzoate/anthranilate, **Citrulline**, **Cystathionine**, Histidinol, Homocysteic acid, **Homocysteine**, Homoserine, Hydroxyphenylpyruvate, **Ornithine**, Phenylpyruvate, 3-Phosphoserine, Prephenate, **Shikimate**, **Shikimate-3-*P***, **Taurine**
Nucleoside bases	**Adenine**, **Cytosine**, **Guanine**, **Hypoxanthine**, **Thymine**, **Uracil**, Xanthine
Nucleosides	**Adenosine**, **Cytidine**, **Deoxyadenosine**, Deoxyguanosine, Deoxyinosine, **Deoxyuridine**,Guanosine, Inosine, **Thymidine**, **Uridine**, **Xanthosine**
Nucleoside monophosphates	**AMP**, CMP, cyclic-AMP, **dAMP**, **dCMP**, dGMP, dTMP, dUMP, **GMP**, **IMP**, UMP, Xanthosine-5-*P*
Nucleoside di/triphosphates	**ADP**, **ATP**, CDP, CTP, dATP, dCDP, dCTP, dGDP, dGTP, dUTP, GDP, GTP, IDP, **ITP**, dTDP, TTP, UDP, UTP
Nucleoside precursor derivatives	**Carbamoyl-L-aspartate**, **Carbamoyl-*P***, Dihydroorotate, Guanosine 5′-*PP*, 3′-*PP*, *N*-Acetyl-glucosamine-1-*P*, Orotate, Orotidine-*P*, Phosphoribosyl pyrophosphate
CoAs	Acetoacetyl-CoA, Acetyl-CoA, CoA, Dephospho-CoA, 3-Hydroxy-3-methylglutaryl-CoA, Malonyl-CoA, Propionyl-CoA, Succinyl-CoA
Carbohydrate derivatives/ precursors	Acetylphosphate, ADP-D-glucose, **Allantoate**, Allantoin, 2-Dehydro-D-gluconate, Deoxyribose-*P*, **Dihydroxyacetone-*P***, Erythrose-4-*P*, **Fructose 1,6-bis-*P***, D-Glucarate, Gluconate, Glucosamine, Glucosamine-1-*P*, Glucosamine-6-*P*, **D-Glucose**, Glycerate, **Glycerol-3-*P***, D-Hexose-*P*, **Phenylpyruvate**, **Phosphoenolpyruvate**, 6-Phospho-D-gluconate, **3-Phosphoglycerate**, **D-Rib(ul)ose-5-*P***, Trehalose, UDP-D-glucose, UDP-D-glucuronate, UDP-*N*-acetyl-D-glucosamine
Vitamins and derivatives	Biotin, **Carnitine**, 7,8-Dihydrofolate, **Folate**, **5-Methyltetrahydrofolate**, **Nicotinamide**, Nicotinate, Pantothenate, **Pyridoxine**, **Riboflavin**, **Thiamine**, Thiamine-*P*
Carboxylic acids	**Acetoacetate**, **Aconitate**, **Citrate**, **Fumarate**, α-Ketoglutarate, **Lactate**, **Malate**, **Oxaloacetate**, **Succinate**
Redox electron carriers and precursors	**FAD**, FMN, **NAD**$^{+}$, NADH, **NADP**$^{+}$, NADPH, **Oxidized glutathione**, Quinolinate, **Reduced glutathione**
Miscellaneous metabolites	Adenosine 5-phosphosulfate, Agmatine, **Choline**, **2,3-Dihydroxybenzoate**, *trans*, *trans*-Farnesyl-*PP*, Geranyl-*PP*, ***p*-Hydroxybenzoate**, ***myo*-Inositol,** Putrescine, Spermidine

More than 700 components in the positive ion polarity mode and over 400 components in the negative ion polarity mode were detected.

We have used a metabolic fingerprinting approach to characterize the progression of malignancy in human breast epithelial cells. Samples were profiled using Gas Chromatography–Mass Spectrometry (GC-MS) and LC-MS. Data were analyzed with supervised and non-supervised algorithms. Genetic Algorithm-Discriminant Function Analysis (GA-DFA) identified a unique set of ions that can discriminate between different stages of malignant transformation in human breast epithelial cells. Many of these ions were identified using both commercially available, and extensive custom libraries of mass spectra. We used a similar approach to study drug responses in malarial parasites. This work has shown that drug treatment results in unique metabolic response patterns that are specific to particular drugs.

Metabolomics is a powerful "omics" approach that complements transcriptomics and proteomics for large-scale measurement of cellular components. It can provide a holistic view of a biological system when combined with mathematical modeling. In the future, we envision further integration of non-targeted and targeted analysis leading to truly quantitative investigation of the majority, if not all, of cellular metabolites.

Conferences and workshops

Nikolau B, Dickerson J, Dixon P, Fiehn O, Lange BM, Rhee SY, Shulaev V, Sumner LW, Welti R, Wurtele E (2006) *Using metabolomics to decipher functions of Arabidopsis genes in the context of metabolic and regulatory networks.* 4th International Conference on Plant Metabolomics, Reading, Berkshire, United Kingdom, 7-10 April, 2006.

Shulaev V (2006) *Insertional mutagenesis as a functional genomics tool in diploid strawberry (Fragaria vesca)*. 3rd International Rosaceae Genomics Conference, War Memorial Conference Centre, Marine Parade, Napier, New Zealand, 19-22 March, 2006.

Shulaev V (2006) *The time-dependent response of S. cerevisiae to oxidative stress.* The International Specialized Symposium on Yeast ISSY25: Systems Biology of Yeast – From Models to Application, Hanasaari, Espoo, Finland, June 18-21, 2006.

Shulaev V (2006) *When defense pathways collide: The response of Arabidopsis to a combination of drought and heat stress.* Gordon Research Conference on Salt and Water Stress in Plants, Magdalen College, Oxford, UK, September 3-8, 2006.

Shulaev V, Cortes DF, Shuman J, Yang Y, Varbanova M, Pichersky E (2006) *Metabolite profiling of the mutant lines overexpressing SABATH family methyltransferases in Arabidopsis thaliana.* 17th International Mass Spectrometry Conference, Prague, August 26-September 1, 2006.

Tuli L, Martins A, Sha W, Mendes P, Shulaev V (2006) *A systems biology study of the kinetics of Saccharomyces cerevisiae response to oxidative stress induced by cumene hydroperoxide*. Seventh International Conference on Systems Biology, Yokohama, Japan, October 9-13, 2006.

Peer-reviewed publications in reporting period

Bajad S, Shulaev V (2007) Highly-parallel metabolomics approaches using LC-MS^2 for pharmaceutical and environmental analysis. *Trends in Analytical Chemistry* **26**: 625-636.

Coutu1 J, Shulaev V, Mittler, R (2007) Reactive oxygen signaling in plants. In *Annual plant reviews: intracellular signaling in plants*. Zhenbiao Yang (ed), Blackwell Publishing. In press.

Martins AM, Sha W, Evans C, Martino-Catt S, Mendes P, Shulaev V (2007) Comparison of sampling techniques for parallel analysis of transcript and metabolite levels in *Saccharomyces cerevisiae*. *Yeast* **24**: 181-188.

Pisciotta JM, Coppens I, Tripathi AK, Scholl PF, Shuman J, Bajad S, Shulaev V, Sullivan DJ (2007) The role of neutral lipid nanospheres in *Plasmodium falciparum* heme crystallization. *Biochem J.* **402**: 197-204.

Varbanova M, Yamaguchi S, Yang Y, McKelvey K, Hanada A, Borochov R, Yu F, Jikumaru Y, Ross J, Cortes D, Je Ma C, Noel JP, Mander L, Shulaev V, Kamiya Y, Rodermel S, Weiss D, Pichersky E (2007) Methylation of gibberellins by *Arabidopsis* GAMT1 and GAMT2. *Plant Cell.* **19**: 32-45.

Bruno Sobral

sobral@vbi.vt.edu

Professor, Executive and Scientific Director, Virginia Bioinformatics Institute

Professor, Dept. of Plant Pathology, Physiology and Weed Science, Virginia Tech

PathoSystems Biology Research

Group contributors: Allan Dickerman, Timothy Driscoll, Matt Dyer, Raymie Equi, Isabel Osorio da Fonseca, Jocelyn Kemp, James Lester, Shenghua Li, Chunhong Mao, Xiaoyan Sheng, Chunxia Wang, Kelly Williams.

Abstract. Our major effort in this reporting period comprised functional genomic studies of key genes for carbon metabolite and nodulation/nitrogen fixation in *Sinorhizobium meliloti*, comparisons of Rhizobia with other α-proteobacteria, and participation in the genome sequencing of *Aedes aegypti* (Nene et al, 2007*)*. We have also performed genome-wide analyses of the ExoS/ChvI two-component regulatory system in *S. meliloti*, investigated the influence of polyhydroxybutyrate granule-associated proteins on polyhydroxybutyrate accumulation and symbiotic nitrogen fixation in *S. meliloti* Rm1021, and carried out comparative modeling of the cell cycle (from *Caulobacter crescentus* to *S. meliloti*). In addition, we have further developed the Rhizobiales Bioinformatics Resource Center and the NodMutDB Nodulation Mutant Database, identified novel small non-coding RNAs in *S. meliloti*, probed the regulation of transcription during infection and nodulation, and identified enzymes using Open Reading Frame screening. The group has also identified a tRNA unique to the α-proteobacteria, constructed a robust species tree for the α-proteobacteria, and continued work on the computational prediction of host-pathogen protein-protein interactions.

Keywords: α-proteobacteria; *Sinorhizobium meliloti*; symbiosis; database; curation; small non-coding RNA; transcriptional regulation; host-pathogen interactions; mathematical modeling; open reading frame screening; bioinformatics; computational biology.

Scientific Progress

Genome-wide analyses of the ExoS/ChvI two-component regulatory system in S. meliloti

Bacteria have signaling systems for eliciting a variety of adaptive responses to their environments. Two-component regulatory systems (TCSs) are one of the most important mechanisms for these responses. The TCS ExoS/ChvI in *S. meliloti*, *Brucella* sp. and *Agrobacterium* sp. are involved in symbiosis, virulence, and tumorigenesis respectively. Though this system plays a crucial role in establishing the symbiosis between *S. meliloti* and its host plant, it remains incompletely characterized. We constructed a *chvI* unmarked, completely deleted open reading frame (ORF). The loss of ChvI resulted in a decrease of more than 300 annotated genes. Phenotypic assays showed that the *chvI* mutant: 1) failed to form nodules on *Medicago truncatula*; 2) exhibited lower tolerance to acidic conditions, and 3) produced significantly less polyhydroxybutyrate (PHB) than wild type. Our results demonstrate that ChvI is strongly involved not only in exopolysaccharide I (EPSI) and flagellum, but also in regulatory networks involving cell envelope and metabolism. However, its precise role within the regulatory network remains to be determined. We will use proteomics to identify and analyze the ExoS/ChvI system in *S. meliloti* further.

Influence of PHB granule-associated proteins on PHB accumulation and symbiotic nitrogen fixation in S. meliloti Rm1021

Poly(3-hydroxybutyrate) (PHB) is a major carbon storage compound in *S. meliloti*. Through a series of genetic and phenotypic studies of PHB synthesis mutants, we demonstrated PHB functions in the symbiotic interactions of *S. meliloti* with *Medicago* spp. (Wang et al, 2007). PHB granule-associated proteins (phasins) are proposed to regulate PHB synthesis and granule formation in bacteria. However, phasins have not yet been described in *S. meliloti*. A previous study in our laboratory has shown that *SMc00777* and *SMc02111* encoded two major phasins in *S. meliloti* Rm1021. The functions of these two genes were further investigated via mutation and phenotypic analyses. We demonstrated that the double mutant was defective in PHB production but had higher exopolysaccharide production levels. Although the mutants exhibited significant acetylene reduction activity on *M. truncatula*, the symbiotic phenotype of the mutants appeared to be Fix$^-$. Our results demonstrated that phasins play important roles in PHB biosynthesis and symbiotic nitrogen fixation.

Comparative modeling of the cell cycle: from Caulobacter crescentus to S. meliloti

The cell cycle of eukaryotes has been described in molecular detail while the bacterial cell cycle remains poorly understood. *C. crescentus*, a relative of *S. meliloti* (both are α -proteobacteria), is an attractive model for examining cell cycle regulation in bacteria because it is easily synchronized and its replication pattern is temporally distinguished. During nodule formation, *S. meliloti* cells enter the host plant where they differentiate and form nitrogen-fixing bacteroids. While the genetics and biochemistry of nitrogen fixation have been studied extensively the genes and factors involved in and controlling the morphological changes during bacteroid differentiation have not been well characterized. We have designed experiments to investigate several important genes potentially involved in cell differentiation in *S. meliloti* based on the cell cycle model of *Caulobacter*. Through density centrifugation, we have succeeded in isolating synchronized populations of *S. meliloti* cells. Our major focus will be to analyze specific mutants and their phenotypes with respect to cell cycle.

Development of the Rhizobiales Bioinformatics Resource Center

The goal of the Rhizobiales Bioinformatics Resource Center (RhizobialesBRC) is to provide a comprehensive and accurate web-based resource for genomic and associated information on the sequenced Rhizobiales. In the reporting period, we focused on improving the annotation of the *S. meliloti* genome. We updated the annotation of pSymB of *S. meliloti* at the genome sequence level. Based on the first round of our curation, a total of 1644 coding sequences (CDSs) were predicted including 84 new CDSs in addition to what was published in RefSeq. By comparing with the RefSeq data, our annotation agreed with 93% of RefSeq CDSs at the sequence level. We shortened or extended 6.7% of the RefSeq CDSs. We are in process of annotating the chromosome of *S. meliloti*. Once the sequence level annotation is complete, protein level curation will be performed.

Identification of novel small RNAs in S. meliloti

Small non-coding RNAs (sRNAs) are an emerging class of gene expression regulators in bacteria. Recent studies have demonstrated the abundance and importance of the regulatory small RNAs. However, sRNAs are not yet included in genome annotation. We have developed a pipeline called CIGfinder to identify and analyze the conserved intergenic regions (CIGs) of Rhizobiales and we are searching for novel sRNAs in *S. meliloti.* We identified 27 novel sRNA candidates. 14 of these sRNAs were expressed according to our microarray data. Several selected sRNA candidates were verified experimentally and more candidates are being tested.

NodMutDB - Nodulation Mutant Database

NodMutDB is a web-accessible relational database to provide a comprehensive and accurate resource for depositing, organizing and retrieving information on symbiosis-related genes, mutants and published literature. To date we have collected and curated symbiosis genes and mutants from Rhizobia, *S. meliloti* and *Bradyrhizobium japonicum*, and their host plants, *Medicago truncatula* and *Glycine max*, respectively. In the reporting period, we collected and updated over 200 genes and mutants in the database. We continue to update and annotate the database as new studies are published.

Regulation of transcription during infection and nodulation

We are interested in the changes in transcriptional regulation within *S. meliloti* that accompany and help drive the transformation between free-living and bacteroid states. We are using genome-wide location assays (GWLAs) to identify the DNA-binding sites of transcriptional regulators (TRs) known to play a role in symbiosis. In GWLA, tagged mutants of a TR are reversibly cross-linked to DNA *in vivo*. TR-DNA complexes of interest are isolated from the cell lysate and the bound DNA fragments are subsequently identified. Mapping these fragments back to the *S. meliloti* genome will allow us to identify the regulatory targets for the TR. We have developed a stable strain of *S. meliloti* harboring a tagged form of *ctrA*, which functions as a master regulator of cell cycle control in *C. crescentus*. To complement our transcriptional regulation project, we are building AlphaRegs, a database of transcriptional regulation in the Rhizobiales. AlphaRegs will incorporate experimentally verified data as well as predictions related to transcriptional regulation in *S. meliloti* and other sequenced Rhizobiales. We will use the AlphaRegs database as a repository of Rhizobiales transcriptional regulation information, as well as a platform for predicting regulatory elements across the Rhizobiales.

Enzyme identification using ORF screening

Gene functions are assigned after homology comparisons by automated data bank searching. A limitation to a genome annotation procedure based solely upon homology comparisons is that it does not always provide rigorous evidence for the biological functions of genes and their products. To overcome these problems, we are developing a method – direct assay *in vitro* to identify target ORFs (considered unknown, putative and hypothetical). The first step is to survey the interested enzymes (e.g., dehydrogenases) that have been characterized by direct assay. These protein sequences will be used in a BLAST search of the *S. meliloti* database. The ORFs that possibly code for proteins that catalyze the considered reaction will be cloned, expressed and assayed. The target ORF will be functionally annotated based on experimental evidence.

A tRNA unique to the α-proteobacteria

All previously known $tRNA^{His}$ have an extra nucleotide at their 5' end, encoded in the gene in bacteria but added post-transcriptionally in eukaryotes. By following bioinformatic observations with molecular experiments, we showed that $tRNA^{His}$ in a group of α -proteobacteria including *Sinorhizobium* and *Caulobacter* lack this extra nucleotide (Wang, Sobral and Williams, 2007). This was correlated with changes at several key sites in the enzyme that adds histidine to tRNA. This altered tRNA charging system may have affected other tRNA charging systems in these bacteria; for example, a site in $tRNA^{Glu}$ sequences was found to covary with $tRNA^{His}$ among α-proteobacteria.

A robust species tree for the α-proteobacteria

Innumerable biological questions are better addressed when the evolutionary history of a study organism is well understood with respect to its relatives. Our interests in members of the Rhizobiales and in *Rickettsia* prompted us to determine the evolutionary history of the entire bacterial class α-proteobacteria. Collecting over a hundred protein families shared by 72 species, we generated a tree that was very well supported by several measures (Williams, Sobral and Dickerman, 2007). Including homologous proteins from several primitive mitochondria, we placed the mitochondrial branch in the α-proteobacterial tree, showing that the mitochondria are a sister group to the two families containing intracellular parasites such as *Rickettsia* and *Anaplasma*, and are also closely related to the most abundant of marine organisms, *Pelagibacter*.

Computational prediction of host-pathogen protein-protein interactions

Infectious diseases result in millions of deaths each year and millions of dollars are spent annually to better understand how pathogens infect their hosts and to identify potential targets for therapeutics, diagnostics and vaccines. Host-pathogen protein-protein interactions (PPIs) play a vital role in initiating infection. Identifying which PPIs enable a pathogen to invade its host may provide potential targets for therapeutics. As far as we know, no systematic methods have been reported for predicting physical interactions between host and pathogen proteins. We have developed a novel framework for predicting and studying host-pathogen PPI networks (Dyer, Murali and Sobral, 2007). We use intra-species PPIs and domain-domain profiles to compute statistics on how often proteins containing specific pairs of domains interact. When applied to the *Homo sapiens–Plasmodium falciparum* system, we show that we are able to make predictions that are biologically significant.

Conferences and workshops

Dyer MD, Murali TM, Sobral BWS (2006) Computational prediction of host-pathogen protein-protein interactions. *Ninth Annual Conference on Computational Genomics.* Baltimore, MD.

Li S, Paul B, Sobral BWS, Tyson J (2006) Kinetic analysis of the mechanism of the cell division cycle in *Caulobacter crescentus. SIAM Conference on Life Sciences.* Raleigh, NC.

Li S**,** Paul B, Sobral BWS, Tyson J (2007) Quantitative modeling of the asymmetric cell division cycle in *Caulobacter crescentus*. *Computational Cell Biology Meeting.* Cold Spring Harbor Laboratory, NY.

Mao C, Qiu J, Wang C, Charles TC, Sobral BWS (2006) NodMutDB: A comprehensive database for genes and mutants involved in rhizobia-legume symbiosis. *The 3rd International Conference on Legume Genomics and Genetics.* Brisbane, Australia.

Mao C (2006) Comparative analysis of conserved intergenic regions in Rhizobiales. *International Symposium on the Comparative Biology of the Alpha-Proteobacteria.* Blacksburg, VA.

Wang C, Sheng X, Mao C, Equi R, Charles TC, Sobral BWS (2006) Investigation of the early stage symbiotic interaction of *Medicago truncatula-Sinorhizobium meliloti* by proteomics. *The 3rd International Conference on Legume Genomics and Genetics.* Brisbane, Australia.

Peer-reviewed publications in reporting period

Dyer MD, Murali TM, Sobral BWS (2007) Computational prediction of host-pathogen protein-protein interactions. *Bioinformatics* (in press).

Nene V, Wortman JR, Lawson D, and 92 other authors (2007) Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science* **316**: 1718-1723.

Wang C, Saldanha M, Sheng X, Shelswell KJ, Walsh KY, Sobral BWS, Charles TC (2007) Roles of poly-3-hydroxybutyrate (PHB) and glycogen in symbiosis of *Sinorhizobium meliloti* with *Medicago sp*. *Microbiology* **153**: 388-398.

Wang C, Sobral BWS, Williams KP (2007) Loss of a universal tRNA feature. *Journal of Bacteriology* **189**: 1954-1962.

Williams KP, Sobral BWS, Dickerman AW (2007) A robust species tree for the Alphaproteobacteria. *Journal of Bacteriology* (in press) **189**: 4578-4586.

Cyberinfrastructure For Biodefense And Emerging/Re-emerging Infectious Diseases

Group contributors: George Abramochkin, Susan Baker, Suren Baliji, Chris Bowns, David Bynum, Cory Byrd, Stephen Cammer, Oswald Crasta, Mike Czar, Chitti Dharmanolla, Allan Dickerman, William Donnell, Matt Dyer, Darius Dzuida, Zhangjun Fei, Herman Formadi, Joe Gabbard, Joseph Gillespie, Mark Hance, Debby Hix, Ranjan Jha, Nithiwat Kampanya, Ron Kenyon, Chaitanya Kommidi, Konstantinos Krampas, Christine Lee, Jian Li, Tania Lima, Dan Liu, Jian Lu, Dustin Machi, Lucas Mackasmiel, Shrinivasrao Mane, Doug McMaster, Bharat Mehrotra, Saroj Mohapatra, Eric Nordberg, Anjan Purkayastha, Harsha Rajasimha, Daphne Rainey, Vincent Samaco, Graciela Santopietro, Mark Scott, Patricia Seeley, João Setubal, Joshua Shallom, Shamira Shallom, Bruce Sharp, Maulik Shukla, Eric Snyder, Bruno Sobral, Jeetendra Soneja, Dan Sullivan, Wei Sun, Satish Tadepalli, Yuying Tian, Nirali Vaghela, Nishantsinh Vaghela, Nataraj Vishnubhat, Sally Waldon, Rebecca Wattam, Rebecca Will, Kelly Williams, Tian Xue, Boyu Yang, Hyuneung Yoo, Gongxin Yu, Qiang Yu, Chengdong Zhang, Fengkai Zhang, Yan Zhang, Jing Zhao.

Abstract. The Cyberinfrastructure Group applies the principles of cyberinfrastructure (Atkins et al., 2003) to integrate data, computational infrastructure and people for the purpose of scientific discovery, primarily in the area of infectious diseases. In the reporting period, the Cyberinfrastructure Group has curated genomic, microarray, proteomic and literature data from various infectious disease systems, and implemented the databases required to support them. The Cyberinfrastructure Group also develops tools for analysis and visualization of the data across its systems. The Cyberinfrastructure Group is bringing these resources together through a common portal architecture for acquiring, storing, accessing, analyzing and portraying the data, enhancing the capability for knowledge discovery and facilitating collaboration among researchers. Through its education and outreach activities, the Cyberinfrastructure Group has developed sessions for and collaborations with external researchers, resulting in discoveries and publications. The Cyberinfrastructure Group also offers Collaborative Research Team resources for the design and analysis of experiments and the discovery of new knowledge, user-centered software design and development, and bioinformatics training materials and session deployment focused on research goals and data.

Keywords: cyberinfrastructure; infectious disease; genomics; gene expression; proteomics; web services; interoperability.

Scientific Progress

Pathosystem biology, data curation and integration

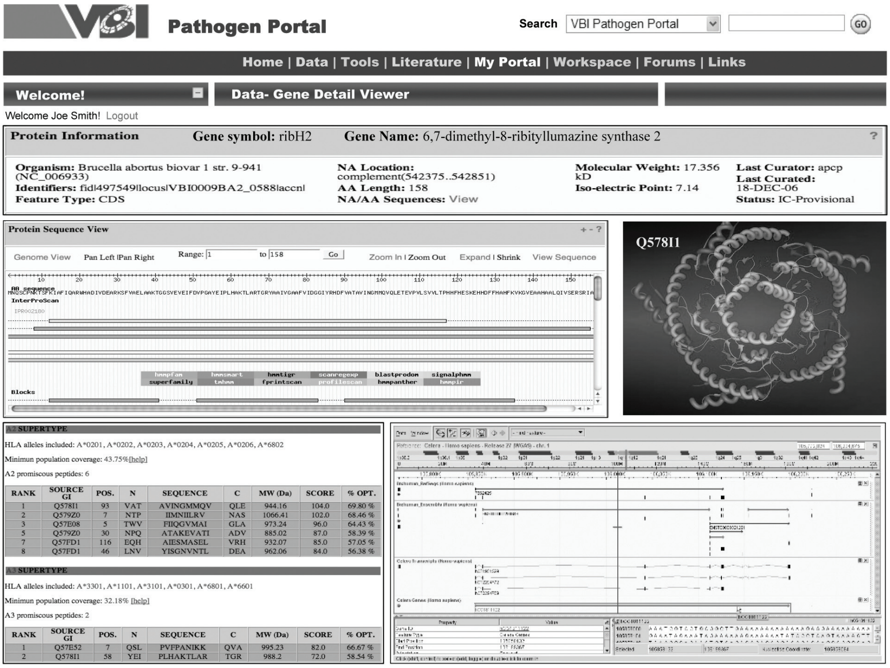
The Pathosystems Resource Integration Center (PATRIC, http://patric.vbi.vt.edu/) project has allowed the Cyberinfrastructure Group (CIG) to improve the quality of genome gene/feature and protein product predictions and annotations, and to develop literature-based curation for biodefense and emerging infectious disease pathosystems. Enhancements to the PATRIC curation infrastructure (Snyder et al, 2007) include the development of automated tools for identifying frameshifts and missed genes (Yu et al, 2007). These enhancements have increased genome annotation throughput and quality. The PATRIC database now contains 20 bacterial and 490 viral genomes. PATRIC scored highest among the eight National Institute of Allergy and Infectious Diseases (NIAID)-funded Bioinformatics Resource Centers (BRCs) for curation metrics accuracy, consistency, and completeness (http://www.brc-central.org/brc-central/VAM/index.html). PATRIC has also established collaboration with the Swiss Institute of Bioinformatics to provide high quality functional annotations for selected target proteins. PATRIC provides groups of orthologous proteins among closely related organisms and metabolic pathways and improved data availability and presentation through formal user-centered design and evaluation methods. This functionality is available to the Sobral PathoSystems Biology group, which is also using the system for curation of select *Rhizobiales* (http://rhizobia.vbi.vt.edu).

CIG's research output includes prioritization of all proteins in *Brucella* that are relevant to countermeasures (modified from Hasan et al, 2006) by applying computational analysis of the genome to predict secreted and surface-attached proteins, druggable domains, metabolic chokepoints, literature data on essentiality for survival and virulence in the host, and experimental data on virulence response. Comparative genomic analysis of *Rickettsia* genomes revealed the presence of a single plasmid in *R. felis*, horizontal inheritance of the plasmid from other organisms, and that the primitive *Rickettsia* ancestor itself harbored plasmids of this type (Gillespie et al, 2007).

CIG is developing and integrating data from PathInfo (He et al, 2005) documents into both PATRIC and the Resource Center for Biodefense Proteomics Research (RCBPR, http://www.proteomicsresource.org) for their corresponding organisms. We have also implemented Islander: Database of Genomic Islands (http://islander.vbi.vt.edu) as part of CIG's genomic infrastructure (based on Mantri and Williams, 2004). In addition, we have developed a pipeline that produces high-quality phylogenetic trees, either of individual genes or of genomes, to aid in understanding biomolecular functions or biological processes by examining evolutionary history. This approach was applied to produce a robust species tree for the α-proteobacteria (Williams, Sobral and Dickerman, 2007).

CIG has enhanced RCBPR by providing improvements in automation for building three-dimensional structure representations, database searches, data exports, and overall website unification. In the reporting period, new data sets from the Proteomics Research Centers (http://www.niaid.nih.gov/dmid/genomes/prc/centers.htm) have included microarray, mass spectrometry, clones, structures, and yeast two-hybrid. This has resulted in an approximate doubling of the amount of experimental data in the database released to the community.

CIG is developing a web portal to provide a single point of access to an array of tools and data sources in pathosystems biology, including data from PATRIC and the RCBPR, as well as external data providers, such as the National Center for Biotechnology Information (NCBI) and the European Molecular Biology Laboratory (EMBL). The portal will support scientific workflows by integrating web services and portlets (portal components). The design is being driven by interviews with representative user groups and formal usability engineering



Welcome Joe Smith! Logout

Protein Information **Gene symbol:** ribH2 **Gene Name:** 6,7-dimethyl-8-ribityllumazine synthase 2

Organism: Brucella abortus biovar 1 str. 9-941 (NC_006933)
Identifiers: fid|497549|locus|VBI0009BA2_0588|accn|
Feature Type: CDS
NA Location: complement(542375..542851)
AA Length: 158
NA/AA Sequences: View
Molecular Weight: 17.356 kD
Iso-electric Point: 7.14
Last Curator: apcp
Last Curated: 18-DEC-06
Status: IC-Provisional

MNQSCPNKTSFKIAFIQARWHADIVDEARKSFVAELAAKTGGSVEVEIFDVPGAYEIPLHAKTLARTGRYAAIVGAAFVIDGGIYRHDFVATAVINGMMQVQLETEVPVLSVVLTPHHFHESKEHHDFFHAHFKVKGVEAAHAALQIVSERSRIA

A2 SUPERTYPE

HLA alleles included: A*0201, A*0202, A*0203, A*0204, A*0205, A*0206, A*6802

Minimun population coverage: 43.75%[help]

A2 promiscous peptides: 6

RANK	SOURCE GI	POS.	N	SEQUENCE	C	MW (Da)	SCORE	% OPT.
1	Q578I1	93	VAT	AVINGMMQV	QLE	944.16	104.0	69.80 %
2	Q579Z0	7	NTP	IIMNIILRV	NAS	1066.41	102.0	68.46 %
3	Q57E08	5	TWV	FIIQGVMAI	GLA	973.24	96.0	64.43 %
5	Q579Z0	30	NPQ	ATAKEVATI	ADV	885.02	87.0	58.39 %
7	Q57FD1	116	EQH	AIESMASEL	VRH	932.07	85.0	57.05 %
8	Q57FD1	46	LNV	YISGNVNTL	DEA	962.06	84.0	56.38 %

A3 SUPERTYPE

HLA alleles included: A*3301, A*1101, A*3101, A*0301, A*6801, A*6601

Minimun population coverage: 32.18% [help]

A3 promiscous peptides: 2

RANK	SOURCE GI	POS.	N	SEQUENCE	C	MW (Da)	SCORE	% OPT.
1	Q57E52	7	QSL	PVFPANIKK	QVA	995.23	82.0	66.67 %
2	Q578I1	58	YEI	PLHAKTLAR	TGR	988.2	72.0	58.54 %

Figure 1. Portal Infrastructure Concept. The Cyberinfrastructure Group is developing a web-based portal to integrate heterogeneous data types for pathosystems biology.

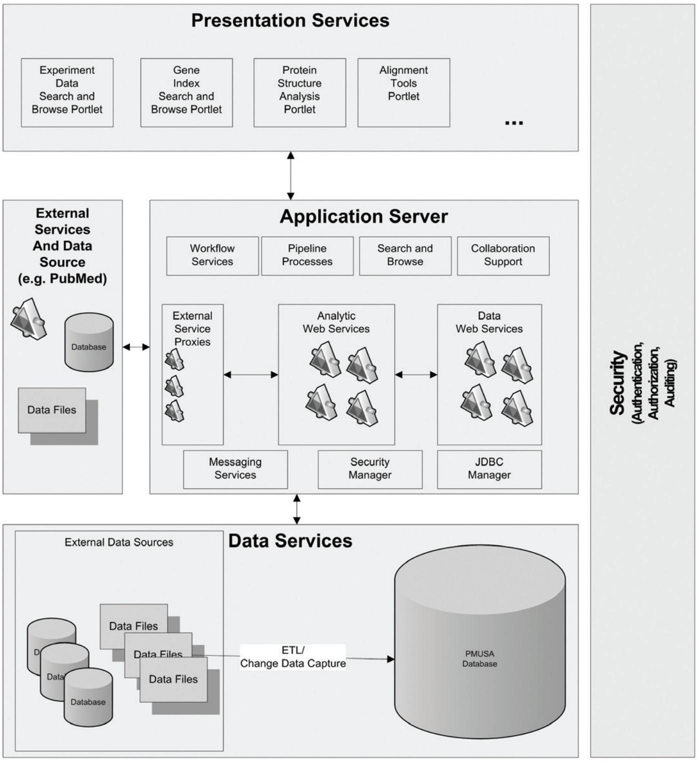


Figure 2. Implementation of Portal Infrastructure. The implementation of the web-based portal shown in Figure 1 is through a modular tiered system of portlet-based presentation services supported by web services and underlying databases.

techniques. We have implemented a proof-of-concept working prototype and have begun populating the infrastructure (see Figures 1 and 2).

CIG's new portal-compatible Gene Set Explorer (GSE) provides efficient data sifting and exploration through grouping across rows and columns, group creation based on various criteria, filtering and sorting to allow group identification and differentiation, set operations like intersection and union, and statistical operations, such as correlations, among data sets.

Collaborative research

CIG delivered research-centric bioinformatics training to the research communities in the Middle-Atlantic Regional Center of Excellence (MARCE, http://marcebiodefense.org/), including two *Francisella*-centric sessions for researchers at Virginia Tech and the University of Virginia. The use of genomics concepts and bioinformatics tools provided perspectives toward countermeasure development and has led to collaborations with *Francisella* researchers.

CIG has engaged in 14 collaborative research projects encompassing host (such as mouse, monkey, and human) and pathogen (including *Bacillus*, *Entamoeba*, and *Escherichia coli*) responses and toxicological responses. Three of the projects, including one with the PathoSystems Biology group, resulted in publications (Gilchrist et al, 2006; Djavani et al, 2007; Wang et al, 2007) and other publications are pending. For example, we have used our web-based microarray analysis pipeline developed in collaboration with GraphLogic to support gene expression analysis.

In the Cyberinfrastructure TEAM (CI-TEAM) demonstration project (http://ci.vbi.vt.edu/CITEAM/), CIG and faculty at Bluefield State College and Galileo Magnet High School have developed and deployed course materials to prepare the future workforce to use cyberinfrastructure. Evaluation of the project is being undertaken to aid in the development of a sustainable outreach program in cyberinfrastructure.

Conferences and workshops

Purkayastha A (2006) Brucella *and* Rickettsia *Genome Analyses.* The International Symposium on the Comparative Biology of Alpha-Proteobacteria, April 2006, Virginia Bioinformatics Institute.

Crasta O (2006) *Comparative Genomic Analysis of Attenuated and Virulent Strains of* Brucella abortus. The International Symposium on the Comparative Biology of Alpha-Proteobacteria, April 2006, Virginia Bioinformatics Institute.

Snyder E, Purkayastha A (2006) *PATRIC.* American Society for Microbiology General Meeting, May 2006.

Yang B, Xue T, Zhao J, Kommidi C, Soneja J, Li J, Will R, Sharp B, Kenyon R, Crasta O, Sobral BW (2006) *Bioinformatics Web Services Provided by VBI.* BIOCOMP'06, June 2006.

Zhao J, Xue T, Yang B, Williams K, Wattam R, Will R, Sharp B, Kenyon R, Crasta O, Sobral BW (2006) *VBI Genome Annotation and Comparison System*. BIOCOMP'06, June 2006.

Crasta O (2006) *Bioinformatics Resources for PathoSystems Biology.* The Bioinformatics Tools for the Beginner/Intermediate Researcher workshop. American Phytopathological Society/Canadian Phytopathological Society/Mycological Society of America
Joint Meeting, July 2006.

Shallom J (2006) *The PathoSystems Resource Integration Center: Implications for Viral Pathogens.* 12th International Symposium on Viral Hepatitis and Liver Disease, July 2006.

Mackasmiel L (2006) *The PathoSystems Resource Integration Center: Implications for Biodefense Viral Pathogens.* The American Society for Virology 25th Annual Meeting, July 2006.

Krampis K (2006) *MyGrid and ToolBus/PathPort Interoperability.* 14th Annual International Conference on Intelligent Systems for Molecular Biology, August 2006.

Shallom J (2006) *Prediction of Secreted Proteins in* Coxiella burnetii *(RSA 493). Using Bioinformatics Tools.* 20th Meeting - Joint meeting of the American Society of Rickettsiology and The 5th International Conference on *Bartonella* as Emerging Pathogens, September 2006.

Cammer S (2006) *The Proteomics Data Center (PDC): NIAID Administrative Resource for Biodefense Research Centers.* HUPO 5th Annual World Congress, October 2006.

Sullivan D (2006) *An Architectural Framework for Integrated Bioinformatics Applications.* Microsoft eScience Workshop, October 2006.

Snyder E (2006) *High-Fidelity Prediction of Orthologous Proteins in Bacterial Genomes.* 9th Annual Computational Genomics 2006, October 2006.

Snyder E (2006) *A Novel Bacterial Genome Analysis Methodology for Annotation Improvement and Comparative Analysis, and Its Application to Brucella.* 9th Annual Computational Genomics 2006, October 2006.

Crasta OR, Zhang Y, Kommidi C, Fei Z, Dziuda DM, Cammer S, Sharp B, Sobral BW (2006) *A Web-Based Array Data Analysis and Management System.* 9th Annual Computational Genomics 2006, October 2006.

Vaghela, N., M. Scott, C. Lee, E. Nordberg, B. Tyler, S. Tripathy, R. Will, R. Kenyon, O. Crasta, and B. Sobral (2006) *Gene Set Explorer A specialized application for creating, managing and mining statistical and gene expression data.* 9th Annual Computational Genomics 2006, October 2006.

Sobral B (2006) *Cyberinfrastructure for Pathosystems Biology.* Global Dialogues on Emerging Science and Technology, November 2006.

Sullivan D (2006) *An Architectural Framework for Integrated Bioinformatics Applications.* Data Warehousing Technologies in Bioinformatics, December 2006.

Sobral B (2007) *Discovery of Vaccines, Diagnostics and Therapeutic Targets Using Bioinformatics Resources. Biodefense Informatics and Proteomics focus session.* American Society for Microbiology Biodefense Meeting, February 2007.

Cammer S, Zhang C (2007) *An Integrated Post-Genome Data Resource for Host-Pathogen Interaction Profiles.* US HUPO 3rd Annual Conference, March 2007.

Crasta OR, Kenyon RW, Sobral BW (2007) *Collaborative Bioinformatics Approaches to Countermeasure Target Discovery at the Middle-Atlantic Regional Center of Excellence (MARCE).* 4th Annual NIAID RCE Research Meeting, April 2007.

Crasta OR, Mane SP, Evans C, Dominguez-Bello MG, Sobral BW (2007) *Whole Genome Sequence of* Helicobacter pylori *Amerindian Strain, V225.* 107th Annual Society of Microbiology General Meeting, May 2007.

Intellectual property

Brucella abortus genome sequencing and virulence factors (11/24/06) *Brucella abortus* genome sequencing and virulence factors, US provisional.

Peer-reviewed publications in reporting period

Cammer S (2007) SChiSM2: creating interactive web page annotations of molecular structure models using Jmol. *Bioinformatics* **23**: 383-384.

Crasta OR, Wattam AR, Purkayastha A, Sobral BW (2007) Genome Sequencing and Bioinformatics Resources. In *Genome Mapping in Animals and Microbes*, Volume 8 - Animal Associated Microbes, Cole C, Nene V (eds), Springer (in press).

Crasta OR, Yu GX, Snyder EE, Williams K, Sobral BW (2007) Comparative Genomics of Brucella. In *Genome Mapping in Animals and Microbes* Volume 8 - Animal Associated Microbes, Cole C, Nene V (eds), Springer (in press).

Deans AR, Gillespie JJ, Yoder MJ (2006) An evaluation of ensign wasp classification (*Hymenoptera: Evaniidae*) based on molecular data and insights from ribosomal RNA secondary structure. *Syst. Entomol.* **31**: 517-528.

Djavani M, Crasta O, Zapata J, Fei Z, Folkerts O, Sobral B, Bryant J, Pauza C, Lukashevich I, Salvato M (2007) Early blood profiles of virus infection in a monkey model for Lassa Fever. *J. Virol.* (in press).

Fei Z, Tang X, Alba R, Giovannoni J (2006) Tomato Expression Database (TED): A suite of data presentation and analysis tools. *Nucleic Acids Research* **34**: D766-770.

Gilchrist CA, Houpt E, Trapaidze N, Fei Z, Crasta O, Asgharpour A, Evans C, Martino-Catt S, Baba DJ, Stroup S, Hamano S, Ehrenkaufer G, Okada M, Singh U, Nozaki T, Mann BJ, Petri WA Jr (2006) Impact of intestinal colonization and invasion on the *E. histolytica* transcriptome. *Molecular and Biochemical Parasitology* **147**:163-176.

Gillespie JJ, Beier MS, Rahman MS, Ammerman NC, Shallom JM, Purkayastha A, Sobral BS, Azad AF (2007) Plasmids and *Rickettsial* Evolution: Insight from *Rickettsia felis*. *PLoS ONE* **2**(3): e266. doi:10.1371/journal.pone.0000266

Gillespie JJ, Johnston JS, Cannone JJ, Gutell RR (2006) Characteristics of the nuclear (18S, 5.8S, 28S, and 5S) and *mitochondrial (16S and 12S) rRNA genes of Apis mellifera (Insecta*: Hymenoptera): Structure, organization and retrotransposable elements. *Insect Mol. Biol.* **15**: 657-686.

Guo M, Rupe MA, Yang X, Crasta O, Zinselmeier C, Smith OS, Bowen B (2006) Genome-wide transcript analysis of maize hybrids: allelic additive gene expression and yield heterosis. *Theor. Appl. Genet.* **113**(5): 831-845.

Hines HM, Hunt JH, O'Connor TK, Gillespie JJ, Cameron SA (2007) A multi-gene phylogeny reveals eusociality evolved twice in vespid wasps. *Proc. Natl. Acad. Sci.* **104**(9): 3295-3299.

Kjer KM, Gillespie JJ, Ober KA (2006) Structural homology in ribosomal RNA, and a deliberation on POY. *Arth. Syst. Phylogenet.* **64**: 159-164.

Kjer KM, Gillespie JJ, Ober KA (2007) Opinions on alignment, and an empirical comparison of repeatability and accuracy between POY and structural alignment. *Syst. Biol.* **56**(1): 133-46.

Ouyang B, Yang T, Li H, Zhang L, Zhang Y, Zhang J, Fei Z, Ye Z (2007) Identification of early salt stress response genes in tomato root by suppression subtractive hybridization and microarray analysis. *Journal of Experimental Botany* **58**(3): 507-520.

Purkayastha A, Crasta O, Eckart JD, Czar M, Meng XJ, Setubal J, Sobral B (2007), *Bioinformatics Resources For The Study Of Viruses At The Virginia Bioinformatics Institute.* In: Structure-based Study of Virus Infection. Cheng RH (ed), Stockholm, World Scientific Publishing Company. ISBN 981-270-405-1.

Snyder EE, Kampanya N, Lu J, Nordberg EK, Karur HR, Shukla M, Soneja J, Tian Y, Xue T, Yoo H, Zhang F, Dharmanolla C, Dongre NV, Gillespie JJ, Hamelius J, Hance M, Huntington KI, Jukneliene D, Koziski J, Mackasmiel L, Mane SP, Nguyen V, Purkayastha A, Shallom J, Yu G, Guo Y, Gabbard J, Hix D, Azad AF, Baker SC, Boyle SM, Khudyakov Y, Meng XJ, Rupprecht C, Vinje J, Crasta OR, Czar KJ, Dickerman A, Eckart JD, Kenyon R, Will R, Setubal JC, Sobral BWS (2007) PATRIC: The VBI PathoSystems Resource Integration Center. *Nucleic Acids Research* **35:** D401-D406 (Database issue).

Wang C, Sobral BW, Williams KP (2007) Loss of a universal tRNA feature. *J. Bacteriol.* **189**(5): 1954-1962.

Wharton RA, Yoder MJ, Gillespie JJ, Patton JC, Honeycutt RL (2006) Relationships of *Exodontiella*, a non-alysiine, exodont member of the family *Braconidae* (*Insecta, Hymenoptera*). *Zool. Scripta* **35**: 323-340.

Williams KP, Sobral BW, Dickerman AW (2007) A robust species tree for the Alphaproteobacteria. *J. Bacteriol.* **189**: 4578-4586.

Yu GX, Snyder EE, Boyle SM, Crasta OR, Czar M, Mane SP, Purkayastha A, Sobral B, Setubal JC (2007) A Versatile Computational Pipeline for Bacterial Genome Annotation Improvement and Comparative Analysis, with *Brucella* as a Use Case *Nucl. Acids Res.* (in press).

Tian Y, Dickerman AW (2007) GeneTrees: a phylogenomics resource for prokaryotes. *Nucleic Acids Research* **35**: D328-D331 (Database Issue).

Zhou CE, Smith J, Lam M, Zemla A, Dyer MD, Slezak T (2006) MvirDB: a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic Acids Research* **35**: D391-D395 (Database Issue).

Zhou CE, Smith J, Lam M, Zemla A, Dyer MD, Kuczmarski T, Slezak T (2006) MannDB: a microbial annotation database. *BMC Bioinformatics* **7**:459.

Zhao Z, Zhang F (2006) Sequence context analysis in the mouse genome: Single nucleotide polymorphisms and CpG island sequences. *Genomics* **87**(1): 68-74.

Other references

Atkins D, Droegemeier K, Feldman S, Garcia-Molina H, Klein M, Messerschmitt DG, Messina P, Ostriker J, Wright M (2003) Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure, National Science Foundation.

Hasan S, Daugelat S, Rao PS, Schreiber M (2006) Prioritizing Genomic Drug Targets in Pathogens: Application to *Mycobacterium tuberculosis*. *PLoS Comput. Biol.* **2**(6):e61.

He YQ, Vines RR, Wattam AR, Abramochkin GV, Dickerman AW, Eckart JD, Sobral BWS (2005) PIML: The Pathogen Information Markup Language. *Bioinformatics 21*, 116-121.

Mantri Y, Williams KP (2004) Islander: a database of integrative islands in prokaryotic genomes, the associated integrases and their DNA site specificities. *Nucleic Acids Res.* **32**:D55-D58.

Brett Tyler

bmtyler@vt.edu

Professor, Virginia Bioinformatics Institute

Professor, Dept. of Plant Pathology, Physiology and Weed Science, Virginia Tech

Group contributors: Felipe Arredondo, Nathan Bruce, Marcus Chibucos, Daolong Dou, Bryndan Durham, Lee Falin, Adriana Fereirra, Nik Galloway, Regina Hanlon, Rays Jiang, Shiv Kale, Konstantinos Krampis, Robert Presler, Brian Smith, Vignesh Sundararajan, Ken Tian, Trudy Torto-Alalibo, Sucheta Tripathy, Lachelle Waller, Xia Wang, Lecong Zhou.

Plant-Pathogen Interactions: From Genome Sequences to Genetic Networks

Abstract. Interconnected genetic regulatory networks govern the interactions of hosts and pathogens as a result of an ongoing co-evolutionary battle between the organisms. We are building data collections and tool sets for dissecting host-pathogen genetic networks, with a principal focus on oomycete pathogens of plants. To catalog the interacting genes we have sequenced the genomes of the oomycetes *Phytophthora sojae*, *Phytophthora ramorum* and *Hyaloperonospora parasitica*, and the fungus *Alternaria brassicicola*. Comparison of these genome sequences has revealed that many genes are evolving unusually rapidly, including a large, diverse set of genes that encode virulence proteins that can enter plant cells. This work has included creating standardized Gene Ontology terms for describing genes of plant pathogens, and using functional linkage graphs to infer the functions of uncharacterized genes. To determine the dynamics of how the pathogen and plant genes interact, we are carrying out gene expression profiling of soybean and its pathogen *P. sojae* using microarrays. To infer genetic regulatory networks from these data, we are using a combination of genetic mapping and statistical inference called genetical genomics. In addition, we are developing novel methods for inferring and modeling biological processes based on functional genomics data, using dimensional reduction and sequential dynamical methods.

Keywords: *Phytophthora*; *Hyaloperonospora*; oomycete; soybean; genome sequences; microarray analysis; regulatory networks; systems genetics; effector proteins; quantitative resistance.

Scientific Progress

Genome sequences of oomycete and fungal plant pathogens

We have completed the determination and analysis of the genome sequences of the oomycetes *Phytophthora sojae* (a soybean pathogen) and *Phytophthora ramorum* (sudden oak death pathogen). Furthermore, we have completed draft genome sequences of the oomycete *Hyaloperonospora parasitica* and the fungus *Alternaria brassicicola*, both pathogens of the model plant *Arabidopsis thaliana*. The genome sequences of these pathogens suggest that the oomycetes rely principally on virulence proteins to attack their hosts, while the fungus relies heavily on toxic metabolites. Both genomes contain significant numbers of genes that appear to be of recent bacterial origin. Approximately 45% of the genes in the oomycete pathogens are diverging so rapidly that the genes cannot be matched between species, presumably as a result of co-evolutionary pressures from different host plant species.

Table 1. Collaborators for genome sequences of oomycete and fungal plant pathogens project.

Collaborators	Research Institution
Jim Beynon	Warwick University, United Kingdom
Jeffrey Boore, Daniel Rokhsar, Igor Grigoriev	Department of Energy Joint Genome Institute
Sandra Clifton	Washington University Genome Sequencing Center
Alan Collmer	Cornell University
Candace Collmer	Wells College
Ralph Dean, David Bird	North Carolina State University
Matteo Garbelotto	University of California, Berkeley
Nik Grunwald	United States Department of Agriculture-Agriculture Research Service, Corvallis, Oregon
Kurt Lamour	University of Tennessee, Knoxville
Frank Martin	United States Department of Agriculture-Agriculture Research Service, Salinas, California
Hayes McDonald	Oak Ridge National Laboratory
John McDowell	Plant Pathology, Physiology and Weed Science at Virginia Tech
TM Murali	Computer Science at Virginia Tech
Nicole Perna	University Wisconsin, Madison
Jane Rogers	Sanger Sequencing Center, Cambridge, United Kingdom
Bruno Sobral, João Setubal, Allan Dickerman, Chris Lawrence	Virginia Bioinformatics Institute
Owen White	The Institute for Genomic Research (TIGR), J Craig Venter Institute

We have used the *P. ramorum* sequence to create an Affymetrix Single Nucleotide Polymorphism (SNP) GeneChip® for the purpose of determining the genetic structure of *P. ramorum* populations. The results showed that there are three major genotypes of *P. ramorum* in the United States that appear to have originated from a sexually reproducing population. However, the pathogen does not appear to be reproducing sexually in the United States.

To enable pathogenicity genes from oomycete and fungal plant pathogens to be compared with genes from other kinds of plant pathogens such as bacteria and nematodes, we are creating standardized Gene Ontology terms to describe the biological processes that these organisms use in their interactions with plants. The terms are also designed to be useful for researchers studying animal pathogens, and to enable genes from plant and animal pathogens to be compared. This year 450 terms were deposited into the Gene Ontology (www.geneontology.org).

Functional genomics of oomycete virulence effector proteins

Pathogens use effector molecules to manipulate the physiology of their hosts in an effort to make the host more susceptible to infection. Effector proteins secreted by oomycete and fungal pathogens have been inferred to enter host cells, but the mechanism of entry is unknown. The genome sequences of *P. sojae* and *P. ramorum* contain over 350 genes having similarity to known oomycete effector genes such as the *P. sojae Avr1b*-1 gene. We have shown that a motif, RXLR, that is conserved among all these candidate effector genes is required for the Avr1b protein to enter plant cells and that the motif functions in the absence of the pathogen. The RXLR motif closely resembles the erythrocyte targeting signal

Table 2. Collaborators for functional genomics of oomycete virulence effector proteins project.

Collaborators	Research Institution
Judith Fliegmann, Jurgen Ebel	Ludwig Maximilian University, Munich, Germany
Mark Gijzen	Agriculture and Agri-Food Canada, London, Ontario
Jonathon Jones	Sainsbury Laboratories, Norwich, United Kingdom
Don Maclean	University of Queensland, Australia
John McDowell	Plant Pathology, Physiology and Weed Science at Virginia Tech
Richard Michelmore	University of California, Davis
Chris Taylor	Donald Danforth Plant Science Center, St. Louis
Yuanchao Wang	Nanjing Agricultural University, China

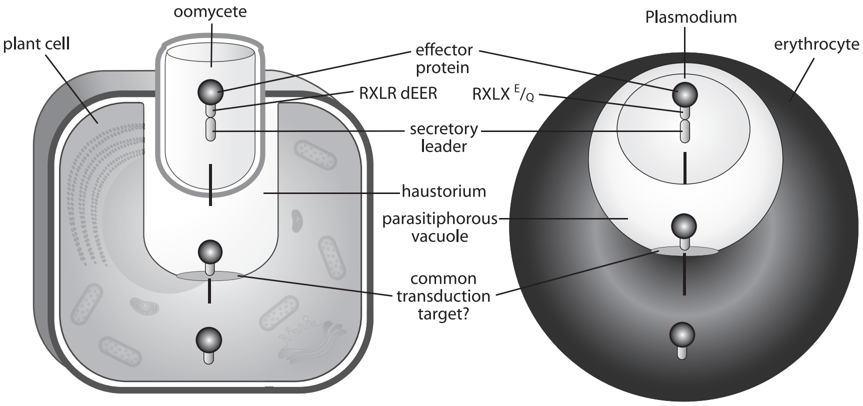


Figure. Comparison of effector protein mechanisms for *Plasmodium* **and** *Phytophthora***.**

of *Plasmodium* effector proteins, which suggests that the machinery of the hosts (soybean and human) targeted by the pathogens is very ancient. We have also shown that approximately half of the RXLR effectors in *P. sojae* contain a C-terminal motif that in *Avr1b*-1 is required for its function.

The genomes of *Phytophthora* and of the malaria parasite *Plasmodium* encode 300-400 proteins with the demonstrated or predicted ability to pass through the surrounding host membrane after secretion. For *Phytophthora* the surrounding host membrane is the haustorial membrane and for *Plasmodium* it is the parastiphorous vacuolar membrane. Both membranes are derived from the host plasma cell membrane as the parasite enters the host cell. The *Phytophthora* and *Plasmodium* proteins require a very similar sequence motif to cross the host membrane. This similarity between two very different pathogens and hosts suggests that the host machinery targeted by the pathogen proteins is very ancient.

Transcriptional interplay between soybean and Phytophthora sojae during infection

We are using Affymetrix GeneChips® to determine the transcriptional profiles of *P. sojae* prior to and during infection, and to simultaneously determine the transcriptional profiles during infection of soybean cultivars having different levels of quantitative resistance against the pathogen. The goals of the study are also to investigate the possible mechanisms of quantitative resistance. The soybean transcriptional profiles reveal that very large numbers (>60%) of genes show significant variation in response to infection.

Furthermore, some genes were uniquely up-regulated in particular resistant cultivars, which suggests that they might be associated with the mechanism of quantitative resistance in those cultivars. The project is currently assaying transcriptional changes in a set of 300 recombinant inbred lines that is segregating for quantitative resistance in order to map the soybean genetic loci responsible for resistance and for any associated transcriptional changes. The project recently completed the assay of 2600 RNA samples on Affymetrix GeneChips® for this mapping.

Genetic network inference

We are using two approaches for inferring genetic networks from systems information such as microarray data. The first is genetical genomics, which exploits randomization of genetic factors e.g. genetic crosses, to make inferences about regulatory genes that influence sets of expressed genes. This approach is being used to infer regulatory networks associated with quantitative disease resistance in soybean. For this, we have developed new methodologies in structural equation modeling, capable of inferring network structures for hundreds of genes and DNA polymorphisms, and have validated them using artificial gene networks, and using published yeast data.

The second approach uses dimensional reduction, such as principal components analysis, combined with multiple regression to build sequential dynamical models of biological systems using time course systems data. We are

Table 3. Collaborators for project on transcriptional interplay between soybean and *Phytophthora sojae* **during infection.**

Collaborators	Research Institution
Anne Dorrance	Ohio State University
Ina Hoeschele	Virginia Bioinformatics Institute
Kurt Lamour	University of Tennessee, Knoxville
Saghai Maroof	Crop, Soil, and Environmental Sciences, Virginia Tech
Hayes McDonald	Oak Ridge National Laboratory
TM Murali	Computer Science, Virginia Tech
Steven St. Martin	Plant Pathology, Physiology, and Weed Science at Virginia Tech

Table 4. Collaborators for genetic network inference project.

Collaborators	Research Institution
Ina Hoeschele, Reinhard Laubenbacher, and Henning Mortveit	Virginia Bioinformatics Institute
Pedro Mendes	Manchester Centre for Integrative Systems Biology, United Kingdom and Virginia Bioinformatics Institute

validating this approach by using microarray data from oxidative stress treatments of yeast and *Arabidopsis* cells.

Conferences and workshops

Rays Jiang RHY, Tyler BM (2006) *Comparative genomics of the oomycete plant pathogen Phytophthora*. Invited talk, 2nd International Fungal Proteomics Symposium, Baltimore, MD, July 30-August 3, 2006.

Tseng TT, Setubal JC, and the PAMGO consortium (Biehl B, Bird D, Chibucos MC, Collmer A, Collmer C, Dean R, Glasner JD, Gwinn-Giglio M, Hannick L,Ireland A, Lindeberg M, Lomax J, Mitchell TK, Perna N, Torto-Alalibo T, Tyler BM, White O) (2006) *Gene Ontology for Plant-Associated Microbes: A. tumefaciens C58*, 27th Annual Crown Gall meeting, Williams College, Williamstown, MA, November 18-19, 2006.

Tyler BM (2006) *Functional Genomics of Phytophthora Effector Proteins*. Invited talk, Warwick University, May 3, 2006.

Tyler BM (2006) *Role of RXLR effectors in Phytophthora sojae virulence and avirulence.* Invited talk, Oomycete Molecular Genetics Workshop, Wageningen, The Netherlands. May 5, 2006.

Tyler BM (2006) *Functional Genomics of Phytophthora Effector Proteins*. Invited talk. Sainsbury Laboratory, John Innes Centre, Norwich, United Kingdom. May 8, 2006.

Tyler BM (2006) *Destroyers from a Distant Kingdom, Systems Biology of Oomycete Pathogens*. Biological Sciences Distinguished Lecture. National Science Foundation. July 31, 2006.

Tyler BM (2006) *Functional Genomics of the Phytophthora sojae-Soybean Interaction*. Invited talk, Mid-Atlantic Plant Molecular Biology Society Conference, August 17-18, 2006.

Tyler BM (2006) *Functional Genomics and Bioinformatics of the Soybean - Phytophthora Interaction*. Invited talk. 24th Stadler Genetics Symposium, Genomics of Disease, University of Missouri, Columbia, October 2-4, 2006.

Tyler BM, Zhang X, Scheuring C, Tripathy S, Xu Z, Wu C, Ko A, Tian K, Arredondo F, Lee MK, Santos FA, Jiang RHY, Govers F, Dehal P, Boore J, Zhang H (2007) *Integrated physical map and genome sequence of the soybean pathogen Phytophthora sojae reveals clustering of pathogenicity genes into regions of elevated genome fluidity.* Invited talk, International Plant and Animal Genome Conference XV, San Diego, January 13-17, 2007.

Hoeschele H, Liu B, de la Fuente A, Mao Y, Zhou L, Hanlon R, Dorrance AE, St. Martin S, Maroof S, Tyler BM (2007) *Inference of Gene Networks Controlling Quantitative Resistance in Soybean against Phytophthora sojae.* Invited talk, International Plant and Animal Genome Conference XV, San Diego, January 13-17, 2007.

Tyler BM, Jiang RHY, Ferreira A, Wang X, Bruce N, Kale SD, Arredondo FD, Dou D (2007) *Bioinformatics and functional genomics of pathogenicity in the soybean pathogen Phytophthora sojae.* Invited talk, International Plant and Animal Genome Conference XV, San Diego, January 13-17, 2007.

Torto-Alalibo T, Collmer C, Gwinn Giglio M, Biehl B, Bird D, Chibucos M, Collmer A, Dean R, Glasner J, Ireland A, Lindeberg M, Lomax J, Mitchell T, Perna N, Setubal J, Tyler BM, White O (2007) *Plant-Associated Microbe Gene Ontology (PAMGO): A Community Resource Of Gene Ontology Terms Describing Gene Products Involved In Microbe-Host Interactions.* Training Workshop, International Plant and Animal Genome Conference XV, San Diego, January 13-17, 2007.

Peer-reviewed publications in reporting period

Jiang RHY, Tyler BM, Govers F (2006) Comparative analysis of *Phytophthora* genes encoding secreted proteins reveals conserved synteny and lineage-specific gene duplications and deletions. *Molecular Plant-Microbe Interactions* **19:** 1311-1321.

Krampis K, Tyler BM, Boore JL (2006) Extensive variation in nuclear mitochondrial DNA content between the genomes of *Phytophthora sojae* and *Phytophthora ramorum*. *Molecular Plant-Microbe Interactions* **19:** 1329-1336.

Tripathy S, Pandey VN, Fang B, Salas F, Tyler BM (2006) VMD: a community annotation database for oomycetes and microbial genomes. *Nucleic Acids Research* **34:** D379-D381.

Tripathy S, Tyler BM (2006) The repertoire of transfer RNA genes is tuned to codon usage bias in the genomes of *Phytophthora sojae* and *Phytophthora ramorum*. *Molecular Plant-Microbe Interactions* **19:** 1322-1328.

Tyler BM (2006) Genomics of fungal plant pathogens. In *Encyclopedia of plant and crop science*, Goodman RM (ed). New York: Marcell Dekker. DOI: 10.1081/E-EPCS-120019942.

Tyler BM, Tripathy S, Zhang XM, Dehal P, Jiang RHY, Aerts A, Arredondo FD, Baxter L, Bensasson D, Beynon JL, Chapman J, Damasceno CMB, Dorrance AE, Dou DL, Dickerman AW, Dubchak IL, Garbelotto M, Gijzen M, Gordon SG, Govers F, Grunwald NJ, Huang W, Ivors KL, Jones RW, Kamoun S, Krampis K, Lamour KH, Lee MK, McDonald WH, Medina M, Meijer HJG, Nordberg EK, Maclean DJ, Ospina-Giraldo MD, Morris PF, Phuntumart V, Putnam NH, Rash S, Rose JKC, Sakihama Y, Salamov AA, Savidor A, Scheuring CF, Smith BM, Sobral BWS, Terry A, Torto-Alalibo TA, Win J, Xu ZY, Zhang HB, Grigoriev IV, Rokhsar DS, Boore JL (2006) *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science* **313:** 1261-1266.

Valer K, Fliegmann J, Frohlich A, Tyler BM, Ebel J (2006) Spatial and temporal expression patterns of *Avr1b*-1 and defense-related genes in soybean plants upon infection with *Phytophthora sojae*. *FEMS Microbiology Letters* **265:** 60-68.

Zhang X, Scheuring C, Tripathy S, Xu Z, Wu C, Ko A, Tian SK, Arredondo F, Lee MK, Santos FA, Jiang RHY, Zhang HB, Tyler BM (2006) An integrated BAC and genome sequence physical map of *Phytophthora sojae*. *Molecular Plant-Microbe Interactions* **19:** 1302-1310.

Martin FN, Bensasson D, Tyler BM, Boore JL. (2007) Mitochondrial genome sequences and comparative genomics of *Phytophthora ramorum* and *P. sojae*. *Current Genetics* **51:** 285-296.

Torto-Alalibo T, Tripathy S, Smith BM, Arredondo FD, Zhou L, Li H, Chibucos MC, Qutob D, Gijzen M, Mao C, Sobral BWS, Waugh ME, Mitchell TK, Dean RA, Tyler BM. (2007) Expressed sequence tags from *Phytophthora sojae* reveal genes specific to development and infection. *Molecular Plant-Microbe Interactions* **20**: 781-793.

Tyler BM (2007) *Phytophthora sojae*: root rot pathogen of soybean and model oomycete. *Molecular Plant Pathology* **8:** 1-8.

2007 Research Reports

from the
Virginia Bioinformatics Institute's
Faculty Fellows

John Tyson

tyson@vt.edu

University Distinguished Professor, Dept. of Biological Sciences,
Virginia Tech

Group contributors: Paul Brazhnik, Shenghua Li, Tongli Zhang, Rajat Singhania.

Simulation and Analysis of Molecular Regulatory Systems in Cell Biology

Abstract. The grand challenge of post-genomic cellular biology is to assemble a working model of a living cell, a model that gives a reliable account of how the physiological properties of a cell derive from its underlying molecular machinery. Complex networks of interacting proteins control the physiological properties of a cell (metabolism, reproduction, motility, signaling, etc.). Using basic principles of biochemical kinetics, we convert network diagrams into sets of nonlinear ordinary differential equations (chemical rate equations) and then explore the models by analytical and computational methods. Of particular interest to us are the mechanisms that control cell growth, division and death in eukaryotes (yeasts, plants, insects, vertebrates). During this reporting period, we published a mathematical model of the protein interaction network (PIN) that is thought to direct DNA synthesis and nuclear division in all eukaryotic cells. By numerical simulations and theoretical analysis, we showed that this model successfully accounts for the properties of cell growth and division in many types of cells. A second paper investigated the cell's response to DNA damage and how the decision is made to attempt repair or to trigger cell death. A third paper studied improvements of the mathematical representation of PINs.

Keywords: network dynamics; cell division cycle; bifurcation analysis.

Scientific Progress

A generic model of the eukaryotic cell cycle

DNA synthesis and nuclear division are controlled in all eukaryotic cells by a universal molecular mechanism, a mechanism that is based on the activation and inactivation of powerful enzymes, the cyclin-dependent kinases. By combining information from biochemical and genetic studies of a wide range of eukaryotic organisms, we have proposed a generic network of interacting genes and proteins that can account for the commonalities and idiosyncrasies of the cell division cycles of all these organisms. The mechanism is common, we claim, but the specific settings of the reaction rate constants vary from one organism to another. These settings are determined by the genetic makeup of the organism.

To support these claims, we convert the generic network into a set of governing differential equations and provide values for the reaction rate constants specific to budding yeast (see figure), fission yeast, frog eggs, and cultured mammalian cells. Numerical simulations of the kinetic equations are compared in detail with the observed properties of cell division in these various cell types and in mutant cells that are derived from the normal cells by knocking out or over-expressing the genes that govern the control system.

Using one-parameter bifurcation diagrams, we show how overall cell growth drives progression through the cell cycle, how the cyclin-dependent kinase network regulates cell size, and how mutations change the dynamics of the control system and create unusual cell-cycle phenotypes. The relation between gene dosage and cell phenotype can be summarized compactly in two-parameter bifurcation diagrams.

Currently we are pursuing studies of mammalian cell cycle regulation in collaboration with the Integrative Cancer Biology Centers at the Massachusetts Institute of Technology (MIT), Vanderbilt University and Duke University.

Cellular responses to DNA damage

After exposure to an agent that damages DNA, a cell must respond by blocking progression through the cell cycle, by attempting to repair the damage, and (if unsuccessful in repair) by initiating a sequence of programmed cell death. The p53 protein coordinates these responses. If the p53 gene is destroyed by mutation, then the mutant cell cannot respond properly to DNA damage and, consequently, may pass the damage to its progeny, which may transform the damaged cell into a malignant cancer.

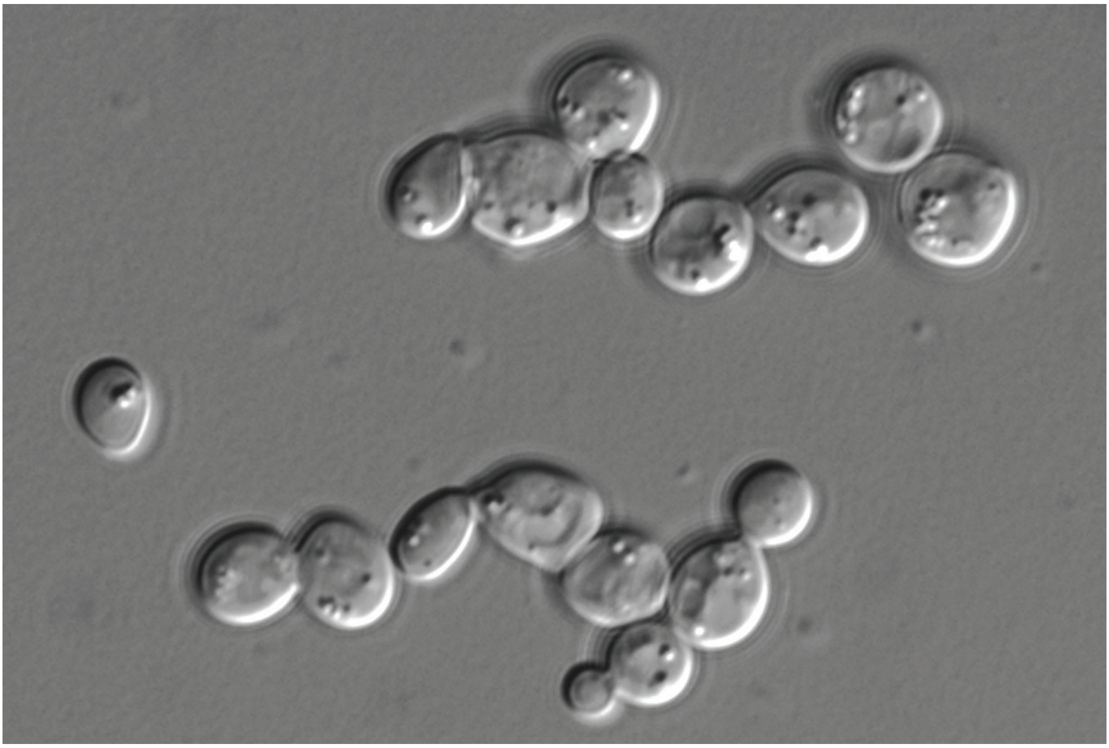


Figure. Differential interference contrast microscopy image of budding yeast (courtesy of Jean Peccoud, Virginia Bioinformatics Institute).

A complex gene-protein interaction network that is not yet well understood controls the functions of p53 protein. We have proposed several possible network topologies and studied the dynamics of the hypothetical control systems by numerical simulation and bifurcation analysis of their governing differential equations. Our models describe several possible mechanisms by which cells might generate periodic pulses of p53 activity, as observed recently by Uri Alon's laboratory in Israel (Geva-Zatorsky et al, 2006). We have also proposed mechanisms whereby the cell can 'count' p53 pulses and decide whether to arrest cell division and repair the damage (low count) or to abort the repair and trigger programmed cell death (high count).

Using the same theoretical approach, we are currently building models of the cell-death response. We plan to combine our models of cell-cycle regulation, DNA-damage signaling, p53 oscillations, and cell death to produce a comprehensive theoretical understanding of the regulation of mammalian cell replication under normal and stressed conditions. Ultimately these models are intended to shed light on the mechanisms of carcinogenesis and chemotherapy.

Simplifying kinetic equations by the quasi-steady state approximation

Our approach to molecular systems biology (to study mathematical models of complex reaction networks) leads to a large number of differential equations (one for every chemical species in the network) with many terms on the right-hand-side of the equations (one for every reaction in which the species participates). Some of these reactions are fast, some are slow, and some are in-between. Large differences of time scales in the network create huge difficulties for simulating the temporal evolution of the network and for understanding the basic principles of its operation.

To sidestep these problems, theoreticians often use the quasi-steady state approximation (QSSA) to eliminate the fastest and slowest variables in the system of differential equations. For example, the Michaelis-Menten rate law provides an elegant and useful description of the kinetics of enzyme-catalyzed reactions ($E + S \leftrightarrow C \rightarrow E + P$), but it is valid only when total enzyme concentration ($E + C = E_T$ = constant) is much less than the initial substrate concentration (S_0). For protein interaction networks, it is unjustified to use Michaelis-Menten kinetics because the components of the network often swap roles as enzymes and substrates. For instance, two kinases may phosphorylate each other, in which case it cannot possibly be true that $E_T << S_0$ for both reactions simultaneously. In this work, we show how to formulate the QSSA properly for protein interaction networks, and we apply the alternative formalism, the total QSSA, to a model of the G2/M transition in the eukaryotic cell cycle. Whereas the classical Michaelis-Menten rate law fails to represent the dynamics of this network correctly, the total QSSA captures the behavior with economy and accuracy.

Currently, we are reformulating our cell cycle control models in terms of the total QSSA equations. We will use the reformulated models to compute stochastic properties of cell cycle control networks. The outcome will be compared with experimental measurements of fluctuating control proteins in single yeast cells, work which is being carried out in Jean Peccoud's lab at the Virginia Bioinformatics Institute.

Conferences and workshops

Tyson JJ (2006) *Understanding complex systems*. University of Illinois, Urbana IL, 15-18 May, 2006.

Tyson JJ (2006) Society for Research on Biological Rhythms, Sandestin FL, 21-25 May, 2006.

Tyson JJ (2006) *Predictive models of complex systems*, NIGMS Centers of Excellence in Complex Systems Biology, Friday Harbor WA, 3-6 June, 2006.

Tyson JJ (2006) International Conference on Systems Biology, Yokohama, Japan, October 9-13, 2006.

Tyson JJ (2007) EMBO/FEBS Workshop on Systems Dynamics of Intracellular Communication, Jerusalem, Israel, 18 – 22 March, 2007.

Invited talks were also given at the University of British Columbia, University of Washington, University of Toronto, Max Planck Institute for Molecular Cell Biology (Dresden), Institute for Advanced Biosciences (Japan), University of Tokyo, Massachusetts Institute of Technology (MIT), Princeton University, Vanderbilt University School of Medicine, Duke University School of Medicine.

Peer-reviewed publications in the reporting period

Csikasz-Nagy A, Battogtokh D, Chen KC, Novak B, Tyson JJ (2006) Analysis of a generic model of eukaryotic cell cycle regulation. *Biophys. J.* **90**(12): 4361-4379.

Ciliberto A, Capuani F, Tyson JJ (2007) Modeling networks of coupled enzymatic reactions using the total quasi-steady state approximation. *PLoS Comp. Biol.* **3**(3): e45.

Csikasz-Nagy A, Gyorffy B, Kapuy O, Tyson JJ, Novak B (2007) Modeling the septation initiation network (SIN) in fission yeast cells. *Curr. Genet.* **51**(4): 245-255.

Zhang T, Brazhnik P, Tyson JJ (2007) Exploring mechanisms of the DNA damage response: p53 pulses and their possible relevance to apoptosis. *Cell Cycle* **6**(1): 85-94.

Other cited publications

Geva-Zatorsky N, Rosenfeld N, Itzkovitz S, Milo R, Sigal A, Dekel E, Yarnitzky T, Polak P, Liron Y, Kam Z, Lahav G, Alon U (2006) Oscillations and variability in the p53 system. *Molecular Systems Biology*. doi:10.1038/msb4100068