Scientific Annual Report
2008

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.

Faculty Fellow Report

The Virginia Bioinformatics Institute (VBI) was founded in 2000 to become a premier institute for internationally competitive research, education, and outreach. In just under eight years, we have developed strong research capabilities in the life sciences and achieved a solid foundation for further growth. Much of this success has been due to our ability to shape VBI into a problem-solving entity on the move, an institute focused on transdisciplinary scientific collaboration.

In this report, readers will be able to share some of the exciting scientific achievements from VBI faculty that have taken place in the past twelve months. The reports presented in this publication give an indication of the depth and breadth of research underway at the institute. VBI continues to develop high quality, internationally competitive research areas as diverse as bioinformatics, systems biology, high-performance computing, complexity science, and policy informatics. We have provided a strong framework for the future development of the institute by nurturing scientific innovation, recruiting talented faculty and staff, and establishing a portfolio of competitive research programs.

Today, VBI is at the cusp of an expansion — a transition into programmatic areas that will help develop and create new opportunities for our unique environment of transdisciplinary research. Plans are being finalized for the construction of a new 50,000-square-feet addition to our main building on the Virginia Tech campus. The three-story extension will include office space that supports the relocation of VBI's Network Dynamics and Simulation Science Laboratory from the Corporate Research Center to the VBI main building as well as the anticipated expansion of cyberinfrastructure and other research group activities.

We will continue to set ambitious goals as we enter the next stage of the institute's development. I firmly believe that VBI is ideally poised to take advantage of the new opportunities that will arise in infectious disease research, public policy, economic development, and science education.

I would like to take this opportunity to thank you for your interest in the activities of the Virginia Bioinformatics Institute.

Sincerely,

Bruno Sobral
Executive and Scientific Director
Virginia Bioinformatics Institute

2008 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
Director, Network Dynamics and Simulation Science Laboratory at
Virginia Bioinformatics Institute

Network Dynamics and Simulation

Abstract. The Network Dynamics and Simulation Science Laboratory (NDSSL) has made substantial progress in basic science and the development of usable tools to study large complex systems. This has resulted in funded programs of more than $15 million in the past three years and includes programs with the National Institutes of Health (NIH), the Centers for Disease Control and Prevention (CDC), the Department of Defense (DoD) and the National Science Foundation (NSF). We have established a presence in the National Capital Region and play a leading role in a new institutional initiative in Policy Informatics for Complex Systems. We are pursuing new programs in wireless networks, commodity markets and high-performance computing (HPC). NDSSL continues to develop advanced HPC-based computational tools and methods for reasoning about complex systems. These resources are integrated into web services providing synthetic databases, national scale interaction-based simulations, and analysis tools. We have used these tools in several stakeholder-designed studies supporting policy planning for pandemics. Two large studies to support pandemic planning for military preparedness were completed for DoD's Defense Threat Reduction Agency (DTRA) using a prototype system we have built — the Comprehensive National Incident Management System (CNIMS). CNIMS integrates surveillance, simulation-assisted hypothesis testing, and decision support for use in situational awareness and planning in complex systems. CNIMS was presented at a technology forum on March 2008, organized by the Senate Armed Services Committee's Subcommittee on Emerging Threats and Capabilities; CNIMS was also featured in a workshop sponsored by DTRA on 15-16 April, 2008. We have continued to develop highly efficient algorithms for several combinatorial and dynamical problems arising from the study of complex networks.

Keywords: interaction-based computing; policy informatics; sustainable planning; 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.

Group contributors: Karla Atkins, Kofi Adasi, Ashwin Aji, Andrea Apolloni, Rahul Bajaj, Manu Bansai, Chris Barrett, Richard Beckman, Sanket Bedare, Keith Bisset, Deepti Chafekar, Karthik Channakeshava, Garima Chaturvedi, Jiangzhuo Chen, Abhijit Deodhar, Tridib Dutta, Stephen Eubank, Annette Feng, Xizhou Feng, Shilpi Gupta, Ginger Hansen, Steven Harris, Ken Kania, Maleq Khan, Chris Kuhlman, Ajit Kulkarni, Jonathan Leidig, Bryan Lewis, Yi Ma, Matthew Macauley, Achla Marathe, Madhav Marathe, Seret Medhani, Henning Mortveit, Kota Nagarjun, Ganesh Narayanaswamy, Elaine Nsoesie, Imran Pirwani, Joyce Randall, Sharon Smyth, Aravind Srinivasan, Paula Stretz, Jyotirmaya Tripathi, Amit Upadhyay, Anil Vullikanti, Xiaoyu Zhang.

Scientific Progress

The Network Dynamics and Simulation Science Laboratory (NDSSL) continues to make significant scientific progress. 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. Simfrastructure forms the basis of a high-performance computing modeling and simulation environment that is the core of the Comprehensive National Incident Management System (CNIMS) system. The CNIMS proof-of-concept prototype is being built for the Defense Threat Reduction Agency (DTRA) and provides the United States military with a unique tool for planning, situational awareness and consequence management in the event of a large-scale crisis affecting urban areas. DIDACTIC is a first prototype tool that we have built as a part of the CNIMS technology. The tool is being configured for use in military preparedness studies in the event of pandemics.

NDSSL has also continued the development of Simdemics - a scalable high-performance computing-based environment for general reaction diffusion systems. New capabilities include handling complex intervention strategies, representing dynamically co-evolving social networks and improved scaling performance. EpiFast and EpiSimdemics are two specific codes integrated within the framework. Both these simulations are based on provable theoretical properties of the algorithms and are extremely efficient. We have also developed an extension of EpiFast called Dynamic-EpiFast that is capable of handling certain kinds of adaptive interventions.

We have continued to refine the synthetic population of 250 million individuals endowed with daily activity patterns performed at specific, real locations. New methods were developed for refining activities for subpopulations such as military populations and bases. These synthetic data sets played a crucial role in our studies for the Department of Defense (DoD) and National Institutes of Health (NIH). We have also enhanced two tools that we developed over the last few years: TIGS - a tool for finding optimal sequestration group sizes in the event of pandemics, and GALIB - a network analysis toolkit, which is designed to handle large complex networks.

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 under more realistic models of interference. The work also yields, as corollaries, new cross-layer protocols for wireless networks

- A new, quantifiable definition of locational market power was proposed for a market with physical constraints such as the electricity market. A detailed analysis of the topological cause of the market power shows that strategic alliances among small generators with geographical advantage can lead to significant amounts of market power

- New stability results for finite dynamical systems used in functional gene annotation, as well as new results characterizing long term behavior in stochastic dynamical systems

Programmatic progress

The work of NDSSL has resulted in progress in several key program areas. NDSSL received a new NSF grant to study the co-evolution of individual behavior, social networks and public policy to support public health epidemiology. We continue to work with DTRA on modeling the spread of infectious diseases related to civil and military planning for situation assessment and response analysis. We also continue to work on a joint project as a part of the Centers for Disease Control and Prevention (CDC), Center of Excellence in Public Health Informatics, led by the University of Utah School of Medicine. NDSSL continues to contribute as a principal institution in the NIH Modeling Infectious Disease Agent Study (MIDAS) project.

In March 2008, we submitted disclosures for eleven new, interrelated inventions to Virginia Tech Intellectual Property (VTIP).   These disclosures formed the basis for an application filed by VTIP for a single comprehensive provisional patent.

National and International Leadership

In September 2007, the Virginia Tech Complexity Science Academy and NDSSL organized and chaired the "Policy Informatics in an Interdependent World" conference in the National Capital Region. This conference brought together national and international invited experts from academia and government to discuss the emerging role of science and computing to support important global policy questions in public health, energy and sustainable urban development.

NDSSL has also hosted a training workshop for DTRA and personnel from operational military commands on the DIDACTIC tool developed by our team. In addition, our group has also been focusing on modeling malaria and hosted a workshop "Malaria Models for Policy Informatics - Taking Research into Practice" with participation from international experts.

We have been advising a NIH Steering committee for developing modeling methods for contagious diseases. Members of our group have participated in a European Union program review and planning committee on complexity science. We have served on an external advisory committee for the Institute of Scientific Exchange in Torino, Italy and on the External Advisory Board of the National Center for Advance Secure Systems at the National Center for Supercomputing Applications. A member of the NDSSL team also served on the Language Foundation Selection Committee for the Lagrange Prize.

The NDSSL team members participate on various program committees including, but not limited to the following: International Conference of Distributed Computing and Networking; International Conference on High Performance Computing; International Workshop on Foundations of Mobile Computing; Fourth Workshop on Approximation and Online Algorithms; International Conference on Parallel Processing; International Conference Parallel and Distributed Systems; IEEE 22nd International Conference on Advanced Information Networking and Applications; IEEE VTC 2008: 67th Vehicular Technology Conference; 28th International Conference on Computer Communications; 5th International Workshop on Foundations of Mobile Computing; 10th International Conference on Bio-Inspired Models of Network; Information and Computing Systems; and the 3rd IEEE International Symposium on Pervasive Computing and Ad Hoc Communication.

NDSSL team members have participated in a wide variety of panels, including the following: NIH Blue Ribbon Panel for the Risk Assessment of the National Emerging Infectious Diseases Laboratory; Virginia Tech College of Science Complexity Group; NIH Director's New Innovator Award Review Panel; Editorial Board

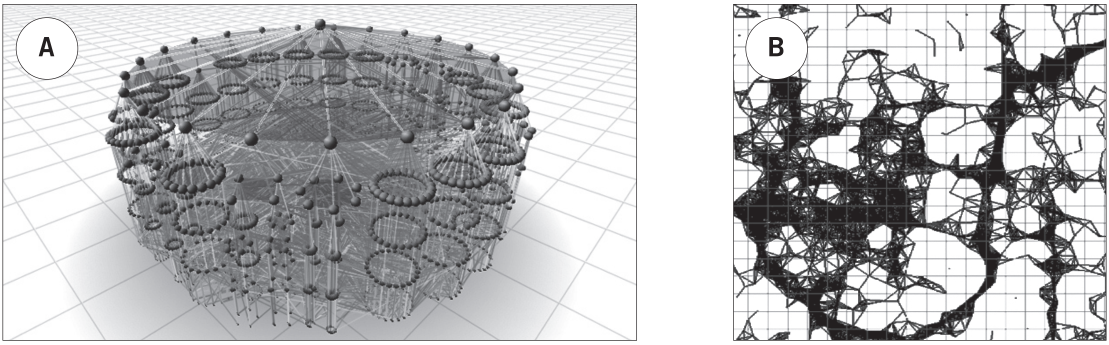


Figure. Two different networks derived from urban mobility models. (A) A social contact network of a synthetic individual in Chicago. Credit: San Diego Supercomputer Center. (B) A synthetic vehicular adhoc network formed by assigning digital devices to individual vehicles on the road.

for The Open Environmental Journal; American Mathematical Society; NSF review panels; Society for Industrial and Applied Mathematics; and External Advisory Board of the National Center for Advanced Secure Systems Research; and the National Center for Supercomputing Applications.

Outreach and Education

NDSSL had significant achievements in Education and Outreach in the reporting period. One of the goals has been to foster interest in science as well as expanding our capabilities in research. Our group has been involved in approximately thirty outreach and educational activities and presentations. These presentations have included involvement in the 2007 Virginia Tech Symposium on Wireless Personal Communications Summer School, talks to the senior management at DTRA on a new vision to support DTRA Reachback that is based on high-performance ubiquitous computing, public lecture on Transportation and Infectious Disease at the Marian Koshland Science Museum of the National Academy of Sciences, a podcast for the American Society of Microbiology, and other invited presentations to places such as Eastern Virginia Medical School, Center for Discrete Mathematics and Theoretical Computer Science (DIMACS) and the Gates Foundation.

NDSSL participated in VBI High School Days providing background talks on theory and epidemiology with hands-on experiments for pandemic flu in the New River Valley. In addition, one of our group mentored Virginia Tech students at the Mathematical Contest in Modeling 2008 competition at Virginia Tech. We also co-organized a mini-symposium within CanaDAM 2007 (1st Canadian Discrete and Algorithmic Mathematics Conference Center, Alberta, Canada) on the topic of Discrete Dynamical Systems over Graphs.

Over the past year, we have mentored and supervised approximately 25 students ranging from high school interns, undergraduate and graduate students pursuing both Master's and PhD degrees. The group has served on committees for several Master's students and PhD students.

Conferences and workshops

Adasi K (2007) Modeling Malaria and other Vector-borne Diseases: A Complex Systems Perspective, *VBI Annual Research Symposium*, Pembroke, VA, September 6-7, 2007.

Adasi K, Eubank S, Marathe A, Mortveit H (2008) Malaria Modeling - A Global Perspective, Invited presentation to *MIDAS Network Meeting*, Bolger Center, Potomac, MD, January 15-16, 2008.

Barrett C (2007) An Analysis of Layered Public Health Interventions of Ft. Lewis and Ft. Hood During a Pandemic Influenza Event, Invited Presentation to Defense Threat Reduction Agency, July 30, 2007.

Barrett C (2007) Interaction Based Models for Epidemiology, Invited presentation at Virginia Tech Department of Entomology, Blacksburg, VA, September 20, 2007.

Barrett C (2007) Interaction-Based Modeling of Population Dynamics and Infectious Disease, Invited Presentation to *Frontiers in Transportation: Social Interactions* Amsterdam, Netherlands, October 12-18, 2007.

Barrett C (2007) Issues of National Guard Force Readiness in Alabama during an Influenza Epidemic, Invited presentation to Defense Threat Reduction Agency in Washington, DC, November 1, 2007.

Barrett C (2007) Transdisciplinary Science and Interaction Based Modeling of Very Large Complex Systems, IEEE Virginia Mountain Section at Holiday Inn, Roanoke Airport, Roanoke, VA, November 15, 2007.

Barrett C, Hosig K, Marathe A, Stretz P (2007) Informatics-Enabled Analysis of the Effects of Built Environment on Obesity, *Deans' Forum on Health, Food and Nutrition*, Virginia Tech, Blacksburg, VA, November 5, 2007.

Barrett C, Marathe M, Fox G, Myers D (2008) Complex Information Integration Institute, Invited presentation to Office of Director of Defense Research and Engineering, Washington, DC, January 15, 2008.

Bisset K (2007) A Scalable Data Management tool to Support Epidemiological Modeling of Large Urban Regions, Poster Presentation to *11th European Conference on Research and Advanced Technology for Digital Libraries*, Budapest, Hungary, September 16-21, 2007.

Chafekar D (2007) Cross-layer latency minimization in wireless networks with SINR constraints, *8th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, Montreal, Quebec, Canada, September 9-14, 2007.

Chen J (2007) EpiFast: Fast Simulation for Epidemics on Networks, *VBI Annual Research Symposium*, Pembroke, VA, September 6-7, 2007.

Chen J, Macauley M, Marathe A (2007) Role of Network and Production Capacity in Allocating Market Power, *Trans-Atlantic INFRADAY Conference on Applied Infrastructure Modeling and Policy Analysis*, College Park, MD, November 2, 2007.

Eubank S (2007) Network Based Epidemiology or Epidemiologically Based Networks? Invited Presentation at *BioWire '07*, Cambridge England, April 5, 2007.

Eubank S (2007) Comparing Individual and Collective Models of Infectious Disease Spread, Invited Presentation at *1st Atlantic Coast Conference on Mathematics in the Life and Biological Sciences*, Blacksburg, VA, May 6, 2007.

Eubank S (2007) The Role of Biological Metaphor in Non-biological Systems, Invited presentation to *First Annual Conference of the International Technology Alliance*, University of Maryland, College Park, MD, September 26, 2007.

Eubank S (2007) Lessons Learned in Developing Scalable Agent-Based Epidemiological Simulation, Eastern Virginia Medical School in Norfolk, VA, November 19, 2007.

Eubank S (2007) Simulations: What level of complexity is appropriate?
Invited presentation for *Advances in Modelization for Infectious Diseases*, Fondation Merieux, Annecy, France, December 10, 2007.

Eubank S (2008) Comparing Models, Invited presentation to *MIDAS Network Meeting*, Bolger Center, Potomac, MD, January 15, 2008.

Eubank S (2008) Spatial Dynamic Modeling, *Gates Foundation Malaria Modeling meeting*, Seattle, WA, March 10, 2008.

Eubank S (2008) Transportation and the Spread of Disease or What is the Contribution of Modern Transportation Systems to the Spread of Disease? Marian Koshland Science Museum, National Academy of Science, Washington, DC, March 5, 2008.

Lewis B (2007) In Silico Surveillance: Enhancing Surveillance System Design through Computer Simulation, Invited Presentation to *2007 International Society for Disease Surveillance Sixth Annual Conference*, Indianapolis, IN, October 11-12, 2007.

Lewis B (2007) In Silico Epidemology: Computer Simulations for Public Health Research, *The Genetics, Bioinformatics, and Computational Biology (GBCB) Seminar Series at Virginia Bioinformatics Institute of Virginia Tech*, Blacksburg, VA, August 30, 2007.

Lewis B (2007) High Performance Computing for Interaction-based Complex Systems, *VBI Annual Research Symposium*, Pembroke, VA, September 2007.

Marathe M (2007) High Performance Computing Enabled Modeling of Coupled Co-Evolving Socio-Technical Networks, Invited Presentation to *Frontiers in Transportation: Social Interactions*, Amsterdam, Netherlands, October 12-18, 2007.

Marathe M (2007) An Integrated Approach to Computing Capacity and Developing Cross Layer Protocols for Wireless Networks, *NSF NeTS PI meeting*, Chicago, IL, July 11-13, 2007.

Marathe M (2007) Deriving Mobility Models from First Principles and their Applications, Invited Presentation at *NSF Mobility in Wireless Networks Workshop*, Rutgers University, North Brunswick, NJ, July 31-August 1, 2007.

Marathe M, Lewis B (2007) Effect of Public Policies and Individual Behavior on the Co-evolution of Social Networks and Infectious Disease Dynamics, *DIMACS/DyDAn Workshop on Computational Methods for Dynamic Interaction*, Rutgers University, NJ, September 24-26, 2007.

Marathe M (2008) Computational Epidemiology: An Interaction Based Approach, Invited presentation at National Institute of Virology, Pune, India, February 2008.

Marathe M (2008) Building Virtual Cities: Computational Aspects of Modeling Large Societal Infrastructure Systems, Invited Colloquium, School of Computing and Indian Institute of Technology (IIT), Bombay, India, February 2008.

Mortveit H (2007) Graph Dynamical Systems and their Applications to Modeling, Invited presentation at the seminar series at the Center for Computational Science, Tulane University, October 2007.

Mortveit H (2007) Discrete Dynamical Systems over Graphs, Organizer of mini-symposium at *CanaDAM'07*, Banff, Canada, May 28-31, 2007.

Mortveit H, Macauley M (2007) Phase Space Equivalences of Sequential Dynamical Systems, Invited presentation to *13th International Workshop on Cellular Automata* in Toronto, Canada, August 27-29, 2007.

NDSSL faculty and staff (2008) DIDACTIC Training for DTRA at the Network Dynamics and Simulation Science Laboratory, NDSSL Blacksburg, VA, February 26-28, 2008.

Srinivasan A (2007) Packet-Scheduling for Latency Minimization, *Battery-Efficient, Robust, Lightweight, Ubiquitous Sensors for Cost-Effective Infrastructures*, Bertinoro, Italy, August 19-24, 2007.

Srinivasan A (2007) Randomized Algorithms and Probabilistic Analysis in Wireless Networking, Invited lecture, *Fourth Symposium on Stochastic Algorithms, Foundations, and Applications (SAGA)*, ETH Zurich, Switzerland, September 2007.

Srinivasan A (2007) Improved Algorithmic Versions of the Lovàsz Local Lemma, Invited lecture, Indian Institute of Technology, Madras, India. December 2007.

Srinivasan A (2008) Network Design and Optimization in Wireless Networking, Invited Lecture, Network Design Workshop, *Ninth INFORMS Telecommunications Conference*, University of Maryland, College Park, MD, March 27-29, 2008.

Vullikanti A, Marathe M (2007) Coupled Co-evolving Social and Communication Networks, Summer School *at 2007 Virginia Tech Symposium on Wireless Personal Communications*, Virginia Tech, Blacksburg, VA, June 6-8, 2007.

Vullikanti A (2007) The Capacity of Wireless Networks: Provable Algorithms for Joint Optimization of Transport, Routing and MAC layers in Wireless Ad Hoc Meeting, Invited presentation to DIAL M-PMOC 2007, Portland, OR, August 16, 2007.

Vullikanti A (2007) Algorithmic Aspects of Capacity of Wireless Networks, Invited presentation to *Battery-Efficient, Robust, Lightweight, Ubiquitous Sensors for Cost-Effective Infrastructures*, Bertinoro, Italy, August 19-24, 2007.

Peer-reviewed publications in reporting period

Atkins K, Marathe A, Barrett C (2007) A computational approach to modeling commodity markets. *Computational Economics* **30**(2): 125-142.

Atkins K, Barrett C, Beckman R, Bisset K, Chen J, Eubank S, Lewis B, Marathe A, Marathe M, Mortveit H, Stretz P, Vullikanti A (2007) An analysis of layered public health interventions at Ft. Lewis and Ft. Hood during a pandemic influenza event. *NDSSL Technical Report No. 07-019*.

Atkins K, Chen J, Kumar A, Marathe A (2008) Structure of electrical networks: A graph theory based analysis. *International Journal of Critical Infrastructure*. In press.

Atkins K, Barrett C, Beckman R, Bisset K, Chen J, Eubank S, Feng A, Feng X, Harris S, Lewis B, Kumar A, Marathe M, Marathe A, Mortveit H, Stretz P (2008) An interaction based composable architecture for building scalable models of large social, biological, information and technical systems. *CT Watch* **4**(1): 46-53.

Atkins K, Chen J, Kumar VS, Macauley M, Marathe A (2008) Locational market power in network constrained markets. *Journal of Economic Behavior and Organization*. In press.

Barrett C, Hunt III H, Marathe M, Ravi S, Rosenkrantz D, Stearns R, Thakur MR (2007) Predecessor existence problems for finite discrete dynamical systems. *Theoretical Computer Science* **386**(1-2): 3-37.

Barrett C, Bisset K, Konjevod G, Marathe M, Wagner D (2008) Engineering label-constrained shortest-path algorithms. In *Proceedings of the Ninth DIMACS Implementation Challenge on Shortest Paths (DIMACS 2006),* AMS.

Barrett C, Bisset K, Konjevod G, Marathe M, Wagner D (2008) Engineering label-constrained shortest-path algorithms. In *Proceedings of the Fourth International Conference on Algorithmic Aspects in Information and Management (AAIM 2008),* June 2008, Lecture Notes in Computer Science, Springer Verlag.

Barrett C, Eubank S, Marathe M (2008) An interaction-based approach to computational epidemiology. *23^rd^ Association for the Advancement of Artificial Intelligence*, Chicago, IL, July 13-17. In press.

Barrett C, Eubank S, Lewis B, Marathe M (2008) Information systems for detection and management of pandemics. In *Encyclopedia of Geographic Information Systems*. Shkhar S, Xiong X (eds), Springer Verlag. In press.

Chafekar D, Kumar VS, Marathe M, Parthasarathy S, Srinivasan A (2007) Cross-layer latency minimization in wireless networks with SINR constraints. In *Proceedings of the 8th ACM International symposium on Mobile ad hoc networking and computing*, ACM, New York, NY, USA, Montreal, Quebec, Canada.

Chafekar D, Kumar VS, Marathe M, Parthasarathy S, Srinivasan A (2008) Approximating the capacity of wireless networks with SINR constraints. *27th IEEE International Conference on Computer Communications (INFOCOM)*. In press.

Chafekar D, Levin D, Parthasarathy S, Kumar VS, Marathe M, Srinivasan A (2008) On the capacity of asynchronous random-access wireless networks. *27th IEEE International Conference on Computer Communications (INFOCOM)*. In press.

Chen J, Macauley M, Marathe A (2007) Role of network and production capacity in allocating market power, *Trans-Atlantic INFRADAY, Conference on Applied Infrastructure Modeling and Policy Analysis, Critical Infrastructures in Energy and Other Networked Industries,* College Park, MD, November 2, 2007.

Choi Y, Khan M, Kumar A, Pandurangan G (2008) Work-efficient distributed Euclidean minimum panning tree construction: tight bounds and algorithms. *20th ACM Symposium on Parallelism in Algorithms and Architectures*, Munich, Germany, June 14-16. In press.

Del Valle SY, Hyman JM, Hethcote HW, Eubank SG (2007) Mixing patterns between age groups in social networks. *Social Networks* **29**: 539-554.

Eubank S, Kumar VS, Marathe M (2008) Epidemiology and wireless communication: Tight analogy or loose metaphor. In *Bio-Inspired Computing and Communication,* Lecture Notes in Computer Science, Springer Verlag.

Feng W, Feng X, Ge R (2008) Green supercomputing comes of age*, IT Professional*, January/February 2008.

Gopalakrishnan V, Morselli R, Bhattacharjee B, Keleher P, Srinivasan A (2007) Distributed Ranked Search. *Proceedings of the Annual International Conference on High Performance Computing (HiPC),* pp. 7-20.

Halloran M, Ferguson N, Eubank S, Longini Jr I, Cummings D, Lewis B, Xu S, Fraser C, Vullikanti A, Germann T, Wagener D, Beckman R, Kadau K, Barrett C, Macken C, Burke D, Cooley P (2008) Modeling targeted layered containment of an influenza pandemic in the United States. *Proceedings of the National Academy of Sciences USA* **105**(12): 4639-4644.

Istrate G, Marathe M, Ravi S (2008) Adversarial scheduling analysis of game-theoretic models of norm diffusion, *Computability in Europe 2008, Logic and Theory of Algorithms,* University of Athens, Greece, June 15-20, 2008.

Khan M, Pandurangan G, Kumar AVS (2007) A simple randomized scheme for constructing low-weight k-connected spanning subgraphs with applications to distributed algorithms. *Theoretical Computer Science* **385**(1-3): 101–114.

Khan M, Pandurrangan G, Kumar VS (2008) Distributed algorithms for constructing approximate minimum spanning trees with applications to wireless sensor networks. *IEEE Transaction on Parallel and Distributed Systems (TPDS).* In press.

Kumar A, Marathe M, Parthasarathy S, Srinivasan A (2007) Provable algorithms for joint optimization of transport, routing and MAC layers in wireless ad hoc networks. *Proceedings of the DialM-POMC Workshop on Foundations of Mobile Computing*. Invited and refereed.

Kumar AVS, Marathe M, Parthasarthy S, Srinivasan A (2008) Minimum weighted completion time. *Springer Encyclopedia of Algorithms. Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc)*. Invited article. In press.

Laubenbacher R, Jarrah A, Mortveit H, Ravi S (2008) A mathematical formalism for agent based modeling. *Encyclopedia of Complexity and System Science*. In press.

Lee S, Bhattacharjee B, Srinivasan A, Khuller S (2008) Efficient and resilient backbones for multihop wireless networks. *IEEE Transactions on Mobile Computing.* In press.

Macauley M, Mortveit H (2007) On enumeration of cojugacy classes of coxeter elements. *Proceedings of the American Mathematical Society.* In press.

Macauley M, McCammond J, Mortveit H (2008) Order independence in asynchronous cellular automata. *Journal of Cellular Automata* **3**: 37-56.

Mortveit HS, Reidys C (2008) *An Introduction to Sequential Dynamical Systems*. 1st Edn, Springer Verlag, p. 264.

Mozumder P, Marathe A (2007) Papel de las redes de informacion y comunicaciones en el desenlace de la malaria. *Pan American Journal of Public Health* **22**(5), November issue, p. 368.

Mozumder P, Marathe A (2007) Role of information and communication networks in malaria survival. *Malaria Journal* **6**:136. http://www.malariajournal.com/ http://www.malariajournal.com/content/6/1/136

Srinivasan A (2007) Randomized algorithms and probabilistic analysis in wireless networking. *Proceedings of the Symposium on Stochastic Algorithms, Foundations, and Applications (SAGA*). Hromkovic J, Kralovic R, Nunkesser M, Widmayer P (eds), *Lecture Notes in Computer Science*, Springer Verlag, pp. 54-57. (Invited paper)

Srinivasan A (2008) Improved algorithmic versions of the Lovàsz local lemma. *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA*), pp. 611-620.

Josep Bassaganya-Riera

jbassaga@vt.edu

Associate Professor, Virginia Bioinformatics Institute

Group contributors: Ashlee Carter, Elisa Duran, Amir Guri, Nikki Harvey, Raquel Hontecillas, Chris Moore, Sarah Mysiak, Marcel Orpi, Rong Song.

Development of Novel Therapeutic Approaches for Modulating Inflammation and Immunity

Abstract. The Nutritional Immunology Group at the Virginia Bioinformatics Institute is leading three research programs on nutraceutical discovery, gut health, and prevention of obesity-related inflammation. Developing novel approaches for modulating inflammatory responses is the central integrative theme amongst these active areas of research and discovery. Our effort in this reporting year comprised screening of naturally occurring compounds for nuclear receptor activity by using reporter expression assays, and assessment of *in vivo* efficacy in mouse models of chronic inflammation, diabetes, and cardiovascular disease, including testing their clinical efficacy in human patients. We identified a naturally occurring compound capable of activating peroxisome proliferator-activated receptor γ *in vitro*: the phytohormone abscisic acid. We found that abscisic acid modulates both macrophage and T cell function and decreases obesity-related inflammation (Guri et al, 2007 and 2008). We have also discovered that regulatory T cells, a population of immune cells with potent anti-inflammatory activity, require peroxisome proliferator-activated receptor γ for protecting against intestinal inflammation (Hontecillas and Bassaganya-Riera, 2007). In addition, we have investigated the ability of novel infant formulas enriched with arachidonic acid and docosahexaenoic acid to modulate immune responses to influenza virus vaccination (Bassaganya-Riera et al, 2007). These efforts have received support from the National Center for Complementary and Alternative Medicine at the National Institutes of Health, Bristol-Myers Squibb and Cognis Nutrition and Health GmbH.

Keywords: obesity; diabetes; inflammatory bowel disease; peroxisome proliferator-activated receptor; abscisic acid; nutraceutical; inflammation; regulatory T cell (Treg); monocyte chemotactic protein-1; macrophage.

Scientific Progress

Characterization of abscisic acid activity in vitro *and* in vivo

Abscisic acid, a carotenoid derivative, is an important phytohormone found in all photosynthetic organisms and a regulator of plant growth involved in stress responses and differentiation from the juvenile to reproductive stage. The receptors for abscisic acid in plants, which predate the nuclear receptors, have been recently identified as the Flowering Time Control Locus A and the nuclear-located auxin receptor. We found that abscisic acid induced transactivation of peroxisome proliferator-activated receptor γ (PPAR-γ) in 3T3-L1 pre-adipocytes *in vitro*. Dietary abscisic acid supplementation for 36 days decreased fasting blood glucose concentrations, ameliorated glucose tolerance, and increased mRNA expression of PPAR-γ and its responsive genes in white adipose tissue of obese mice (Guri et al, 2007).

We next used a loss-of-function approach to investigate whether abscisic acid ameliorates insulin resistance through a mechanism dependent on immune cell PPAR-γ. We characterized two phenotypically distinct adipose tissue macrophage subsets in obese db/db mice based on their surface expression of the glycoprotein and macrophage maturation marker F4/80. The F4/80^{hi} adipose tissue macrophages were more abundant and expressed greater concentrations of chemokine receptor 2 (CCR2) and CCR5 when compared to F4/80^{lo} adipose tissue microphages. Abscisic acid significantly decreased CCR2^{+}F4/80^{hi} infiltration into white adipose tissue and suppressed monocyte chemoattractant protein-1 (MCP-1) expression in white adipose tissue and MCP-1 plasma concentrations. Furthermore, the deficiency of PPAR-γ in immune cells impaired the ability of abscisic acid to suppress infiltration of F4/80^{hi} adipose tissue macrophages into white adipose tissue, repressed white adipose tissue MCP-1 expression, and improved glucose tolerance. We provide molecular evidence *in vivo* demonstrating that abscisic acid improves insulin sensitivity and obesity-related inflammation by inhibiting MCP-1 expression and F4/80^{hi} adipose tissue macrophages infiltration through a PPAR-γ -dependent mechanism (Guri et al, 2008). A United States Patent (USPT) application and a Patent Cooperation Treaty (PCT) application are pending for using abscisic acid to treat and prevent obesity, diabetes and cardiovascular disease.

Discovery of the importance of PPAR-γ for regulatory T Cell (Treg) anti-inflammatory activity

PPAR-γ activation has been implicated in the prevention of immunoinflammatory disorders, however the regulation mechanisms of effector

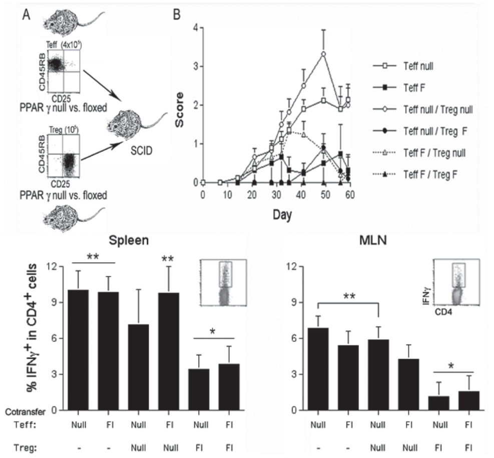


Figure. PPAR-γ-expressing, but not PPAR-γ null Treg, ameliorate disease and decrease interferon-γ expression by pathogenic CD4$^+$ T cells in a mouse model of experimental colitis.

and regulatory CD4$^+$ T cell functions by endogenously activated PPAR-γ remain unclear. We have used PPAR-γ-deficient CD4$^+$ T cells obtained from tissue-specific PPAR-γ null mice (i.e., PPAR-γ fl/fl; MMTV-Cre+) to investigate the role of endogenous PPAR-γ on regulatory (Treg) and effector CD4$^+$ T cell function. Overall, we show that the loss of PPAR-γ results in enhanced antigen-specific proliferation and overproduction of interferon-γ (IFN-γ) in response to interleukin-12. These findings correlate *in vivo* with enhanced susceptibility of tissue-specific PPAR-γ null mice to chemically induced colitis. Furthermore, the transfer of purified PPAR-γ null CD4$^+$ T cells into severe combined immunodeficient (SCID) recipients results in enteric disease.

To test the assertion that the deficiency of PPAR-γ in regulatory T cells (Treg) impairs its ability to prevent effector T cell-induced colitis, we performed co-transfer studies. These studies demonstrate that PPAR-γ-expressing but not PPAR-γ null Treg prevent colitis induced by the transfer of naïve CD4$^+$ T cells into SCID recipients. In line with these findings, the production of IFN-γ by spleen and mesenteric lymph node-derived CD4$^+$ T cells was downregulated following transfer of PPAR-γ-expressing, but not PPAR-γ null Treg. In conclusion, our data suggest that endogenous PPAR-γ activation represents a Treg-intrinsic mechanism of downregulation of effector CD4$^+$ T cell function and prevention of colitis (Hontecillas and Bassaganya-Riera, 2007). Future studies will characterize the endogenous agonists that activate PPAR-γ in Treg and develop novel Treg-based therapies for treating inflammatory conditions.

Modulation of antigen-specific responses to influenza virus by arachidonic acid- and docosahexaenoic acid-enriched infant formulas

In addition to essential fatty acids (i.e., linoleic and linolenic acids), human breast milk contains other polyunsaturated fatty acids such as arachidonic acid, docosahexaenoic acid, and conjugated linoleic acid. The presence of these polyunsaturated fatty acids in human milk suggests that they may be required for optimal health of infants and children. Accordingly, infant formulas enriched in docosahexaenoic acid and arachidonic acid have entered the global infant food markets in recent years. We sought to investigate the ability of arachidonic acid/docosahexaenoic acid-enriched infant formula to modulate immune responses in the neonate. To achieve this goal we used a neonatal pig model of immune modulation in response to influenza virus vaccination. We found that dietary arachidonic acid/docosahexaenoic acid supplementation and sow's milk modulated influenza virus-specific CD4$^+$ and CD8$^+$ T cell *ex vivo* lymphoproliferative responses on day 23 post-immunization in comparison to the unsupplemented control formula. The immunomodulatory effects of arachidonic acid/docosahexaenoic acid-enriched formulas were consistent with upregulation of interleukin-10 in peripheral blood mononuclear cells. Overall, the arachidonic acid/docosahexaenoic acid-enriched formula modulated antigen-specific T cell responses, in part, through an interleukin-10-dependent mechanism (Bassaganya-Riera et al, 2007).

Effect of Helicobacter pylori *colonization on obesity-related inflammation*

Helicobacter pylori is known to decrease the production of ghrelin – a hormone produced by the stomach that stimulates appetite. Based on this background, we hypothesized that gastric colonization of db/db mice by *H. pylori* ameliorates glucose tolerance and obesity-related inflammation by modulating gastric ghrelin expression. To test this hypothesis, we colonized the gastric mucosa of obese db/db mice fed high-fat diets with two isogenic strains of *H. pylori* [strain 98-325 (wild-type, HpcagA+) and strain 99-305 (mutant, HpcagA-)]. A third group of mice was used as a non-colonized control. The mice were weighed on a weekly basis. We found that colonization by the *H. pylori* 99-305 mutant strain decreased fasting blood glucose levels, improved glucose tolerance and suppressed obesity in db/db mice fed high-fat diets. In addition, these improvements in markers of obesity and type 2 diabetes (T2D) were accompanied by suppressed infiltration of macrophages and enhanced influx of Treg in white adipose tissue of mice colonized by the *H. pylori* 99-305 mutant strain. In addition, gastric mRNA leptin expression and plasma leptin concentrations were upregulated by both strains. These data suggest a possible role of colonization, by certain *H. pylori* strains or other gastric bacteria, as a therapy against the obesity and type 2 diabetes pandemics (Bassaganya-Riera et al, 2007, Virginia Tech Intellectual Properties Disclosure 07.048 and provisional United States Patent Application 60/942,016). This work is being developed in collaboration with Bruno Sobral's group at VBI and Maria Gloria Dominguez-Bello's laboratory at the University of Puerto Rico.

Conferences and workshops

Bassaganya-Riera J (2007) Mechanisms of modulation of inflammation-induced colorectal cancer by CLA, Invited speaker for *II International Congress on Conjugated Linoleic Acid (CLA)*, Corsica, Italy, September 19-22, 2007.

Bassaganya-Riera J (2007) Nutritional modulation of chronic inflammation, Invited speaker at Michigan State University, East Lansing, MI.

Bassaganya-Riera J (2007) Nutritional modulation of inflammation and immunity, Invited presentation for the *Nutrition Seminar*, University of Connecticut, Storrs, CT.

Bassaganya-Riera J (2007) Nutritional modulation of mucosal immunity, Invited speaker at University of London, London, UK.

Bassaganya-Riera J, Guri AJ, Noble AM, Reynolds KA, King J, Wood C, Ashby M, Rai D, Hontecillas R (2007) Docosahexaenoic and arachidonic acid-enriched infant formulas modulate antigen-specific T cell responses in the neonatal piglet, *Federation of American Societies for Experimental Biology Meetings*, Washington, DC, abstract #2530.

Bassaganya-Riera J, Hontecillas R (2007) Immunoregulatory actions of Treg PPAR gamma in the gut mucosa, 94th Annual Meeting of the American Association of Immunologists, abstract #59, Miami, FL, May 18-22, 2007.

Evans NP, Guri AJ, Bassaganya-Riera J, Grange RW (2008) Quantification of muscle pathology and inflammation during early disease time course of mdx mice, *New Directions in Skeletal Muscle Pathology and Disease*, New Orleans, LA.

Guri AJ, Hontecillas R, Ferrer G, Casagran O, Wankhade U, Noble AM, Bassaganya-Riera J (2007) The loss of PPAR γ in immune cells abrogates the ability of abscisic acid to improve insulin sensitivity through a mechanism involving suppression of MCP-1 expression and macrophage infiltration into white adipose tissue, *Federation of American Societies for Experimental Biology Meetings*, Washington, DC, abstract #934.

Hontecillas R, Guri AJ, Wankhade U, Lopez-Velasco, G McCall ET, Bassaganya-Riera (2007) Dietary abscisic acid modulates T cell proliferation and cytokine production, *Federation of American Societies for Experimental Biology Meetings*, Washington, DC, abstract #4269.

Intellectual property

Bassaganya-Riera J, Guri AJ, Hontecillas R Method of Using Abscisic Acid to Treat and Prevent Diabetes and Inflammation, US Patent Application No: 60/821,720.

Bassaganya-Riera J, Guri AJ, Hontecillas R Method of Using Abscisic Acid to Treat and Prevent Diseases and Disorders, Patent Cooperation Treaty, Application No: 11/672,223.

Bassaganya-Riera J, Guri AJ, Hontecillas R Method of Using Nutraceuticals for Ameliorating Hypertension and Vascular Inflammation, 08.094 VTIP Disclosure.

Bassaganya-Riera J, Hontecillas R, Dominguez-Bello MG, Sobral B Method of using probiotic bacteria to prevent obesity and type 2 diabetes, US Provisional Patent Application No: 60/942,016.

Bassaganya-Riera J, Hontecillas R, Dominguez-Bello MG, Sobral B Method of using probiotic bacteria to prevent insulin resistance and obesity-related inflammation, 07.048 VTIP Disclosure.

Peer-reviewed publications in reporting period

Barbeau W, Bassaganya-Riera J, Hontecillas R (2007) Putting the pieces of the puzzle together: a series of hypotheses on the etiology and pathogenesis of type 1 diabetes. *Medical Hypotheses* **68**: 608-619.

Bassaganya-Riera J, Guri AJ, Noble AM, Reynolds KA, King J, Wood C, Ashby M, Rai D, Hontecillas R (2007) Docosahexaenoic and arachidonic acid-enriched infant formulas modulate antigen-specific T cell responses to influenza virus in neonatal piglets. *American Journal of Clinical Nutrition* **85**: 824-836.

Guri AJ, Hontecillas R, Si H, Liu D, Bassaganya-Riera J (2007) Abscisic acid ameliorates glucose tolerance and obesity-related inflammation in db/db mice fed high fat diets. *Clinical Nutrition* **26**: 107-116.

Hontecillas R, Bassaganya-Riera J (2007) Peroxisome proliferator-activated receptor γ is required for regulatory CD4$^+$ T cell-mediated protection against colitis. *Journal of Immunology* **178**: 2940-2949.

Guri AJ, Hontecillas R, Ferrer G, Casagran O, Wankhade U, Noble AM, Eizirik D, Ortis F, Cnop M, Liu D, Si H, Bassaganya-Riera J (2008) The loss of PPAR γ in immune cells abrogates the ability of abscisic acid to improve insulin sensitivity through a mechanism involving suppression of MCP-1 expression and macrophage infiltration into white adipose tissue. *Journal of Nutritional Biochemistry* **19**: 216-228.

Guri AJ, Hontecillas R, Bassaganya-Riera J (2008) Dietary modulators of PPARs: Implications for the prevention and treatment of metabolic syndrome. *Nutrigenetics and Nutrigenomics* **1**: 126-135.

Thakare K, Shi W, Barbeau WE, Bassaganya-Riera J, Hontecillas R, Scott F (2008) Investigation of chloroform-methanol soluble wheat proteins and sphingolipids as potential dietary triggers of diabetes in BBdp rats. *Food and Agricultural Immunology*. In press.

Allan Dickerman

dickerman@vt.edu

Assistant Professor, Virginia Bioinformatics Institute

Group contributors:

Elena Shulaeva, Kelly Williams.

Phylogenomics and Systems Biology of Plants and Plant Pathogens

Abstract. Work in the 2007-2008 reporting period spanned several distinct projects. The SeedGenes database of essential genes in Arabidopsis continued to grow. The arrival of Dr. Kelly Williams into the group enabled extended work on bacterial phylogeny as well as proposal development for two metagenomics opportunities. A newly funded project with Eric Beers and Amy Brunner on the Virginia Tech campus provides an opportunity to develop protein-network and proteomics informatics support. A major effort in the period was the design of an Affymetrix microarray to identify pathogens in plant tissue.

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

Scientific Progress

Plant pathogen microarray

The Highly Parallel Microarray for Plant Pathogen Identification project is funded by a United States Department of Agriculture (USDA) Cooperative State Research, Education, and Extension Service (CSREES) grant to Allan Dickerman and Chris Lawrence, with Stephen Goodwin of Purdue University serving as a collaborator. The primary goal is to design and validate an Affymetrix microarray with probes to a large diversity of bacterial and eukaryotic plant pathogens. In the first year of this project, we utilized the so-called "Phylochip" (Brodie et al, 2007), which covers a diversity of bacteria but is not focused on plant pathogens and does not cover eukaryotic pathogens. Consequently, we considered this chip to be useful for exploring the feasibility of labeling and processing joint plant and pathogen samples. Results in the previous reporting period supported the efficiency of labeling bacterial rRNA using a pool of conserved anti-16S-rRNA primers for first-strand cDNA synthesis in a modified version of the Affymetrix reverse-transcriptase labeling method typically applied to mRNA samples.

Designing a new microarray to both bacterial and eukaryotic plant pathogens occupied a significant portion of the reporting period. This design work requires assembling the known ribosomal RNA sequence data from all species of interest. Complete lists of plant pathogens at the species level were not readily available at the time we began our work, though the Comprehensive Phytopathogen Genomics Resource is now an excellent source for this. Our approach was to compile a list of 100 genera, which included the most serious plant pathogens from lists available from the USDA and American Phytopathological Society. All ribosomal sequences from any taxon within any of these 100 genera were sought by a combination of scripted and ad hoc search methods and a set of 11 plant species emphasizing major crops was added. This resulted in the gathering of ribosomal RNA sequence data from 7551 distinct GenBank taxa that often included distinct strains within species. Many of these were represented by full-length ribosomal sequences of both large and small subunits, although many others were sampled by short regions amplified by PCR. Mitochondrial ribosomal sequences were searched for and excluded using sequence search against known examples.

Approximately 14% of bacterial species have at least one rRNA gene with an intervening segment (IVS) whose sequence is spliced from the mature rRNA. Such intervening segments are diverse and therefore might appear as attractive candidates for species-specific rRNA probes, but because their sequences are absent from rRNA samples, it is useful to identify

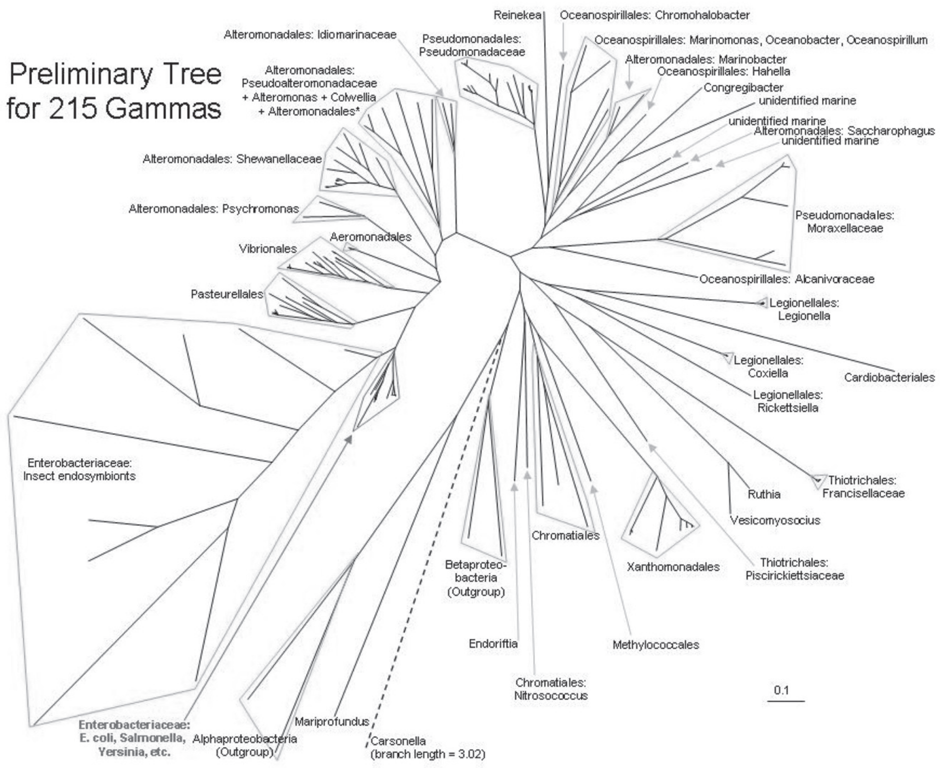

Preliminary Tree for 215 Gammas
Reinekea
Oceanospirillales: Chromohalobacter
Alteromonadales: Idiomarinaceae
Pseudomonadales: Pseudomonadaceae
Oceanospirillales: Marinomonas, Oceanobacter, Oceanospirillum
Alteromonadales: Marinobacter
Oceanospirillales: Hahella
Alteromonadales: Pseudoalteromonadaceae + Alteromonas + Colwellia + Alteromonadales*
Congregibacter
unidentified marine
unidentified marine
Alteromonadales: Saccharophagus
unidentified marine
Alteromonadales: Shewanellaceae
Alteromonadales: Psychromonas
Pseudomonadales: Moraxellaceae
Aeromonadales
Vibrionales
Oceanospirillales: Alcanivoraceae
Pasteurellales
Legionellales: Legionella
Legionellales: Coxiella
Cardiobacteriales
Legionellales: Rickettsiella
Enterobacteriaceae: Insect endosymbionts
Thiotrichales: Francisellaceae
Ruthia
Vesicomyosocius
Chromatiales
Xanthomonadales
Thiotrichales: Piscirickiettsiaceae
Betaproteo-bacteria (Outgroup)
Endoriftia
Methylococcales
Chromatiales: Nitrosococcus
0.1
Mariprofundus
Enterobacteriaceae: E. coli, Salmonella, Yersinia, etc.
Alphaproteobacteria (Outgroup)
Carsonella (branch length = 3.02)

Figure. An outline tree for the Gammaproteobacteria based on 30 aligned genes. Ongoing work will extend this data set to over 200 genes to increase resolution of the tree.

intervening segments and purge them from the list of candidate probes. We analyzed the special features of rRNA intervening segments and designed software for their detection, applying the system to remove these spurious sequences from our list of rRNA probes.

The challenge for chip design was using this rRNA sequence data to find a set of around 32,000 25-mer probes that will optimally distinguish among the taxa of interest. We first looked for taxon-specific probes at hierarchical levels of the GenBank taxonomy from isolates/strains up through species, genera, families, etc. While perfect taxon-specific probes are most desirable, finding them was rare, which was in part due to the incomplete sampling of the rRNA sequence regions. To evaluate probes on a graded scale we used the product of a specificity measure (in-group probe matches divided by in-group size) and a selectivity measure (in-group matches divided by total matches). Scoring higher-scoring probes provided taxon specificity at higher levels, but there still remained a lack of resolution among isolates and strains within species, particularly in densely sampled species. We addressed the need for additional resolution at the lowest levels by searching for additional probes that could distinguish between isolates and species within a genus, allowing for lower taxa to have unique combinations of probes hybridizing even where no single probe was perfectly specific.

Additional constraints on probe selection included probe scores provided by Affymetrix on a zero to 1.0 scale, with 0.36 being a cut-off for suitable predicted performance and the number of rounds of nucleotide extensions needed to synthesize the 25-mers, which was capped at 74. Another aspect of probe selection is that single-base mismatches can still hybridize at close to the perfect-match signal when the mismatch is near the ends of the probe. For this reason, we scored probes using exact string matches to only the central 21 bases out of 25.

The selection of a set of 32,000 out of hundreds of thousands of potential probes to achieve optimal taxon-identification properties is not trivial with so many distinct criteria. We developed a numerical score to evaluate any candidate chip design which allowed us to track improvements in our algorithms. This taxon non-discrimination score summed up how many pairs of taxa (using the GenBank taxonomy) exhibited identical probe match patterns for the proposed set of probes. This score is plagued by the incomplete sampling problem but should provide a relative indication of design improvement. All data manipulations performed for this chip design were done in a MySQL database using the Perl programming language.

The final design contained 31,405 probes and conceptually targets 7930 taxa in 97 pathogen genera plus 11 plant species. Affymetrix has synthesized the chip and we have received our initial shipment of 90 chips. Experiments to test the validation and sensitivity of the chip are under way.

Poplar biomass protein interaction network

A collaboration with Eric Beers in Virginia Tech's Department of Horticulture and Amy Brunner in the Department of Forestry received funding from the United States Department of Energy (DOE). The goal of this project is to determine important protein interaction patterns relevant to wood development in poplar by exploiting the genome sequence that was recently completed.

Proteins must interact with other molecules, such as carbohydrates, lipids, nucleic acids, and other proteins, to function. Very little is known about the protein-protein interactions that plant cells use to control cell-wall-related biomass production. The identification of protein-protein interaction networks associated with biomass production in the woody tissues of poplar, a model biomass crop, will lead to a more detailed understanding of the molecular biology and genomics of plant biomass production and ultimately contribute to strategies for biomass crop improvement.

The objective of the proposal is to map protein-protein interactions relevant to biomass production by focusing on proteins expressed in poplar wood, which is the site of the majority of secondary cell wall synthesis and hence biomass accumulation in poplar. We have identified approximately 250 poplar genes specifically associated with wood formation and have designated these as the poplar biomass gene set. Each of these genes will be cloned and expressed as protein in a yeast expression system known as the yeast two-hybrid system for protein-protein interaction. In the yeast two-hybrid system, individual proteins can be tested for their ability to interact with other proteins expressed in the same yeast cell. We will perform a 250 x 250 yeast two-hybrid matrix assay that will identify interactions occurring between any two members of the poplar biomass protein set. Selected members of the biomass protein set will also be tested in the yeast two-hybrid system for their ability to interact with any protein found in a library of proteins derived from all the genes expressed in poplar wood-forming tissues. Finally, a small number of biomass proteins will be expressed in plants and tested for their ability to form complexes with other proteins present in woody tissue. Members of protein complexes isolated from plants will be identified using mass spectrometry. All of the interactions revealed as a result of the three types of protein-protein interaction studies will be combined to produce a biomass protein-protein interaction map to aid in the comprehensive understanding of how protein-protein interactions contribute to biomass production by plants.

Gammaproteobacteria species phylogeny

We have spent a significant amount of time in the 2007-2008 reporting period working with members of the VBI Cyberinfrastructure Group to develop an evolutionary tree for the Gammaproteobacteria. This is a large and important bacterial class, containing the familiar infectious genera *Escherichia*, *Salmonella*, *Vibrio*, *Haemophilus* and *Pseudomonas*, together with a large number of additional human, plant and animal pathogens, endosymbionts of eukaryotic cells, and marine bacteria. Much of the current understanding of evolutionary relationships among these bacteria is based on 16S rRNA sequences, which may not contain sufficient information to resolve these relationships properly. We are preparing a robust phylogenetic tree in order to resolve the evolutionary history of the Gammaproteobacteria. The tree is based on our collection of a large number of protein families from the numerous genomic sequences that are available for the group. An initial problem was that there were too many genomes available, leading us to develop an algorithm for selecting the most phylogenetically diverse representatives for the target number of genomes. The protein sequences in each family are processed to identify and remove impostor proteins, align the sequences, and remove the less reliable portions of the alignment. The data from all families are then combined, resulting in an extremely information-rich data set.

Our preliminary trees provide several interesting results. One extremely divergent species that had been difficult to place in phylogenetic trees can now be included in a group based on its ecology. However, two species that had previously been placed in the Gammaproteobacteria we can now reject from the class. There are several cases where order-level groupings (and even some family-level groupings) based on 16S rRNA trees are now split. Our tree will provide important insights into the evolution of this large and important group of bacteria.

Gene regulation by acquisition of RNA binding sites

Dr. Kelly Williams joined the Dickerman group in the fall of 2007, bringing with him an interest in all aspects of prokaryote RNA biology. One project he developed single-handedly in the reporting period describes a striking instance of genes for RNA-binding proteins developing binding sites for their own products. The results are in press in *RNA Biology*.

There are numerous cases where a ribosomal protein autoregulates by binding to its own mRNA, which is usually explained as the mRNA site mimicking the rRNA site for the protein, although the resemblance is usually limited. Here an mRNA segment is described that very strongly resembles a segment of 23S rRNA at the primary sequence and secondary structure levels in a gammaproteobacterial clade comprising Pseudomonadaceae and certain marine bacteria. Moreover, this segment from rRNA is known to bind to either or both of the two proteins, L4 and L24, that are encoded downstream of the mRNA segment. Analysis of *Pseudomonas aeruginosa* gene regulatory signals in the region suggests that the L4-encoding operon extends to include the L24 gene. Thus two proteins may co-autoregulate their shared operon with interesting combinatorial possibilities. This implies that two independent autoregulatory mechanisms have arisen for this operon in the Gammaproteobacteria, since the L4-encoding operon of *Escherichia coli* does not include the L24 gene and is autoregulated by L4 protein alone at a different site that bears a weak resemblance to its binding site in 23S rRNA.

Conferences and workshops

Dickerman A (2007) Highly Parallel Pathogen Microarrays for Plant Biosecurity, Project Directors Meeting for the USDA-CSREES Plant Biosecurity Competitive Program, *American Phytopathological Society Meeting*, San Diego, CA, August 4, 2007.

Peer-reviewed publications in reporting period

Fisher M, Miller D, Brewster C, Husseneder C, Dickerman A (2007) Diversity of gut bacteria of *Reticulitermes flavipes* as examined by 16S rRNA gene sequencing and amplified rDNA restriction analysis. *Current Microbiology* **55**(3): 254-259.

Muralla R, Chen E, Sweeney C, Gray JA, Dickerman A, Nikolau BJ, Meinke D (2008) A bifunctional locus (BIO3-BIO1) required for biotin biosynthesis in Arabidopsis. *Plant Physiology* **146**(1): 60-73.

Williams KP (2008) Strong mimicry of an rRNA binding site for two proteins by the mRNA encoding both proteins. *RNA Biology*. In press.

Other cited references

Brodie EL, DeSantis TZ, Parker JP, Zubietta IX, Piceno YM, Andersen GL (2007) Urban aerosols harbor diverse and dynamic bacterial populations. *Proceedings of the National Academy of Sciences USA* **104**: 299-304.

Ina Hoeschele

inah@vbi.vt.edu

Professor, Virginia Bioinformatics Institute

Professor, Department of Statistics, Virginia Tech

Adjunct Professor, Department of Cancer Biology, Wake Forest University School of Medicine, Winston-Salem, NC

Group contributors: Alberto de la Fuente, Bing Liu, Lei Bao, Lucia Gan, Fei Zou.

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 of plants, animals and humans. At the present time, our main research focus is on developing, implementing and evaluating methods for very high-dimensional Quantitative Trait Locus (QTL) mapping and global gene regulatory network inference in systems genetics experiments. For this we focus on Bayesian variable selection using parametric and nonparametric approaches and with and without dimension reduction. We are analyzing a large systems genetics experiment investigating the genetic basis of resistance to a major pathogen in soybean.

Keywords: systems genetics; statistical genetics; gene network inference; design and analysis of microarray experiments; linkage and linkage disequilibrium QTL mapping; haplotype inference in complex pedigrees.

Scientific Progress

Analysis of expression and proteomics profiling experiments

We have analyzed, in collaboration with Brett Tyler's research group at VBI, three microarray experiments that were performed as part of the soybean–*Phytophthora sojae* systems genetics project. In this project, researchers have been interested in evaluating the effects of multiple factors and their interactions on the expression of 37,420 soybean genes. We have fitted several linear mixed models accounting for a complex covariance structure of the experimental designs in one- and two-step approaches (gene-specific *F*-tests followed by analysis of individual contrasts) using Proc Mixed in the Statistical Analysis System (SAS). Multiple testing was accounted for by controlling the false discovery rate (FDR and positive FDR), and several controlling methods were investigated. The two-step method of Benjamini et al (2006) was found to be most reliable for our situation that is characterized by the majority of null hypotheses being truly false.

We performed quality control analysis for the Recombinant Inbred Line soybean systems genetics experiment that comprises 2519 Affymetrix GeneChip® microarrays (Bao and Hoeschele, 2008) and evaluated its effect (via removal of outlier arrays) on the subsequent high-level analyses (mixed model, Quantitative Trait Loci analyses).

We have also analyzed the proteome profile of the MCF-7 breast cancer cell line in response to estradiol and tamoxifen treatment in a quantitative proteomic study. The experiments, which use isobaric tags for relative and absolute quantification (iTRAQ labeling), were conducted in Iuliana Lazar's laboratory at VBI. Biomarkers were selected based on a method developed for data filtering, normalization (including quantile normalization), and statistical analysis of differential expression with control of the FDR (Armenta J, Lazar I, Hoeschele I, submitted).

High-dimensional mapping of Quantitative Trait Loci and inference of gene regulatory networks

We performed basic expression QTL mapping for the soybean systems genetics project at VBI. For this purpose, we used single marker analysis of each individual etrait (expression trait, i.e. expression profile of an individual gene in the recombinant inbred line population), computing *P*-values adjusted for genome-wide marker testing, retaining the top 1, 2, 3 or 5 markers for each of the 28,395 etraits and applying FDR control to the resulting lists of 28,395, 2×28,395, 3×28,395 and 5×28,395 *P*-values, respectively, to adjust for multiple testing across etraits. The numbers of identified eQTL at an FDR level of 0.01 (0.001) were

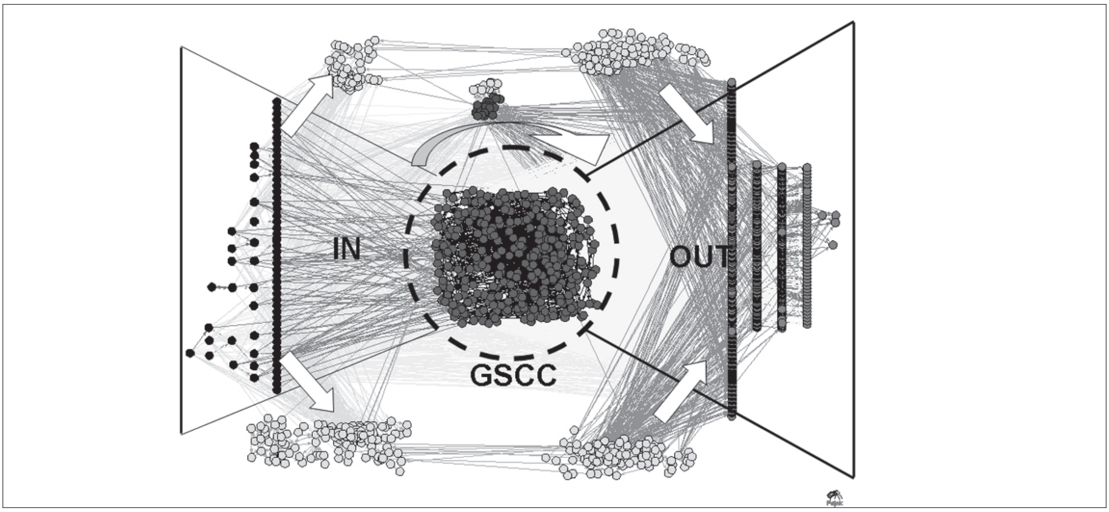


Figure. Gene regulatory network recently inferred for the yeast *S. cerevisiae* (Mancosu, Pieroni, Maggio, Fotia, Liu, Hoeschele, de la Fuente, 2008; Gianmaria Mancosu, CRS4 Bioinformatica). The network consists of 4239 nodes and 14,723 directed edges. The layout was obtained according to the network's "bow tie" structure and by combining several layout algorithms implemented in Pajek (Batagelj and Mrvar, 2003).

21,361, 30,026, 23,313, and 16,970 (18,520, 18,147, 15,632, and 12,851), respectively (Bao L, unpublished results). While these results show that more than two thirds of all etraits appear to be under genetic control, the decrease in the number of detected eQTLs with increasing size of the list of *P*-values indicates, as expected, that the power of this individual etrait–individual marker analysis is unsatisfactorily low. We are now investigating multivariate methods and methods that borrow information across etraits and markers, which are expected to provide a more complete picture of the genetic control of gene expression. The methods are implemented in a Bayesian framework, and different parametric and nonparametric approaches to high-dimensional variable selection are being explored. We have also investigated a method for eQTL mapping which increases the power of eQTL detection by utilizing DNA sequence information to jointly perform *cis*-, *cistrans*- and *trans*-eQTL mapping, and we have applied this method to a yeast dataset (Liu, de la Fuente and Hoeschele, 2008). Finally, using the sequence-based eQTL mapping method and local structural models we have reconstructed a gene regulatory network using the complete set of yeast genes and yeast genetic markers (Mancosu, Pieroni, Maggio, Fotia, Liu, Hoeschele, de la Fuente, submitted).

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

As an alternative to the use of local structural models for gene regulatory network inference mentioned in the previous section, we have implemented global Structural Equation Modeling (SEM) to search for a set of well-fitting gene regulatory networks. Typically, SEM has been implemented for at most tens of variables. Based on a factorization of the likelihood and a strongly constrained search space defined by the results from the eQTL analysis, our algorithm can currently construct networks of several hundred genes and eQTL (Liu, de la Fuente and Hoeschele, 2008). 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 advantage of the SEM over the currently popular Bayesian network analysis is that it can reconstruct cyclic networks. Gene regulatory networks represent a projection of all cellular regulatory processes onto the gene space and such networks are expected and found to be cyclic. A Bayesian SEM analysis capable of incorporating prior knowledge remains to be fully developed. Lastly, we have written a book chapter on inferring gene regulatory networks from genetical genomics data (Liu et al, in review).

Haplotyping in complex human pedigrees

Haplotyping in pedigrees is an essential component of genetic studies on pedigrees (e.g., linkage analyses and association studies). We have written a review paper on haplotyping in pedigrees (Gao, Allison and Hoeschele, under revision), revised a book chapter (Hoeschele, 2007), and have continued to work towards incorporating missing marker data and marker linkage disequilibrium into our efficient deterministic haplotyping method for large pedigrees and large numbers of markers (Gao and Hoeschele, 2007).

Conferences and workshops

Tyler BM, Zhou L, Mideros SX, Bao L, Tripathy S, Torto-Alalibo T, Mao Y, Liu B, Krampis K, Tucker DM, Li H, Waller L, Hanlon R, Arredondo F, Smith BM, Jerauld AC, Evans C, St. Martin SK, Hoeschele I, Dorrance AE, Maroof SMA (2008) Functional genomics of quantitative resistance against *Phytophthora sojae* in soybean, *International Plant and Animal Genome Conference XVI*, San Diego, January 2008.

Waller L, Zhou L, Mideros SX, Tripathy S, Mao Y, Hanlon R, Torto-Alalibo TA, St. Martin SK, Maroof SMA, Dorrance AE, Hoeschele I, Tyler BM (2007) Transcriptional profiling of potential regulatory factors modulating defense mechanisms in soybean under *P. sojae* infection, *The 8th International Conference on Systems Biology*, October 2007.

Zhou L, Mideros SX, Bao L, Tripathy S, Torto-Alalibo TA, Mao Y, Liu B, Krampis K, Tucker DM, Li H, Waller L, Hanlon R, Arredondo F, Smith BM, Jerauld AC, Evans C, St. Martin SK, Maroof SMA, Hoeschele I, Dorrance AE, Tyler BM (2008) Dissecting soybean resistance to *Phytophthora by* QTL analysis of host and pathogen expression profiles, *International Plant and Animal Genome Conference XVI*, San Diego, January 2008.

Zou F, Chen F, Zhou H, Hoeschele I, Huang H (2008) Bayesian semiparametric multiple quantitative trait loci mapping, *Eastern North American Regional Meeting of the International Biometric Society,* March 2008.

Peer-reviewed publications in reporting period

Bao L, Hoeschele I (2008) Quality assessment of a very large microarray experiment. *Technometrics*. In press.

Gao G, Hoeschele I (2007) A note on a haplotyping method in pedigrees. *Genetics, Selection, Evolution* **40**: 25-36.

Hoeschele I (2007) Mapping quantitative trait loci in outbred populations. In *Handbook of Statistical Genetics*, Balding DJ, Bishop M, Cannings C (eds), Wiley, NY, pp. 623-677.

Liu B, de la Fuente A, Hoeschele I (2008) Gene network inference via structural equation modeling in genetical genomics experiments. *Genetics* **178**: 1763-1776.

Stock KF, Hoeschele I, Distl O (2007) Bayesian estimation of genetic parameters for multivariate threshold and continuous phenotypes and molecular genetic data in simulated horse populations using Gibbs sampling. *BMC Genetics* **8**:19.

Stock KF, Hoeschele I, Distl O (2007) Estimation of genetic parameters and prediction of breeding values for multivariate threshold and continuous data in a simulated horse population using Gibbs sampling and residual maximum likelihood. *Journal of Animal Breeding and Genetics* **124**: 308-319.

Other cited references

Batagelj V, Mrvar A (2003) Pajek - Analysis and visualization of large networks. In *Graph drawing software*, Jünger M, Mutzel P (eds), Springer Verlag, pp. 77-103.

Benjamini Y, Krieger AM, Yekutieli D (2006) Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* **93**: 491-507.

Reinhard Laubenbacher

reinhard@vbi.vt.edu

Professor, Deputy Director of Education and Outreach, Virginia Bioinformatics Institute

Professor, Dept. of Mathematics, Virginia Tech

Group contributors: Valerie Hower, Abdul Jarrah, Reinhard Laubenbacher, Alan Veliz-Cuba.

Mathematical Systems Biology

Abstract. Mathematical systems biology aims to understand complex biological networks through the use of mathematical models and data analysis techniques. An important focus of our work is on the development and application of mathematical techniques for the inference, or reverse-engineering, of biochemical networks from "-omics" data in the form of time course measurements of RNA, protein, and metabolite concentrations. Our chosen modeling framework is multi-state, time-discrete dynamical systems, which can be viewed as generalizations of Boolean networks that have been used in biology for several decades. One application area we have pursued is yeast systems biology, looking specifically at the oxidative stress response network. As part of this effort, we are carrying out a research program to develop further the mathematical foundation for this family of dynamical systems. Other systems biology projects we are pursuing involve the use of breast cancer metabolomics to identify metabolic markers for breast cancer diagnosis and the investigation of the role of the immune system in the pathogenesis of Duchenne muscular dystrophy using a mathematical model consisting of a system of ordinary differential equations.

Keywords: mathematical systems biology; reverse-engineering of biochemical networks; finite dynamical systems; breast cancer metabolomics; Duchenne muscular dystrophy.

Scientific Progress

Introduction

A variety of mathematical model types are being used in systems biology, including systems of differential equations, stochastic models such as Markov chains, and statistical models such as Bayesian networks. We have focused on the use of finite dynamical systems and the development of algebraic tools for its analysis. Algebraic biology is emerging as a new approach for modeling and analyzing biological systems using tools from algebra, algebraic geometry, and discrete mathematics. Application areas cover a wide range of molecular biology, from the analysis of DNA and protein sequence data to the study of secondary RNA structures, assembly of viruses, modeling of cellular biochemical networks, and algebraic model checking for metabolic networks, to name a few.

Reverse-engineering of biochemical networks

The reverse-engineering of biochemical networks is one of the central problems in systems biology. We have continued our research program to develop and apply mathematical algorithms for this purpose. The majority of existing methods have been focused on the inference of a "wiring diagram" of the network from a collection of time courses of mRNA expression measurements. We are using the framework of finite dynamical systems as a modeling paradigm. Relevant publications include Jarrah et al (2007), Dimitrova et al (2007), Stigler et al (2007). We have also investigated different methods and compared performance on a benchmark data set (Camacho et al, 2007). Since biochemical networks are dynamic rather than static, we continue to develop methods that result in a dynamical system model of the network, which provides more information than just a static wiring diagram. This project is a collaboration with the Mendes and Shulaev groups at VBI and researchers at Clemson University, Cornell University, The Ohio State University, and Rutgers University.

Finite dynamical systems

Finite dynamical systems represent a very general mathematical family of dynamical systems that provide a general framework for the modeling of a variety of phenomena, including gene regulatory networks (mentioned above), immunology, and a variety of social networks and their use in epidemiology. We are continuing our work to develop a mathematical foundation for the analysis of such systems, with a key focus on the relationship between the structure of such systems and the resulting dynamics. Another central focus of this investigation is the relationship between the structure of the wiring diagram of finite dynamical systems and its dynamics. The primary importance is the possibility of inferring general design principles of biological networks.

As part of the study (Sontag et al, 2008), we provided computational evidence for the conjecture that biological networks are close to being "monotone," a concept from dynamical systems theory that relates to the structure of negative feedback loops in the wiring diagram of the network. Feedback loops play an important role in determining the dynamics of biological networks. In order to study the role of negative feedback loops, we introduced the notion of "distance to positive feedback (PF-distance)," which in essence captures the number of "independent" negative feedback loops in the network, a property inherent in the network topology. Through a computational study using Boolean networks, we show that PF-distance has a strong influence on network dynamics and correlates very well with the number and length of limit cycles in the phase space of the network. To be precise we show that as the number of independent negative feedback loops increases, the number (length) of limit cycles tends to decrease (increase). These conclusions are consistent with the fact that certain natural biological networks exhibit generally regular behavior and have fewer negative feedback loops than randomized networks with the same number of nodes and same connectivity. This project is a collaboration with researchers at Rutgers University.

Breast cancer metabolomics

This project is a collaboration with the Mendes and Shulaev groups at VBI and the Wake Forest University Comprehensive Cancer Center. Metabolomics is the study of cells by measuring profiles of all or a large number of metabolites. Metabolomics was originally proposed as a method of functional genomics, but its utility extends well beyond that. Metabolomics approaches are being increasingly utilized for biomarker discovery in many human diseases. As a global approach it is useful in identifying overall metabolic changes associated with breast cancer development and the most affected metabolites and metabolic networks. In this project we attempt to identify small molecule biomarkers that can aid in breast cancer diagnosis. These molecular markers are being developed using a combination of metabolic fingerprinting and mathematical, statistical, and machine learning algorithms.

As a first step the project is focused on *in vitro* cell lines, with the goal of expanding the study to serum samples from patients. The mathematical and statistical challenge here is that the number of samples is typically very small compared to the number of metabolites being considered. We have developed a prototype of a new algorithm to identify a small number of markers that have the ability to distinguish between normal and malignant samples in a small initial study. We have validated algorithm performance by comparing it to the results of an algorithm relying on different principles developed by the Mendes group.

The project has several novel aspects: (1) the use of metabolic markers for the study of breast cancer, (2) a novel approach to metabolic fingerprinting, and (3) the use of new data analysis techniques.

Duchenne muscular dystrophy

Duchenne muscular dystrophy (DMD) is a common and debilitating genetic disease. At present, therapy options for DMD patients are limited and only supportive, not curative. The precise mechanisms that initiate and/or contribute to the relentless progression of DMD are not well understood. The goal of this project is to develop and apply quantitative experimental observations from dystrophic mouse models to a new generation of modeling and computer simulation technology, which will provide a better understanding of the role the immune/inflammatory response plays in the mechanisms that initiate the pathophysiology associated with DMD.

Using existing time course measurements of the interplay between muscle tissue damage and immune system response, we have constructed a detailed predictive mathematical model of the interplay between muscle tissue damage and immune system response. The long-term goal of the project is to identify and validate potential therapeutic targets that can be used to develop treatment strategies. This project is a collaboration with researchers at Virginia Tech and the University of California, Los Angeles.

Conferences and workshops

Laubenbacher R (2007) Biochemical network inference*, FLAD Computational Biology Collaboratorium (FCBC)*, Instituto Gulbenkian de Ciência (IGC), Lisbon, Portugal, April 30, 2007.

Laubenbacher R (2007) Algebraic methods in systems biology, *Applied Mathematical Sciences Summer Institute (AMSSI)*, California State Polytechnic University, Pomona, CA, June 25, 2007.

Laubenbacher R (2007) The role of mathematics in systems biology*, Second Argentinian Summer School in Biomathematics (BIOMAT)*, Córdoba, Argentina, June 28-July 7, 2007.

Laubenbacher R (2007) Discrete models of biochemical networks: The toric variety of nested canalyzing functions, *Second International Conference on Algebraic Biology 2007*, Research Institute for Symbolic Computation (RISC), Linz, Austria, July 2-4, 2007.

Laubenbacher R (2007) A mathematical formalism for agent-based modeling, *22nd Mini-Conference on Discrete Mathematics and Algorithms*, Clemson University, Clemson, SC, October 11-12, 2007.

Laubenbacher R (2007) Complexity in biological systems, *Cyber-Enabled Discovery and Innovation (CDI) Workshop*, Statistical and Applied Mathematical Sciences Institute (SAMSI), Research Triangle Park, NC, November 1, 2007.

Laubenbacher R (2007) Graduate studies in computational biology*,* The Institute for Math Biology Education and Research (TIMBER), Appalachian State University, Boone, NC, November 3-5, 2007.

Laubenbacher R (2008) Metabolic fingerprinting of breast cancer, Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University, Washington, DC.

Laubenbacher R (2008) Design of experiments and biochemical network inference, Department of Statistics, University of Kentucky, Lexington, KY.

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* **1115**: 73-89.

Choi V, Huang Y, Lam V, Potter D, Laubenbacher R, Duca K (2008) Using formal concept analysis for microarray data comparison. *Journal of Bioinformatics and Computational Biology* **6**(1): 65-75.

Dimitrova E, Jarrah A, Stigler B, Laubenbacher R (2007) A Gröbner fan method for biochemical network modeling. *Proceedings of the 18th International Symposium on Algorithms and Computation (ISSAC 2007)*, ACM Press, New York, pp.122-126.

Duca KA, Shapiro M, Delgado-Eckert E, Hadinoto V, Jarrah A, Laubenbacher R, Thorley-Lawson DA (2007) A Virtual Look at Epstein–Barr virus infection: biological interpretations. *PLoS Pathogens* **3**(10): 1388-1400.

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*, Lecture Series in Computer Science, **4545**: 15-22.

Jarrah A, Laubenbacher R, Stigler B, Stillman M (2007) Reverse-engineering polynomial dynamical systems. *Advances In Applied Mathematics* **39**: 477-489.

Jarrah A, Raposa B, Laubenbacher R (2007) Nested canalyzing, unate cascade, and polynomial functions. *Physica D: Nonlinear* **233**(2): 167-174.

Shapiro M, Duca K, Delgado-Eckert E, Hadinoto V, Jarrah A, Laubenbacher R, Lee K, Polys N, Thorley-Lawson D (2008) A virtual look at Epstein–Barr virus infection: simulation mechanism. *Journal of Theoretical Biology,* February 16, 2008, Epub ahead of print, PMID: 18371986.

Sontag E, Veliz-Cuba A, Laubenbacher R, Jarrah A (2008) The effect of negative feedback loops on the dynamics of Boolean networks. *Biophysics Journal,* March 28, 2008, Epub ahead of print.

Stigler B, Jarrah A, Stillman M, Laubenbacher R (2007) Reverse engineering of dynamic networks. *Annals of the New York Academy of Sciences* **1115**: 168-177.

Thorley-Lawson DA, Hadinto V, Luzariaga K, Jarrah AS, Laubenbacher R, Polys N, Delgado-Eckert E, Shapiro M, Duca KA (2007) Simulating Epstein-Barr virus infection with C-ImmSim. *Bioinformatics* **23**(11): 1371-1377.

Chris Lawrence

lawrence@vbi.vt.edu

Associate Professor, Virginia Bioinformatics Institute

Associate Professor, Dept. of Biological Sciences, Virginia Tech

Group contributors: Mihaela Babiceanu, Jinkii Cho, Yangrae Cho, Amanda Cronin, Ha Dang, Kwang-Hyung Kim, Mauricio La Rota, Graciela Santopietro, Derrick Scott.

Fungal Pathogenomics

Abstract. The genus Alternaria contains many economically important fungal species that impact plant and human health. Two model systems are being developed in our laboratory that allow 1) identification and characterization of fungal virulence factors influencing plant pathogenesis and 2) characterization of secreted proteins associated with fungal innate immunity and inflammation in mammals. We are using the *Alternaria brassicicola*-Brassicaceae interaction as a system to study fungal pathogenesis of cultivated Brassicas and the model flowering plant Arabidopsis. We are using *Alternaria alternata*, the airborne, ubiquitous species clinically associated with human airway disorders including allergy, asthma, and chronic rhinosinusitis, for our studies involving human cells and mice. Our group has developed and/or optimized robust methods for functional genomics including high-throughput gene knockout methodologies, gene overexpression, and an Alternaria-based recombinant protein production platform used in both research projects. Several virulence factors of plants, including signal transduction components (kinases and transcription factors), secondary metabolite biosynthetic genes, and other completely novel factors, have been identified using these molecular approaches coupled with bioinformatic analyses of the recently completed and annotated *A. brassicicola* genome. Moreover, the genome sequence has already proven valuable for the identification of fungal proteins that cause inflammatory responses in humans.

Keywords: Alternaria-Brassicaceae pathosystem; necrotrophic fungal-plant interactions; genome sequence; *Alternaria brassicicola*; *Alternaria alternata* chronic rhinosinusitis; allergy; asthma.

Scientific Progress

The Alternaria brassicicola *genome sequencing project*

Our laboratory has been the lead group involved in the *A. brassicicola* genome project. The United States Department of Agriculture Microbial Genome Sequencing Program previously funded random shotgun sequencing of the *A. brassicicola* genome (isolate ATCC 96836) at 6.4x coverage, generation of a physical map [fingerprinting of Bacterial Artificial Chromosomes (BACs)], sequencing of BAC and fosmid ends to help facilitate assembly (0.4x coverage), and a Massively Parallel Signature Sequencing (MPSS) experiment as a novel means for genome annotation. The 6.4x shotgun sequencing, BAC and fosmid end sequencing, physical map construction, and MPSS experiments were completed at Washington University Genome Center (St. Louis, MO) and Solexa Inc., Hayward, CA. A genome assembly (V1.0) was generated and is publicly available (http://www.genome.wustl.edu/genome.cgi?Genome=Alternaria%20 brassicola).

The assembly is composed of 838 supercontiguous sequences averaging 36,147bp in length and having an *N*50 value of 2,400,717 bp (the length such that 50% of all nucleotides contained in supercontiguous sequences are at least this size). The supercontiguous sequences are composed of 4,039 contiguous sequences with 84% of the contigs longer than 1000 bp and 98.8% of the bases having a phred quality above 20. The total length of the sequenced portion is ~31 Mb, consistent with the previous estimate of 29.6 Mb for genome size using pulse-field gel electrophoresis. Approximately 80% of the assembled genome (25 Mb) is distributed on 12 supercontiguous sequences, which suggests a relatively robust assembly for only 6x coverage. Bioinformatic analyses of the genome were carried out at the Virginia Bioinformatics Institute. Version 2.4 of the FGENESH software (http://www.softberry.com) was used with an *Alternaria* trained parameter matrix for the initial prediction of genes in the genome assembly.

10,688 genes were predicted using the contiguous sequences as input, or 9,814 using the supercontiguous sequences. Similar values were obtained using other gene prediction programs such as SNAP. MegaBLAST was used to map a set of 6,430 Expressed Sequence Tags (ESTs from fungus grown *in vitro* and plant infection libraries) to support the *ab-initio* gene prediction and included BLASTX results from similarity searches against the Uniprot protein database and against other related gene models from taxonomically-related fungi. Results of MPSS of an mRNA library derived from late stage Alternaria-infected cabbage leaves have provided additional experimental evidence of transcription for a subset of the gene models (~4,500). HMMER analysis of predicted proteins based on gene models has been performed utilizing protein functional domain databases individually (Pfam, Tigrfam, Superfam) and part of the Interpro suite of tools (http://www.ebi.ac.uk/interpro/). All annotation data are currently deposited into a local SQL-queryable database and are being transferred to the web-accessible, community annotation database system that is in the final stages of development. Public release and access to the annotated genome are expected to occur in late 2008.

Alternaria brassicicola *functional genomics: production and optimization of linear disruption and replacement knockout mutants*

A transformation method has been established for *A. brassicicola* to generate targeted gene disruption mutants at 80-100% knockout efficiency using linear minimal element (LME) constructs (Cho et al, 2006). This was the first study in which polyethylene glycol-mediated protoplast transformation of this fungus was reported to be successful. Targeted gene disruption was accomplished using a single homologous recombination event. The

Table 1. Collaborators for Alternaria research projects.

Collaborators	Research Institution	Project
Nalvo Almeida	Virginia Bioinformatics Institute	Alternaria Genome
Sandra Clifton	Washington University, St. Louis, MO	Alternaria Genome
Robert Cramer	Montana State University, Bozeman, MT	Functional Genomics
Allan Dickerman	Virginia Bioinformatics Institute	Pathogen Microarrays
Joe Falkinham	Virginia Tech	Antifungal Drug Discovery
Otto Folkerts	Virginia Bioinformatics Institute	Fungal Biotechnology
Hirohito Kita	Mayo Clinic, Rochester, MN	Chronic Rhinosinusitis and Asthma
Yong-Hwan Lee	Seoul National University	Alternaria Genome
Liwu Li	Virginia Tech	Fungal Innate Immunity
Thomas K. Mitchell	Ohio State University, Columbus, OH	Alternaria Genome and Functional Genomics
Nick Oberlies	Research Triangle Institute, Raleigh, NC	Antifungal Drug Discovery
Tobin Peever	Washington State University, Pullman, WA	Fungal Mating Genes
Barry Pryor	University of Arizona, Tucson, AZ	Alternaria Systematics
João Setubal	Virginia Bioinformatics Institute	Alternaria Genome
Vladimir Shulaev	Virginia Bioinformatics Institute	Alternaria Secondary Metabolites
Brett Tyler	Virginia Bioinformatics Institute	Alternaria Genome
Jonathan Walton	Michigan State University, East Lansing, MI	Alternaria Secondary Metabolites (Histone Deacetylase Inhibitors)

targeting efficiency was consistently high for the constructs harboring 250-bp homologous sequences corresponding to target genes. Shorter sequences have not been evaluated. In a recent regulatory gene study described in more detail in this report, the transformation method was used not for disruption but for deletion of target genes by replacing them with the Hygromycin B phosphotransferase (*HygB*) cassette (selectable marker) using linear replacement constructs containing ~1-kb genomic sequences that flank the target gene. The efficiency of targeted gene knockout (gene disruption or gene replacement) in *A. brassicicola* is high when compared to that observed in other fungi with sequenced genomes, and this high efficiency has allowed for targeted functional genomics experiments in *A. brassicicola*.

The targeted gene knockout via disruption and/or replacement by homologous recombination is high enough for a high-throughput approach. However, as expected, most of the targeted gene mutants produced so far have shown little or no reduction of virulence (Lawrence et al, unpublished results). The genes studied thus far were selected based upon plant expression levels or machine annotated features of interest (e.g. histidine kinases, secondary metabolite biosynthesis, transcription factors, cell wall degrading enzymes, etc.). Interestingly, some genes identified in the current studies can be considered exciting exceptions as described later in this report.

Identification of A. brassicicola *pathogenicity factors*

One area of interest regarding *A. brassicicola* pathogenicity lies in the area of secondary metabolite biosynthesis. We have identified every gene in the *A. brassicicola* genome predicted to encode a polyketide synthase (PKS) or nonribosomal peptide synthase (NPS) using HMMER analysis of the predicted set of proteins in the genome in conjunction with protein functional domain databases such as Pfam, Tigrfam, Smartfam, and Panther. We have functionally analyzed all of these genes through gene knockout studies. One gene (*AbNPS2*) was found to be important for cell wall integrity, conidial viability, and virulence of aged spores of *A. brassicicola* (Kim et al, 2007). The secondary metabolite corresponding to or synthesized via *AbNPS2* has yet to be characterized. More research is needed to further characterize secondary metabolite biosynthetic genes and their role in pathogenicity and fungal development.

Another area ripe for exploration in the *A. brassicicola*-Brassicaceae pathosystem is fungal signal transduction mechanisms. Disruption of the Fus3/Kss1 MAP kinase homolog (*Amk1*) in *A. brassicicola* resulted in a complete loss of pathogenicity as observed in other fungi (Cho et al, 2007). Interestingly, our studies showed that addition of long polypeptide nutrients partially restored pathogenicity to the mutants.

In contrast to the null mutants of other phytopathogenic fungi, *A. brassicicola amk1* mutants were capable of partially infecting wounded tissues. Targeted gene deletion mutants for 21 putative regulatory genes were produced and characterized in *A. brassicicola* (Cho et al, 2008). In this study, the SNF1 (sucrose non-fermenting 1) kinase, a collection of putative histidine kinases, and several transcription factors such as Ste12 and Pro1 were subjectively selected from the machine annotated *A. brassicicola* genome for analysis. Deletion of the *A. brassicicola SNF1* homolog did not significantly affect virulence in contrast to the observations made in several other pathogenic fungi (Ospina-Giraldo et al, 2003; Tonukari et al, 2000). Deletion of the STE12 homolog, a transcription factor downstream of Amk1, resulted in loss of pathogenicity as was described in other pathogenic fungi (for review see Xu, 2000). However, Δ*abste12* mutants were capable of infecting wounded plants. In addition, two novel virulence factors were discovered as part of the study by Cho et al. (2008) predicted to encode a transcription factor (*AbPro1*) and a two-component histidine kinase gene (*AbNIK1*). Deletion of *AbPro1* resulted in a 70% reduction in virulence and also exhibited a 25% reduction in vegetative growth rates *in vitro*. Deletion of *AbNIK1* resulted in a near complete loss of virulence without changes in vegetative growth rates *in vitro*. Importantly, the addition of long polypeptides to spores of both Δ*abste12* and Δ*abnik1* during plant inoculations resulted in a complete restoration of pathogenicity. These results strongly confirm the presence of a previously undescribed nutrient- or polypeptide-sensing pathway downstream of Amk1/AbSte12 signaling pathways and a putative AbNIK1 osmoregulation pathway. In addition to *AbNIK1* and *Amk1*, the *A. brassicicola* Slt2 MAP kinase and HOG MAP kinase homologs have been knockedout and characterized (Lawrence, unpublished results). Both of these kinases are pathogenicity factors in phytopathogenic fungi (for review see Xu, 2000). Slt2 is associated with cell wall integrity and HOG is associated with oxidative stress tolerance (for review see Xu, 2000). The Slt2 homolog is a major virulence determinant in *A. brassicicola* (Cho, Scott, and Lawrence, unpublished results). However, knockout of the HOG kinase homolog in *A. brassicicola* did not result in reduced virulence, which suggests that this fungus has evolved or developed alternative mechanisms for oxidative stress tolerance. In another recent study, disruption of *Aso-1*, a gene required for hyphal fusion (anastomosis) is also required for pathogenicity in *A. brassicicola* (Craven et al, 2008). Interestingly, it was also shown in this study that the *amk1* mutants also failed at hyphal fusion suggesting a link between MAP kinase signaling and anastomosis.

In contrast to these studies, the Δ*abste12* mutant exhibited a hyper-anastomosis phenotype but was non-pathogenic on plants. In summary, considerable progress has been made over the last several years regarding identification of *A. brassicicola* virulence genes. More than a hundred genes have been functionally analyzed through gene knockout and overexpression experiments making *A. brassicicola* the species of choice for functional genomics research to define conserved virulence mechanisms for this important genus of fungi.

Alternaria *pathogenomics and human airway disorders*

In medical mycology, Alternaria species are gaining importance as emerging human invasive pathogens, particularly in immunocompromised patients (Morrison and Weisdorf, 1993; Vartivarian et al, 1993). Several Alternaria species and numerous uncharacterized Alternaria taxa have been found associated with infections of the cornea, oral and sinus cavities, respiratory tract, skin, and nails (Arrese et al, 1996; Barbasso et al, 2005; Barnes et al, 2007; Romano et al, 2001; Mirkin, 1994; Machet et al, 1994; Hazouard et al, 1999). In a number of cases, these infections have been fatal or associated with certain cancers (Liu et al, 1992; Yekeler et al, 2001; Brugger et al, 2006; Hazouard et al, 1999; Neumeister et al, 1994). Perhaps more importantly from an economic and chronic health perspective, *A. alternata* spores are one of the most common and potent airborne sources of allergens yet very little is known regarding the immunological properties of and human host response to Alternaria proteins (Black et al, 2000; Bush and Prochnau, 2004). In the United States alone, over 3 billion dollars are spent annually for the relief of allergic rhinitis, much of which is due to sensitization to Alternaria spores (Bush and Prochnau, 2004). Additionally, *A. alternata* sensitization has been determined to be one of the most important factors in the onset of childhood and fatal asthma (O'Hollaren et al, 1991; Black et al, 2000; Bush and Prochnau, 2004).

Currently eight major and minor proteinaceous allergens have been identified in *A. alternata* including the major allergen Alta1 (Yunginger and Jones, 1978). In a recent study in our laboratory, it was demonstrated that *A. brassicicola* expresses *Altb1*, a highly conserved homolog of *Alta1 in vitro*, and it was found to be highly upregulated during pathogenesis of Arabidopsis (Cramer and Lawrence, 2003; Cramer and Lawrence, 2004). This study was the first to show that this major allergen gene was present in another species of Alternaria besides *A. alternata*. Since this study, it has now been shown by our laboratory and collaborators that over 52 species of Alternaria and very closely related fungi possess highly conserved Alta1 homologs, which suggests that virtually every species within the genus is potentially allergenic (Hong et al, 2005). Alta1 has been recently shown to possess phosphatase activity (Saenz-de-Santamaria et al, 2006). Preliminary data generated in our laboratory also suggest that this protein has phytotoxic properties and is a putative virulence factor. However, little is known about how Alta1 affects the mammalian immune system (Lawrence, unpublished results).

In several recent studies related to chronic rhinosinusitis, antigen extracts derived from *A. alternata* appear to have more potent immunostimulatory activity towards chronic rhinosinusitis patient lymphocytes, eosinophilis, and in a nasal epithelium explant model than other fungi examined including *Aspergillus*, *Cladosporium*, and *Penicillium* (Shin et al, 2006; Shin et al, 2004). Over the last two years, we have established a consortium with researchers at the Mayo Clinic, Allergic Diseases Laboratory, directed by Dr. Hirohito Kita, and now have two, active National Institutes of Health-funded projects. One project is directed towards furthering our understanding of the pathogenesis of chronic rhinosinusitis and the role of Alternaria proteins in airway inflammation in humans. The other project is centered on furthering our understanding of how secreted Alternaria proteins influence the innate and adaptive immune systems in a mouse model of allergic airway inflammation and using patient cells. Several candidate proteins have already been identified using a combination of proteomics and immunology. These candidate proteins exhibit potent eosinophil degranulating activity or cause maturation of naïve T-cells into the T helper 2 type (Th_2). Eosinophils are a type of granulocyte associated with airway inflammation and Th_2 cells have been associated with allergic airway inflammation due to their ability to secrete specific cytokines, such as interleukin-5 and interleukin-13, involved in eosinophil recruitment and survival. Extensive immunological and biochemical analysis of these identified proteins is now underway. Interestingly, many of these Alternaria proteins are predicted to be hydrolytic enzymes such as proteases or glycosyl hydrolases.

In a separate project also in collaboration with Mayo, we have demonstrated in the reporting period that degranulation of human eosinophils is dependent upon the recognition of fungal cell wall β glucan by the β2-integrin CD11b and not through known fungal glucan receptors such as Dectin 1 (Yoon et al, 2008). We have started amassing *A. alternata* genomic sequence information through an expressed sequence tag project funded by the Mayo Research Foundation. In this pilot study, we have obtained and analyzed over 20,000 expressed sequence tags derived from a cDNA library created from mRNA isolated from spores germinating in the presence of mucin. Preliminary analysis of these sequences has revealed an overabundance of hydrolytic enzyme coding genes such as secreted proteases and known fungal allergens.

Finally, we have initiated studies in the reporting period in collaboration with Dr. Liwu Li at Virginia Tech focused on studying innate immunity to fungi and how this may affect fungal-associated allergic airway inflammation. In these studies we are focusing on the role of interleukin 1 receptor associated kinases (IRAK) in innate immune responses to fungi by using knockout mice and various cell types derived from the knockouts that are deficient in various IRAK isoforms.

Acknowledgements

Many of the studies described in this report were supported by the Virginia Bioinformatics Institute, National Science Foundation award number DBI-0443991 to C.B.L., and the National Research Initiative of the United States Department of Agriculture Cooperative State Research, Education and Extension Service, grant number #2004-35600-15030 to C.B.L.

Conferences and workshops

Lawrence CB (2007) Alternaria genomics: functional approaches for dissecting plant pathogenicity and fungal associated airway disorders in humans, Invited talk at *Korean Conference on Innovation in Science and Technology*, Jeju, South Korea, July 11-13, 2007.

Lawrence CB (2007) The *Alternaria brassicicola* genome sequencing project, Invited talk at *24th Fungal Genetics Conference*, Asilomar, CA, March 20-27, 2007.

Lawrence CB (2007) Identification of pathogenicity factors among secondary metabolite related genes in *Alternaria brassicicola*, Invited talk at *24th Fungal Genetics Conference*, Asilomar, CA, March 20-27, 2007.

Lawrence CB (2008) The role of the ubiquitous fungus Alternaria in chronic respiratory disorders of humans, Invited lecture, Allergy Residents Training Program, SW Texas Medical Center, Department of Internal Medicine, Division of Allergy, Pulmonary, Immunology, Critical Care and Sleep Disorders (APICS), Galveston, TX.

Lawrence CB (2008) Alternaria pathogenomics and biotechnology. Invited talk at the Department of Veterinary Molecular Biology, Montana State University, Bozeman, MT.

Peer-reviewed publications in the reporting period

Cho Y, Cramer RC, Kim K-H, Pryor BM, Mitchell TK, Lawrence CB (2007) The Amk1 map kinase regulates hydrolytic enzyme gene expression in *Alternaria brassicicola*. *Fungal Genetics and Biology* **44**: 543-553.

Cho Y, Kim K-H, Scott D, Santopietro G, Mitchell TK, Lawrence CB (2008) Identification of novel virulence factors associated with signal transduction pathways in *Alternaria brassicicola*. *Molecular Microbiology*. In press.

Craven K, Valez H, Cho Y, Lawrence CB, Mitchell, TK (2008) Anastomosis is required for pathogenicity of *Alternaria brassiciola*. *Eukaryotic Cell* **7**: 675-683.

Kim K-H, Cho Y, La Rota CM, Cramer RC, Lawrence CB (2007) Functional analysis of the *Alternaria brassicicola* non-ribosomal peptide synthetase gene *AbNRPS2* reveals a role in conidial cell wall construction. *Molecular Plant Pathology* **8**: 23-29.

Lawrence CB, Mitchell TK, Cramer RC, Craven KD, Cho Y, Kim K-H (2008) At death's door: Alternaria pathogenicity mechanisms. *The Plant Pathology Journal* **24**: 101-111.

Yoon J, Ponikau JU, Lawrence CB, Kita H (2008) Innate anti-fungal immunity of human eosinophils mediated by a β2-integrin, CD11b. *Journal of Immunology*. In press.

Other Cited References

Arrese JE, PierardFranchimont C, Pierard GE (1996) Onychomycosis and keratomycosis caused by *Alternaria* sp. - A bipolar opportunistic infection in a wood-pulp worker on chronic steroid therapy. *American Journal of Dermatopathology* **18**: 611-613.

Barbasso E, Sforza F, Stoppini R, Peruccio C (2005) Mycotic flora in the conjunctival fornix of horses in Northern Italy. *Ippologia* **16**: 21-29.

Barnes SD, Dohlman CH, Durand ML (2007) Fungal colonization and infection in Boston keratoprosthesis. *Cornea* **26**: 9-15.

Black PN, Udy AA, Brodie SM (2000) Sensitivity to fungal allergens is a risk factor for life-threatening asthma. *Allergy* **55**: 501-504.

Brugger EM, Wagner J, Schumacher DM, Koch K, Podlech J, Metzler M, *Lehmann L (*2006) Mutagenicity of the mycotoxin alternariol in cultured mammalian cells. *Toxicology Letters* **164**: 221-230.

Bush RK, Prochnau JJ (2004) Alternaria-induced asthma. *Journal of Allergy and Clinical Immunology* **113**: 227-234.

Cho Y, Davis JW, Kim KH, Wang J, Sun QH, 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 Interactions* **19**: 7-15.

Cramer RA, Lawrence CB (2003) Cloning of a gene encoding an Alt a 1 isoallergen differentially expressed by the necrotrophic fungus *Alternaria brassicicola* during Arabidopsis infection. *Applied Environmental Microbiology* **69**: 2361-2364.

Cramer RA, Lawrence CB (2004) Identification of *Alternaria brassicicola* genes expressed in planta during pathogenesis of *Arabidopsis thaliana*. *Fungal Genetics and Biology* **41**: 115-128.

Hazouard E, Segalen J, Belleguic C, Quinquenel ML, Kernec J, Diot P, Delaval P (1999) Adult bronchiectasis revealing familial ciliary anomaly. Revue des maladies respiratoires 16(5): 839-841.

Hong SG, Cramer RA, Lawrence CB, Pryor BM (2005) Alt a 1 allergen homologs from Alternaria and related taxa: analysis of phylogenetic content and secondary structure. Fungal Genetics and Biology 42(2): 119-129.

Liu GT, Qian YZ, Zhang P, Dong WH, Qi YM, Guo HT (1992) Etiologic role of *Alternaria alternata* in human esophageal cancer. *Chinese Medical Journal* **105**: 394-400.

Machet MC, Stephanov E, Esteve E, Declosets F, Barrabes A, Therizolferly M, Lebret G, Grangeponte MC, Vaillant L (1994) Cutaneous alternariosis occurring in the course of pemphigus - report on 2 cases. *Annals of Pathology* **14**: 186-191.

Mirkin LD (1994) *Alternaria alternata* infection of skin in a 6-year-old boy with aplastic anemia. *Pediatric Pathology* **14**: 757-761.

Morrison VA, Weisdorf DJ (1993) Alternaria - a sinonasal pathogen of immunocompromised hosts. *Clinical Infectious Diseases* **16**: 265-270.

Neumeister B, Hartmann W, Oethinger M, Heymer B, Marre R (1994) A fatal infection with Alternaria alternata and Aspergillus terreus in a child with agranulocytosis of unknown origin. Mycoses 37(5-6): 181-185.

O'Hollaren MT, Yunginger JW, Offord KP, Somers MJ, O'Connell EJ, Ballard DJ, Sachs MI (1991) Exposure to an aeroallergen as a possible precipitating factor in respiratory arrest in young-patients with asthma. *New England Journal of Medicine* **324**: 359-363.

Ospina-Giraldo MD, Mullins E, Kang S (2003) Loss of function of the *Fusarium oxysporum* SNF1 gene reduces virulence on cabbage and Arabidopsis. *Current Genetics* **44**: 49-57.

Romano C, Paccagnini E, Difonzo EM (2001) Onychomycosis caused by Alternaria spp. in Tuscany, Italy from 1985 to 1999. Mycoses 44(3-4): 73-76.

Saenz-de-Santamaria M, Guisantes JA, Martinez J (2006) Enzymatic activities of *Alternaria alternata* allergenic extracts and its major allergen (Alt a 1). *Mycoses* **49**: 288-292.

Shin SH, Lee YH, Jeon CH (2006) Protease-dependent activation of nasal polyp epithelial cells by airborne fungi leads to migration of eosinophils and neutrophils. *Acta Otolaryngologica* **126**: 1286-1294.

Shin SH, Ponikau JU, Sherris DA, Congdon D, Frigas E, Homburger HA, Swanson MC, Gleich GJ, Kita H (2004) Chronic rhinosinusitis: An enhanced immune response to ubiquitous airborne fungi. *Journal of Allergy and Clinical Immunology* **114**: 1369-1375.

Tonukari NJ, Scott-Craig JS, Walton JD (2000) The *Cochliobolus carbonum SNF1* gene is required for cell wall-degrading enzyme expression and virulence on maize. *Plant Cell* **12**: 237-248.

Vartivarian SE, Anaissie EJ, Bodey GP (1993) Emerging fungal pathogens in immunocompromised patients - classification, diagnosis, and management. *Clinical Infectious Diseases* **17**: S487-S491.

Xu JR (2000) MAP kinases in fungal pathogens. *Fungal Genetics and Biology* **31**: 137-152.

Yekeler H, Bitmis K, Ozcelik N, Doymaz MZ, Calta M (2001) Analysis of toxic effects of Alternaria toxins on esophagus of mice by light and electron microscopy. *Toxicologic Pathology* **29**: 492-497.

Yunginger JW, Jones RT (1978) Isolation of Alternaria allergens. *Federation Proceedings* **37**: 1553-1553.

Iuliana M. Lazar

lazar@vbi.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, Yang Xu.

Quantitative Proteomic Studies of MCF-7 Breast Cancer Cells

Abstract. Monitoring quantitative changes in the cellular proteome in response to stress inducing conditions is essential to decipher complex signaling pathways. The discovery of novel-diagnostic protein patterns and therapeutic targets could have tremendous impact on public health, which highlights the importance of facilitating such discoveries. As a result, the scientific community has invested extensive efforts into the development of adequate technologies for monitoring the up/down regulation of proteins in the complex cellular milieu. We have developed and implemented two methods based on stable isotope labeling and spectral counting for protein differential expression profiling. Both conventional and microfluidic platforms were used to demonstrate the applicability of the two techniques. MCF-7 breast cancer cells were cultured in the presence of estradiol, the most abundant circulating estrogen in humans, and tamoxifen, a non-steroidal drug commonly prescribed in hormonal breast cancer therapy. Over 500 proteins ($P<0.001$) were identified in the two extracts – approximately 255 proteins were matched by two peptides and 16 proteins exhibited a greater than two-fold change in expression level as a result of estrogen stimulation or tamoxifen inhibition of cell proliferation. The stand-alone microfluidic liquid chromatography system has enabled the identification of ~50 proteins, along with several previously reported biomarkers. In future work, the label-free approach will be compared with the stable isotope labeling method, and the differential protein expression pattern will be evaluated for biological relevance.

Keywords: oncoproteomics, mass spectrometry, protein differential expression analysis.

Scientific Progress

Protein quantitation by stable isotope labeling

Breast cancer is the most common type of cancer among women worldwide. As a result, there has been increased interest in the elucidation of pathways that lead to its development and of how estrogens work in the stimulation of breast cancer. Among innovative approaches for cancer treatment, tamoxifen has proven very effective in improving survival rates.

Two-dimensional liquid chromatography tandem mass spectrometry (2DLC-MS/MS) is a technique that has found widespread applicability in quantitative proteomic studies. MS approaches for protein quantitation typically involve label-free and stable isotope labeling techniques. More recently, the strategy of isobaric tags for relative and absolute quantitation (iTRAQ) has captured the attention of many scientists. For this technique, isobaric compounds are used for tagging peptides at the N-terminus and the lysine side chains for MS quantitation. Upon MS/MS fragmentation, signature ions are produced that can be used to obtain quantitative information. The advantages of this approach are its simplicity and multiplexing capability, which enables the simultaneous analysis of 4-8 samples, reducing analysis times.

Typically, iTRAQ-LC/MS/MS is conducted on time-of-flight (TOF)/TOF-MS instruments. In order to extend the applicability of this technology to ion trap mass spectrometers, a unique technique termed pulse Q dissociation (PQD), which is used to trap low *m*/*z* ions, has been implemented. Very few studies related to the performance of this technique have been reported. We have conducted a quantitative proteomic study on the MCF-7 breast cancer cell line using iTRAQ-LC/MS/MS with PQD detection with the goal of discovering biomarkers. The effect of estradiol and tamoxifen on the proteomic profile of these cells was investigated. In an extended application, the potential of a microfluidic LC-MS platform for differential expression analysis of MCF-7 breast cancer cells using the iTRAQ-PQD technology was also evaluated.

MCF-7 cells were cultured under four different conditions in the presence of estradiol and/or tamoxifen. The cultured cells were lysed and the protein concentration of each sample was measured with the Bradford assay. Each cell extract was spiked with a standard protein mix solution consisting of eight proteins. The protein extracts were separately digested with trypsin, cleaned-up with SPEC-PTC18 solid-phase extraction pipette tips, and labeled with iTRAQ reagents (that generate the reporter ions 114, 115, 116, 117). The iTRAQ-labeled samples

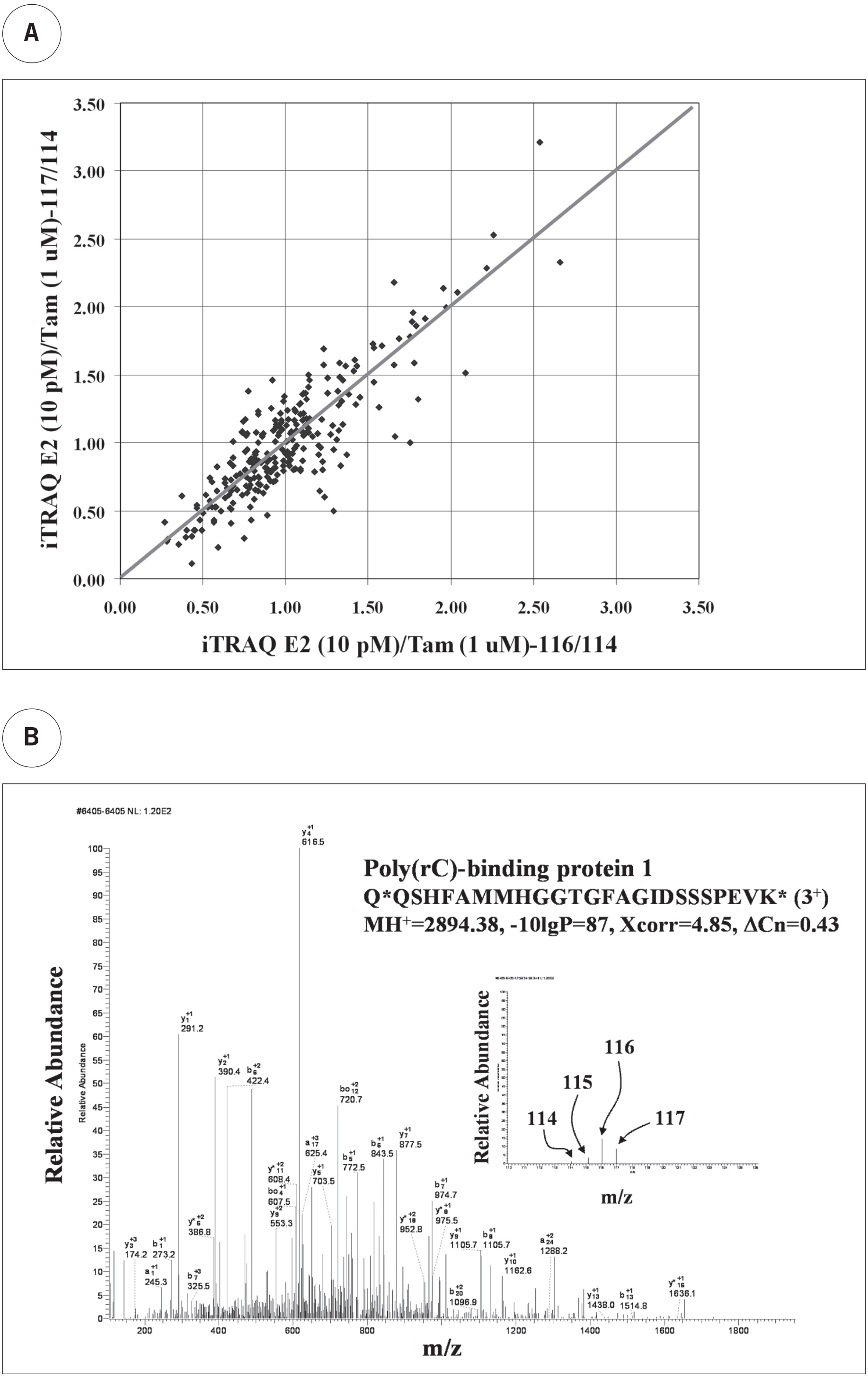


Figure. Protein differential expression analysis of MCF-7 extracts using iTRAQ reagents.

were combined in a 1:1:1:1 ratio and cleaned-up with SPEC-PTSCX solid-phase extraction cartridges. RPLC-MS/MS analysis using both bench top instruments and microfluidic LC chips was conducted on the iTRAQ-labeled whole cellular extract.

After extensive data filtering, a list of 255 proteins (all with $P<0.001$, P being the probability of a random match) was generated for differential expression profiling. The sample complexity, protein concentration levels, and number of peptides used for protein identifications were the factors that most affected the accuracy of the quantitation. For example, >85% of the proteins quantified in the MCF7 extracts by 5 unique peptides had percentage variations of iTRAQ ratios within ±30 % of the theoretical values. The relative standard deviation of the experimental global average iTRAQ ratios calculated for five LC/MS/MS runs was <8%, demonstrating the reproducibility of the analysis. A scatter plot of iTRAQ ratios for two experimental replicates of iTRAQ-labeled MCF-7 extracts is shown in Figure A. A list of 16 proteins exhibited >two-fold change in expression level – two of these proteins were up-regulated, and 14 down-regulated. The tandem mass spectrum of an up-regulated peptide is shown in Figure B. The inset displays the m/z region of the iTRAQ reporter ions: 116 and 117 up-regulated vs. 114 and 115 control. The up/down-regulated proteins were involved in biological processes such as cell proliferation, apoptosis, oncogenesis and metastasis, which all play major roles in cancer development. The differential expression of these proteins was confirmed by means of statistical tools (in collaboration with Dr. Ina Hoeschele at VBI). Microchip-LC-MS/MS analysis of iTRAQ-labeled MCF-7 breast cancer cell extracts enabled rapid identification of 40-50 proteins and several previously reported human putative cancer biomarkers. The advantages of applying microfluidic-LC to quantitative proteomic studies include high-throughput capabilities, no sample carryover, and disposability. The generated results demonstrate the applicability of iTRAQ-LC/MS/MS with PQD ion trap detection, and of microfluidic-LC chips, for differential expression analysis of complex samples and biomarker discovery.

Protein quantitation by spectral counting

Recently, label-free methods have emerged as alternative approaches to stable isotope labeling for protein quantitation by LC-MS. Label-free methods rely on measuring peak areas and intensities or spectral counts, and benefit from not having to chemically alter the sample. As a result of its simplicity, spectral counting has proven very useful for quantifying relative changes in protein expression levels. Spectral counting is based on the assumption that there is a linear relationship between the abundance of a given protein and its spectral count.

We have implemented a spectral counting approach to systematically evaluate the linear dynamic range, accuracy and variation of this method when applied to the quantitation of a complex cellular extract. MCF-7 breast cancer cells were cultured in the presence of insulin and the protein soluble fraction was digested with trypsin. The resulting peptide mixture was cleaned up with C18 SPEC-PTSCX solid-phase extraction pipette tips and a series of dilutions on the original sample (2, 3, 5, 10 X) were performed. RPLC-MS/MS analysis was conducted on each dilution sample.

For low abundance proteins, the relationship between spectral count and abundance was, indeed, linear. However for high abundance proteins, the linear relationship was lost. For example, for a 10-fold change in concentration (by dilution), proteins with a spectral count of <20 exhibited a linear relationship between spectral count and abundance. The linearity cut-off count, however, remains to be determined. Overall, the spectral count experiments have revealed that ~85% of the proteins that are matched by at least three-to-five spectra will change their spectral count at a two-fold change in concentration, while ~98% of proteins will change their spectral count only at a three-fold change in concentration. To support these observations, we hypothesize that spectral counting is not only dependent on protein abundance, but also on protein length (molecular weight), the number of protease cleavage sites, and the mass of the generated peptides. Further experiments need to be performed to determine the linear dynamic range of this method.

Conferences and workshops

Armenta JM, Yang X, Dawoud AA, Lazar IM (2007) Comparative Analysis of Non-Labeling and iTRAQ Based Methods for Protein Quantification in the MCF7 Cell Line Using 2DLC/MS/MS, Poster Presentation at *55th Conference on Mass Spectrometry and Allied Topics,* Indianapolis, IN, June 3-7, 2007.

Dawoud AA, Armenta JM, Lazar IM (2007) Phosphoproteomics on a Microfluidic-Mass Spectrometry Analysis Platform, Poster Presented at *55th Conference on Mass Spectrometry and Allied Topics,* Indianapolis, IN, June 3-7, 2007.

Lazar, IM, Armenta JM, Dawoud AA (2007) Microfluidic Technology for Proteomic Profiling and Biomarker Screening Applications, *Montreux Symposium on LC-MS*, Hilton Head Island, SC, October 8-12, 2007.

Lazar IM, Armenta JM, Dawoud AA, Yang X (2007) Microfluidic Strategies for Proteomic Profiling and Biomarker Discovery, Poster presented at *INCEMC/Timisoara Polytechnic Institute-10 Year Anniversary Symposium*, Timisoara, Romania, April 19-20, 2007.

Lazar IM, Armenta JM, Dawoud AA, Yang X (2007) Novel Microfluidic Strategies for Proteomic Profiling of Cancerous Cell Extracts, Poster presented at *55th Conference on Mass Spectrometry and Allied Topics,* Indianapolis, IN, June 3-7, 2007.

Intellectual property

Lazar IM Microfluidic Devices and Methods Facilitating High-Throughput, On-Chip Detection of Separation Techniques, Disclosure (06.057) made to the university in September 2005. US Utility filed in August 2006, PCT filed in August 2007.

Peer-reviewed publications in reporting period

Dawoud AA, Sarvaiya HA, Lazar IM (2007) Microfluidic platform with mass spectrometry detection for the analysis of phosphoproteins. *Electrophoresis* **28**: 4645-4660.

Lazar IM (2007) Microfluidic devices with mass spectrometry detection. In *Handbook of Capillary and Microchip Electrophoresis and Associated Microtechniques*. Landers JP (ed), CRC Press, 3rd edition, pp. 1459-1506.

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: Hui Cheng, Stefan Hoops, Aejaaz Kamal, Ana M. Martins, Bharat Mehrotra, Revonda Pokrzywa.

Computational Systems Biology

Abstract. Computational modeling and simulation of biochemical networks is at the core of systems biology and includes many types of analyses that can aid in the understanding of how these systems work. We have been developing COPASI (Complex Pathway Simulator), a generic software package for modeling and simulation of biochemical networks that provides many of these analyses in convenient ways not requiring the user to program or to have deep knowledge of the numerical algorithms. B-Net is a database schema to organize background knowledge about the biochemical networks, including which reactions transform specific metabolites, as well as the enzymes that catalyze them and the genes that encode them. B-Net was adapted to work for whole-genome metabolic network reconstruction and was used in building a consensus metabolic network of baker's yeast. We have also continued our involvement in the establishment of community standards for systems biology that play a formal role in the definition of the Systems Biology Markup Language.

Keywords: systems biology; biochemical networks; computer simulation; computational modeling; software development; community standards; data analysis; enzyme kinetics.

Scientific Progress

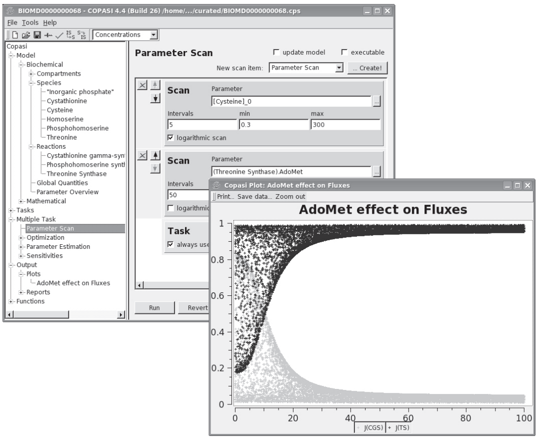
Biochemical networks are intrinsically complex, not only because they encompass a large number of interacting components, but also because the interactions are nonlinear. Like many other nonlinear phenomena in nature, their behavior is often nonintuitive and thus quantitative models are needed to describe and understand their function. Computational modeling is thus a technique of systems biology as important as its experimental counterparts.

COPASI

COPASI (Complex Pathway Simulator) is an open-source software package for the modeling and simulation of biochemical networks. It allows users with limited knowledge of mathematics to construct kinetic models. COPASI, which runs on Windows, Mac, Linux, and Solaris, provides the means to carry out simulations (time course and steady states) and analyze stoichiometric properties and sensitivity to model parameters. It also allows users to optimize models according to specific criteria of arbitrary complexity and fit models to experimental data.

This year we have achieved a very significant milestone in this project – the receipt of an R01 grant from the National Institutes of Health (National Institute of General Medical Sciences and National Center for Research Resources), including a subcontract to Ursula Kummer's group at the University of Heidelberg, which is a long-standing collaboration. This is important because it allows us to continue developing COPASI by adding an array of features that are yet unavailable in any systems biology software. Our four-year project plan includes the development of tools to explore the dynamics of network models, parallelize the software to make it ready for high-performance computing, and make enhancements to the user-interface based on a collection of feedback from users. This award also serves as a recognition of COPASI as one of the major systems biology packages, which is something that is also evident from the large number of downloads and rapid accumulation of citations to the article describing the software (16 per year). Additionally, there are now several commercial licensees of the software.

The new features in COPASI added in the last year have improved the modeling capabilities of the software in several areas. Models now include compartments with volume that is variable and can also include arbitrary variables that can be anything other than chemical species or compartment volumes. It is now also possible to specify dependencies between initial values of various model entities. Another important new modeling tool introduced is an algorithm that checks for possible mistakes in the rate equations, which is completed by



Parameter Scan □ update model □ executable

New scan item: Parameter Scan ... Create!

Scan — Parameter: [Cysteine]_0; Intervals: 5; min: 0.3; max: 300; ☑ logarithmic scan

Scan — Parameter: (Threonine Synthase).AdoMet; Intervals: 50; □ logarithmic

Task — ☑ always use

Figure. COPASI user interface showing a stochastic sampling of parameter space of a biochemical model.

symbolic analysis. These new features appeared over the course of five consecutive new version releases at regular intervals throughout the year. Our policy is to release new versions frequently, as this allows us to maximize collection of user feedback. This feedback is essential in keeping the project relevant to the whole research community and identifying existing problems in the software. It also gives us the opportunity to provide solutions in a timely manner.

Modeling biochemical networks

We have continued our modeling work in specific biochemical networks and developing methods to reverse-engineer such networks directly from data. We have collected data in previous years to construct a detailed model of the pentose-phosphate pathway of yeast by direct measurement of *in vitro* kinetics of its constituent enzymes. This has continued with a comparative analysis of the two isoenzymes of 6-phosphogluconate dehydrogenase.

A collaboration with the Laubenbacher and Shulaev groups at the Virginia Bioinformatics Institute, as well as with colleagues in the Cancer Biology Department at Wake Forest University, has involved reconstructing the human network of iron metabolism. This has been reconstructed directly from published literature and stored in Systems Biology Markup Language (SBML) using Cell Designer software. Specific versions for different human cell types were also created after construction of a generic version of this model including all know reactions in humans. This network is important in some types of cancer, as was revealed in prior microarray results we obtained.

We have continued research in the regulatory network that is responsible for the oxidative stress response in yeast in collaboration with the Laubenbacher and Shulaev groups. Our group has been mostly responsible for data analysis. We have also continued research in methods to identify networks directly from omics data (reverse-engineering) under the auspices of the same project,.

Community standards

SBML is an application-independent file format to specify systems biology models. SBML has been a focal point for the computational systems biology community and a catalyst for new developments in this area. Our involvement with the SBML effort started at its inception and has continued uninterrupted. This year our involvement becomes more formal with the election of Dr. Stefan Hoops to serve as an SBML Editor. SBML Editors are a small group of researchers in charge of writing the specifications of the language, both in terms of new features and corrections of past specifications.

Another activity in this area was participation in the construction of a consensus model of the metabolic network of *Saccharomyces cerevisiae* (baker's yeast). This effort stemmed from the publication of two different metabolic reconstructions of this network by the groups of Bernhard Palsson, University of California, San Diego and Uwe Sauer, Institute of Molecular Systems Biology. However, these two models referred to metabolites through arbitrary abbreviations, which were in many instances different in each model. Unfortunately these arbitrary names were in most cases not sufficiently explicit to identify molecules unequivocally. Additionally, each one of the reconstructions contained reactions not present in the other.

To solve these problems, several researchers assembled at an annotation jamboree in April 2007 at the Manchester Interdisciplinary Biocentre, University of Manchester. The main task of the group was to uniquely identify all molecules using International Chemical Identifier (InCHI) (a string representation of the two-dimensional structure of molecules), annotate them with accession numbers from other databases (e.g. Kyoto Encyclopedia of Genes and Genomes [KEGG] and Chemical Entities of Biological Interest [ChEBI]), and collect as many synonyms as possible for each chemical species. This information is made available in an SBML file and a database using the B-Net schema that our group has developed in earlier years (with slight adaptions that were required here). These data are available at http://www.comp-sys-bio.org/yeastnet and a manuscript is now in press.

Conferences and workshops

Mendes P (2008) New Developments in COPASI, *Workshop on Computational Cell Biology*, Hinxton, United Kingdom, March 26-29, 2008.

Mendes P, De La Fuente A, Camacho D (2007) Adventures in Reverse Engineering of Biochemical Networks, *The Microsoft Research – University of Trento Centre for Computational and Systems Biology*, Trento, Italy, November 15, 2007.

Mendes P, De La Fuente A, Camacho D (2007) Adventures in Reverse Engineering of Biochemical Networks, *Otto Warburg International Summer School and Workshop on Computational Systems Biology*, Berlin, Germany, August 27- September 5, 2007.

Mendes P, Mehrotra B, Li XL, Kamal A, Hoops S (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.

Mendes P, Sahle S, Hoops S (2007) Advanced Model Analysis with COPASI, *International Conference for Systems Biology*, Long Beach, CA, October 1-6, 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* **1115**: 73-89.

Kell DB, Mendes P (2007) The markup is the model: Reasoning about systems biology models in the Semantic Web era. *Journal of Theoretical Biology* **252**: 538-543.

Bandara AB, Contreras A, Contreras-Rodriguez A, Martins AM, Dobrean V, Poff-Reichow S, Rajasekaran P, Sriranganathan N, Schurig GG, Boyle SM (2007) *Brucella suis* urease encoded by *ure1* but not *ure2* is necessary for intestinal infection of BALB/c mice. *BMC Microbiology* **7**: 57.

Biswarup Mukhopadhyay

biswarup@vbi.vt.edu

Assisant Professor, Virginia Bioinformatics Institute

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

Group contributors: Christopher L. Case, Lakshmi Dharmarajan, Eric F. Johnson, Usha Loganathan, Karla Piedl, Endang Purwantini, Jason Rodriguez, Jennifer P. Stieber, Dwi Susanti, Lindsay Von Herbulis, Ban Wang.

Evolution of Methanogenic Archaea, Biological Gasification of Coal, PEPCK Mechanism, and Mycobacterial Coenzyme F_{420} Metabolism

Abstract. From the results of our research on the methanogenic archaea, we have hypothesized that methanogenesis, sulfate reduction and anaerobic oxidation of methane – three of the oldest respiratory energy metabolisms of earth – at one time existed in one organism, which was a strict anaerobe. To gain a competitive advantage, this ancestral organism converted highly diffusible and reactive H_2 to CH_4, which is less reactive under anaerobic conditions. After H_2 was exhausted, it oxidized accumulated CH_4 to CO_2 and reduced sulfate to sulfide. Thus, biological methanogenesis originally developed as a storage metabolism. We have begun a metagenomic investigation for discovering new enzymes and pathways that will allow biological gasification of coal and paraffin. The residue Tyr^{235} of human liver cytosolic GTP-dependent phosphoenolpyruvate carboxykinase (PEPCK) was found to exert a fine control on the activity of the enzyme. We found that the mycobacteria might utilize reduced coenzyme F_{420} ($F_{420}H_2$) in its defense against the bactericidal effect of macrophages and for the biosynthesis of mycolic acids, which are critical components of mycobacterial cell envelope.

Keywords: evolution; archaea; sulfur; coal gasification; metagenomics; phosphoenolpyruvate carboxylase; phosphoenolpyruvate carboxykinase; diabetes; coenzyme F_{420}; tuberculosis.

Scientific Progress

Evolutionary and Archaeal Biology

> Evolution of methanogenesis, sulfate reduction and anaerobic oxidation of methane, three of the oldest respiratory energy metabolisms of earth

Our enzymology (Johnson and Mukhopadhyay, 2005) and genetic and genome analysis (Johnson and Mukhopadhyay, 2007; Johnson and Mukhopadhyay, 2008) research on the methanogenic archaea that inhabit hyperthermophilic and psychrophilic environments has led to the hypothesis that the extant anaerobic archaea that perform methanogenesis, sulfate reduction, and methane oxidation descended from an ancient archaeon. This archaeon embodied these respiratory energy metabolisms, which are three of the oldest on earth, and used methanogenesis as an energy storage mechanism (Susanti and Mukhopadhyay, unpublished results). Sulfite is an obligatory intermediate for the reduction of sulfate to sulfide ($SO_4^{2-} + 2e \rightarrow SO_3^{2-}$; $SO_3^{2-} + 6e \rightarrow S^{2-}$) (LeGall and Fauque, 1988). Methanogens are generally sensitive to sulfite because this oxyanion inhibits methylcoenzyme M reductase (Mcr), which is essential for methanogenesis ($4H_2 + CO_2 \rightarrow CH_4 + 2H_2O$) (1, 17). The same is true for anaerobic oxidation of methane (AOM) ($CH_4 \rightarrow CO_2 + 8e$), which also requires Mcr (Shima and Thauer, 2005). For this reason, on today's earth sulfate reduction coupled AOM is carried out by a consortium, where an archaeon performs the AOM and the electrons generated from this process are transferred to a bacterium that reduces sulfate (Boetius et al, 2000). However, our research has shown that certain methanogens from extremely hot and permanently cold environments and an archaeal member of an AOM consortium carry a novel coenzyme F_{420}-dependent sulfite reductase (Fsr) that enables a methanogen not only to withstand, but also utilize sulfite as a sulfur source (Johnson and Mukhopadhyay, 2005; Johnson and Mukhopadhyay, 2007; Johnson and Mukhopadhyay, 2008). Even more surprisingly, methanogens carry a variety of small dissimilatory sulfite reductase-type ORFs (sDsr) (Johnson and Mukhopadhyay, 2007; Susanti and Mukhopadhyay, unpublished results). Geological modeling suggests that the methanogens that live within the hydrothermal vents are exposed to sulfite and that the early methanogens experienced similar conditions during the early oxygenation of sulfide-rich earth (Johnson and Mukhopadhyay, 2007). Using structural modeling, we have identified two ORFs in *Methanocaldococcus jannaschii* as putative ATP sulfurylase and APS reductase (Sat and Apr) that catalyze the reaction $SO_4^{2-} + 2e \rightarrow SO_3^{2-}$ (Dharmarajan and Mukhopadhyay, unpublished data). Hence, the complete reduction of sulfate probably occurred in an ancient methanogen. To gain a competitive advantage, this methanogen converted highly reactive (diffusible) H_2 to CH_4, which is less reactive under anaerobic conditions. After H_2 got exhausted, it oxidized accumulated CH_4 to CO_2 and reduced sulfate to sulfide. We are investigating the enzymology, *in vivo* roles, and gene regulation of Fsr, sDsrs and putative Sat and Apr and attempting to isolate and/or construct the envisaged ancient archaeon. The enzymes employed by such organisms for the reduction of sulfur are also under investigation (Rodriguez, Case, and Mukhopadhyay, unpublished results).

Microbial conversion of coal and paraffin to natural gas

This work represents a collaboration with Altuda Energy Corporation, San Antonio, TX, and has support from the Natural Gas industry. It is focused on following questions: 1. Do we need a set of new enzymes and pathways for degrading these complex and inert compounds (coal and paraffin)? and 2. Are such systems hard to recognize in available databases or found only in some microorganisms that are yet to be discovered? The investigation has the following objectives: (a) To develop microbial consortia from coalbed well water and core samples that would degrade coal and paraffin to methane, or at least to smaller soluble compounds that can be converted to methane by use of a mixture of known microorganisms; (b) To determine the breadth of microbial diversity in the coalbed by amplifying, cloning and sequencing 16S rRNA genes; and (c) To explore new biodegradation enzymes via a metagenomic analysis of DNA isolated from cores and water samples.

Structural biology of GTP-PEPCK

GTP-dependent PEP carboxykinase (GTP-PEPCK) catalyzes nucleotide-dependent reversible synthesis of PEP from oxaloacetate or OAA ($GTP + OAA \rightarrow GDP + PEP + HCO_3^-$), a critical step for gluconeogenesis and glycerogenesis in certain archaea and bacteria and several eukaryotes, including humans. The gluconeogenic activity of GTP-PEPCK is most likely required for attainment and maintenance of dormancy or latency of *Mycobacterium tuberculosis*, which causes tuberculosis (TB), within a human host (Mukhopadhyay, Concar and Wolfe, 2001; Mukhopadhyay and Purwantini, 2000). In humans, PEPCK helps to maintain normal blood glucose level during fasting (between meals). An untimely and unusually high activity of the enzyme causes unwarranted production of glycerol and glucose and contributes to the development of obesity and type 2 diabetes, respectively (Beale, Harvey and Forest, 2007). An agent that will lower the activity of this enzyme would be useful in the treatment of these diseases, although a complete inactivation will cause fasting hypoglycemia. Our goal is to find the sites on the PEPCK that are distal to the active site but connected to catalytically essential events via long-range interactions. These sites can be exploited for developing therapeutic agents that will lower the activity of the enzyme, but not inactivate the enzyme and cross-react with other PEP, GTP, GDP, OAA or bicarbonate-utilizing enzymes. The distal sites are often not under high selective pressure for conservation. This logic will also help to identify target sites of mycobacterial GTP-PEPCK for developing TB drugs that will not interact with the human enzyme. We have developed a model for the mycobacterial enzymes (Mukhopadhyay, Concar and Wolfe, 2001) and are studying a human GTP-PEPCK (Case and Mukhopadhyay, 2007). Our investigations employing structural modeling, site-directed mutagenesis and kinetic analysis have shown that the fully conserved Tyr^{235} and Asn^{403} of human PEPCK exert fine control on catalysis (Dharmarajan, Case, Dunten and Mukhopadhyay, unpublished results). The aromatic ring of Tyr^{235} interacts with the PEP carboxylate and thereby helps to establish a favorable distance between the PEP phosphate and enzyme-bound Mn^{2+}, which is critical for catalysis. The hydroxyl group of Tyr^{235} keeps this interaction at an optimal level. A loss of either the hydroxyl group or the aromatic ring of Tyr^{235} perturbs this interaction and makes the enzyme less active but not fully inactive. We are currently investigating the residues that are away from the catalytic site yet interact with Tyr^{235}, making them potential targets for the development of therapeutics for type 2 diabetes.

Coenzyme F_{420} metabolism of the mycobacteria and MDR/XDR strains of Mycobacterium tuberculosis

Dr. Endang Purwantini at VBI is the lead for this project. Coenzyme F_{420}, a deazaflavin derivative, is structurally similar to flavins but functionally acts as a hydride transfer-restricted coenzyme similar to the nicotinamides. F_{420} is present in all known methanogenic archaea, but 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} and express an F_{420}-specific glucose-6-phosphate dehydrogenase (Fgd) (Purwantini and Daniels, 1996; Purwantini, Gillis and Daniels, 1997). The use of Fgd-generated reduced F_{420} ($F_{420}H_2$) in the mycobacteria is unknown. We are investigating this aspect to identify new cellular targets for the development of new therapeutics for TB. The emergence of MDR- and XDR-TB as a major threat to world population call for rapid development of new TB drugs. The last effective drug specifically developed for treating TB, was introduced in 1966 (Maggi et al, 1966; Willcox, 2000). We have found that the mycobacteria utilize $F_{420}H_2$ in its defense against the bactericidal effect of macrophages and likely play a role in the biosynthesis of mycolic acids that are critical components of mycobacterial cell envelope. In collaboration with the Rotinsulu Pulmonary Hospital and the Institut Teknologi Bandung, Bandung, Indonesia, we are investigating the genetic and biochemical basis for the development of more virulent MDR- and XDR-*M. tuberculosis*.

Conferences and workshops

Mukhopadhyay B (2007) Sulfite reduction in methanogenic archaea - new enzymes and new hypotheses on two ancient metabolisms. *Gordon Conference on Archaea: Ecology Metabolism & Molecular Biology*. Andover, NH, August 19-24, 2007.

Mukhopadhyay B (2007) Sulfite reduction in methanogens, a clue to an ancient cell performing both methanogenesis and sulphate driven anaerobic methane oxidation, *Thermophiles 2007*, Bergen, Norway, September 24-27, 2007.

Peidl K, Purwantini E, Johnson EF, Mukhopadhyay B (2008) Isolation and characterization of methanotrophic bacteria from plant leaves, Poster presentation to *The Blue Ridge Regional Science Fair*, New River Community College, Dublin, VA, March 1, 2008.

Purwantini E, Stieber JP, Mukhopadhyay B (2007) Involvement of Rv0406c homolog of *Mycobacterium smegmatis* in Mycolic Acid Synthesis, Poster presentation to *108th Annual Meeting of the American Society for Microbiology*, Toronto, Canada, May 21-25, 2007.

Wang B, Purwantini E, Johnson EF, Mukhopadhyay B (2008) Nutritional characteristics and morphology of duck pond soil bacteria, Poster presentation to *The Blue Ridge Regional Science Fair*, New River Community College, Dublin, VA, March 1, 2008.

Peer reviewed publications in reporting period

Anderson I, Rodriguez J, Susanti D, Porat I, Reich C, Ulrich LE, Elkins JG, Mavromatis K, Lykidis A, Kim E, Thompson LS, Nolan M, Land M, Copeland A, Lapidus A, Lucas S, Detter C, Zhulin IB, Olsen GJ, Whitman W, Mukhopadhyay B, Bristow J, Kyrpides N (2008) Genome sequence of *Thermofilum pendens* reveals an exceptional loss of biosynthetic pathways without genome reduction. *Journal of Bacteriology* **190**: 2957-2965.

Case CL, Mukhopadhyay B (2007) Kinetic characterization of recombinant human cytosolic phosphoenolpyruvate carboxykinase with and without a His$_{10}$-tag. *Biochimica Biophysica Acta* **1770**: 1576-1584.

Johnson EF, Mukhopadhyay B (2008) Coenzyme F$_{420}$-dependent sulfite reductase enabled sulfite detoxification and use of sulfite as a sole sulfur source by *Methanococcus maripaludis. Applied Environmental Microbiology* **74**(11): 3591-3595.

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. Journal of Bacteriology* **89**: 7392-7398.

Other cited references

Beale EG, Harvey BJ, Forest C (2007) PCK1 and PCK2 as candidate diabetes and obesity genes. *Cell Biochem Biophys* **48**: 89-95.

Boetius A, Ravenschlag K, Schubert CJ, Rickert D, Widdel F, Gieseke A, Amann R, Jorgensen BB, Witte U, Pfannkuche O (2000) A marine microbial consortium apparently mediating anaerobic oxidation of methane. *Nature* **407**: 623-626.

Johnson EF, Mukhopadhyay B (2005) A new type of sulfite reductase, a novel coenzyme F$_{420}$-dependent enzyme, from the methanarchaeon *Methanocaldococcus jannaschii. Journal of Biological Chemistry* **280**: 38776-38786.

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 Verlag, New York, N.Y.

LeGall J, Fauque G (1988) Dissimilatory reduction of sulfur compounds. In *Biology of anaerobic microorganisms*, Zendher AJB (ed), Wile, N.Y., pp. 587-693.

Maggi N, Pasqualucci CR, Ballotta R, Sensi P (1966) Rifampicin: a new orally active rifamycin. *Chemotherapy* **11**: 285-292.

Mukhopadhyay B, Concar EM, Wolfe RS (2001) A GTP-dependent vertebrate-type phosphoenolpyruvate carboxykinase from *Mycobacterium smegmatis. Journal of Biological Chemistry* **276**: 16137-16145.

Mukhopadhyay B, Purwantini E (2000) Pyruvate carboxylase from *Mycobacterium smegmatis*: stabilization, rapid purification, molecular and biochemical characterization and regulation of the cellular level. *Biochimica Biophysica Acta* **1475**: 191-206.

Purwantini E, Daniels L (1996) Purification of a novel coenzyme F_{420}-dependent glucose-6-phosphate dehydrogenase from *Mycobacterium smegmatis*. *Journal of Bacteriology* **178**: 2861-286.

Purwantini E, Gillis TP, Daniels L (1997) Presence of F420-dependent glucose-6-phosphate dehydrogenase in *Mycobacterium* and *Nocardia* species, but absence from *Streptomyces* and *Corynebacterium* species and methanogenic Archaea. *FEMS Microbiology Letters* **146**: 129-134.

Shima S, Thauer RK (2005) Methyl-coenzyme M reductase and the anaerobic oxidation of methane in methanotrophic Archaea. *Current Opinion in Microbiology* **8**: 643-648.

Willcox PA (2000) Drug-resistant tuberculosis. *Current Opinion in Pulmonary Medicine* **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, William Baumann, Yizhi Cai, Stephen Moody, Jean Peccoud, Rebecca Shelton.

Computer-Assisted Design of Synthetic Genetic Systems

Abstract. We released GenoCAD™, a web-based application that can be regarded as the first Computer Assisted Design system for synthetic DNA sequence. It provides a sequence builder function that guides users through the process of designing a new genetic construct from a database of standard genetic parts. It also includes a sequence verification tool that can be used to ensure that DNA sequences are consistent with a set of predefined design guidelines. In order to build a database of genetic parts for GenoCAD, we conducted an extensive analysis of the MIT Registry of Standard Biological Parts. This work unraveled widespread limitations of the data model upon which this important community resource has been developed. These results suggested a more targeted approach that we are pursuing for the development of genetic part libraries that better reflect the needs of users. Our group has also developed a live cell imaging capability allowing us to observe the dynamics of gene networks in individual cells. We observed that gene expression processes are affected by significant fluctuations that could be taken advantage of to build functional models of genetic parts.

Keywords: computer-assisted design; live cell imaging; synthetic biology; genetic parts; regulatory networks; microscopy.

Scientific Progress

Release of the GenoCAD web site

The theoretical work conducted in our group has allowed us to build a solid foundation for the Computer Assisted Design of genetic systems. In order to allow potential users to benefit from this work we have implemented our algorithms in a web site called GenoCAD (www.genocad.org). The homepage of the web site points toward two distinct sections. The Design link points to the Sequence Builder while the Validate link points to a design verification engine.

The Sequence Builder feature has two panes. The History pane keeps track of the different design steps and allows users to step back in the process. The main pane consists of a graphical representation of the sequence. Each syntactic category is identified by an icon. Below each icon is a list of choices that the user can click on to move to the next step in the design process. Choices in capital letters correspond to the addition of functional blocks that modify the sequence structure while the links in small letters correspond to actual DNA sequences. When the sequence of the design is completely specified, the user has the option of exporting it as a text file that can be used as input to a manufacturing process.

GenoCAD users can create an account on the web site to customize their workspace. This feature allows them to save, retrieve, and copy their designs. This feature greatly enhances the user experience since in many projects designs are derived from one another. In addition, users can add new parts if they need to work with DNA sequences that are not in the central database. They can also make custom libraries of parts in which they can combine selected parts from the parts database along with their personally generated sequence parts.

Sequencing of the MIT Registry of Standard Genetic Parts

The Massachusetts Institute of Technology (MIT) Registry of Standard Genetic Parts, which was founded four years ago and which had 4856 entries in July 2007, is the largest publicly available library of genetic parts. In addition to the database, the Registry includes a repository of 995 bacterial clones containing the DNA fragments corresponding to entries in the Registry database. We have analyzed this resource to identify usage patterns that could help design the next generation of infrastructures hosting libraries of genetic parts.

The abstraction hierarchy utilized by the Registry includes a level called "parts", which contains basic building blocks such as ribosome binding sites, regulatory sequences, coding sequences, RNA, DNA linkers, and transcription terminators. Constructs combining multiple parts correspond to higher levels of abstraction called "devices" and "systems".

We derived a network of relationships between entries in the Registry from their published sequence. First, inclusion relationships between entries were identified by pairwise comparison of the sequences in the database. Entry A is connected to entry B if the sequence of A includes the sequence of B. In a second step, this directed graph was pruned to eliminate redundant relationships. For instance, if A includes B and B includes C, then a relation between A and C can be derived from the previous relations. This operation allowed us to draw a network of 1435 relationships among the 995 entries considered in this analysis. Since parts are atomic building blocks corresponding to the bottom layer of the abstraction hierarchy, it was expected that there would be few if any connections among entries in this group. However, 53 relationships between parts have been identified. Even more surprising, 13 relationships indicated that entries in the design group were present in the sequences of parts. These observations appear to be inconsistent with the Registry abstraction hierarchy. These results are helping us to build the next generation of a database for genetic parts that will be integrated into GenoCAD.

Live cell imaging

Our group is also involved in developing a live cell imaging capability as a way to characterize the contribution of different genetic parts to the phenotype of genetic systems. For instance, in this study a yeast strain expressing yellow fluorescent protein (YFP) under the control of the *GAL1* promoter was grown overnight in media lacking galactose. The cells were washed and resuspended in a 2% galactose media, and placed on a microscope slide where they were observed for 12 hours. In the Figure, at 600 minutes, three distinct levels of fluorescence are visible. The brightest cells are all daughters of cell 1, while the rest are daughters of cell 2. Some of the offspring of cell 2 show dim fluorescence, while some exhibit no fluorescence at all. In this small sample it would be easy to manually identify cell lineage and therefore monitor the inheritance of each phenotype. Manual tracking quickly becomes impractical if not impossible when analyzing a larger cell sample.

This experiment illustrates the stochastic fluctuations affecting gene expression mechanisms. Measuring these fluctuations requires the observation of many individual cells over long periods of time. Unfortunately no commercial instrument is available for this purpose. Flow-cytometry makes it possible to take gene expression data in large samples of cells but does not allow the observation of individual cells over multiple time points.

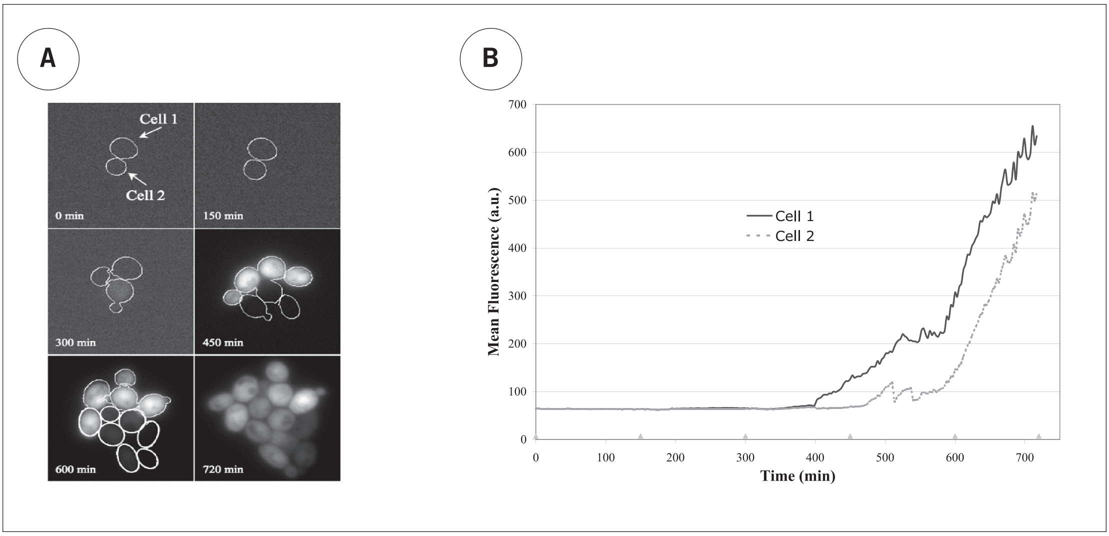


Figure. Live cell imaging. (A) Composite bright-field and *GAL1*pr-yfp images from a 12-hour time-lapse of yeast. (B) Time-course of the mean cell fluorescence for the original cells in the images shown in A. The large peaks in the fluorescence in cell 2 are caused by limitations of the algorithms used to automatically identify cells in the series of images.

Fluorescence microscopy makes it possible to observe a small number of individual cells in time-lapse experiments but does not scale up well. Our group has designed a new generation of instruments that should enable the observations of many individual cells over extended periods of time.

Educational activities

In 2007, we hosted two teachers enrolled in a Research Experience for Teachers program funded as a supplement to the National Science Foundation/National Institutes of Health-funded Bioengineering/Bioinformatics Summer Institute (EEC-0609225). Brennan Kraje from Staunton River High School, VA, developed a complex epidemiological model while Martha Eborall from Bluefield State College, WV, received an introduction to molecular biology.

We also organized Virginia Tech's participation in iGEM, the International Genetically Engineered Machine competition organized by MIT. After promoting the opportunity to students enrolled in departments spanning three colleges, we selected a team of six undergraduates from four different departments (Electrical Engineering, Computer Science, Biochemistry, and Biology). The team designed synthetic DNA molecules to monitor the infection of a population of bacteria by a virus. This work was presented in November 2007 on the MIT campus during the iGEM Jamboree and the team was awarded a gold medal for this work.

Conferences and workshops

Peccoud J (2007) Microsoft eScience Conference, Chapel Hill, NC, October 21-23, 2007.

Peccoud J (2007) MIT Registry of Biological Parts workshop, Cambridge, MA, November 4-6, 2007.

Peccoud J (2007) Chemical and Biological Defense Conference, Baltimore, MD, November 13-15, 2007.

Peccoud J (2008) CHI Peptalk Protein Expression Conference, San Diego, CA, January 12-16, 2008.

Peccoud J (2008) University of Nice, Department of Applied Mathematics, Nice, France, February 26-28, 2008.

Peccoud J (2008) Institute of Biological Engineering Conference, Chapel Hill, NC, March 6-9, 2008.

Intellectual property

Peccoud J, Ball D, Moody S (10/10/07) Adaptive 5D optical tracker for closed loop acquisition of gene network dynamics, US provisional.

Peccoud J, Cai Y (03/30/08) Software for design and verification of synthetic genetic constructs, US utility application 12/058,712.

Peer-reviewed publications in reporting period

Cai Y, Hartnett B, Gustafsson C, Peccoud J (2007) A syntactic model to design and verify synthetic genetic constructs derived from standard biological parts. *Bioinformatics* **23**: 2760-2767.

Peccoud J, Courtney T, Sanders WH (2007) Möbius: an integrated discrete-event modeling environment. *Bioinformatics* **23**: 3412-3414.

Peccoud J, Coulombel L (2007) A competition of synthetic biology or how to create the "water of *E. coli*" and nano-barbies. *M/S-Médecine Sciences (Paris)* **23**: 551-552.

Andy Pereira

pereiraa@vbi.vt.edu

Professor, Virginia Bioinformatics Institute

Adjunct Professor, Dept. of Biological Sciences, Virginia Tech

Adjunct Professor, Crop and Soil Environmental Sciences, Virginia Tech

Group contributors: Madana Ambavaram, Utlwang Batlang, Amal Harb, Michelle Harlow, Arjun Krishnan, Bryan Moretz, Marcella Welch, Peter Wittich.

Plant-Environment Interaction Systems Biology

Abstract. Structural and functional genomics technologies are generating a large amount of molecular data from organisms that are most useful when understood in an integrated manner. We study plant-environment interaction as a system that integrates plant perception, response and resistance to environmental stress. Drought is one stress factor that has emerged as a serious problem due to global climate change and the needs of an increasing world population, especially since crop plants use two-thirds of available freshwater. However, the response and resistance of crop plants to drought or water-deficit is complex, making it difficult to identify the causal traits using classical breeding techniques. To dissect conserved pathways involved in drought stress we use a comparative analysis of the dicot and monocot reference plants *Arabidopsis* and rice, based on the hypothesis that perception and resistance to drought has evolved in land plants through some common conserved mechanisms. We have integrated *Arabidopsis* gene interaction networks with rice ortholog information to analyze comparative gene functions between the two plants as a model to extend to other plants. Transcriptome analysis of *Arabidopsis* and rice responses to soil water deficit simulating field conditions reveals common regulated pathways and genes that are being dissected by reverse genetics approaches. In a forward genetics approach we use a gain-of-function strategy to identify genes that impart improved drought resistance. The stress resistance mechanisms of two *Arabidopsis* transcription factors, *HRD* and *SHN*, are being studied in *Arabidopsis* and rice to determine the network of gene actions, in an effort to provide insights for future applications in crop improvement.

Keywords: gene network; *Arabidopsis*; rice; maize; switchgrass; drought stress; disease resistance; microarray; phosphoproteome; water-use efficiency.

Scientific Progress

Plants respond to external stress factors in a variety of ways, ranging from quick, short-term signals and transcriptional responses to longer-term, metabolic and developmental responses that help avoid and resist the imposed stress. Many of the induced plant responses take energy away from the plant. This can reduce crop yield, which demonstrates the importance of studying the balance between plant defense and crop productivity. The analysis of drought stress responses at the whole plant level and at different developmental stages provides a multi-factorial system to catalogue and dissect these perturbations of normal plant processes. The cross-talk between different abiotic/biotic stress perception and response pathways provide the means to develop a global view of plant-environment interactions.

Drought stress gene interaction network

Drought is the decreased availability of soil water, which can cause cellular dehydration or low water potential in a plant. To dissect conserved genes and pathways involved in drought stress we use the sequenced plant genomes of *Arabidopsis* and rice for functional genomics studies. Controlled soil water deficit drought stress was imposed on *Arabidopsis* and rice at the vegetative leaf stage and the reproductive inflorescence stage, and the isolated RNA was used for Affymetrix GeneChip® analysis. As the reproductive stage of cereals is drought sensitive and has the most significant effect on yield losses, we develop here a model system to compare rice and *Arabidopsis*. A significant set of drought responsive genes are differentially expressed in these two plant-drought systems.

In order to make detailed functional comparisons we constructed an *Arabidopsis* gene interaction network integrated with rice orthologs. The global coexpression and interaction (protein-protein-DNA) network comprises around 20,000 *Arabidopsis* and 15,000 rice genes onto which we superimposed our drought differential expression data to derive a significant network of 1,400 genes, which is shown in the Figure. From this network, genes of different classes such as regulatory (transcription factors, kinases, phosphatases), where the *Arabidopsis* and rice orthologs exhibit similar coexpression patterns can be identified. As perception and resistance to drought has evolved in land plants through some common conserved mechanisms, the comparative network-based analysis provides candidates for conserved gene functions across different plants. Candidate regulatory genes are being analyzed using overexpression and knockout genotypes for alterations in drought response phenotypes.

Around the monocot model rice

Stress transcriptome studies have been extended to include maize and switchgrass. In maize we have used leaf meristem and post-fertilization ovaries that have undergone drought stress to analyze the specific drought responses at the vegetative and critical drought sensitive reproductive stages, which are then compared to rice. For transcriptome analysis we have used maize Affymetrix GeneChips and the Roche GS-FLX™ transcriptome sequencing to identify and quantitate expression of drought responsive genes. In switchgrass, a biomass crop with very little available gene/transcript information, we have also used drought stressed leaf tissue to compare with control grown plants for GS-FLX transcriptome analysis. Using comparative analysis to examine these two monocot plants in addition to rice further extends our list

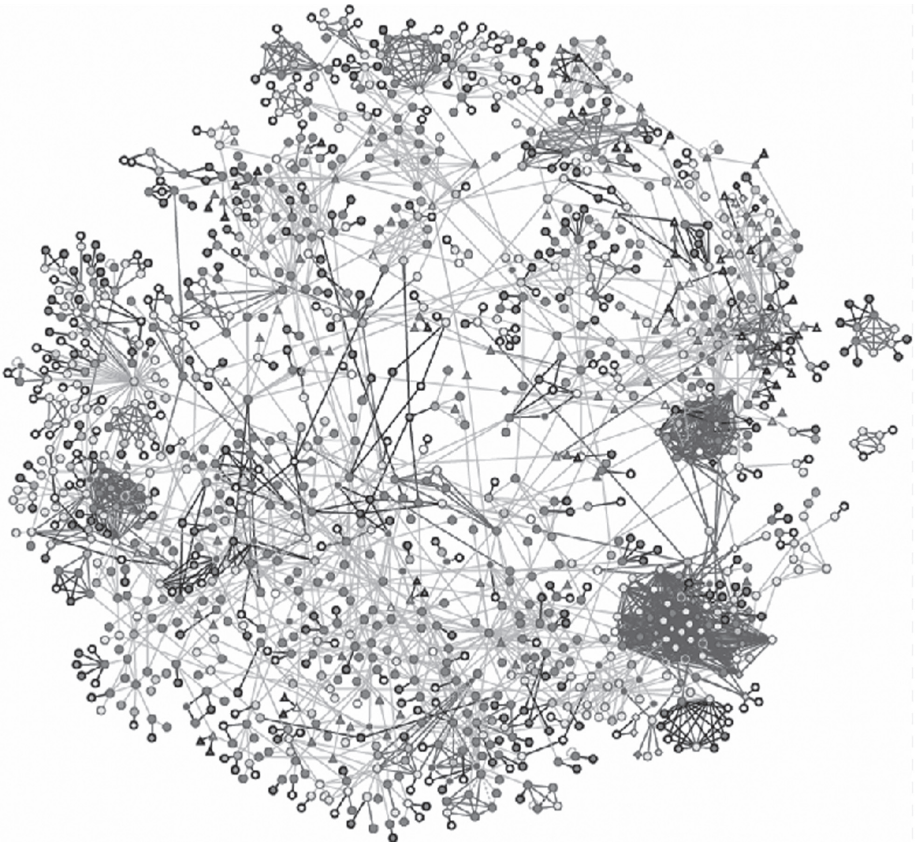


Figure. *Arabidopsis* drought gene interaction network with rice orthologs. The nodes correspond to genes for transcription factors (triangles), kinases/phosphatases (squares/diamonds) and other classes (circles), with the *Arabidopsis* genes outlined in light and the rice genes in dark borders. Genes that are up-regulated are shaded dark and the down-regulated genes shaded light. The edges signifying interactions (co-expression, protein-protein-DNA) are shown in different grey shades and the *Arabidopsis*-rice orthologs are shown as black lines.

of candidate genes to be validated by other means.

The early signaling events resulting from drought stress perception have been studied at the phosphoproteome level. Drought-treated and control tissue from maize vegetative and reproductive stages, as well as from rice vegetative stages, were used to isolate the phosphoprotein fractions that were run on two-dimensional gels, and the proteins were visualized with a phosphoprotein-specific stain. This analysis revealed multiple phosphoproteins differing in intensity, as well as a difference in the presence/absence of phosphoproteins when comparing drought conditions to those featuring well water. The identification of these phosphoproteins and their genes in rice and maize would provide us targets that can be tested for relevance to drought perception and response pathways.

Gain-of-function mutagenesis strategies

An *En/Spm* transposon-based activation tag system is being used to isolate gain-of-function mutant genotypes in *Arabidopsis*. In addition to an earlier collection in ecotype Ws, we are generating an extended mutant population in ecotype Col. Screening of these activation tag lines have revealed novel mutants involved in developmental processes, flowering time, abiotic stress and plant productivity. Mutant lines are being tested for drought resistance phenotypes using a soil water deficit assay system to measure productivity.

In tomato, an Ac-Ds transposon-based activation tag system is being developed to generate a population for the plant community as part of a plant genome project funded by the United States Department of Agriculture. In addition, we are collaborating with other groups to generate rice activation tag lines to contribute to the International Rice Functional Genomics Consortium.

Transcription factors with conserved stress resistance phenotypes

Using the gain-of-function activation tagging strategy in *Arabidopsis*, we have previously identified the transcriptional factor (TF) genes *SHN* and *HRD* that confer drought resistance (DR) when overexpressed in *Arabidopsis* and rice. Most prominently they confer a phenotype of improved water use efficiency (WUE), which is the ratio of biomass produced to water used by the plant. This equates to plants that utilize water more efficiently and can grow in water limiting environments, including drought. These two genes employ different mechanisms to arrive at this WUE trait – *HRD* by an enhanced biomass production and *SHN* by reduced transpiration rate.

Microarray analysis of three AP2/ERF TF (*SHN, HRD, DREB1A*) overexpressing *Arabidopsis* plants conferring DR reveal a network of downstream regulated genes or 'regulon'. The *DREB1A* gene was described previously by others and is used for comparison here. Comparison of the three DR-gene regulons reveal a number of shared genes as well as some specific to each DR gene. The drought transcriptome, which are the genes differentially regulated by progressive drought stress, shows significant overlap with subsets of the DR regulons. We thereby propose that the shared subset of the drought transcriptome and the DR regulon are responsible for the DR phenotype.

The *SHN* and *HRD* overexpression genotypes have been tested in various plants and in addition to drought resistance and WUE, they display significant salinity tolerance as well as resistance to some plant pathogens. The further analysis of the downstream regulated pathways by these transcription factors will reveal the subset of genes responsible for providing different facets of stress resistance mechanisms in plants.

Conferences and workshops

Ambavaram MMR, Utlwang B, Krishnan A, Wittich P, Lee K, Pereira A (2008) Comparative drought associated phosphoproteome analysis of maize and rice, *Plant & Animal Genomes XVI Conference*, San Diego, CA, January 12-16, 2008.

Karaba A, Dixit S, Trijatmiko KR, Nataraja KN, Udayakumar M, Krishnan A, Reddy MM, Pereira A. (2008) Improvement of Water Use Efficiency and Drought Resistance in Rice, Challenge Program Workshop, *Plant & Animal Genomes XVI Conference*, San Diego, CA, January 12-16, 2008.

Krishnan A, Dixit S, Karaba A, Trijatmiko KR, Reddy MM, Batlang U, Wittich P, Pereira A (2008) Systems biology of conserved drought stress resistance, *Plant Animal Genomes XVI Conference*, San Diego, CA, January 12-16, 2008.

Pereira A, Karaba A, Karaba N, Udayakumar M, Trijatmiko KR, Xiong L, Leung H (2007) From genomics to traits for sustainable rice crop production, *International Symposium on "Breeding Rice – Progress and Strategies for the future"*, New Delhi, India, July 17, 2007.

Pereira A, Reddy MM, Wittich P, Veilleux R (2008) Development of a versatile mutant resource in tomato for functional genomics analysis, *USDA-CSREES Plant Genome project workshop*, San Diego, CA, January 11, 2008.

Peer-reviewed publications in reporting period

Karaba A, S Dixit, R Greco, A Aharoni, K Trijatmiko, N Marsch-Martinez, A Krishnan, K Nataraja, M Udayakumar, A Pereira (2007) Improvement of water use efficiency in rice by expression of *HARDY*, an *Arabidopsis* drought and salt tolerance gene. *Proceedings of the National Academy of Sciences USA* **104**: 15270-15275.

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

Dharmendar Rathore

rathore@vbi.vt.edu

Assistant Professor, Virginia Bioinformatics Institute

Group contributors: Rana Nagarkatti, Dewal Jani.

Deciphering Hemozoin Formation in the Malaria Parasite

Abstract. Growth and development of the malaria parasite within its host red blood cell are dependent upon its efficiency to process host hemoglobin. Proteolysis of hemoglobin releases heme, which is toxic to the parasite. To protect itself, the parasite converts free heme into hemozoin, a crystalline material that has played an important role in the discovery of the parasite. We recently identified a novel parasite protein, which plays a critical role in hemozoin formation, and named it Heme Detoxification Protein (HDP). *In vivo*, hemozoin crystal is encapsulated in neutral lipid nanospheres, while *in vitro* these lipids can also produce hemozoin. However, on a mole-for-mole basis, HDP is 2,000-fold more potent in producing hemozoin. We report that a HDP-catalyzed reaction produces a synergistic effect, leading up to a four-fold increase in hemozoin production when neutral lipids found in the food vacuole are introduced into the reaction. HDP plays a critical role in this synergism as this effect is obliterated when the HDP-lipid mix is subjected to a proteinase K treatment. Investigations focused on understanding the involvement of lipids revealed that its fatty acid moieties are responsible for this synergism. Our results not only provide an interesting insight into hemozoin production, but also raise the possibility of such a mechanism being deployed by the parasite to protect itself from the toxic effects of heme.

Keywords: heme; Heme Detoxification Protein (HDP); malaria; hemozoin; neutral lipids; fatty acids.

Scientific Progress

The parasite-specific origin of hemozoin synthesis has made this pathway one of the most attractive targets for antimalarial drugs. Until recently, lipids present in the food vacuole were accepted as major mediators of this activity as a result of several *in vitro* studies that had found them to be capable of converting free heme into hemozoin. However, our recent discovery of Heme Detoxification Protein (HDP), a highly conserved and unique *Plasmodium* protein, has reinvigorated the role of a parasite-specific factor in hemozoin formation. During *Plasmodium* trophozoite, an intraerythrocytic stage of the life cycle during which hemozoin is produced, the food vacuole contains HDP and a wide variety of lipids and fatty acids.

We initially mimicked this environment by including monopalmitoyl glycerol (MPG) in a HDP-mediated hemozoin reaction. MPG has recently been shown to be present in neutral lipid nanospheres that surround the hemozoin crystal. Addition of MPG to a HDP-mediated reaction synergized hemozoin production with the reaction achieving completion, while requiring up to a five-fold less concentration of HDP. The addition of MPG to a HDP-mediated reaction synergized hemozoin production. The reaction reached completion and required up to a five-fold less concentration of HDP. However, this synergism is not due to an enhancement of lipid-mediated activity. Pre-treatment of the reaction with proteinase K obliterated the synergism, producing hemozoin at levels found with MPG alone. We hypothesized that HDP is most likely the primary mediator of the formation of an iron-carboxylate bond between the two free heme molecules and that MPG plays a major role in the stacking of this unit cell that leads to the crystal. In such a scenario, it could be argued that the physical properties of lipids would play a role in determining its compatibility in supporting hemozoin crystal growth, while lipids similar to MPG could also show synergism.

Predictably, we found that the synergism was not limited to MPG, but could also occur when dipalmitoyl glycerol, monostearic glycerol, monooleoyl glycerol or dioleoyl glycerol was added to a HDP-based reaction. These lipids, though capable of synergizing the reaction, showed remarkable differences in their potential to accelerate the reaction. The number of fatty acid chains, as well as chain length and degree of unsaturation serve as some of the deterministic parameters of biological functions in lipids and its associated fatty acids. We found that monoacyl glycerols (MPG, MOG) were more potent than its diacyl counterparts (DPG, DOG) in synergizing the reaction. Not only do the number of fatty acid chains, the length of the chains and the presence of unsaturated bonds also affect the outcome, but the unsaturated lipid MOG was found to be more potent than its saturated form MSG.

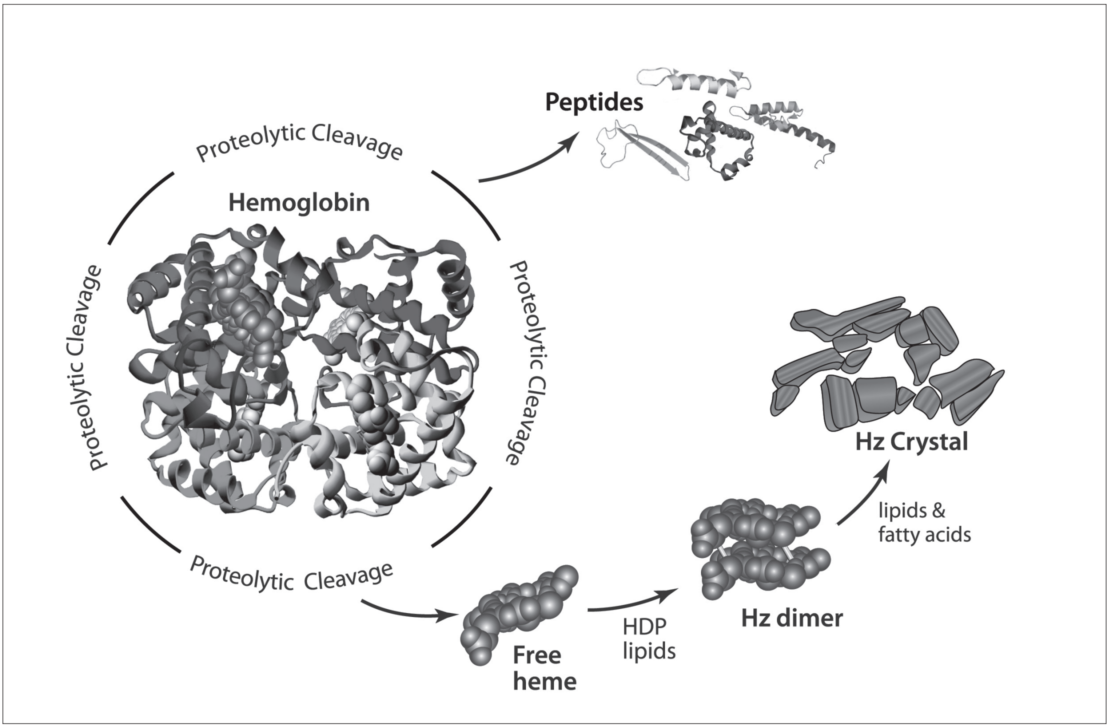


Figure. Parasite specific origin of hemozoin. HDP, Heme Detoxification Protein; Hz, hemozoin.

Interestingly, synergism achieved with neutral lipids can also be replicated using fatty acids in the reaction. We found that saturated fatty acids consisting of a 12-16 carbon chain length were best suited to synergize the reaction, while acids of longer chain lengths were ineffective. An increase in acyl chain lengths causes fatty acids to form self-assembled, insoluble structures and, consequently, higher melting temperatures, which could account for the inability of stearic and arachidic acids to synergize hemozoin production. Though saturated fatty acids of longer chain lengths were ineffective, its unsaturated forms demonstrated synergism. The introduction of unsaturated bonds disrupts molecular packaging and ordering among the fatty acid chains, leading to an increase in its fluidity and lower melting temperatures. Thus, it is likely that the fluidity of fatty acids, maintained by restricting the chain length, the presence of unsaturated bonds, or both, is a strong determinant in the synergistic activity of these molecules. The observation of *in vitro* synergism for a wide variety of neutral lipids and saturated and unsaturated fatty acids, many of which are present in the food vacuole, strongly suggests that *in vivo*, lipids and fatty acids work in concert with HDP, leading to the formation of hemozoin crystal.

Previously, reactions performed with HDP alone showed a maximum of 60-70% conversion, however the introduction of lipids/fatty acids increased it to 90% while simultaneously reducing the amount of protein required by more than half. If extrapolated *in vivo*, it would suggest that the previous reports of the hemozoin production rate could have significantly underestimated the biochemical prowess of the parasite in this undertaking.

Conferences and workshops

Rathore D (2007) Drug Discovery in Academia, The Art of *Team Science* on a Shoestring Budget, *ChemBiotek*, Salt Lake City, India, December 18, 2007.

Rathore D (2008) Deciphering Heme Detoxification in Malaria – Implications for Disease Control, *Biology Research Symposium*, Virginia Tech, Blacksburg, VA, March 28, 2008.

Peer-reviewed publications in reporting period

Rathore D (2007) Targeting parasite-mediated host hemoglobin degradation in malaria. *IDrugs* **10**(2): 877-880.

Jani D, Nagarkatti R, Beatty W, Angel R, Slebodnick C, Andersen J, Kumar S, Rathore D (2008) HDP – A novel heme detoxification protein from the malaria parasite. *PLoS Pathogens* **4**(4): e1000053.V

Intellectual property

Rathore D HDP as a therapeutics agent for the treatment of Toxoplasmosis and Babesiosis, pending US Patent #60/968,384.

David Samuels

dsamuels@vbi.vt.edu

Assistant Professor, Virginia Bioinformatics Institute

Group contributors: Vishal Gandhi, Jonghoon Kang, Sanghoo Lee, Harsha K. Rajasimha, Zhou Song, Katherine Wendelsdorf, Passorn Wonnapinij.

Computational Research on Mitochondrial Disease and Toxicity

Abstract. Research in our group involves the use of computational and mathematical tools applied to research problems in cell biology and medicine. Our current projects are concerned with mitochondrial diseases. These projects arise from research into genetic diseases that are caused by mutations in either nuclear or mitochondrial genes, and from research into the toxicity of nucleoside analogs used as antiviral agents in the treatment of Human Immunodeficiency Virus/ Acquired Immunodeficiency Syndrome (HIV/AIDS). In this scientific progress report, we discuss the results presented in our papers published during the reporting period. These include simulations of mammalian embryogenesis, hematopoietic stem cells, and the enzyme kinetics of small molecule carriers. Several of these papers involve experimental tests of hypotheses that are based on our simulation models. Other papers propose new hypotheses that may be tested in future experiments.

Keywords: antiviral drugs; azidothymidine (AZT); mitochondrial DNA (mtDNA); deletion mechanisms; stem cells; embryogenesis; deoxyribonucleotide metabolism.

Scientific Progress

Modeling and clinical collaborations

The purpose of our modeling work is to better understand clinical and laboratory data. At best, this process results in the prediction of a new hypothesis that can be tested in new experiments. We had put forward the "maintenance of wild-type" hypothesis based on our simulations of mitochondrial DNA (mtDNA) populations within individual cells. From this hypothesis we made an experimentally testable prediction concerning the amount of wild-type mtDNA and total mtDNA in individual cells. In Durham et al. (2007), we tested this hypothesis for three clinically important mtDNA mutations: the 10010T→C mutation; the 3243A→G mutation; and the common deletion. Each of the three mutations is pathogenic. The data on the common deletion and the 10010T→ C point mutation were consistent with our hypothesis. The 3243A→G mutation showed a different behavior. This departure from our hypothesis indicates that the pathogenic mechanism for this particular mutation, which is clinically very important, must be different than the pathogenic mechanism of the other two mutations. The hypothesis is based on an assumption that the mutation causes a loss of function in the mutant mtDNA. The new data for 3243A→G indicate that the presence of this mutation interferes with the function of the wild-type mtDNA. This gives us a fundamental new insight into the pathogenesis of the many disease states caused by this particular mutant, and has implications for potential therapies.

The new experiments in Durham et al. (2007) were reported several years after our modeling work. This reflects the amount of time needed to fund and carry out the difficult clinical sample collection and experimental procedures required to test the hypothesis. In Cree et al. (2008), we designed and carried out a simulation in parallel with the experiments, and on a much faster time scale. Here our collaborators gathered basic data on mtDNA levels at different stages in early embryogenesis in mice. At VBI, our group used these data to design a computational model of the first 15 days of embryo development to show how the dynamical changes in mtDNA content over this period control the rate of neutral drift in mtDNA heteroplasmy. This rate of drift is the main driving factor that determines the large random changes that occur in the levels of pathogenic mtDNA mutations between mother and offspring in mammals.

We have developed a model of hematopoietic stem cells, and used this model to explain the observed variations with respect to patient age in the mtDNA heteroplasmy measured from blood samples, a very important clinical measurement due to the relative ease of blood sample collection from patients (Rajasimha et al, 2008). While it has long been known that these measurements tend to decrease slowly with age, this decrease had always been interpreted as a constant linear decrease, and the measured values varied greatly among the patients. The simulation model was used to show that this decrease should be exponential, and that all the available data from many clinical papers were consistent with each other when interpreted as an exponential decrease. A critical subset of the clinical data was new data collected by our collaborators in order to test the hypothesis based on our simulation results.

Mutation mechanism hypothesis

An important area of our research concerns the mechanisms of mutation formation. In Krishnan et al. (2008), we put forward the hypothesis that double strand breaks are an important mechanism for the formation of deletions in mitochondrial DNA, in contrast to the prevailing view that replication errors drive most formation of deletions. This hypothesis was based on a range of supporting evidence, including the pattern of deletion endpoints observed in patients, the variation of deletion frequency with respect to cell type and an analysis of the distribution of the positions of direct repeats across the human mitochondrial genome.

Deoxyribonucleotide metabolism

We have two National Institutes of Health-funded projects on deoxyribonucleotide metabolism in mitochondria: an R21 project on nucleoside analog toxicity and an R01 project on the genetic diseases of this metabolism. For both of these projects, an important mechanism is the transport of deoxyribonucleotides between the cytoplasm and the mitochondrial inner matrix where these chemicals are used as precursors for mtDNA replication. This transport includes the transport of activated nucleoside analogs such as azidothymidine (AZT) that are used as antiviral agents. These nucleoside analog drugs have serious toxicity problems, which are believed to be due to their interference with mtDNA replication. In 2001 and in later experiments, a particular small molecule transporter - Deoxynucleotide Carrier (DNC) - has been identified as the transporter for deoxyribonucleotides in mitochondria. As part of our project, we modeled the function of this transporter based on the reported transport kinetics. However, we found that in the advanced version of this transport model, in which the measured kinetics for ribonucleotides as substrates were also included, the inhibitions due to the interaction of DNC with the ribonucleotides were so strong under physiological substrate concentrations that this transporter could not possibly be the correct deoxyribonucleotide carrier. At the same time as we were coming to this conclusion, new experimental results from our collaborator in the Department of Biochemistry and Biophysics at Oregon State University, Chris Mathews, also showed that DNC did not function as the deoxyribonucleotide carrier *in vivo*. In Kang and Samuels (2008), we reviewed the data from experiment, clinical studies, bioinformatics comparisons with yeast proteins, and enzyme kinetics that all point to a different *in vivo* function for this protein, namely the transport of thiamine pyrophosphate.

Conferences and workshops

Samuels DC (2007) Modeling the Metabolism of DNA Precursors in Mitochondria, *Atlantic Coast Conference on Mathematics in the Life and Biological Sciences*, VBI, Blacksburg, VA, May 3-5, 2007.

Samuels DC (2007) Virtual Models of Diseases of mtDNA Synthesis, *ENMC International Workshop on Disorders of mitochondrial DNA synthesis: understanding the complex relationship between genotype and phenotype*, The Netherlands, September 21-23, 2007 (Invited talk at a clinical workshop).

Samuels DC (2008) Simulations of Nucleoside Analog Interactions with Polymerase Gamma, *Mitochondria and their Proteomics Symposium*, National Institutes of Health, Bethesda, MD, January 9-11, 2008.

Peer-reviewed publications in the reporting period

Cree LM, Samuels DC, Chuva de S-LS, Rajasimha HK, Wonnapinij P, Mann JR, Dahl H-HM, Chinnery PF (2008) A reduction of mitochondrial DNA molecules during embryogenesis explains the rapid segregation of genotypes. *Nature Genetics* **40**(2): 249-254.

Durham SE, Samuels DC, Cree LM, Chinnery PF (2007) Normal levels of wild-type mitochondrial DNA maintain cytochrome *c* oxidase activity for two pathogenic mitochondrial DNA mutations but not for m.3243A -> G. *American Journal of Human Genetics* **81**(1): 189-195.

Kang J and Samuels DC (2008) The evidence that the DNC (SLC25A19) is not the mitochondrial deoxyribonucleotide carrier. *Mitochondrion* **8**(2): 103-108.

Krishnan KJ, Reeve AK, Samuels DC, Chinnery PF, Blackwood JK, Taylor RW, Wanrooij S, Spelbrink JN, Lightowlers RN, Turnbull DM (2008) What causes mitochondrial DNA deletions in human cells? *Nature Genetics* **40**(3): 275-279.

Rajasimha HK, Chinnery PF, Samuels DC (2008) Selection against pathogenic mtDNA mutations in a stem cell population leads to the loss of the 3243A -> G mutation in blood. *American Journal of Human Genetics* **82**(2): 333-343.

João C. Setubal

Associate Professor, Deputy Director of Faculty Development, Virginia Bioinformatics Institute
Associate Professor, Department of Computer Science, Virginia Tech

Group contributors: Nalvo Almeida, Chris Lasher, Luisa Mesquita, Jian Sun, Tsai-Tien Tseng, Andrew Warren, and the PATRIC team.

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 bioinformaticians working on genome analyses. In addition to work related to specific genomes (which currently cover the genera *Agrobacterium, Brucella, Rickettsia*, *Azotobacter, Coxiella,* and *Pseudomonas*), current topics of interest include automated genome annotation, web-based infrastructure for genome annotation and analysis, exploitation of ortholog groupings for comparative studies, and sequence analysis in cancer experiments.

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

Scientific Progress

Bacterial genome projects

The Setubal group continued its contribution to three separate bacterial genome projects in the reporting period: *Agrobacterium* biovars (in collaboration with the Agrobacterium Consortium), *Azotobacter vinelandii* (with the Azotobacter consortium), and *Pseudomonas syringae* (with Boris Vinatzer, Department of Plant Pathology, Physiology & Weed Science at Virginia Tech). *Agrobacterium* is a genus is the Rhizobiaceae family of the α-proteobacteria that includes the well-known plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58. The goal of the project is to better understand the pathogenicity mechanisms and evolution of *Agrobacterium* species through genome comparisons to *Agrobacterium vitis* S4, a grape pathogen, to *Agrobacterium radiobacter* K84, a biocontrol agent for C58, and to other Rhizobiaceae genomes. *Azotobacter vinelandii* is a γ-proteobacterium free-living species that has nitrogen fixation capabilities and is a well-known model for biochemistry studies. It is the primary model used in the laboratory of Dennis Dean, Director of the Fralin Center for Biotechnology at Virginia Tech, one of our partners in this project. The goal of the project is to obtain a detailed view of *A. vinelandii*'s genome and couple that with the extensive existing knowledge of its biochemical properties. *Pseudomonas syringae* is a plant pathogen that is also in the γ-proteobacteria group and that has many different strains. Each strain has its own set of host specificities, many of them for important agricultural crops, such as tomato and beans. The goal of this project is to study the genomes of several strains, starting with ptoT1, for which we already have draft sequence. Such studies can throw light on the interplay between pathogen evolution and the early stages of human agriculture, an idea advanced by Boris Vinatzer. Our primary roles in all three projects have been to provide the bioinformatics infrastructure for genome annotation and the execution of various genome analyses, either using third party computational tools or tools that our group has developed. Examples of genome analyses are a comprehensive catalog of regulatory genes in *A. vinelandii,* made possible by PFAM tools, and annotation of RNA genes and elements in all genomes, based on RFAM tools. Luisa Mesquita, an undergraduate student in our laboratory, performed both of these analyses.

The first results of the *Agrobacterium* project were published as a chapter in the book *Plant Pathogenic Bacteria: Genomics and Molecular Biology* (Setubal et al, 2008). For the *Pseudomonas syringae* project, Yan et al (2008) reported the results obtained in the Vinatzer laboratory regarding the evolution of this plant pathogen.

Genome analysis tools

Andrew Warren, a Ph.D. student, has continued the development of the Genome Reverse Compiler. The Genome Reverse Compiler is a stand-alone, open-source efficient annotation tool for prokaryotic genomes. It has been used on the *P. syringae* ptoT1 genome, among others. Chris Lasher, a Ph.D. student, is developing OAK, a tool that builds whole genome ortholog alignments for sets of closely related prokaryotic genomes based on back-end ortholog computation (such as that provided by OrthoMCL) using a graph approach. We expect this tool to have numerous applications in whole-genome analysis, ranging from detection of false positive gene predictions and missed gene calls to improvement in the ortholog groupings themselves and improvement in protein function assignment. The foundation of OAK is accurate identification of conserved gene blocks, also known as micro-syntenic regions. We have applied a prototype version of OAK to the Rhizobiaceae genomes being studied in the *Agrobacterium* project. An example of the results is shown (see Figure). Nalvo Almeida, a visiting scholar from the Federal University of Mato Grosso do Sul, in Brazil, has carried out several genome analyses in the group. One of the tools he developed is a complement to OAK and allows comparisons of genomes based on their ortholog families, as computed by orthoMCL. The result of several set operations on the ortholog families may be obtained for a given set of genomes using Nalvo Almeida's tool. One example is as follows: obtain all orthologs present in genomes *A, B,* and *C,* but absent from genomes *D* and *E,* and possibly including genes from genome *F.* This tool has been used for *Pseudomonas* and for *Brucella*, an

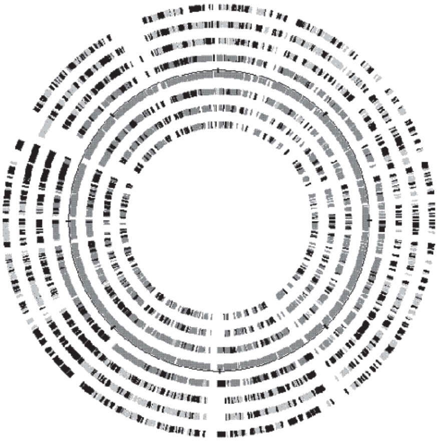


Figure. Ortholog alignment for chromosome 2 of *Agrobacterium vitis* S4 — the *anchor* replicon. The anchor is represented by the circle with a scale. All of its genes are placed in their correct physical position around the circle and painted dark gray. The other circles represent genes from other genomes, in the following order, from external to internal: *Sinorhizobium meliloti*, *Rhizobium leguminosarum, Rhizobium etli, Agrobacterium radiobacter* K84, *Agrobacterium tumefaciens* C58, *Ochrobactrum anthropi*, and *Brucella suis* 1330. In each of these genomes only the orthologs to S4 are placed, and their position is completely dependent on the position of the anchor ortholog. This means that relative position between genes in any of the outer or inner circles may or may not have meaning. At a finer scale (not shown), it is possible to see that many genes in the outer and inner circles do occur consecutively in their respective genomes, thus representing *conserved blocks*. Genes are color-coded according to the replicon they belong to: black, main chromosome; dark grey, second chromosome; light grey, plasmid. The figure was generated by Nalvo Almeida using GenomeViz.

organism studied in the PATRIC project (see next section).

PathoSystems Resource Integration Center (PATRIC)

The PathoSystems Resource Integration Center (PATRIC), which is funded by the National Institutes of Health and the National Institute of Allergy and Infectious Diseases, manages a bioinformatics resource for genomic and other related information on the human pathogens *Brucella, Coxiella,* and *Rickettsia*, as well as on the viral classes Calicivirus, Coronavirus, Hepatitis A and E, and Lyssaviruses. More information on PATRIC can be found at https://patric.vbi.vt.edu. The total number of genomes managed by PATRIC increased for all PathoSystems in the reporting period, which demonstrates the increasing need for effective and efficient data management tools as well as genome comparison tools. For *Rickettsia*, we now manage 14 genomes; for *Brucella,* 10 genomes; and for *Coxiella,* 3 genomes. For the viruses, the numbers are in the several dozen (Lyssa and Hepatitis) or hundreds (Corona and Calici) of genomes. Joseph Gillespie led an effort (Gillespie et al, 2008) in the study of 10 *Rickettsia* genomes centered around the analysis of orthologous groups (OGs). These analyses revealed a set of 752 core rickettsial OGs, plus other OGs characteristic of subgroups of *Rickettsiae*. This led to a much better understanding of the evolutionary history of the various species, which in turn should facilitate the development of drugs, vaccines, and therapeutics for the diseases caused by the different *Rickettsiae*.

Other projects

We collaborate with Sergio Verjovski-Almeida, who works at the Department of Biochemistry at the University of São Paulo, in Brazil. His laboratory focuses on the genomic basis of cancer and on the genomics of the parasite *Schistosoma mansoni.* Thiago Venâncio, a co-advised Ph.D. student, led a study whose goal was to identify by means of bioinformatics *S. mansoni* genes with possible roles in host interactions. That study also uncovered interesting insights into *S. mansoni* evolution (Venâncio et al, 2007). We also collaborate with Kimmen Sjölander, who works at the department of Bioengineering at the University of California-Berkeley. The focus of this collaboration is the recently awarded National Science Foundation grant for the PhyloFacts phylogenomic encyclopedia of microbial protein families. Another collaboration is with Chris Lawrence, from VBI, on genome analyses of the fungi *Alternaria brassicicola* (a plant pathogen) and *Alternaria alternata* (a human pathogen). A collaboration that emerged during the past year is with Emmanuel Dias Neto, from the M.D. Anderson Cancer Center, in Houston, Texas. The current project includes analyses of short nucleotide sequences and the peptides they encode obtained from cancer patients using the technique of phage display, a primary technique used in the laboratory of Wadih Arap and Renata Pasqualini, where E. Dias Neto works. Finally, I contributed a chapter on sequence comparison by similarity to the book *Bioinformatics in Tropical Disease Research: A Practical and Case-Study Approach*, which is published electronically in the NCBI bookshelf (Setubal, 2008).

Conferences and workshops

Setubal J (2007) Bioinformatics Tools That Help Education In Genomics, *Genomics: Opportunities, Tools, and Challenges*, Howard Hughes Medical Institute, Chevy Chase, MD, November 16-18, 2007.

Peer-reviewed publications in reporting period

Gillespie JJ, Williams K, Snyder EE, Nordberg E, Ceraul SM, Dharmanolla C, Rainey D, Soneja J, Shallom JM, Shukla M, Vishnubhat ND, Wattam R, Purkayastha A, Czar M, Crasta O, Setubal JC, Azad AF, Sobral BS (2008) *Rickettsia* phylogenomics: unwinding the intricacies of obligate intracellular life. *PLoS ONE* **3**(4): e2018 doi:10.1371/journal.pone.0002018.

Setubal JC (2008) Similarity Search (theory), chapter A05. In *Bioinformatics in Tropical Disease Research: A Practical and Case-Study Approach*, Gruber A, Durham AM, Huynhtop C, del Portillo H (eds), Bethesda, M.D., National Library of Medicine, National Center for Biotechnology Information.

Setubal JC, Wood D, Burr T, Farrand S, Goldman B, Goodner B, Otten L, Slater S (2008) The Genomics of *Agrobacterium*: Insights into Pathogenicity, Biocontrol, and Evolution. In *Plant Pathogenic Bacteria: Genomics and Molecular Biology*, Jackson R (ed.), Horizon Press.

Venancio TM, DeMarco R, Almeida GT, Oliveira KC, Setubal JC, Verjovski-Almeida S (2007) Analysis of *Schistosoma mansoni* genes shared with Deuterostomia and with possible roles in host interactions. *BMC Genomics* **8**: 407.

Yan S, Liu H, Mohr TJ, Jenrette J, Chiodini R, Zaccardelli M, Setubal JC, Vinatzer B (2008) The role of recombination in the evolution of the model plant pathogen *Pseudomonas syringae* pv. *tomato* DC3000, a very atypical tomato strain. *Applied Environmental Microbiology.* March 31, Epub ahead of print. PMID: 18378665.

Vladimir Shulaev

vshulaev@vbi.vt.edu

Associate Professor, Virginia Bioinformatics Institute

Associate Professor, Dept. of Horticulture, Virginia Tech

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

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

A High-Throughput Metabolomics Platform

Abstract. A high-throughput metabolomics platform is needed as a tool to understand systems biology, discover metabolic biomarkers, and elucidate gene function. The Biochemical Profiling Group at the Virginia Bioinformatics Institute is building a high-throughput metabolomics platform that combines untargeted metabolite profiling, metabolic fingerprinting, and targeted analysis. Analytical techniques like mass spectrometry provide sample analysis for a wide range of metabolites at high sensitivity. This platform has been successfully used at the Virginia Bioinformatics Institute to elucidate early metabolic responses to abiotic stress in plants, to identify unique metabolic signatures associated with the progression of malignancy in human breast epithelium cells, to study the infection of red blood cells by the malaria parasite *Plasmodium falciparum* and investigate the impact of drug treatment on infection, and in a collaborative systems biology project to study the oxidative stress response in the yeast *Saccharomyces cerevisiae*.

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

Scientific Progress

Metabolomics, a powerful "omics" approach, complements other genomics approaches like transcriptomics and metabolomics that aim at a holistic understanding of cellular systems. The Biochemical Profiling Group at the Virginia Bioinformatics Institute (VBI) applies a metabolomics platform, in combination with other genomics platforms, to study the following: infection of human red blood cells by the malarial parasite and its response to antimalarials drugs; the molecular signatures of malignant transformation in human breast epithelial cells; identification of gene functions in the model plant *Arabidopsis thaliana*; discovery of novel human health-promoting compounds in fruits using the woodland strawberry as a model; stress response in plants, micro-organisms and animals; and yeast systems biology research.

Further progress in metabolomics depends upon the development of novel analytical techniques for the identification and quantitation of more metabolites. Our laboratory is constantly working on developing this new technology for metabolomics. One of the approaches used to study biomarkers is called metabolic fingerprinting, which is often performed by direct infusion mass spectrometry. Although this method allows for simpler processing and higher sample throughput, direct infusion mass spectrometry has several limitations. One of these is co-suppression, where the signal of many analytes can be suppressed at the mass spectrometer interface by analytes with higher ionization efficiencies. Recently, we have developed a novel approach to metabolic fingerprinting that can minimize the co-suppression effect. We first carry out a chromatographic separation prior to the mass spectrometry, and then compress a full chromatographic run into a single summary mass spectrum that can be treated as a fingerprint. The summary mass spectrum obtained following direct infusion lacks many ions corresponding to important plant metabolites, including flavonoids and anthocyanins, when compared to the mass spectrum obtained following chromatographic separation (Figure 1). We have a data cube consisting of thousands of mass spectra at different elution times after data acquisition has been performed. This is transformed into a single cumulative mass spectrum that is similar to a direct-infusion mass spectrum, but lacks co-suppression interference. We then use this cumulative spectrum for multivariate analysis to find masses that discriminate between sample groups. We can later identify specific molecules using our custom libraries of spectra without the need for additional experiments since all the original data are kept for further analysis.

Rosaceous fruits are an important component of human diet and a major source of health-promoting metabolites. We apply our metabolomics platform to identify novel biologically active photochemicals in rosaceous fruit. Rosaceae, comprising over 100 genera and 3,000 species, is the third most economically important plant family in temperate regions. The Rosaceae genomics initiative, like those of other crop groups, exploits and extends the fundamental knowledge generated in the *Arabidopsis thaliana* model system, which has dramatically advanced our understanding of the physiology, biochemistry, genetics, and evolution of plants. The genomic information gathered directly from Arabidopsis, the best-established model for plant research, translates well to Brassicaceae. However, Arabidopsis has only partial functional overlaps with many crop species in more phylogenetically remote plant families. For families with valuable crop species, genomics has progressed by selecting a representative model species for the family and focusing research efforts on developing genomic tools for that model. Findings are then translated directly to other species of agricultural importance. Patterns of diversification within the Rosaceae, combined with the need for rapid translation of genomics research into agronomic practice, suggest that multiple rosaceous species must be designated as reference models for acquisition of genomic information. The currently best-developed model species for Rosaceae include apple (*Malus ×domestica* Borkh.), peach (*Prunus persica* L. Batsch), and diploid strawberry (*Fragaria vesca* L.) (Table 1).

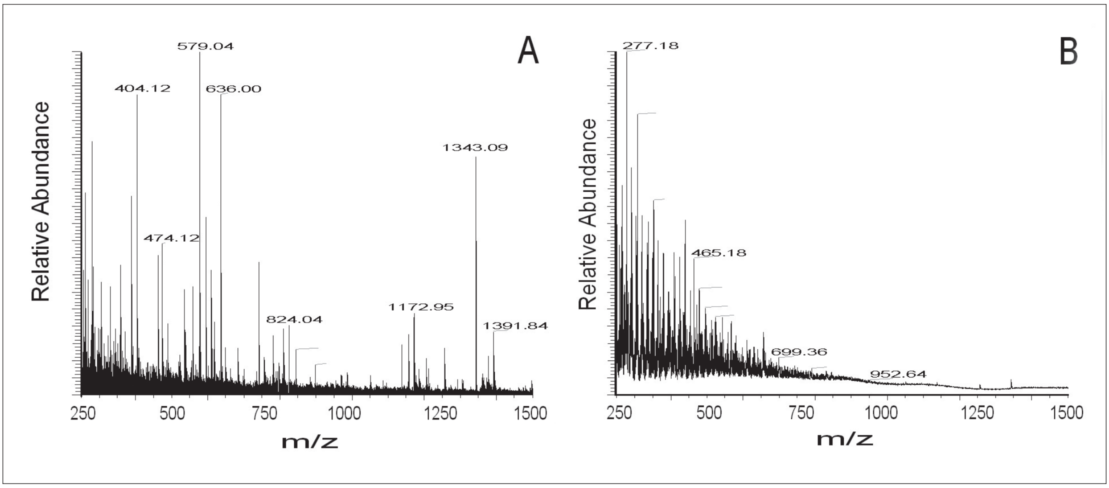


Figure 1. Summary mass spectrum of Arabidopsis leaf extract after either chromatographic separation (A) or direct infusion (B). Ions were detected for positive ionization full-scan mass spectrometry. chromatography was performed on a 0.1x450 mm monolithic C18 column. The summary mass spectrum, which derives from adding up all mass scans over the chromatographic run, shows distribution of *m/z* within the acquisition mass range of 100-1500 a.m.u. exceeding *S/N* > 6.

In collaboration with Dr. Richard Veilleux's group at the Department of Horticulture (Virginia Tech) and Dr. Allan Dickerman's group at VBI, we have been developing a functional genomics platform for Rosaceous crops using the woodland strawberry, *Fragaria vesca*, as a model. We have developed a high-throughput reverse/forward genetics platform in *F. vesca* utilizing an efficient transformation protocol to produce thousands of T-DNA tagged insertional mutant lines. We have already generated a collection of insertional mutant lines, and our initial screens identified a series of mutants with morphological phenotypes (Figure 2). We are now isolating genes responsible for these phenotypes using thermal asymmetric interlaced (TAIL)-PCR technique.

Genomics research in Rosaceae has grown exponentially in the last several years, but future progress is hampered by the lack of the full genome sequence of the representative species

Table 1. Model plant species for Rosaceae genetics and genomic research (modified from Shulaev et al, 2008)

	Malus ×domestica	*Prunus persica*	*Fragaria vesca*
Genome size (Mbp/C)	750	280	206
Chromosome number (2*n*)	2*n* = 2*x* = 34 (some triploid cultivars)	2*n* = 2*x* = 16	2*n* = 2*x* = 14
No. of species in the genus	~35	~53	~20
Generation time (seed to seed)	4-8 years	3-5 years	10-16 weeks
Life cycle	Perennial	Perennial	Perennial
Seed production / plant	~700 (5-10 per fruit)	~300 (1 per fruit)	> 2,500
Plants / m^2	0.67	0.67	~100
Fleshy fruit formation	Pome	Drupe	Receptacle
Juvenile period	3-7 years	1-2 years	None
Vegetative propagation	Yes, hard and softwood cuttings	Yes, hard and softwood cuttings	Yes, runners and crown divisions
Self-compatible	No	Yes	Yes, cultivars and model species
Inbreeding depression	Moderate to severe	Moderate	None to moderate
Transformation	Tissue culture	Tissue culture	Tissue culture
Transformation efficiency (%)	80	<1	100
Expressed Sequence Tags	> 300,000	85,340	>45,000
Genome sequence	Mid-2008	End of 2008	~1%
Physical map	Yes	Yes	In progress
Linkage maps	Yes	Yes	Yes

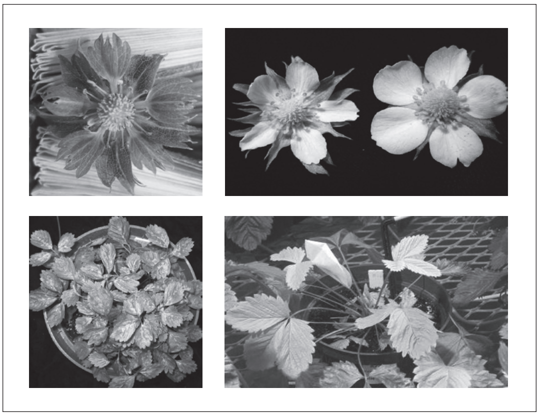


Figure 2. Examples of morphological mutants in *Fragaria vesca* **after transformation with pCAMBIA vector 1304.**

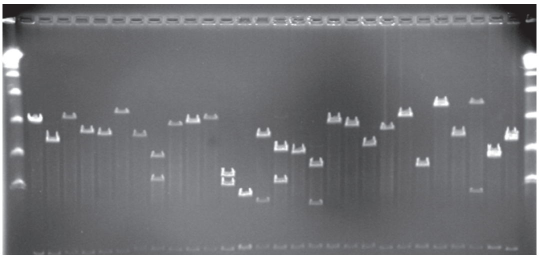


Figure 3. Agarose gel electrophoresis of DNA isolated from 28 randomly selected *F. vesca* BAC clones. DNA was digested with *Not*I enzyme. The marker is Lambda Ladder (with bands every 50 kb). The cloning enzyme is *Mbo*I and the vector is pECBAC1 (*Bam*HI site).

in the family. The Rosaceae community will benefit greatly from sequencing genomes of the representative models for each of the three subfamilies. The woodland strawberry provides the best candidate for full genome sequencing in the Rosaceae family, because it has the smallest genome of any economically important plant species and relatively small numbers of repetitive sequences.

We have organized an International Consortium aimed at sequencing the full genome of the woodland strawberry using the Roche GS-FLX™ technology. The sequencing of the woodland strawberry genome can fully utilize the advantages of the Roche GS-FLX™ pyrosequencing technology and provide an invaluable resource to the plant research community. The sequencing will take place at the VBI Core Laboratory Facility (CLF), where the first Roche GS-FLX™ was installed in January 2007. The availability of the Roche GS-FLX™ genome sequencer at VBI has created unique opportunities for VBI researchers to use next-generation sequencing systems in their projects and several bacterial full-genome sequencing projects have already been completed at the VBI CLF. The woodland strawberry genome will be assembled and annotated as a joint community project. The members of this consortium have significant expertise and experience in all aspects of full-genome sequencing, assembly and annotation.

We constructed a *F. vesca* genomic BAC library in the pECBAC1 vector to help with full-genome sequencing. The average insert size is about 130 kb (Figure 3). The sequencing of the individual Bacterial Artificial Chromosome (BAC) clones will assist with anchoring contiguous sequences derived from the Roche GS-FLX™ sequencing, and will also assist in completing the genome sequencing.

Metabolomics is a global holistic approach for the large-scale measurement of cellular metabolites. It can provide an integrated view of the biological system when combined with transcriptomics, proteomics and mathematical modeling. Our group is working at the forefront of developing and applying this metabolomics technology to various biological systems.

Conferences and workshops

Bajad S, Shuman J, Pisciotta J, Coppens I, Tripathi A, Scholl P, Sullivan D, Shulaev V (2007) Lipidomics approach to the identification of lipids associated with heme crystals (hemozoin) formed by human malarial parasite *Plasmodium falciparum, 2nd Annual Virginia Bioinformatics Institute Research Symposium*, Mountain Lake, VA, September 5-6, 2007.

Bajad S, Shuman J, Pisciotta J, Sha W, Rasolson D, Shi L, Ginsanrin O, Sullivan D, Shulaev V (2007) A metabolomics study of intra-erythrocytic *Plasmodium falciparum* ring and trophozoite developmental stages, *Journal of American Society of Mass Spectrometry* **18**: 37S.

Cortes D, Yang Y, Varbanova M, Wang G, Ross J, Noel J, Pichersky E, Shulaev V (2007) Functional genomics of Arabidopsis SABATH methyltransferases: the identification of enzymes and their substrates and products related to methylation of plant signaling molecules, *2nd Annual Virginia Bioinformatics Institute Research Symposium*, Mountain Lake, VA, September 5-6, 2007.

Pisciotta JM, Tripathi AK, Ogunbiyi O, Rasoloson D, Shuman J, Shulaev V, Sullivan DJ (2007) Changes in global metabolites of *Plasmodium falciparum* with erythrocyte development and antimalarial drugs, *MPM XVII - 2007 Molecular Parasitology Meeting*, Woods Hole, MA, September 16-20, 2007.

Sha W, Shuman J, Torty S, Akman S, Jarrah A, Laubenbacher R, Torti F, Mendes P, Shulaev V (2007) Metabolic fingerprinting of normal and fully malignant human breast epithelial cells, *2nd Annual Virginia Bioinformatics Institute Research Symposium*, Mountain Lake, VA, September 5-6, 2007.

Shulaev V (2007) Systems biology of the yeast oxidative stress response, *BioConvene-2007 - International Conference on Bioinformatics and Drug Discovery*, University of Hyderabad, India, December 2007.

Shulaev V (2007) Metabolic responses of *Plasmodium falciparum* to anti-malarial drugs, *Metabolomics Society 3rd Annual Conference*, The University of Manchester, Manchester, United Kingdom, June 11-14, 2007.

Shulaev V (2007) A systems biology study of the kinetics of *Saccharomyces cerevisiae* response to oxidative stress induced by cumene hydroperoxide, *RoS in Plants 2007*, Ghent University, Ghent, Belgium, September 12-14, 2007.

Shulaev V, Akman S (2007) Metabolic fingerprinting of breast cancer, *Biomarker Discovery Summit*, Philadelphia, PA, September 17-19, 2007.

Tuli L, Martins A, Shuman J, Sha W, Mendes P, Shulaev V (2007) A systems biology study of the kinetics of *Saccharomyces cerevisiae* response to oxidative stress induced by cumene hydroperoxide, *2nd Annual Virginia Bioinformatics Institute Research Symposium*, Mountain Lake, VA, September 5-6, 2007.

Veilleux R, Oosumi T, Baxter A, Holt S, Ruiz-Rojas J, Flinn B, Dan Y, Nessler C, Shulaev V (2007) Insertional mutagenesis in the diploid strawberry *Fragaria vesca, GMO 2007 - ISHS International Symposium "Genetic Transformations - Challenges and Possibilities for Horticulture of the World",* Ski, Norway, September 16-20, 2007.

Cortes DF, Shuman J, Yang, Y, Varbanova M, Pichersky E, Shulaev V (2008) Untargeted LC-MS metabolite profiling with capillary monolithic columns for the identification of substrates and products of the Arabidopsis SABATH family methyltransferases, *The 32nd International Symposium on Capillary Chromatography (ISCC)*, Riva Del Garda, Italy, May 26-30, 2008.

Gasic K, Han Y, Kertbundit S, Shulaev V, Iezzoni AF, Stover EW, Bell RE, Wisniewski ME, Korban SS (2008) Transferability of apple EST-derived SSRs to other Rosaceae species, *XVI Plant & Animal Genomes Conference*, Town and Country Convention Center, San Diego, CA, January 12-16, 2008.

Pantazis C, Flinn BS, Veilleux RE, Pattison J, Nessler C, Shulaev V, Dan Y (2008) Strawberry transformation using kanamycin selection, *The World Congress on In Vitro Biology*, Tucson, AZ, June 14 - 18, 2008.

Ruiz-Rojas JJ, Pattison J, Sargent D, Oosumi T, Shulaev V, Veilleux R (2008) Mapping insertional mutants of the diploid strawberry, *Fragaria vesca*, through SNP discovery in flanking regions, *4th International Rosaceae Genomics Conference*, Pucon, Chile, March 16-19, 2008.

Shulaev V, Laurance L, Lee LL, Ferguson AT, Wilson DL (2008) The effects of drought and heat stress combination on Arabidopsis plants: a metabolomics analysis, *56th ASMS Conference on Mass Spectrometry and Allied Topics*, Denver, CO, June 1-5, 2008.

Peer-reviewed publications in the reporting period

Miller G, Courtu J, Shulaev V, Mittler R (2008) Reactive oxygen signaling in plants. In *Annual Plant Reviews: Intracellular Signaling in Plants*, Zhenbiao Yang (ed), Volume 33, pp. 189-201, Blackwell, Oxford.

Miller G, Shulaev V, Mittler R (2008) Reactive oxygen signaling and abiotic stress. *Physiologia Plantarum* doi: 10.1111/j.1399-3054.2008.01090.x

Shulaev V, Cortes D, Miller G, Mittler R (2008) Metabolomics for plant stress response. *Physiologia Plantarum* **132**: 199-208.

Shulaev V, Korban KS, Sosinski B, Abbott AG, Aldwinckle HS, Folta KM, Iezzoni A, Main D, Arús P, Dandekar AM, Lewers K, Brown SK, Davis TM, Gardiner SE, Potter D, Veilleux RE (2008) Multiple models for Rosaceae genomics. *Plant Physiology* doi: 10.1104/pp.107.115618

Suzuki N, Bajad S, Shuman J, Shulaev V, Mittler R (2008) The transcriptional co-activator MBF1c is a key regulator of thermotolerance in *Arabidopsis thaliana*. *Journal of Biological Chemistry* **283**: 9269-9275.

Tuli L, Martins A, Sha W, Mendes P, Shulaev V (2007) Proteome analysis of oxidative stress response to cumene hydroperoxide in *Saccharomyces cerevisiae*. *Molecular & Cellular Proteomics* 6: 46-46.

Varbanova M, Yamaguchi S, Yang Y, McKelvey K, Hanada A, Borochov, Yu F, Jikumaru Y, Ross J, Cortes D, Ma CJ, Noel J, Mander L, Shulaev V, Kamiya Y, Rodermel S, Weiss D, Pichersky E (2007) Discovery of gibberellin methyltransferases in Arabidopsis. *Plant and Cell Physiology* **48**: S59.

Veilleux R, Oosumi T, Wadl PA, Baxter A, Holt SH, Ruiz-Rojas J, Pattison J, Flinn B, Dan Y, Nessler C, Shulaev V (2008) Insertional mutagenesis in the diploid strawberry, *Fragaria vesca*. *Acta Horticulturae.* In press.

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: Timothy Driscoll, Matt Dyer, Isabel Osorio da Fonseca, James Lester, Chunhong Mao, Nebiyu Shukur, Chunxia Wang, Kelly Williams.

Abstract. The PathoSystems Biology Group studies intracellular symbiotic systems. We typically investigate *Sinorhizobium meliloti-Medicago* interactions. These dinitrogen fixing symbioses share many interesting features with related pathogenic bacteria (e.g. *Brucella*). In the reporting period, we focused on functional genomics, seeking to further elucidate the biological roles for genes annotated as unknown function and others known to play critical roles in symbiosis. We have updated the annotations for the chromosome of *S. meliloti* at the genome sequence level by re-using the PATRIC system and we have continued to update the NodMutDB (Nodulation Mutant Database; curated knowledge from scientific literature). We used Open Reading Frame screening to identify several enzymatic activities for previously unknown functional genes. We discovered 20 new genes using pyrosequencing technology. Furthermore, we demonstrated that GcrA, a master cell cycle regulator in *Caulobacter crescentus*, is present in *S. meliloti* and essential for viability. Our work has confirmed that the upstream tmRNA end in *Sinorhizobium* is the unprocessed start site of transcription. We continued and extended our work on the comparative interactomics of human-pathogen systems, and have initiated several new projects to apply visual analytic techniques to such data.

Keywords: Alphaproteobacteria; *Sinorhizobium meliloti*; functional/genome annotation; symbiosis; database; curation; transfer messenger RNA (tmRNA); host-pathogen protein-protein interactions; visual analytics; pyrosequencing; transcriptome; secretome; interactomics.

Scientific Progress

Enzymatic function identification using Open Reading Frame screening

Functional annotation consists of attaching biological information, for example biochemical and biological functions, to genomic elements, through a series of standard operating procedures that try to maximally leverage different forms of knowledge and evidence. *In silico* analyses play a very important role in the process but do not always provide definitive evidence for biological functions. We have targeted genes annotated as 'hypothetical' activity (inferred electronically) as a major target group for high-throughput laboratory experiments to help define substrates and activities where appropriate. First, our highly curated *S. meliloti* database (http://rhizobia.vbi.vt.edu/curation_v2/) was searched via BLASTp using characterized glyoxylate dehydrogenases (Gdh) as templates. BLASTp results revealed that VBISMc0237, VBISMc3895, VBISMc3376, VBISMa0056, VBISMa1352, and VBISMc3699 are likely Open Reading Frames (ORFs) that code for enzymes with Gdh activity. All targets were cloned, expressed, purified, and tested by direct assay. The assays demonstrated that VBISMc0237, VBISMc3376, and VBISMa1352 ORFs, annotated as 2-hydroxyacid dehydrogenases, have Gdh activity with different k_{cat} and specificity constants. However, VBISMc3895, annotated as glycerate dehydrogenase, did not show the expected activity. In addition, VBISMc3699, annotated as a putative D-3-phosphoglycerate dehydrogenase (3-pgdh), presented Gdh activity but did not have 3-pgdh activity. These results show the importance of performing *in vitro* activity analysis in parallel with *in silico* analysis to obtain a higher degree of confidence in the functional annotation. The process is being adapted for high-throughput and the information will be fed back into the annotation processes on an ongoing basis.

A homolog of GcrA, a cell cycle regulator, is present and essential in S. meliloti *Rm1021*

S. meliloti invades host plant cells and differentiates to form nitrogen-fixing bacteroids in the early stage of nodulation. The bacteroids seem to be non-dividing from this point in the cell cycle onwards (Mergaert et al, 2006). The regulatory genes and factors controlling the cell cycle and the process of bacteroid differentiation during the early stage nodulation have not been well characterized. In contrast, cell cycle control in *Caulobacter crescentus*, a closely related member of the Alphaproteobacteria, is well understood and several regulatory proteins that play essential roles in cell cycle progression have been identified (Laub et al, 2007; Li et al, 2008). For example, GcrA is essential and is a global regulator of multiple cell cycle functions in *C. crescentus*. We have identified the promoter region of *gcrA* using rapid amplification of 5' complementary DNA ends (5'-RACE). The transcript initiation site is mapped to an adenosine residue 122 nucleotides upstream of the start codon. Furthermore, we showed, using knockouts, that *gcrA* is essential in *S. meliloti*.

Identification of new genes in S. meliloti *using pyrosequencing of transcripts*

We used pyrosequencing to search for bacterial genes missed in genome annotation processes (Mao et al, 2008). A pilot sequencing run generated 19,005 reads with an average length of 136 nucleotides, from which we identified 20 new genes. Reverse Transcriptase Polymerase Chain Reaction (RT-PCR) confirmed the transcripts and their possible functions were analyzed. Our results indicate that high-throughput sequence analysis of bacterial transcriptomes is feasible and next-generation sequencing technologies will greatly facilitate the discovery of new genes and improve our genome annotation efforts.

Pyrophosphatase sensitivity of the distal 5´ end of S. meliloti *two-piece tmRNA*

Transfer messenger RNA (tmRNA) rescues stalled bacterial ribosomes through combined tRNA-like and mRNA-like properties. It has been found in a two-piece form in all Alphaproteobacteria, some cyanobacteria and some Betaproteobacteria, due to ancestral gene permutation events. Processing of the internal region of permuted pre-tmRNA, producing two of the four mature ends, is effectively the same as for tRNAs, whereas processes acting on the two distal ends of the precursor are less well-defined. We mapped these two distal ends and the internal ends of an alphaproteobacterial tmRNA through ligation-mediated amplification and found them to be highly uniform, although three differ from those mapped in a previous study. We further demonstrated that the 5´ end of the coding piece is a monophosphate as expected from tRNA-like processing, while the 5´ end of the acceptor piece is sensitive to pyrophosphatase as would be the 5´-triphosphate left upon transcription initiation. This, together with its identity as a purine nucleotide and its appropriate positioning relative to conserved promoter sequences, indicates that the upstream tmRNA end in *Sinorhizobium* and other Alphaproteobacteria is the unprocessed start site of transcription.

Improvement of genome annotation for S. meliloti

We updated the genome sequence annotation for the chromosome of *S. meliloti* using the PATRIC system (http://www.patric.vbi.vt.edu). A total of 3531 coding sequences (CDSs) were predicted including 195 new CDSs that were not present in RefSeq. Our annotation agreed with 93% of RefSeq CDSs at the sequence level when compared with the RefSeq data. We shortened/extended 7% of the RefSeq CDSs. The update of the annotation of *S. meliloti* proteins is in progress.

NodMutDB - Nodulation Mutant Database

NodMutDB (http://nodmutdb.vbi.vt.edu) is a web-accessible relational database developed to provide a comprehensive and accurate resource for depositing, organizing and retrieving information on symbiosis-related genes, mutants and published literature (Mao et al, 2005). This year, we collected, updated and curated over 100 symbiosis genes and mutants from rhizobia, *S. meliloti* and *Bradyrhizobium japonicum*, and their host plants, *Medicago truncatula* and *Glycine max*, respectively.

Comparative interactomics of human-pathogen systems

Mechanisms of infection have been studied in molecular detail for many host-pathogen systems, especially for human hosts. However, many questions are relatively unexplored. For example, what are the properties of human proteins that interact with pathogens? Do pathogens interact with certain functional classes of human proteins? Which infection mechanisms and pathways are commonly triggered by multiple pathogens? Recently, we provided the first study of the landscape of human proteins interacting with pathogens. To do so, we integrated human-pathogen protein-protein interactions (PPIs) from seven public databases. Nearly all of the PPIs were for viral systems, with the majority belonging to the human-HIV system. We found that both viral and bacterial pathogens tend to interact with hubs (proteins with many interacting partners) and bottlenecks (proteins that are central to many paths in the network) in the human-PPI network. We constructed separate sets of human proteins interacting with multiple bacteria and with multiple viruses. Gene Ontology functions enriched in these sets reveal a number of processes, such as cell cycle regulation, nuclear transport, and immune response that participate in interactions with different pathogens. Our results provided the first global view of strategies used by pathogens to subvert human cellular processes and infect human cells. We are currently extending these comparative models to identify and prioritize antiviral and antibacterial targets.

Conferences and workshops

Driscoll T, Sobral BWS (2007) Genomics of transcriptional regulation in the Rhizobiales, *2nd VBI Annual Research Symposium*, Mountain Lake, VA, September 5-6, 2007.

Dyer MD, Murali TM, Sobral BWS (2007) Computational prediction of host-pathogen protein, *11th Annual International Conference on Research in Computational Molecular Biology*, Oakland, CA, April 2007.

Dyer MD (2007) Computational prediction of host-pathogen protein-protein interactions, Invited talk, *15th Annual Conference on Intelligent Systems for Molecular Biology and Sixth European Conference on Computational Biology*, Vienna, Austria, July 2007.

Dyer MD (2007) Pathosystems biology: using systems biology to study host-pathogen systems, Invited talk, *Chile, United States, and Brazil Bioinformatics Workshop* hosted by the US Department of State, Santiago, Chile, October 2007.

Fonseca IO, Martins AMM, Mendes P, Sobral BWS (2007) Characterization of unknown, putative and hypothetical open reading frames (ORFs) using substrate screening, *2nd VBI Annual Research Symposium*, Mountain Lake, VA, September 5-6, 2007.

Mao C, Driscoll T, Dyer M, Wang C, Williams K, Sobral BWS (2007) Identification of novel small RNAs in *Sinorhizobium meliloti, 2nd VBI Annual Research Symposium*, Mountain Lake, VA, September 5-6, 2007.

Wang C, Sheng X, Equi R, Charles TC, Sobral BWS (2007) Influence of the PHB granule-associated proteins (PhaP) on PHB accumulation and symbiotic nitrogen fixation in *Sinorhizobium meliloti* Rm1021, *107th General Meeting of the American Society for Microbiology*, Toronto, Canada, May 21-25, 2007.

Wang C, Sheng X, Equi R, Charles TC, Sobral BWS (2007) PHB granule-associated proteins in *Sinorhizobium meliloti* Rm1021, *20th North American Symbiotic Nitrogen Fixation Conference.* Marquette University, WI, July 10-14, 2007.

Wang C, Kemp J, Mao C, Sheng X, Equi R, Sobral, BWS (2007) Genome-wide analyses of *Sinorhizobium meliloti* 1021 gene expression responsive to the ExoS/ChvI two-component regulatory system, *2nd VBI Annual Research Symposium*, Montain Lake, VA, September 5-6, 2007.

Peer-reviewed publications in reporting period

Dyer MD, Murali TM, Sobral BWS (2007) Computational prediction of host-pathogen protein protein interactions. *Bioinformatics* **23**(13): i159-166.

Dyer MD, Murali TM, Sobral BWS (2008) The landscape of human proteins targeted by viruses and other pathogens. *PLoS Pathogens* **4**(2): e32.

Mao C, Evans C, Jensen RV, Sobral BWS (2008) Identification of new genes in *Sinorhizobium meliloti* using the genome sequencer FLX system. *BMC Microbiology* **8**: 72.

Wang C, Sheng X, Equi RC, Trainer MA, Charles TC, Sobral BWS (2007) Influence of the PHB granule-associated proteins (phaP) on PHB accumulation and symbiotic nitrogen fixation in *Sinorhizobium meliloti* Rm1021. *Journal of Bacteriology* **189**: 9050-9056.

Other cited references

Laub MT, Shapiro L, McAdams HH (2007) Systems biology of *Caulobacter. Annual Review of Genetics* **41**: 429-441.

Li S, Brazhnik P, Sobral B, Tyson JJ (2008) A quantitative study of the division cycle of *Caulobacter crescentus* stalked cells. *PLoS Computational Biology* **4**: e9.

Mao C, Qiu J, Wang C, Charles TC, Sobral BW (2005) NodMutDB: a database for genes and mutants involved in symbiosis. *Bioinformatics* **21**(12): 2927-2929.

Mergaert, P, Uchiumi T, Alunni B, Evanno G, Cheron A, Catrice O, Mausset A-E, Barloy-Hubler F, Galibert F, Kondorosi A, Kondorosi E (2006) Eukaryotic control on bacterial cell cycle and differentiation in the Rhizobium-legume symbiosis. *Proceedings of the National Academy of Sciences USA* **103**: 5230-5235.

Cyberinfrastructure for Biodefense and Emerging and Re-emerging Infectious Diseases

Group contributors: Chris Bowns, Cory Byrd, Stephen Cammer, Sarah Cousins, Oswald Crasta, Mike Czar, Chitti Dharmanolla, Nataraj Dongre, William Donnell, Matt Dyer, Herman Formadi, Joe Gabbard, James Gardner, Joe Gillespie, Debby Hix, Ranjan Jha, Ron Kenyon, Christine Lee, Jian Li, Zhiyi Li, Dan Liu, Jian Lu, Shrinivasrao Mane, Bharat Mehrotra, Saroj Mohaptra, Eric Nordberg, Anjan Purkayastha, Daphne Rainey, Harsha Rajasimha, Vincent Samaco, Mark Scott, João Setubal, Joshua Shallom, Shamira Shallom, Bruce Sharp, Maulik Shukla, Eric Snyder, Bruno Sobral, Jeetendra Soneja, Dan Sullivan, Wei Sun, Yuying Tian, Nirali Vaghela, Nishantsinh Vaghela, Rebecca Wattam, Rebecca Will, Kelly Williams, Tian Xue, Boyu Yang, Hyunseung Yoo, Qiang Yu, Chengdong Zhang, Yan Zhang.

Abstract. The Cyberinfrastructure Group (CIG) develops and uses methods, infrastructure, and resources to enable scientific discoveries in infectious disease research by applying the principles of cyberinfrastructure to integrate data, computational infrastructure, and people (Atkins, 2003). CIG has developed many public resources for curated, diverse molecular and literature data from various infectious disease systems, and implemented the processes, systems, and databases required to support them. It also conducts research applying its methods, infrastructure and data to make new discoveries of its own. CIG participates in education and outreach activities, resulting in scientific discoveries and publications, and an outreach program involving development of project-centric cyberinfrastructure courses for the educators from high schools and undergraduate institutions as well as graduates and postgraduates. In the reporting period, key accomplishments include publication of the *Brucella abortus* S19 genome, deployment of a pipeline to improve genome annotations, phylogenomic analysis of ten rickettsial genomes, publication of an Alphaproteobacteria phylogenetic tree, development of a program to generate oligonucleotide sequences from whole genome sequences, continuing progress in numerous collaborative research projects, and development of an online self-guided bioinformatics tutorial.

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

Scientific Progress

The Cyberinfrastructure Group (CIG) has undertaken several significant scientific projects in the reporting period. In addition to ongoing enhancements to the National Institute of Allergy and Infectious Diseases (NIAID)-funded Pathosystems Resource Integration Center (PATRIC) (Snyder, et al, 2007), one of eight Bioinformatics Resource Centers (Greene et al, 2007), the Resource Center for Biodefense Proteomics Research (Zhang et al, 2008), and work as part of the Middle Atlantic Regional Center of Excellence for Biodefense and Emerging Infectious Diseases (MARCE; www.marcebiodefence.org), they include the following:

Sequencing and comparative genomic analysis of Brucella abortus *S19*

Brucella abortus S19, a spontaneously attenuated strain, has been used as a vaccine strain in vaccination of cattle against brucellosis for six decades. CIG used pyrosequencing to determine the genome sequence of S19 (Crasta et al, 2008). The main goal was to identify candidate virulence genes by comparative analysis against published genome sequences of virulent, closely related strains of *B. abortus*, 9-941 and 2308. 3062 genes were identified, annotated and compared, resulting in 263 genes that were non-identical between the S19 genome and either of the two virulent strains. Functional analyses of the differences have revealed 24 genes that may be associated with the loss of virulence in S19. Of particular relevance are four genes with more than 60 bp consistent difference in S19 compared to both the virulent strains. In the virulent strains, these genes encode an outer membrane protein and three proteins involved in erythritol uptake or metabolism (Crasta et al, 2008).

GenVar: Bacterial genome computational analysis pipeline

The GenVar pipeline (Yu et al, 2007) was 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). CIG used GenVar to analyze four genomes of *Brucella* revealing hundreds of missed gene calls, new split genes and indels, several of which were species-specific and hence provided clues to the understanding of *Brucella* pathogenicity and host specificity. CIG has also used GenVar on all bacterial genomes in PATRIC and the results are available on the PATRIC website. GenVar software is available to the community at http://patric.vbi.vt.edu/download/software/GenVar.

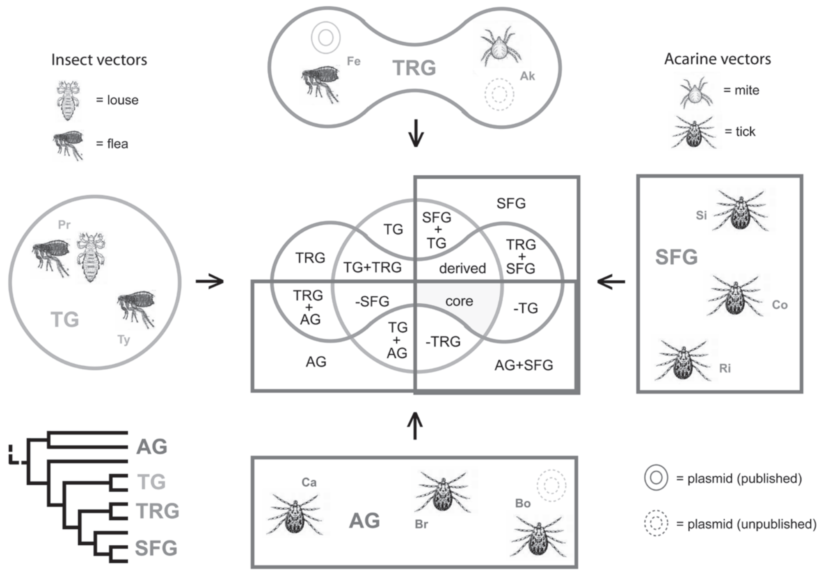


Figure. Venn diagram depicting 15 intersections for the four rickettsial groups. The classification scheme is based on molecular phylogeny estimation (see Gillespie JJ, Beier MS, Rahman MS, Ammerman NC, Shallom JM, et al. (2007) Plasmids and rickettsial evolution: insight from *Rickettsia felis*. PLoS ONE 2: e266), the topology of which is shown in the lower left; AG = ancestral group, TG = typhus group, TRG = transitional group, SFG = spotted fever group. Genome codes are as follows: Br = *R. bellii* str. RML369-C, Bo = *R. bellii* str. OSU 85 389, Ca = *R. canadensis* str. McKiel, Pr = *R. prowazekii* str. Madrid E, Ty = *R. typhi* str. Wilmington, Ak = *R. akari* str. Hartford, Fe = *R. felis* str. URRWXCal2, Ri = *R. rickettsii* str. Sheila Smith CWPP, Co = *R. conorii* str. Malish 7, and Si = *R. sibirica* str. 246. Arthropod hosts are illustrated for each genome, and strains known to harbor plasmids are depicted. The figure is reproduced from the following paper: Gillespie JJ, Williams K, Shukla M, Snyder EE, Nordberg EK, et al. (2008) Rickettsia Phylogenomics: Unwinding the Intricacies of Obligate Intracellular Life. PLoS ONE 3(4): e2018 doi:10.1371/journal.pone.0002018.

Phylogenomics

CIG undertook a phylogenomic study analyzing 14,354 PATRIC-predicted open reading frames (ORFs) from 10 rickettsial genomes (Gillespie et al, 2008). This analysis was used to generate orthologous groups (OGs) of proteins defining the core *Rickettsia* genome (752 OGs), and the accessory genome (1330 OGs and 1467 ungrouped singletons). Signature proteins that define the four major rickettsial groups as well as species infecting common arthropod hosts and species harboring plasmids were generated by mapping the distribution of the accessory genome over a robust phylogeny estimate (see Figure). The remaining OGs of the accessory genome contained the majority of duplicated genes, as well as ORFs coding for gene families typically involved in microbial pathogenicity and/or the lateral acquirement of virulence factors. These results are consistent with the recent large increase in the number of identified plasmids in *Rickettsia*, and are not in agreement with the traditional dogma for low amounts of lateral gene transfer in obligate intracellular bacteria. The work presents novel avenues of research for solving the riddle of *Rickettsia* pathogenesis.

CIG built a robust phylogenetic tree for the Alphaproteobacteria using 104 carefully selected and curated protein families (Williams et al, 2007). The mitochondrial branch could be placed in the tree within the *Rickettsiales*. Although some of the closest bacterial relatives to the mitochondria likewise live within eukaryotic cells, it was surprising to find that another of the closest relatives to mitochondria is an abundant marine bacterium. Several innovative bioinformatic methods were developed during this project (e.g., for phylogenetic detection of mis-annotated genes), and the whole approach is being wrapped into a large pipeline to help explore the evolutionary history of other bacterial groups. At present, the work-in-progress pipeline has been used to generate a phylogenetic tree for Gammaproteobacteria.

Oligonucleotide subsequence generation for diagnostics

CIG has developed a computer program called OligoId that starts with a collection of whole genome sequences and generates a grouped output of all possible oligonucleotide subsequences of a fixed length (Scott and Nordberg, 2007). The grouping provides a complete listing of all genome subsets that share a fixed-length subsequence that is absent in all non-subset genomes. The program will also use this output along with physical characteristics of the subsequences to generate desirable probes or primer pairs to be used for Polymerase Chain Reaction or microarray identification experiments. The computation time required by the program scales according to $O(n\log(n))$, making it practical to use even on non-high performance computing hardware. Data generation also produces information helpful in phylogenetic analysis. As a test, we applied the program to all of the genomes available to the organisms mentioned in Williams et al. (2007), and the phylogenetic trees it generated were in general agreement with those of the paper.

Collaborative research and outreach

The Collaborative Research Team (CRT) of CIG worked on 29 projects largely as part of MARCE. Almost all the work included analysis of the high-throughput data from sequencing, gene expression, or proteomics analyses of microbial or host samples. CRT completed the analyses of five newly sequenced genomes and over fifteen microarray experiments. Three papers were published, which describe the results from high-throughput sequencing of the *Brucella abortus* vaccine strain S19 bacterial genome (Crasta et al, 2008); analyses of the high-throughput gene expression data in hosts affected by pathogen or toxins (Koterski et al, 2007; Djavani et al, 2007). Similarly, the high-throughput microarray or proteomics data generated and analyzed in hosts or pathogens have lead to the submission of five manuscripts and the preparation of six additional manuscripts for submission to peer-reviewed journals.

CIG is developing and deploying a self-guided, online bioinformatics tutorial, using *Francisella tularensis* as the model organism, a MARCE-funded project. The course is an encapsulation of a CIG-developed, day-long, instructor-lead tutorial, providing investigators with background information on the resources and tools available for bioinformatics research. The tutorial consists of stand-alone modules with instruction in the areas of sequence comparison, sequence alignment, phylogeny, hidden Markov models, orthologous clusters, and whole genome comparison with hands-on exercises following each module. The tutorial is a multimedia-based application, and will be available from the CIG-developed MARCE website in the Summer of 2008 (http://marce.vbi.vt.edu/).

Conferences and workshops

Cammer S (2008) Protein structure visualization, illustration, and annotation, *First Workshop Proteomics in the New World,* Petropolis, Brazil, May 12-16.

Crasta O*, Rainey D, Mutter B, Craddock L, Faulkner S, Hart F, Eborall M, Foster L, Cammer S, Tretola B, Sobral B (2007) AC 2007-1200: A project-centric approach for cyberinfrastructure in bioinformatics, *American Society for Engineering Education*, 2007. http://www.asee.org/acPapers/AC%202007Full1200.pdf, June 24-27, 2007. *Presenting author.

Crasta O, Rainey D, Mutter B, Craddock L, Faulkner S, Hart F, Eborall M, Foster L, Cammer S, Tretola B, Sobral B (2007) AC 2007-1200: A project-centric approach for cyberinfrastructure in bioinformatics, *Proceedings of the 2007 ASEE Annual Conference & Exposition*, Hilton Hawaiian Village, Honolulu, HI, June 24-27, 2007.

Gillespie JJ* (2007) From bugs to bioinformatics: anything is possible, *Career Opportunity Seminar*, University of Maryland (Baltimore County campus), Baltimore, MD. *Presenting author.

Gillespie JJ*, Williams K, Snyder EE, Nordberg E, Ceraul SM, Dharmanolla C, Rainey D, Soneja J, Vishnubhat ND, Wattam R, Purkayastha A, Shallom JM, Shukla M, Czar M, Crasta O, Setubal J, Azad AF, Sobral B (2007) Rickettsia phylogenomics: unwinding the intricacies of obligate intracellular life, *21st Meeting of The American Society For Rickettsiology*, Colorado Springs, CO, September 8-11, 2007. *Presenting author.

Mutter B, Craddock L, Rainey D, Faulkner S, Hart F, Eborall M, Foster L, Cammer S, Tretola B, Sobral B, Crasta O* (2007) AC 2007-1179: Developing a multidisciplinary online cyberinfrastructure course through project-centric bioinformatics, *American Society for Engineering Education*, 2007, http://www.asee.org/acPapers/AC%202007Full1179.pdf, June 24-27, 2007. *Presenting author.

Purkayastha A* et al. (2007) How do I…? A tutorial on the Pathosytems Resource Integration Center V1.0, *12th International Conference of Emerging Infectious Diseases in the Pacific Rim*. Haikou, Hainan Island, China, December 4-7, 2007. *Presenting author.

Purkayastha A, Crasta OR, Eckart JD, Czar MJ, Meng XJ, Setubal JC, Sobral BW Bioinformatics resources for the study of viruses at the Virginia Bioinformatics Institute. In *Structure-based Study of Virus Infection*. Holland Cheng R, Miyamura T (eds), World Scientific Publishing Company, Stockholm, Sweden, pp. 267-288. ISBN: 981270406X.

Rainey D*, Cammer S, Faulkner S, Tretola B, Sobral B, Crasta O (2007) A project-centric approach to cyberinfrastructure education, *TeraGrid workshop*, Madison WI, June 2007. *Presenting author.

Sakamoto JM, Gillespie JJ*, Beier MS, Azad AF (2007) Investigating the role of *Rickettsia felis* plasmid pRF genes in pathogenicity, *6th Cold Spring Harbor Meeting on Microbial Pathogenesis and Host Response*, September 15-19, 2007. *Presenting author.

Setubal JC, Verjovski-Almeida S (2007) Best papers from the Brazilian Symposium on Bioinformatics, BSB 2005 - Guest editorial for a special issue of *Computers in Biology and Medicine*. **37**(2): 113-114.

Shallom J*, Dongre N, Crasta O, Gillespie J, Azad A, Setubal J (2007) Bioinformatics data and tools used to identify outer membrane proteins in *Coxiella* and *Rickettsia*. Oklahoma State University, Stillwater, OK, October 15, 2007. *Presenting author.

Shallom JM*, Dongre NV, Crasta OR, Gillespie JJ, Azad A, Setubal JC, Sobral BS (2007) Prediction of Sec-dependent secreted proteins in *Rickettsia conorii* (Malish 7) using bioinformatics tools, *21st Meeting of The American Society For Rickettsiology*, Colorado Springs, CO, September 8-11, 2007. *Presenting author.

Shallom JM (2008) Annotation of *Rickettsia africae* and *Rickettsia massiliae* at the PathoSystems Resource Center (PATRIC), *5th International Conference on Rickettsiae and Rickettsial Diseases*, Marseille, France, May 18-20, 2008.

Sobral B* et al. (2007) Cyberinfrastructure for Pathosystems Biology, Manchester, England, October 2007. *Presenting author.

Sobral B (2008) The PathoSystems Resource Integration Center: Implications for Rickettsiology.

Sobral B* et al. (2008) Identification of Countermeasure Targets using Computational Methods and their Experimental Validation, *Fifth National RCE Meeting Program*, Chicago, IL, April 2008.*Presenting author.

Sobral B* et al. (2008) PATRIC BRC: Resource for *Brucella*, *Rickettsia*, and *Coxiella* Countermeasure Developers, *Western Regional Center of Excellence for Biodefense and Emerging Infectious Diseases Meeting*, Galveston, TX, January 2008. *Presenting author.

Xue T, Yang B*, Will R, Sharp B, Kenyon R, Crasta O, Sobral B (2007) A Generalized Framework for Pathosystems Informatics & Bioinformatics Web Services, *The 2007 International Conference on Bioinformatics & Computational Biology* (BIOCOMP), Las Vegas, NV, June 2007. *Presenting author.

Intellectual property

Crasta OR, Folkerts O, Fei Z, Evans C, Yu G, Martino-Catt S, Mane S, Bricker B, DU L, Sobral BW (2007) *Brucella abortus* sequences and strains and uses thereof. The USPTO application was accorded a filing date of 15 November 2007 and assigned Application No. 60/996,390.

Scott M and Nordberg E (2007) Method for nucleic acid sequence identification. The USPTO application was accorded a filing date of 17 October 2007 and assigned Application No. 60/980,564.

Peer-reviewed publications in reporting period

Cammer SA, Czar MJ (2008) Web resources for protein analysis. In *Medical protein engineering*, Khudyakov Y (ed), CRC Press, Boca Raton, FL. In press.

Ceraul SM, Dreher-Lesnick SM, Gillespie JJ, Rahman MS, Azad AF (2007) A new tick defensin isoform and antimicrobial gene expression in response to *Rickettsia montanensis* challenge. *Infection and Immunity* **74**: 1973-1983.

Crasta OR, Folkerts O, Fei Z, Mane SP, Evans C, Martino-Catt S, Bricker B, Yu G, Du L, Sobral BW (2008) Genome sequence of *Brucella abortus* vaccine strain S19 compared to virulent strains yields candidate virulence genes. *PLoS ONE* **3**(5): e2193. doi:10.1371/journal.pone.0002193

Crasta OR, Wattam AR, Purkayastha A, Sobral BW (2007) Genome sequencing and bioinformatics resources. Animal associated microbes. In *Genome Mapping in Animals and Microbes*, Volume 8, Cole C, Nene V (eds), Springer Verlag. In press.

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

Djavani M, Crasta O, Zapata J, Fei Z, Folkerts O, Sobral BW, Bryant J, Pauza C, Lukashevich I, Salvato M (2007) Early blood profiles of virus infection in a monkey model for Lassa Fever. *Journal of Virology* **81**(15): 7960-7963.

Gabriels SH, Vossen JH, Ekregen SK, Van Ooien G, Abd-el-Halim AM, Van den Berg GCM, Rainey DY, Martin GB, Takken FL, de Wit PJ, Joosten MH (2007) An NB-LRR protein required for plant signaling mediated by both extra- and intracellular resistance proteins. *Plant Journal* **50**(1): 14-28.

Gillespie JJ, Williams K, Shukla M, Snyder EE, Nordberg E, Ceraul SM, Dharmanolla C, Rainey D, Soneja J, Vishnubhat ND, Wattam R, Purkayastha A, Shallom JM, Czar M, Crasta O, Setubal J, Azad A, Sobral B (2008) Rickettsia phylogenomics: unwinding the intricacies of obligate intracellular life. *PLoS One* **3**(4): e2018 doi:10.1371/journal.pone.0002018

Gillespie JJ, Beier MS, Rahman S, Ammerman NC, Shallom JM, Purkayastha A, Sobral BS, Azad AF (2007) Plasmids and rickettsial evolution: Insight from *Rickettsia felis*. *PLoS One* **2**(3): e266.

Greene JM, Collins F, Lefkowitz EJ, Roos D, Scheuermann RH, Sobral B, Stevens R, White O, Di Francesco V (2007) NIAID Bioinformatics Resource Centers - New Assets for Pathogen Informatics. *Infection and Immunity* **75**(7): 3212-3219.

Koterski J, Twenhafel N, Porter A, Reed DS, Martino-Catt S, Sobral BW, Crasta O, Downey T, DaSilva L (2007) Gene expression profiling of non-human primates exposed to aerosolized Venezuelan equine encephalitis virus (VEEV). *FEMS Immunology & Medical Microbiology* **51**(3): 462-472.

Li S, Brazhnik P, Sobral B, Tyson JJ (2008) A quantitative study of the division cycle of *Caulobacter crescentus* stalked cells. *PLoS Computaional Biology* **4**(1): p. e9.

Williams KP, Sobral BW, Dickerman AW (2007) A robust species tree for the Alphaproteobacteria. *Journal of Bacteriology* **189**(13): 4578-4586.

Yu GX, Snyder EE, Boyle SM, Crasta OR, Czar MJ, Mane SP, Purkayastha A, Sobral BW, Setubal JC (2007) A versatile computational pipeline for bacterial genome annotation improvement and comparative analysis, with Brucella as a use case. *Nucleic Acids Research* **35**(12): 3953-3962.

Zhang C, Crasta O, Cammer S, Will R, Kenyon R, Sullivan D, Yu Q, Sun W, Jha R, Liu D, Xue T, Zhang Y, Moore M, McGarvey P, Huang H, Chen Y, Zhang J, Mazumder R, Wu C, Sobral B (2008) An emerging cyberinfrastructure for biodefense pathogen and pathogen:host data. *Nucleic Acids Research* **36**(Database issue): D884-D891.

Other cited 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.

Snyder EE, Kampanya N, Lu J, Nordberg EK, Rajasimha HK, Shukla M, Soneja J, Tian Y, Xue T, Yoo H, Zhang F, Dharmanolla C, Dongre NV, Gillespie J, 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 A, 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 BW (2007) PATRIC: The VBI PathoSystems Resource Integration Center. *Nucleic Acids Resesearch* **35**: D401-D406.

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 Ferreira, Nick Galloway, Biao Gu, Regina Hanlon, Shiv Kale, Konstantinos Krampis, Grace Martin, Robert Presler, Trudy Torto-Alalibo, Sucheta Tripathy, Ken Tian, 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. Understanding the structure of these networks will enable more sophisticated approaches to disease prevention and control. We are building data sets and tool sets to dissect host-pathogen genetic networks, with a principal focus on oomycete pathogens of plants. This year we have completed the draft genome sequence of the *Arabidopsis* pathogen *Hyaloperonospora parasitica* and began perfecting the sequence of the soybean pathogen *Phytophthora sojae.* Bioinformatic analyses of these genome sequences have identified a large, diverse set of genes that encode virulence proteins that can enter plant cells to suppress plant defense reactions such as programmed cell death. Mathematical modeling and genome sequence analysis predict that these genes evolve by a birth-and-death process. To determine the dynamics of how the pathogen and plant genes interact, we are carrying out gene expression profiling of soybean and its pathogen *Phytophthora sojae* using microarrays. This work has included the construction of a very high-density genetic map of soybean. To infer genetic regulatory networks from these data, we are developing novel methods for inferring and modeling biological processes using dimensional reduction and sequential dynamical methods. This work has included developing a novel method for quantitating and minimizing uncertainty in the inferred models.

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

Scientific Progress

Genome sequences of oomycete plant pathogens

We are sequencing the genomes of several oomycete plant pathogens in order to characterize their genetic repertoire. This year we completed the analysis of the draft genome sequence of the obligate parasite of *Arabidopsis*, *Hyaloperonospora parasitica*. This included hosting a community sequence annotation workshop in August 2007. Major findings from this study are that the genome of this parasite has fewer virulence genes in almost every category than its saprophytic relatives, *Phytophthora sojae* and *Phytophthora ramorum*, suggesting that it has become highly specialized to its host. Furthermore, it has lost genes for inorganic nitrogen and sulfur utilization and has lost most genes involved in zoospore production (this pathogen disseminates by windborne conidia, not by waterborne zoospores).

We have also begun a project to finish the genome sequence of *P. sojae*, taking it from draft quality to near-perfect quality. Major findings resulting from this work so far are that many virulence effector genes (see next section) occur in multiple tandem copies, a feature that was overlooked by the sequence assembly software that produced the draft sequence. This provides evidence that these genes are evolving via a birth-and-death model (see next section).

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

Collaborators	Research Institution
Jim Beynon	Warwick University, United Kingdom
Jeffrey Boore	Genome Project Solutions, Inc.
Sandra Clifton	Washington University Genome Sequencing Center
Alan Collmer	Cornell University
Candace Collmer	Wells College
Ralph Dean, David Bird	North Carolina State University
Allan Dickerman, João Setubal	Virginia Bioinformatics Institute
Igor Grigoriev	United States Department of Energy Joint Genome Institute
John McDowell	Plant Pathology, Physiology and Weed Science at Virginia Tech
Richard Meyers, Jeremy Schmutz	Stanford University
TM Murali	Computer Science at Virginia Tech
Nicole Perna	University Wisconsin, Madison
Jane Rogers	Sanger Sequencing Center, Cambridge, United Kingdom
Owen White	The Institute for Genomic Research (TIGR), J Craig Venter Institute

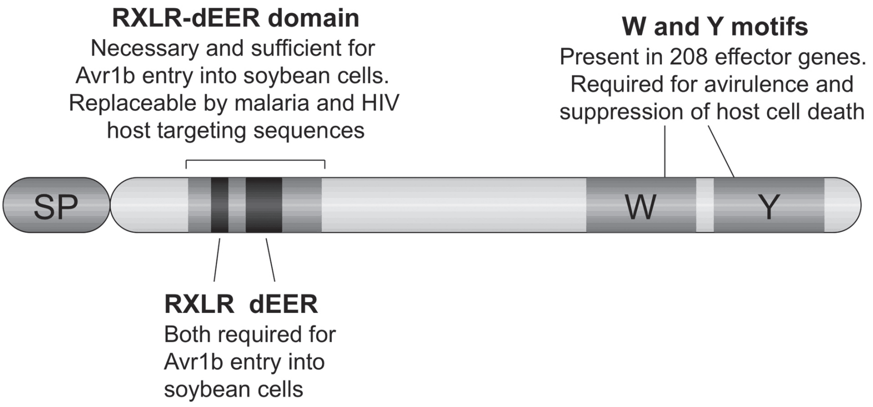

RXLR-dEER domain
Necessary and sufficient for Avr1b entry into soybean cells. Replaceable by malaria and HIV host targeting sequences
W and Y motifs
Present in 208 effector genes. Required for avirulence and suppression of host cell death
SP
W
Y
RXLR dEER
Both required for Avr1b entry into soybean cells

Figure 1. Structure and functional motifs of *Phytophthora sojae* effector protein Avr1b.

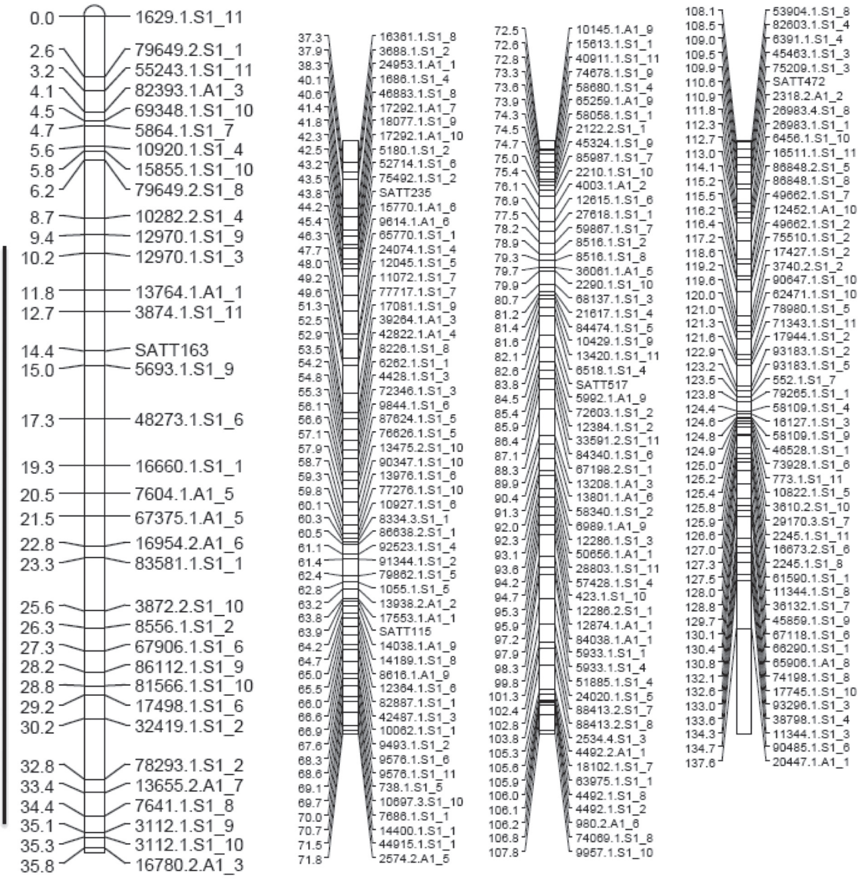


Figure 2. High-density genetic map of soybean linkage group G showing the location of a *P. sojae* resistance QTL.

Associated with the genome sequencing projects, we have been developing standardized Gene Ontology terms to describe the biological processes that plant-associated microbes use in their interactions with plants (www.geneontology.org). This year we have been refining our previously developed terms and have been using them to annotate the *P. sojae* genome sequence. Furthermore, we conducted two training workshops for the scientific community in the use of the terms in the summer of 2007.

Functional genomics of oomycete virulence effector proteins

Pathogens use effector molecules to manipulate the physiology of their hosts, making them more susceptible to infection. Some of these effectors can be recognized by plant resistance gene products, triggering an effective defense response; in this case the effectors are called avirulence proteins. Based on the sequence of the *P. sojae* avirulence proteins Avr1b, we used bioinformatic methods to predict nearly 400 effector genes in the genomes of *P. sojae* and *P. ramorum* and nearly 200 in the *H. parasitica* genome. Furthermore, we showed that these proteins all evolved from a singe ancestor. All of the proteins share two motifs at the N-terminus called RXLR and dEER. We showed that these two motifs, together with surrounding sequences, are necessary and sufficient for the effector proteins to enter plant cells in the absence of the pathogen. We also showed that around half of the proteins contain several C-terminal motifs (called W, Y and L), and that one function of these motifs is to suppress a key plant defense process called programmed cell death.

We mathematically modeled the evolution of the effector genes to address the question of how 400 genes with similar functions can be preserved. The results suggest that the genes evolve by a birth-and-death model, which is when genes are constantly being lost due to mutations, but natural selection favors

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

Collaborators	Research Institution
Mark Gijzen	Agriculture Canada, London, Ontario
Rays Jiang	Broad Institute
Jonathon Jones	Sainsbury Laboratories, Norwich, United Kingdom
John McDowell	Plant Pathology, Physiology and Weed Science at Virginia Tech
Weixing Shan	Northwest Agricultural and Forestry University, China
Chris Taylor	Donald Danforth Plant Science Center, St. Louis
Yuanchao Wang	Nanjing Agricultural University, China

pathogen strains in which the lost genes have been replaced by duplications of remaining functional copies. The model predicts that the genomes of oomycetes should contain large numbers of duplications of effector genes and also large numbers of effector pseudogenes that have been damaged by mutations; both predictions could be verified by an analysis of the genome sequences.

Transcriptional interplay between soybean and Phytophthora sojae *during infection*

To investigate the possible mechanisms of quantitative resistance, we have used Affymetrix GeneChips® to determine the transcriptional profiles of *P. sojae* during infection of soybean cultivars that have different levels of quantitative resistance against the pathogen. This year we completed the assay of transcriptional changes in a set of 300 recombinant inbred lines that is segregated for quantitative resistance in order to map the soybean genetic loci responsible for resistance and any associated transcriptional changes. As part of this project, we created a high-density soybean genetic map with more than 3500 genetic markers and aligned it with the soybean draft genome sequence.

Genetic network inference

To infer genetic regulatory networks from transcriptional data such as microarrays, we are using a dimensional reduction approach, which includes principal components analysis combined with multiple regression to build sequential dynamical models of biological systems using time course systems data. A major problem that arises with such experiments is the sparsity and uneven placing of the time points. We have used interpolation to solve this problem, and have developed an algorithm

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

Collaborators	Research Institution
Anne Dorrance and Steven St. Martin	Ohio State University
Ina Hoeschele	Virginia Bioinformatics Institute
Saghai Maroof	Crop, Soil, and Environmental Sciences, Virginia Tech
TM Murali	Computer Science, 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

to quantitate the uncertainty that arises from interpolation. From this we can predict which additional time points of real data would best improve the inferred model, leading to a new cost-effective approach to designing microarray time course experiments.

Conferences and Workshops

Tyler BM (2008) Bioinformatics and functional genomics of the soybean-*Phytophthora sojae* interaction, *Biotic Plant Interactions*, Brisbane, Australia, March 27-29, 2008.

Tyler BM (2008) Bioinformatics and functional genomics of the soybean-*Phytophthora sojae* interaction, *Genetics Program*, Texas A&M University, TX, February 21, 2008.

Tyler BM (2008) Dissection of quantitative resistance in soybean against *Phytophthora sojae* using genetical genomics, *International Plant and Animal Genome Conference XVI*, San Diego, CA, January 15, 2008.

Tyler BM (2008) Bioinformatics and functional genomics of *Phytophthora* infection, Department of Plant Pathology, University of California, Davis, CA, January 14, 2008.

Tyler BM (2008) Synthesizing genome sequence information from diverse host-associated microbes using the Gene Ontology, *International Plant and Animal Genome Conference XVI*, San Diego, CA, January 12, 2008.

Tyler BM (2007) Inferring genetic regulatory networks in host-pathogen interactions, *7th Annual ORFeome Meeting, "ORFeomes and Systems"*, Harvard Medical School, MA, October 24-26, 2007.

Tyler BM (2007) Genome sequences of oomycete pathogens reveal photosynthetic origins and mechanisms of pathogenesis, *2nd ESF/EMBO Symposium on Comparative Genomics of Eukaryotic Micro-organisms*, St Feliu de Guixols, Spain, October 20-25, 2007.

Tyler BM (2007) Effector repertoire of *Phytophthora sojae*: structural and functional genomics, *Annual Meeting of Amerian Phytopathological Society (APS)*, San Diego, CA, July 28-August 1, 2007.

Tyler BM (2007) Effector repertoire of *Phytophthora sojae*: structural and functional genomics, *13th International Congress Society of Plant-Microbe Interactions*, Sorrento, Italy, July 21-27, 2007.

Tyler BM (2007) Bioinformatics and functional genomics of pathogenicity in the soybean pathogen *Phytophthora sojae*, Oregon State University, Corvallis, OR, May 9, 2007.

Peer-reviewed publications in reporting period

Dou D, Kale SD, Wang XL, Chen Y, Wang Q, Wang X, Jiang RHY, Arredondo FD, Anderson RG, Thakur PB, McDowell JM, Wang YC, Tyler BM (2008) Conserved C-terminal motifs required for avirulence and suppression of cell death by *Phytophthora sojae* effector Avr1b. *Plant Cell* **20**(4): 1118-1133.

Jiang RHY, Tripathy S, Govers F, Tyler BM (2008) RXLR effector reservoir in two *Phytophthora* species is dominated by a single rapidly evolving super-family with more than 700 members. *Proceedings of the National Academy of Sciences USA* **105**(12): 4874-4879.

Tyler BM (2008) Genomics of fungal- and oomycete-soybean interactions. In *Soybean Genomics*. Stacey G (ed), Springer Verlag, New York. In press.

Tyler BM, Dou D, Kale SD, Jiang RHY, Chen Y, Wang Q, Wang X, Arredondo FD, Wang Y (2008) The effector repertoire of *Phytophthora sojae*: structure, function and evolution. In *Biology of Molecular Plant-Microbe Interactions*, Volume 6. Lorito M, Woo S, Scala F (eds), International Society for Molecular Plant-Microbe Interactions, St. Paul, Minnesota. In press.

Tyler BM, Jiang RHY, Zhou L, Tripathy S, Dou D, Torto-Alalibo T, Li H, Mao Y, Liu B, Vega-Sanchez M, Mideros SX, Hanlon R, Smith BM, Krampis K, Ye K, Martin SS, Dorrance AE, Hoeschele I, Maroof MAS (2008) Functional genomics and bioinformatics of the *Phytophthora sojae*-soybean interaction. In *The Genomics of Disease*. Gustafson P, Stacey G, Taylor J (eds), Kluwer Academic/Plenum Publisher, New York, 67-78.

2008 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.

Simulation and Analysis of Molecular Regulatory Systems in Cell Biology

Abstract. Complex networks of interacting proteins control the physiological properties of a cell (for example, metabolism, reproduction, motility, signaling, etc.). We convert network diagrams into sets of nonlinear ordinary differential equations (chemical rate equations) using basic principles of biochemical kinetics and then explore the models by analytical and computational methods. Of particular interest are the mechanisms that control cell division in prokaryotes (bacteria) and eukaryotes (yeasts, plants, insects, vertebrates). We published a comprehensive model of the genetic regulatory network that controls DNA synthesis, division and differentiation in a class of bacterial cells in the reporting period. A second paper included an exploration of the mechanisms controlling nuclear division during the earliest stage of development of fruit fly embryos. A third paper described modeling of spatial patterns of growth in fission yeast cells.

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

Scientific Progress

Cell cycle control in Caulobacter crescentus

Caulobacter crescentus is a dimorphic bacterium inhabiting freshwater, seawater and soils, where it plays an important role in global carbon cycling by mineralizing dissolved organic material. *Caulobacter* normally undergoes an asymmetric cell division cycle producing two different progeny: a motile swarmer cell with a flagellum and a sessile stalked cell (Figure 1). The two cell types undergo different developmental programs. The nascent stalked cell immediately enters into a new round of cell division and produces, about 90-120 min later, another swarmer cell. The nascent swarmer cell swims around for 30-45 min before it differentiates into a stalked cell and initiates division.

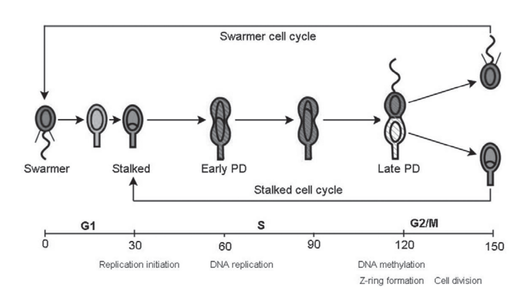


Figure 1. Cell division cycle of *Caulobacter crescentus*.

We have proposed a detailed molecular mechanism controlling DNA replication and asymmetric division in this organism by following the suggestions of experimental groups studying the cell cycle regulatory genes of *Caulobacter*. The mechanism is converted into a set of nonlinear ordinary differential equations that are simulated for the case of wild-type cell cycles and the aberrant phenotypes of dozens of mutations of the genetic system. Our simulations of the stalked cell cycle were published this year in *PLoS Computational Biology* (Li et al, 2008), and a manuscript describing the swarmer cell cycle is in preparation. This project is in collaboration with Dr. Bruno Sobral's group at the Virginia Bioinformatics Institute.

Cell cycle control in Drosophila melanogaster

Immediately after fertilization, the fruit fly egg undergoes 13 rapid nuclear divisions to produce 8192 nuclei in a common cytoplasmic zone just beneath the egg membrane (Figure 2). These nuclear division cycles are noteworthy for their speed: the early cycles require only eight minutes to replicate the genome and divide the sister chromatids into two daughter nuclei. The cycle can run quickly because the regulatory protein, a dimer of Cdk1 and cyclin B, is preformed in the cytoplasm and needs only to be transported into the nucleus to trigger DNA synthesis and nuclear division. At the end of each cycle, cyclin B is cleared from the nuclei by proteolysis, and a new cycle can begin. This mode of regulation of Cdk1/cyclin B is unknown in any other types of cells. By modifying our model of the generic mechanisms of Cdk1/cyclin B regulation, we were able to account for all the curious features

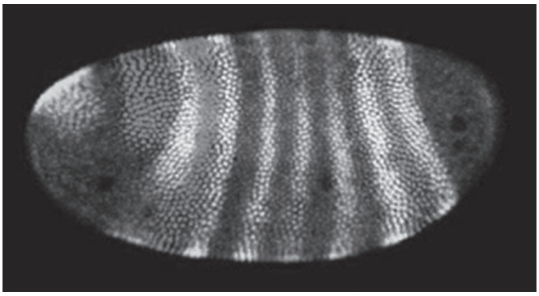


Figure 2. Rapidly dividing fruitfly egg.

of the rapid mitotic cycles in the fruit fly embryo, as described in a paper by Calzone et al (2007), published in *Molecular Systems Biology*. This research is performed in collaboration with Dr. Bela Novak's group at the Technical University of Budapest, Hungary.

Cell growth patterns in Schizosaccharomyces pombe

Fission yeast normally grow as rod shaped cells that elongate first from one end only and later (in the cell cycle) from both ends. A complex network of interacting proteins involved in cell cycle regulation, microtubule transport, and actin polymerization controls 'New End Take Off'. In collaboration with Dr. Bela Novak's group at the Technical University of Budapest in Hungary, Dr. Tyson contributed to a mathematical model that successfully explains the growth dynamics of wild-type cells and a collection of mutants with very bizarre shapes (see Figure 3) because of misregulated growth zones. This work is described in a paper by Csikasz-Nagy et al (2008) in the journal *Yeast*.

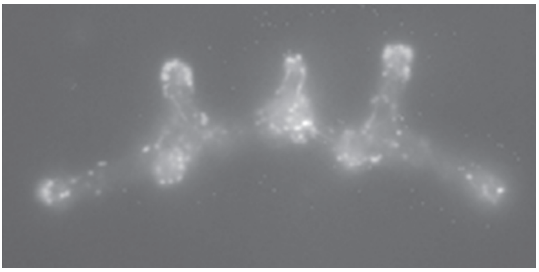


Figure 3. Cell growth pattern of mutant *Schizosaccharomyces pombe.*

Other notable achievements and activities

In July and August 2007, Dr. Tyson was the lead organizer of a six-week workshop on "Biological Switches and Clocks" at the Kavli Institute for Theoretical Physics at the University of California at Santa Barbara. About 50 senior scientists from around the world attended this workshop, from the disciplines of physics, chemistry, biology and mathematics. During the Fall Semester, 2007, Dr. Tyson held the Aisenstadt Chair of Mathematical Research at the University of Montreal, and from April-June, 2008, he served as a Visiting Research Fellow at Merton College, Oxford UK.

In the reporting period, Dr. Tyson gave invited talks at the following institutions: San Raffaele Institute (Milan, Italy), the Center for Computational & Systems Biology (Trento, Italy), Harvard Medical School

(Systems Biology), University of Michigan (Mathematics), University of Arizona (Molecular Cell Biology), Cornell University (Center for Applied Mathematics), and University of North Carolina (Department of Pharmacology). Dr. Tyson also provided tutorials on systems biology at the following institutions: Boulder School for Condensed Matter Physics, Otto Warburg International Summer School on Computational Cell Biology (Berlin), University of Montreal, Institute for Mathematics and Applications (Minneapolis).

Conferences and workshops

Tyson JJ (2007) *Symposium on Systems Biology*, Experimental Biology Annual Meeting, Washington, DC, April 28 – May 2, 2007.

Tyson JJ (2007) National Centers of Integrative and Systems Biology, Cambridge, MA, June 12-14, 2007.

Tyson JJ (2007) British Society for Developmental Biology, Sheffield, United Kingdom, September 5-7. 2007.

Tyson JJ (2007) *Workshop on Deconstructing Biochemical Networks*, University of Montreal, Canada, September 24-28, 2007.

Tyson JJ (2008) *Workshop on Molecular/ Biological Communications Technology*, National Science Foundation, Arlington, VA, February 20-21, 2008.

Tyson JJ (2008) *Computational Cell Biology*, Cold Spring Harbor-Wellcome Trust Conference Center, Hinxton, United Kingdom, March 26-29, 2008.

Peer-reviewed publications in the reporting period

Albert R, Goldbeter A, Ruoff P, Sible J, Tyson JJ (2008) Biological switches and clocks. *Journal of the Royal Society Interface.* In press.

Calzone L, Thieffry D, Tyson JJ, Novak B (2007) Dynamical modeling of syncytial mitotic cycles in *Drosophila* embryos. *Molecular Systems Biology* **3**: 131.

Csikasz-Nagy A, Gyorffy B, Alt W, Tyson JJ, Novak B (2008) Spatial controls for growth zone formation during the fission yeast cell cycle. *Yeast* **25**: 59-69.

Li S, Brazhnik P, Sobral B, Tyson JJ (2008) A quantitative study of the division cycle of *Caulobacter crescentus* stalked cells. *PLoS Computational Biology* **4**: e9.

Novak B, Tyson JJ, Gyorffy B, Csikasz-Nagy A (2007) Irreversible cell-cycle transitions are due to systems-level feedback. *Nature Cell Biology* **9**: 724-728.

Sabouri-Ghomi M, Ciliberto A, Novak B, Tyson JJ (2008) Antagonism and bistability in protein interaction networks. *Journal of Theoretical Biology* **250**: 209-218.