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

## Director’s Report

## Research Reports

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Network Dynamics and Simulation Science Laboratory</td>
<td>6</td>
</tr>
<tr>
<td>Chris Barrett</td>
<td></td>
</tr>
<tr>
<td>Development of Novel Approaches for Preventing Inflammation in Infectious and Chronic Diseases</td>
<td>21</td>
</tr>
<tr>
<td>Josep Bassaganya-Riera</td>
<td></td>
</tr>
<tr>
<td>Phylogenomics and Systems Biology of Plants and Plant Pathogens</td>
<td>33</td>
</tr>
<tr>
<td>Allan Dickerman</td>
<td></td>
</tr>
<tr>
<td>Genetic Architecture of Complex Diseases</td>
<td>37</td>
</tr>
<tr>
<td>Ina Hoeschele</td>
<td></td>
</tr>
<tr>
<td>Computational Systems Biology</td>
<td>42</td>
</tr>
<tr>
<td>Reinhard Laubenbacher</td>
<td></td>
</tr>
<tr>
<td>Alternaria Pathogenomics</td>
<td>49</td>
</tr>
<tr>
<td>Chris Lawrence</td>
<td></td>
</tr>
<tr>
<td>Fast Proteomic Fingerprinting in Cancerous Cells</td>
<td>61</td>
</tr>
<tr>
<td>Iuliana Lazar</td>
<td></td>
</tr>
<tr>
<td>Software, Standards and Methods for Computational Systems Biology</td>
<td>67</td>
</tr>
<tr>
<td>Pedro Mendes</td>
<td></td>
</tr>
<tr>
<td>Redox Biology of Methanogenic Archaea and Mycobacteria, Microbial Gasification of Coal and PEPCK Mechanism</td>
<td>72</td>
</tr>
<tr>
<td>Biswarup Mukhopadhyay</td>
<td></td>
</tr>
<tr>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Genetic Design Automation</td>
<td>88</td>
</tr>
<tr>
<td>Jean Peccoud</td>
<td></td>
</tr>
<tr>
<td>Gene Interaction Networks for Functional Analysis of Complex Biological Processes</td>
<td>96</td>
</tr>
<tr>
<td>Andy Pereira</td>
<td></td>
</tr>
<tr>
<td>Bacterial Genomics and Bioinformatics</td>
<td>102</td>
</tr>
<tr>
<td>João Setubal</td>
<td></td>
</tr>
<tr>
<td>Metabolomics for Systems Biology</td>
<td>108</td>
</tr>
<tr>
<td>Vladimir Shulaev</td>
<td></td>
</tr>
<tr>
<td>Cyberinfrastructure for Infectious Diseases</td>
<td>118</td>
</tr>
<tr>
<td>Bruno Sobral</td>
<td></td>
</tr>
<tr>
<td>Plant-Pathogen Interactions: From Genome Sequences to Genetic Networks</td>
<td>130</td>
</tr>
<tr>
<td>Brett Tyler</td>
<td></td>
</tr>
<tr>
<td>Faculty Fellow Report</td>
<td></td>
</tr>
<tr>
<td>Deterministic and Stochastic Models of Cell Cycle Regulation in Budding Yeast</td>
<td>139</td>
</tr>
<tr>
<td>John Tyson</td>
<td></td>
</tr>
</tbody>
</table>
Nine years ago, Virginia Tech established a new environment for research at the university. An initial concept of a collaborative scientific institute that would bring together researchers from diverse disciplines in the life sciences became the Virginia Bioinformatics Institute (VBI). In my short time here, I have come to appreciate the expertise and commitment that has been behind the development of VBI into the internationally competitive research institute that it is today.

The 2009 Scientific Annual Report showcases the scientific achievements of the past 12 months. It is organized around progress reports from our research groups. Readers of the report will discover the scientific advances that are helping to broaden our understanding of the natural world and impact the transformation of scientific knowledge into tools, technologies, and solutions for the life sciences.

The past year has seen changes at VBI. Bruno Sobral, the founding director of the Institute, returned to research in March. Bruno has been a driving force behind the growth of VBI, transforming it from a start-up enterprise to an internationally renowned research institution with a track record of successful scientific achievements. He continues to lead the Cyberinfrastructure and PathoSystems Biology research groups at VBI. A search is underway for a new Executive Director who will lead VBI into the next phase of its development.

Meanwhile, we anticipate further exciting contributions from VBI to the scientific landscape. The Institute’s strengths in infectious disease research make it a natural partner for research collaborations in the Virginia Tech and wider scientific communities. VBI is playing a lead role in Virginia Tech’s initiative to establish a major research center in the Ballston area of the National Capital Region. The state-of-the-art facility will offer a broad translational research effort in high-performance computing applied to social computing, decision and policy informatics, which constitutes a major part of the research agenda at VBI’s Network Dynamics and Simulation Science Laboratory. VBI will continue to pursue the national and international research partnerships that have been behind its successful growth.

I would like to take this opportunity to thank everyone who has contributed to the many accomplishments of VBI.

Sincerely,

Paul Knox
Interim Director, Virginia Bioinformatics Institute
University Distinguished Professor
2009 Research Reports

from the Virginia Bioinformatics Institute
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 $20 million in the past four years and includes programs with the National Institutes of Health, the Department of Transportation (through AECOM), the Centers for Disease Control and Prevention, the Department of Defense, the National Science Foundation, and the Bill & Melinda Gates Foundation. 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, computational economics, energy systems, sustainable interdependent infrastructure design and analysis, and high-performance computing. We continue to develop advanced high-performance computing 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. Four large studies to support pandemic planning for military preparedness were completed for the Defense Threat Reduction Agency of the Department of Defense using a prototype system we have built, called the Comprehensive National Incident Management System. The Comprehensive National Incident Management System integrates surveillance, simulation-assisted hypothesis testing, and decision support for use in situational awareness and planning in complex systems. For the first time, we transitioned our technology so that public health analysts can now do certain kinds of case studies by using our system, via a simple web interface. The novel use of high-performance computing assets for addressing practical societal problems represents a unique capability developed by NDSSL.

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:** Kofi Adasi, Alexander Alon, Andrea Apolloni, Chris Barrett, Richard Beckman, Sanket Bedare, Katelyn Bitely, Keith Bisset, Deepti Chafekar, Karthik Channakeshava, Jiangzhuo Chen, Danielle Choi, Ha Dang, Tom DuBois, Lisa Durbeck, Tridib Dutta, Stephen Eubank, Annette Feng, Xizhou Feng, Joshua Ferrier, Emily Gooding, Kai Han, Ginger Hansen, Steve Harris, Lauren Hoops, Fei Huang, Rahul Kanna Narayanan Jayaraman, Ken Kania, Maleq Khan, Sanjay Kishore, Pavan Konanki, Chris Kuhlman, Ajit Kulkarni, Jonathan Leidig, Bryan Lewis, Suzanne Locascio, Yifei Ma, Matthew Macauley, Achla Marathe, Madhav Marathe, Gabriel Mateescu, Khristi Moore, Henning Mortveit, Ganesh Narayanaswamy, Elaine Nsoesie, Guanhong Pei, Joyce Randall, Tom Scogland, Naresh Shenoy, Sharon Smyth, Shvetha Sondrararajan, Prateek Srivastava, Paula Stretz, Zhifeng Sun, Samarth Swarup, Anil Vullikanti, Kate Wendelsdorf, Mary Williams, Shrirang Yardi, Zhao Zhao, 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). 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. High-performance computing (HPC) systems and software that form the backend of DIDACTIC are completely invisible to the end user, allowing the analyst to focus on simulation-based epidemiological case studies without becoming a computing expert. The tool is being configured for use in military preparedness studies in the event of pandemics. This is a unique and novel capability – it is based on combining novel methods in service-oriented computing, grid computing, high-performance parallel simulations, network science, and statistical methods for large simulation-based experimental designs. To our knowledge this is the first time such a technology has been developed and deployed in the context of public health epidemiology.
We have also continued the development of Simdemics – a scalable HPC-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 of 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. A novel feature of our tools is that they combine new algorithmic techniques with fast parallel implementations resulting in the ability to compute the spread of infectious diseases and their interactions with individual behavior and social contact networks on large urban regions in a matter of a few minutes per run (for example, it takes approximately two minutes for one run on networks the size of Chicago that comprise nine million nodes and 500 million edges). Moreover, the underlying mathematical model is general and we fully anticipate that it will be useful in understanding other diffusion processes such as fads, norms, and others.

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. The TIGS tool can now map on various parallel clusters. The specific choice of a cluster is determined automatically at run time based on the perceived system load. We have added a new module to the GALIB toolkit for extracting information on small patterns in large relational networks (also known as graph mining). Our group has acquired significant new computing hardware that makes it one of the few places to have and use such diverse HPC resources at a sustained level of utilization. This includes a large shared memory machine, a 750-core SGI parallel cluster, and a GPGPU-based cluster.

We are extending our modeling frameworks and tools to address other problems in systems biology and medical and health informatics. We have initiated new research programs in developing HPC-based computing for highly resolved models for systems biology. We are currently concentrating on developing prototype models to understand Inflammatory Bowel...
Disease and Human Immunodeficiency Virus/Acquired Immunodeficiency Syndrome (HIV/AIDS). We are also extending our methods in social networks and modeling diffusion processes for developing models of individual behaviors within a networked environment as they pertain to non-infectious health issues such as smoking and obesity.

Finally, new programs are being developed to integrate modeling tools for energy systems. This is a timely project that can leverage our expertise in individual-based demand modeling, commodity markets, and service-oriented systems. The importance of developing new technologies to meet the future requirements on production, distribution, and storage of renewable energy are consistent with the national and international focus on this important societal problem.

In the reporting period, we obtained several new theoretical and simulation-based analysis results pertaining to the study of large complex systems:

- Extension of results on estimating performance limits of wireless networks; the new results were obtained using a more detailed model of interference radio transmissions
- A new, distributed algorithm for embedding in tree metrics and using these embeddings to derive efficient approximation algorithms
- New results on graph dynamical systems, which are the theoretical underpinnings of our models and simulations, allow us to classify the possible long-term system behaviors in terms of the network structure and the update sequence
- New dynamical measures related to diffusion processes; including vulnerability and criticality, their importance, efficient computation, and their role in epidemic planning. Our results show that vulnerability and criticality are important dynamical measures; moreover classical structural measures that purely graph theoretic or purely individual agent based (e.g. betweenness centrality) are often inadequate
- New methods for representing and analyzing the co-evolution between individual behaviors, public policies, socio-technical networks, and dynamical processes (e.g. disease spread, spread of malware on wireless networks) and the impact of this co-evolution on various design and control strategies. For example, in a recent study pertaining to epidemics, we looked into the fairness of different intervention strategies by examining the economic impact within specific demographic classes. The results identify population strata by demographics that are likely to win or lose from the implementation of such policies. The research incorporates not only the cost of the disease but also the cost of disease avoidance
Figure. Number of people infected on each day of an influenza epidemic, based on a single run of an NDSSL contagious disease simulator. Each line represents the number of people infected while performing a particular type of activity (attending primary school, at home, attending college, working, and all others).

**Programmatic Progress**

The work of the NDSSL has resulted in progress in several key program areas. We received a new National Science Foundation grant to develop Multi-Theory Multi-Level (MTML) systems, operating over extremely large dynamic socio-technical networks, in which multiple contagions and behaviors are simultaneously coevolving by repeated interactions, utilizing a computational agent-based model resolved down to the level of the individual. Anil Vullikanti, one of our team members, received a National Science Foundation Faculty Early Career Development Award to examine the theoretical foundations of cross-layer optimization in cognitive radio networks in the physical interference model. We also received a National Science Foundation grant to study efficient market-based frequency allocation mechanisms in wireless radio networks. We have received a basic research and development grant from the Defense Threat Reduction Agency for a project to use HPC-based representation methods for inference and state assessment of complex socio-technical networks. Our group has also been invited to be a part of the Department of Homeland Security’s new Center of Excellence for...
Command, Control and Interoperability. We will be a part of a team led by Purdue University to conduct research and develop technologies, tools, and advanced methods for information analysis. Finally, we received a grant from the Bill & Melinda Gates Foundation to design, specify, and implement the computational infrastructure for the malaria model developed at the Swiss Tropical Institute. Our work will develop informatics tools for supporting large-scale experimental designs and analysis using distributed computing backends for researchers involved in malaria eradication.

We continue to work on the Defense Threat Reduction Agency-funded project 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, Center of Excellence in Public Health Informatics, led by the University of Utah School of Medicine. We continue to contribute as a principal institution in the NIH Modeling Infectious Disease Agent Study (MIDAS) project. The Centers for Disease Control and Prevention project has led to new modeling environments to study how public policy personnel would use simulation-based tools for effective response in the event of an epidemic. In April 2008, we submitted disclosures for eleven new, interrelated inventions to Virginia Tech Intellectual Property. These disclosures formed the basis for an application filed by Virginia Tech Intellectual Property for a single comprehensive provisional patent.

National and International Leadership

In the reporting period, we have also hosted a training workshop for the Defense Threat Reduction Agency and personnel from operational military commands on the DIDACTIC tool developed by our team. We continue to advise the National Institutes of Health Steering committee for developing modeling methods for contagious diseases. We also continue to advise DTRA and the Department of Defense on modeling and policy planning related to interdependent infrastructure analysis and pandemic preparedness. Our recent efforts supporting DTRA and the National Institutes of Health have been in the context of H1N1. 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 and panels including, but not limited to: the National Institutes of Health Blue Ribbon Panel for Risk Assessment of the National Emerging Infectious Disease Laboratory at Boston University Medical Center; the 28th Institute of Electrical and Electronics Engineers (IEEE) Conference on Computer Communications (INFOCOM); the 5th International Workshop on Foundations of Mobile Computing; the 3rd International Conference on Bio-Inspired Models of Networks, Information and Computing Systems; the National Institutes of Health Director’s New Innovator Award Review Panel; National Science Foundation review panels; the editorial board for the Open Environmental Journal; plenary talks at the Society for Industrial and Applied Mathematics (SIAM) annual conference on Computational Science and Engineering; invited talk at the General Electric Whitney Symposium; 18th International symposium on Mathematical Theory of Networks and Systems (MTNS2008); 10th International Conference on Distributed Computing and Networking; 3rd IEEE International Symposium on Pervasive Computing and Ad Hoc Communications.

Outreach and Education

NDSSL had significant achievements in Education and Outreach in the reporting period. This included activities for students ranging from high school all the way to doctoral students. We participated in a number of new programs. This included Research Experiences for Undergraduates (REU) programs, summer internship program instituted by the Department of Computer Science at Virginia Tech for talented undergraduate students from the Indian Institutes of Technology, India, and local high school students.

We have mentored four high school interns over the past year. The interns have been involved in a project that assesses the contact network of high schools in the New River Valley. This work will allow us to obtain very detailed information about disease transmission pathways in these types of settings. Sanjay Kishore won the first prize in a science competition held recently at the Southwest Virginia Governor’s school for his poster titled “Remodeling the Spread of Disease at Radford High School” which was based on work performed at the NDSSL.

We organized a half-day workshop on our models of epidemics and their simulations that was attended by faculty and students in July 2008. We also organized a minisymposium on complex networks as part of the SIAM Computational Science and Engineering (CSE 2009) conference in March 2009.

Members of our group have taught courses for the Department of Mathematics and the Department of Computer Science at
Virginia Tech during the spring and fall of 2008. They have also mentored two teams of undergraduate students from Virginia Tech in the Mathematical Contest in Modeling which is a worldwide competition. One of the teams mentored received the meritorious award in this competition.

We have also mentored and supervised approximately 25 students at the college level, ranging from undergraduate to graduate students pursuing both Masters and Ph.D. degrees. One of our students, Deepti Chafekar, successfully completed her Ph.D. while working with our group and has since joined the Nokia Research Center. She has been selected as a recipient of an outstanding Ph.D. student award by the Department of Computer Science at Virginia Tech. Another student, Matthew Macauley, was jointly advised by members of the NDSSL group and finished his Ph.D. at the University of California, Santa Barbara. The topic of his thesis was discrete dynamical systems and was based on the work he performed at Virginia Tech. The group has served on committees for several Masters and Ph.D. students.

Conferences and workshops


Barrett C (2009) Social Networks and Travel Connection and Mobility in an Information Age, Panel Discussion at the 88th Annual Meeting of the Transportation Research Board, January 11-15.

Barrett C (2009) Spiral R&D program in computational social and network science to support of sustainable societal infrastructures, Presentation to Australian university delegates, March 11.


**Intellectual property**

**Peer-reviewed publications in reporting period**


Abstract. The Nutritional Immunology Group at the Virginia Bioinformatics Institute is leading research programs on infectious disease, gut health and obesity-related inflammatory complications. Our effort in this reporting period comprised the following: the development of novel immunotherapeutics and vaccines that regulate immune responses and mitigate pulmonary inflammation associated with respiratory virus infections such as influenza virus; the sorting and characterization of adipose tissue macrophage subsets involved in the pathogenesis of obesity and diabetes; the creation of a virtual mucosal immune system for studying enteric infectious and autoimmune diseases; and the sequencing and characterization of an Amerindian Helicobacter pylori strain (V225) associated with low cancer risk and decreased obesity. The central integrative theme amongst these active areas of research and discovery is understanding the inflammatory processes that underlie various diseases and developing novel approaches for modulating inflammatory responses. We have continued to develop novel preventive and therapeutic approaches for widespread and debilitating diseases. These efforts have received support from the National Institutes of Health, Cognis Nutrition and Health GmbH, the Virginia Bioinformatics Institute and the European Commission.

Keywords: obesity; diabetes; inflammatory bowel disease; peroxisome proliferator-activated receptor; abscisic acid; inflammation; CD4+ T cells; monocyte chemotactic protein-1; macrophage; mucosal immunity; influenza virus infection.
Scientific Progress

Immune modulatory actions of abscisic acid during influenza virus infection

According to the World Health Organization, respiratory infections are the third leading cause of death in humans worldwide. Among these infections, influenza virus represents a major challenge to the public health system due to the high incidence and morbidity of seasonal influenza. In addition, the 2009 outbreak of swine flu highlights the relevance of cross-species transmission, which has resulted in the emergence of new and more virulent strains for humans and coalescence into a global pandemic. Influenza-related severe disease and death are associated with a "cytokine storm" in the lungs, which leads to excessive pulmonary inflammation and immune cell infiltration. Excessive inflammatory responses have been linked to the death of young people in Mexico during the 2009 H1N1 swine influenza pandemic.

Our group, led by Dr. Raquel Hontecillas-Magarzo, assistant professor at the Virginia Bioinformatics Institute (VBI), in collaboration with Dr. Chris Roberts at the Virginia-Maryland Regional College of Veterinary Medicine and Dr. Bruno Sobral and the VBI Cyberinfrastructure Group (CIG), is studying the ability of abscisic acid (ABA), an isoprenoid phytohormone, to attenuate the excessive inflammatory response that develops at the respiratory tract of mice following infection. We have established that ABA treatment diminishes lung inflammatory lesions induced by a lethal dose (100 LD$_{50}$) of A/Puerto Rico/8/34 (H1N1) influenza virus. We found that arginase-1 (Arg1) and 2,3-indolamine dioxygenase are significantly upregulated in the lung by infection. Both of these genes are involved in amino acid metabolism by immune cells, and by depleting arginine and tryptophan they limit the expansion capacity of inflammatory cells. Arg1 is produced by alternatively activated macrophages, a cell type induced by interleukin-4 (IL-4) and IL-13 that suppresses inflammation. However, in our model of influenza-induced viral pneumonia there is a massive influx of neutrophils into the lungs. Neutrophils, which also express Arg1, are part of the acute innate immune response driven by IL-17, a cytokine produced by Th17 cells. Our microarray data show that both IL-4 and IL-17 pathways are upregulated by infection. Figure 1 shows the types of immune cells infiltrating the lungs of influenza virus-infected mice, the histopathologic lesions, the gene network involved in influenza-associated disease and how it is modulated by ABA. In follow-up experiments, we will identify the cell type specificity involved in the shift in arginase activity and the exact role that it has in the immunopathogenesis of the pulmonary inflammatory disease induced by influenza virus.
In collaboration with Dr. Javier Dominguez at the Instituto Nacional de Investigaciones Agronomicas in Madrid, Spain, we have conducted a study in pigs aimed at characterizing the host response to intranasal vaccination, intramuscular vaccination and intranasal infection with porcine respiratory and reproductive virus in pigs. We found that live infection, but not nasal or systemic vaccination, induced the production of antigen-specific immunoglobulin A in the respiratory mucosa. In contrast, only intramuscular vaccination induced detectable systemic antigen-specific T cell proliferative responses, suggesting that there is a need for developing better nasal vaccines capable of providing mucosal antigen-specific immune protection against respiratory pathogens, including influenza virus.

**Characterization of adipose tissue macrophage subsets connecting obesity, type II diabetes and inflammatory bowel disease**

Over the past twenty years the prevalence of obesity in the United States has risen considerably. According to recent estimates, approximately one-third of Americans are overweight or obese (Ogden et al, 2006). Obesity is strongly associated with chronic diseases such as cardiovascular disease, type II diabetes, stroke, obstructive sleep apnea, and colon cancer. Obesity is projected to be one
of the most costly health conditions for future generations (Wang et al, 2008).

Over the past year our laboratory has been studying the immunological aspects underlying obesity and its association with systemic insulin resistance. It is well established that chronic inflammation is integral to the development of obesity-induced insulin resistance and, more specifically, it has been shown that macrophages infiltrating white adipose tissue are influential in this response. Our group has been studying two specific types of adipose tissue macrophages (ATMs), which differ in protein expression of the macrophage maturation marker and glycoprotein F4/80 (F4/80\textsuperscript{hi} vs F4/80\textsuperscript{lo}) and the C-C motif chemokine receptor 2 (CCR2) – the receptor for monocyte chemoattractant protein-1 (MCP-1). In our previous work, we showed that F4/80\textsuperscript{hi} ATMs decrease in adipose tissue of obese mice treated with abscisic acid, a natural botanical that activates the nuclear receptor peroxisome proliferator-activated receptor γ (PPAR-γ) (Guri et al, 2007). To follow up on this finding, we analyzed these two populations in both lean and obese mice to gain a better understanding of their phenotypic and functional distinctions. We found that F4/80\textsuperscript{hi} ATMs express more PPAR-γ and PPAR-δ than F4/80\textsuperscript{lo}. In contrast, F4/80\textsuperscript{lo} ATMs expressed more IL-4 (Bassaganya-Riera, 2009a). Both PPAR-γ and -δ have been shown to be crucial inflammatory regulators in macrophages, while IL-4 is a marker characteristic of alternative, anti-inflammatory macrophages. More notably, both PPAR-γ and -δ represent important molecular targets for anti-diabetic drugs and novel anti-inflammatory botanicals. We have also demonstrated that intestinal inflammation, in combination with increased F4/80\textsuperscript{hi}CD11b\textsuperscript{hi}CCR2\textsuperscript{hi} macrophage infiltration into the omental adipose tissue during obesity, results in increased severity of experimental inflammatory bowel disease and worsening of diabetes (Bassaganya-Riera, 2009b). In collaboration with Dr. T. A. Lucktong and other bariatric surgeons at the Carilion Medical School, we will examine the changes in macrophage subset accumulation in omental fat of obese individuals prior to and after bariatric surgery. This along with a mouse model of bariatric surgery will allow us to identify putative molecular targets for intervention against obesity and its co-morbidities.

Our group has also investigated novel PPAR-γ agonists as candidates for treatment of cardiovascular disease. Our data demonstrate that ABA treatment elicits an anti-hypertensive effect and decreases immune cell infiltration in the wall of aortas from mice with atherosclerosis (Bassaganya-Riera, 2008, US Provisional Patent
Application No: 61/095662). We also have data demonstrating synergisms between botanicals and pharmaceuticals for enhanced modulation of obesity-related inflammation (Bassaganya-Riera, 2009, US Provisional Patent Application No: 61/147773). Collaborations are in place with Dr. Jesmin at the University of Dhaka and the Diabetes Association of Bangladesh to develop naturally occurring and cost-effective interventions against diabetes and heart disease that are affordable to millions of people in South East Asia. Virginia Tech supported the launch of the spin-off company BioTherapeutics Inc. to commercialize some of these promising discoveries.

Development of a virtual mucosal immune system for studying enteric disease

Inflammatory bowel disease (IBD) is a chronic, recurring, immunoinflammatory illness of unknown cause that afflicts more than 1 million Americans and is characterized by two clinical and histopathologic manifestations – Crohn’s disease and ulcerative colitis. In addition to some identified genetic factors such as mutations in the Card15 gene encoding NOD2 (Podolsky, 1999) and association with rare alleles of the nuclear receptor PPAR-γ (Sugawara et al, 2005), the prevailing notion is that IBD is maintained by a defective down-regulation of mucosal immunoinflammatory responses to gut microbiota (Strober et al, 2002). We sought to develop a novel approach capable of integrating diverse datasets from host response, gene polymorphism and gut microbiota by creating a virtual mucosal immune system.

In collaboration with CIG, we discovered a total of 214 colonic genes that were differentially expressed, depending on expression of PPAR-γ and stage of IBD. These genes clustered in three groups based on their pattern of expression: 1) genes up-regulated by day 7 in colons of mice lacking PPAR-γ in epithelial cells or CD4+ T cells; 2) genes up-regulated on either day 2 or 7 in colons of mice lacking PPAR-γ in CD4+ T cells; and 3) genes up-regulated in colons of mice lacking PPAR-γ in epithelial cells at any given time. Some representative clustering trends are shown (Figure 2). The latter pattern is associated with increased severity of colonic inflammatory lesions.

In collaboration with the Network Dynamics and Simulation Science Laboratory (NDSSL) at VBI we are building an agent-based model representing a virtual mucosal immune system composed of mesenteric lymph nodes (inductive site) and a colon (effector site) with the following architecture: lumen, epithelial cell layer, and lamina propria. The entities included are: 1) commensal (resident) or non-commensal (invading) bacteria, 2) epithelial cells that may or may not secrete pro-inflammatory cytokines, 3) naïve CD4+ T cells, 4) memory CD4+ T cells that may be effector or central memory T-cells, 5) effector CD4+ T cells that may have a Th1,
Figure 2. Differential expression of colon genes and their clustering patterns.

Th2, Th17, or induced T-regulatory (iTreg) phenotype, 6) natural regulatory T cells, 7) dendritic cells that may be immature or mature with a tolerogenic or effector phenotype, and 8) macrophages that may have an M0, M1 or M2 phenotype. The rules encompass conditions under which each cell type can secrete one or a combination of cytokines and be modulated by PPAR-γ ligands. Cells are assigned phenotypes based on interactions with other cells upon occupying the same location and the cytokine levels within the site of interaction. In this manner, the model can simulate an inflammatory or tolerogenic response.

In collaboration with Dr. Abdul Jarrah at VBI and Alan Veliz-Cuba in the Department of Mathematics at Virginia Tech, we have also developed a mechanistic mathematical logical model representing the CD4+ T cell...
differentiation process. Previous logical models focused only on the Th1/Th2 paradigm. Herein, we develop a model that more faithfully represents the heterogeneity of CD4+ T cell subsets, including Treg and Th17. Computationally, the logical network could be embedded within each cell of the agent-based model through a truth table. We found that Th1 and Th2 are the only non-naive cells that exist after withdrawing the cytokine stimuli and PPAR-γ favors T cell differentiation towards Th2 or Treg. Moreover, a novel finding from the model is that PPAR-γ activation can induce Th2 differentiation in the absence of IL-4.

On the translational side, in collaboration with Dr. Richard Bloomfeld at Wake Forest University, Dr. Kim Isaacs at University of North Carolina at Chapel Hill, and Dr. Ulf Helwig of the Federal Research Center of Nutrition and Food in Kiel, Germany, we are testing the efficacy of a novel PPAR-γ agonist and immune modulator in patients with Crohn’s disease. We have also established clinical collaborations with Drs. Kinshore, Hart and Rubio at the Carilion Medical School to obtain access to human clinical specimens.

Characterization of a Helicobacter pylori Amerindian strain

*Helicobacter pylori* persistently colonizes the stomach of about half of the human population and it usually does not cause adverse effects, but it is associated with an increased risk of noncardia gastric adenocarcinoma, gastric lymphoma, and peptic ulcer. Its prevalence ranges from 80 to 90% in developing countries and below 40% in developed countries. However, the risk of gastric cancer is not proportional to *H. pylori* prevalence. For instance, rural areas in Africa and the Amazonas are regions with very high prevalence of *H. pylori* and very low gastric cancer. Also, *H. pylori*-colonized individuals exhibit a high degree of leanness that cannot be explained based on nutritional status and physical activity. Furthermore, *H. pylori* eradication programs are associated with increased obesity rates. In collaboration with Dr. Bruno Sobral and CIG, Dr. Maria Gloria Dominguez-Bello at the University of Puerto Rico, and Dr. Martin Blaser at the New York University School of Medicine, we have sequenced the whole genome of strain V225, a *CagA*-positive *vacAs1c*, EPIYA-ABCC Amerindian strain cultured from a Venezuelan Piaroa Amerindian with acute superficial gastritis. To gain insight into the evolution and host adaptation of this bacterium, we undertook comparative genomic analyses of geographically distant *H. pylori* strains. Several European strains (HPAG1, G27, P12, 26695), an African strain (J99), and an Asian strain (Shi470) have been sequenced. Hence, dissecting the genetic differences among geographically distant strains may shed new light on the *H. pylori* paradox (e.g., why some strains cause disease whereas others may prevent it). We uncovered duplication of the *cagA* gene
with an insertion in the N-terminal end of CagA, two changes in EPIYA B motif, 8 and 9 base pair deletions in EPIYA C1 and C2 motifs. These deletions wiped out CRPIA domains that are involved in phosphorylation-independent interaction of CagA with host proteins, all of which may alter the function of V225 CagA.

We have previously demonstrated that colonization of stomachs from obese mice, with the 99-305 (mutant, HpcagA−) strain of *H. pylori*, which is a knockout of the whole *cag* pathogenicity island (PAI) made by transformation of strain 98-325 with a PCR product from a strain in which the chloramphenicol resistance cassette replaced the entire *cag*PAI, is associated with suppressed infiltration of macrophages, enhanced influx of regulatory T cells in white adipose tissue, and improved glucose tolerance when compared with the isogenic 98-325 (wild-type, mouse-passed European 26695 HpcagA+). Gastric colonization with the HpcagA+ strain altered the expression of genes in an ‘insulin signaling pathway’ in the gastric mucosa of obese mice. From that pathway, the expression of PGC-1α, a nuclear coactivator of transcription, was upregulated after colonization with the HpcagA+ strain, but was dramatically downregulated following colonization with the HpcagA− strain. PGC-1α is involved in the regulation of several pathways involved in energy metabolism, including the maintenance of plasma glucose levels (Liang and Ward, 2006). Liver PGC-1α levels are increased in models of diabetes and potentially contribute to hyperglycemia (Yoon et al, 2001; Puigserver et al, 2003). We also found that expression of genes in the Hedgehog signaling pathway is altered following colonization with *H. pylori*. Sonic hedgehog (Shh) normally regulates epithelial cellular proliferation (El-Zaatari et al, 2009). Shh is expressed at high levels in gastric carcinoma cell lines and is necessary for tumor growth in other tissues (El-Zaatari et al, 2009). It has also been reported that the hedgehog signaling pathway may be regulated by the ERK/MAPK pathway (Seto et al, 2009). This provides a potential link between *H. pylori* infection as CagA/SHP-2 signaling activates the ERK/MAPK pathway.

Recent evidence (e.g., benefits in obesity, diabetes and metabolic disorders) questions the paradigm in which *H. pylori* is viewed as a pathogen. Interestingly, disorders such as esophageal diseases and childhood-onset asthma or certain infectious diseases have been reported to occur more frequently in uncolonized individuals. Thus, better understanding the relation between *H. pylori* and humans may result in novel therapeutic applications against both chronic and infectious diseases.
**Conferences and workshops**


**Intellectual property**

Peer-reviewed publications in reporting period


**Other cited references**


Phylogenomics and Systems Biology of Plants and Plant Pathogens

Abstract. Genomics is playing a greater role in understanding processes important for human health, ecology, and economy. Agricultural pathogens, the commensal gut microbiome, and biofuels are societally important topics that our group has touched on in recent work. Our crop pathogen microarray, PhytPath, is in testing. Early experiments with known infections have provided probe signal patterns that implicate the correct pathogen. The Virginia Cooperative Extension Plant Disease Clinic at Virginia Tech is providing a large number of naturally infected plants for extensive testing of the microarray. Multiple phylogenomic projects involved us with the evolution of plant and animal pathogens. The large analysis of the Gammaproteobacteria is complete and the manuscript is in preparation. Collaborations with other teams provided opportunities to perform phylogenomics analyses of Agrobacterium, Coxiella, and Liberibacter. Eric Nordberg is building a system to automate large-scale phylogenomic analyses as part of his PhD dissertation research. Microbial ecology has a direct impact on human health through the environmental bacteria we contact and by interaction with individual human microbiome. Our analysis of a large pyrosequencing data set on small subunit (SSU) ribosomal DNA amplicons from aerial, plant-surface, and termite gut bacteria continues. The group is starting to enter the realm of microbial ecology of the mammalian gut flora. A grant proposal was developed to study the functional genomics of bacteria during food digestion in a pig model of obesity. Other proposals were developed to study evolution and dispersal of antibiotic resistance in bacteria of pigs on farms. The ecological and functional genomics of microbes present many opportunities for research and will form an important component of future work in our group.

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

PhytPath phytopathogen microarray

In previous years we designed and had synthesized the PhytPath Affymetrix microarray for identification of plant pathogens. The chip has 31,202 25-mer probes with perfect sequence match to ribosomal RNA sequences in GenBank. These perfect hits correspond to 76,362 distinct GenBank taxa, for an average of 2.4 taxa per probe. Given that significant hybridization may occur with a single base mismatch, we can consider over 110,000 distinct GenBank taxa that are represented with 24 or 25 matches. We have developed software for interpreting the signals from these probes, based on the GenBank taxonomy. A table of all sequence matches between the 25-mer probes on PhytPath and any sequence in the GenBank nucleotide BLAST data set is used to interpret signal strength patterns for each chip. This table of BLAST hits can be updated to keep current as the public data grows. To date, we have tested the PhytPath chip on laboratory-infected plants (Arabidopsis and cabbage) infected with Pseudomonas (two strains), Xanthomonas carotivora, Sclerotinia sclerotiorum, and Botrytis cinerea. We have recently initiated a collaboration with the Virginia Cooperative Extension Plant Disease Clinic at Virginia Tech to obtain samples of naturally infected plants to further test the chip on many more plant species and pathogen types.

Phylogenomics

Genome-scale phylogenetic analysis, or phylogenomics, provides for very accurate reconstructions of the vertical component of inheritance from common ancestors, which is used to predict shared biology and analyze novel adaptations. It can also identify genomic regions that were acquired by horizontal transfer. We participated in several large phylogeny projects with diverse collaborators. We are analyzing the phylogeny of all completely sequenced Gammaproteobacteria. After eliminating redundancy in the form of very closely related genomes and searching extensively for single-copy genes, we have settled on a dataset of 107 genomes, 356 protein families, and 91,903 characters. Among our findings are that three of the current taxonomic orders are polyphyletic; one clade in our tree both mixes members of these three orders and separates them from other members of each order. Collaboration with João Setubal of the Virginia Bioinformatics Institute and a team sequencing two new Rhizobium genomes provided an opportunity to gather and analyze a large data set of single-copy genes from the Agrobacterium/Rhizobium lineage and related outgroups spanning over 100,000 aligned amino acid positions. This data set was restricted to genes on the main chromosome to provide the stable vertical inheritance background used to infer the very
active evolution of plasmids and the origin of a second true chromosome (Slater et al., 2009).

For the genus *Brucella*, our group helped the Virginia Bioinformatics Institute Cyberinfrastructure Group assemble a database of 2,246 protein families, 671,030 characters. The resulting tree provided strong support that modern *Brucella* strains have emerged from an explosive radiation into four main clades.

A United States Department of Agriculture scientist, Duan Yong-Ping, requested our help to place *Liberibacter*, a new genome he had sequenced, within the phylogenetic tree of the alphaproteobacteria. We found that *Liberibacter* is by far the most divergent genome known among Rhizobiales, yet can be placed as an early-branching member of the Rhizobiaceae. The *Liberibacter* manuscript has been accepted and is in press at *Molecular Plant-Microbe Interactions*.

Finally, Dr. Kelly Williams participated in a phylogenetic analysis of the genus *Coxiella* using a data set of 1,402 protein families with 425,592 aligned amino acid positions. The tree showed that the *Coxiella* strain previously thought to be ancestral was not ancestral, but was primitive in the sense of having the shortest distance to the root.

In the 2008-2009 period, the group began to pursue our interests in bacterial evolution, horizontal transfer and functional genomics in a natural and medically relevant microbial ecosystem: the mammalian gut. We started this effort with metagenomics investigations into the pig gut microbiome. This experimental system will provide opportunities to study phage-

**Figure. Consensus tree from Bayesian phylogenetic analysis based on 94 proteins.** Branches for the 46 strains most distant to *L. asiaticus* have been removed for clarity. All nodes had 100% support.
mediated horizontal transfer, host-microbiome functional interactions, and evolution of antibiotic resistance and pathogenicity.

**Peer-reviewed publications in reporting period**


Ina Hoeschele

inah@vbi.vt.edu
Professor, Virginia Bioinformatics Institute
Professor, Department of Statistics, Virginia Tech
Adjunct Professor, Department of Cancer Biology, Member of
Biostatistics Core, Comprehensive Cancer Center, Wake Forest
University School of Medicine, Winston-Salem, NC
Adjunct Professor, Virginia Tech Carilion School of Medicine

Group contributors: Alberto de la Fuente, Guimin Gao, Inyoung Kim.

Genetic Architecture of Complex Diseases

Abstract. The common theme of our research is the statistical design and analysis of genome-wide linkage and association studies and systems genetics experiments to further our understanding of how the joint action and interaction of multiple genes determines complex diseases and biomedical phenotypes of humans, animals and plants. At the present time, our main research focus is on developing, implementing and evaluating methods for Quantitative Trait Locus mapping of very high-dimensional ‘omics phenotypes, gene regulatory and causal network inference in systems genetics experiments, efficient linkage and association mapping in the presence of interactions among genes and among genes and environmental factors, and identification of pathways from gene expression data affecting biomedical traits of interest.

Keywords: systems genetics; statistical genetics; gene network inference; design and analysis of microarray experiments; linkage analysis; association analysis; statistical haplotyping; complex pedigrees.
Scientific Progress

High-dimensional quantitative trait loci analysis of ’omics phenotypes

Traditional analyses considering a single Quantitative Trait Locus (QTL)/marker and phenotype lack sufficient power and are hence able to detect only a few of the strongest signals. We are investigating multivariate, dimension-reduction based methods such as biclustering algorithms, Bayesian and penalized sparse canonical correlation analyses, and sparse partial least-squares approaches via simulation of systems genetics data and real data analyses to detect the weaker signals of a subset of QTL jointly regulating a subset of phenotypes. This project is a collaboration with Dr. Alberto de la Fuente, CRS4 Bioinformatica, Italy, and Dr. Inyoung Kim, Department of Statistics, Virginia Tech.

Nonparametric Bayesian method for association and linkage analyses and identification of pathways

Association analyses of human populations are still often performed using single marker analysis, which can only detect a few individual, large signals, but not the joint action of multiple Quantitative Trait Locus/Single Nucleotide Polymorphism (QTL/SNP) with weaker signals and interactions. For linkage analysis, we propose to jointly analyze all candidate QTL, while for association analysis we propose to jointly analyze a subset of all SNP that are associated with a set of candidate genes and/or the genes and regulatory regions belonging to a set of candidate pathways. The joint analysis of multiple QTL/SNP is based on a model containing an arbitrary function of all QTL/SNP, as well as environmental factors/covariates. This function is given a Gaussian process prior and is then able to capture any multi-way interaction and non-linear interaction among QTL/SNP and with environmental factors. Instead of having to perform a very difficult multi-dimensional search for interactions and requiring many parameters, this approach identifies QTL/SNP irrespective of whether they act through main or interaction effects and performs model selection by requiring only a single parameter per QTL/SNP (Huang et al, 2009; Zou et al, 2009). The same type of approach can be used to identify which pathways affect a continuous biomedical trait or a clinical outcome of interest and can discover any interaction between the genes in a single or in multiple candidate pathways. This project is a collaboration with Dr. Fei Zou, Department of Biostatistics, The University of North Carolina at Chapel Hill.
Gene regulatory/causal network inference in systems genetics

We have written a book chapter on inferring gene regulatory networks from genetical genomics data (Liu et al, 2009), which describes current methods including our own approach based on global Structural Equation Modeling. We have reconstructed a gene regulatory network using the complete set of yeast genes and yeast genetic markers using a sequence-based method for expression QTL mapping previously developed by our group and local structural models (Mancosu et al, 2009). This project is a collaboration with Dr. Alberto de la Fuente and Dr. Bing Liu, The Monsanto Company.

Analysis of genome-wide expression and DNA methylation data using Illumina platforms

In preparation for a genome-wide investigation of expression and DNA methylation in atherosclerosis, we have developed an analysis pipeline for the preprocessing, including normalization and summarization, and quality control of gene expression data from the Illumina HumanHT-12 BeadChip and of DNA methylation data from the Illumina Infinium HumanMethylation27 platform. We use both Illumina’s proprietary BeadStudio software (www.illumina.com) and the open-source Bioconductor software (www.bioconductor.org), which runs in the statistical analysis environment R (www.r-project.org). The use of Bioconductor for data preprocessing and statistical analysis gives us substantial

Figure. Gene regulatory network 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).
flexibility and allows us to use and adapt the many tools available in Bioconductor that were developed for other platforms (e.g. Affymetrix GeneChip®). This project is a collaboration with Dr. Yongmei Liu, Wake Forest University School of Medicine.

**Haplotyping in complex human pedigrees**

Statistical haplotyping in pedigrees is an essential component of QTL linkage analysis and other genetic studies of complex human pedigrees. We have written a review paper on statistical haplotyping in pedigrees (Gao et al, 2008), and we have continued to work towards incorporating missing marker and marker linkage disequilibrium data into our efficient deterministic haplotyping method for large pedigrees and large numbers of markers. This project is a collaboration with Dr. Guimin Gao, University of Alabama at Birmingham.

**Conferences and workshops**


**Peer-reviewed publications in reporting period**


**Other cited references**


**Group contributors:** Abdul Jarrah, Alan Veliz-Cuba, Valerie Hower, Franziska Hinkelmann, David Murrugarra.

## Computational Systems Biology

**Abstract.** The current focus of the Applied Discrete Mathematics Group at the Virginia Bioinformatics Institute encompasses theoretical and applied projects. The aim of the theoretical work is to develop new mathematical methods in computational systems biology, primarily using techniques from discrete mathematics. The main application focus is the systems biology of breast cancer, particularly to develop an understanding of malignant changes in the metabolic profile of cells and changes in their iron metabolism.

**Keywords:** systems biology; network modeling; breast cancer.
Scientific Progress

*Modeling and simulation of gene regulatory networks*

During the last decade, finite dynamical systems, that is, discrete dynamical systems with a finite phase space, have been used increasingly in systems biology to model a variety of biochemical networks, such as metabolic, gene regulatory and signal transduction networks. In many cases, the available data quantity and quality are not sufficient to build detailed quantitative models of systems, such as ordinary differential equations, which require many parameters that are frequently unknown. In addition, discrete models tend to be more intuitive and more easily accessible to life scientists. Boolean networks and their generalization, the so-called multistate logical models, are the main types of finite dynamical systems that have been used successfully in modeling biological networks.

Boolean models require less detailed information about the system to be modeled, so they can be used in cases where quantitative information is missing. They are also useful if qualitative predictions from the model are desired, such as whether a T cell becomes pro- or anti-inflammatory. Finally, Boolean models are very intuitive compared to models based on differential equations or other more sophisticated formalisms and it is easier to explore their dynamics, at least for reasonably small models. On the other hand, an important disadvantage of Boolean models, and algebraic models in general, is that there are very few theoretical tools available for their construction. Typically, Boolean models are built by translating information from the literature into logical statements about the interactions of the different molecular species involved in the network. In many cases, however, the biological information about a particular network node might not be sufficient to construct a logical function governing regulation. In the case of a continuous model, the remedy would be to insert a differential equation of specified form (e.g., mass action kinetics) with unspecified parameters. If experimental time course data are available, several parameter estimation methods can then be used to determine those unspecified model parameters so that the model fits the given data. Data fit is determined by model simulation, using numerical integration of the equations in the model. The software package described in this paper addresses the need for a discrete analog of this process.

We have developed a software package, *Polynome* (http://polymath.vbi.vt.edu/polynome), which carries out discrete analogs of parameter estimation for Boolean network. The software integrates several packages for network inference that we have developed in recent years. When information is missing about a particular node
couples parameter estimation with extensive simulation capabilities. For instance, it is increasingly likely that Boolean models using sequential update of the variables are more realistic than parallel update systems and, moreover, it has been shown that stochastic models are sometimes more appropriate than deterministic ones. The package described here has the capability of simulating models deterministically as well as stochastically. The stochastic features can arise either through a random update schedule choice or random choice of functions at each update.

Figure. Feedback regulatory loops in iron metabolism. Species depicted are Monensin sensitivity protein homolog 1 (mon1a); ferroportin (Fpn); hepcidin (Hepc); Janus kinase 2 (JAK2); heme oxygenase 1 (HO1); reactive oxygen species (ROS); aminolevulinate synthase 1 (ALAS1); transferrin receptor 1 (TfR1); ferritin H and ferritin L (Ft); divalent metal transporter 1 (DMT1); iron regulatory proteins 1 and 2 (IRPs). Solid lines indicate positive or negative regulation; dotted lines indicate reactions that consume or produce the indicated species.
Cancer systems biology

There is increasing evidence that differences in iron metabolism play an important role in cancer risk, survival, and clinical prognosis. Indeed, a recent editorial in *Lancet* argues that differences in iron metabolism may be key to both development and recurrence of breast cancer in women. For example, increases in transferrin receptor 1, a protein responsible for iron import, has been linked to a greater likelihood of developing breast and other cancers since the 1980s and the transferrin receptor is frequently used as a targeting ligand in the design of anti-cancer drugs. In a cell culture system, repression of ferritin, an iron storage protein, along with the activation of IRP2, an iron regulatory protein, were required for the transforming and proliferative effects of the c-*myc* oncogene. However, why these differences in iron metabolism exist, as well as how they contribute to malignant processes and whether they can be successfully exploited to therapeutic advantage, remains unknown. This uncertainty is because the current approach to identifying whether these or other proteins in the complex network of iron metabolism should be targeted therapeutically is an empiric, protein-by-protein validation exercise. The intent of this project is to use the power of systems biology to identify the key regulatory points in this complex pathway and determine how they change in malignancy.

In work published in February of 2009, our group described an iron network in non-cancer cells (Hower et al, 2009). We distilled this complex network into a system of feedback control loops that represent important points of potential control of the network (see Figure). The goal of this project is to build on this network to create predictive models of iron metabolism in normal and cancer cells, which will provide more information about key nodal points in iron metabolism and how they become different during malignant change. Ultimately, we hope to be able to leverage knowledge gained to therapeutic or diagnostic advantage.

Our work on breast cancer metabolomics is currently supported by two grants. As part of the work on an RO1 grant from the National Cancer Institute awarded to co-investigators Drs. Vladimir Shulaev and Laubenbacher from VBI and Drs. Steven Akman and Suzy Torti from the Wake Forest University School of Medicine (and Dr. Abdul Jarrah from VBI as senior personnel), we have developed new statistical techniques to identify metabolic biomarkers from gas chromatography-mass spectrometry (GC-MS) data derived from cell culture samples of normal and transformed epithelial breast cells. We have also developed and validated methods for metabolite extraction and GC-MS, as well as liquid chromatography (LC)-MS metabolite profiling of human plasma samples using Ultra High Pressure Liquid Chromatography (UPLC). We are now applying...
the extraction techniques and novel statistical algorithms to plasma samples obtained from patients who present for a mammogram to the Wake Forest University Comprehensive Cancer Center, which is part of a study funded by a Team Science Award from the Wake Forest University Translational Sciences Institute.

Conferences and workshops


Peer-reviewed publications in reporting period


Abstract. The genus Alternaria contains many economically important fungal species from plant and human health perspectives. Two model systems continue to be developed in our laboratory with the aims of 1) identification and characterization of fungal virulence factors influencing plant pathogenesis and 2) characterization of fungal proteins and products associated with innate immunity and inflammation in mammals. In this regard, we are using the *Alternaria brassicicola*-Brassicaceae interaction as a system to study fungal pathogenesis of cultivated Brassicas and the model flowering plant Arabidopsis. For our studies involving human cells and murine models of allergic airway inflammation, we are using *Alternaria alternata*, the airborne, ubiquitous, species clinically associated with human airway disorders including allergy, asthma, and chronic rhinosinusitis. For both of these research areas, we have 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. Using these molecular approaches coupled with bioinformatic analyses of the recently completed and annotated *A. brassicicola*, *A. tenuissima*, and *A. alternata* genomes, we have identified several virulence factors of plants, including signal transduction components (kinases and transcription factors), secondary metabolite biosynthetic genes, and other completely novel ones. Moreover, these genome sequences have already proven valuable for the identification of allergens and other fungal proteins that cause inflammatory responses in mammals.

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

Alternaria comparative genomics

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%20brassicicola).

The assembly is composed of 838 supercontigs averaging 36,147 bp in length and having an N50 value of 2,400,717 bp (the length such that 50% of all nucleotides contained in supercontigs are of at least this size). The supercontigs are composed of 4,039 contigs with 84% of the sequences 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 supercontigs, which suggests a relatively robust assembly for only 6x coverage. Bioinformatic analyses of the genome were carried out at the Virginia Bioinformatics Institute (VBI). 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 contig sequences as input, or 9,814 using the supercontigs. 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 has been developed (www.alternaria.org).

In addition to the *A. brassicicola* genome project, we have also completed sequencing two additional Alternaria genomes (*A. alternata* and *A. tenuissima*) this year, using the Roche 454 GS-FLX™ Titanium platform in the VBI Core Laboratory Facility. In contrast to the plant pathogen *A. brassicicola*, the ubiquitous, saprophytic *Alternaria* spp., especially *A. alternata*, have been clinically linked for almost a century to human respiratory disorders. Both of these genomes were sequenced to approximate 20x coverage. The *A. alternata* genome assembly consists of 325 large contigs making up ~33.2 Mb. The average contig size was 102 kb, N50 contiguous size 304 kb, and the largest contig size of 1,046,214 bp. The *A. tenuissima* genome assembly consists of 337 large contigs making up ~33.6 Mb. The average contig size was 99.6 kb, N50 contig size 451 kb, and the largest contig was 1,735,711 bp. Both of these assemblies were machine annotated with FgeneSH *ab initio* gene predictions (Alternaria training matrix) and resulting sets of predicted genes and proteins subjected to BLAST analyses against GenBank NR and Uniprot databases, Interpro analysis, Signal P 3.0, and TMHMM. The annotation data are currently being integrated into the Alternaria Genome Project website at VBI.

The data have already been useful to identify the unique sets of genes corresponding to each species. For example, approximately 800 genes appear to be specific to the plant pathogen *A. brassicicola* and are primarily associated with secondary metabolite biosynthesis and secretion. We anticipate that these new genomes will be critical for studying the pathophysiology of species like *A. alternata* that are clinically relevant to human airway health. For example, we will be able to use the predicted sets of genes to identify pathologically relevant proteins such as novel allergens, immunogenic factors, and possibly small molecules that trigger or exacerbate allergic airway inflammation.
Identification of A. brassicicola pathogenicity factors

One area of interest regarding A. brassicicola pathogenicity lies in the area of secondary metabolite biosynthesis. 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 identified every gene in the A. brassicicola genome predicted to encode a polyketide synthase (PKS) or nonribosomal peptide synthase (NPS). 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. In addition, this year we successfully identified the PKS gene and biosynthetic cluster responsible for the production of the histone deacetylase inhibitor, depudecin (Wight et al, 2009). In this study we found that depudecin is a minor virulence factor of Brassicas. Depudecin is of biomedical interest because of its anticancer and antiparasitic activities. We have also identified several other PKS and NPS genes that play roles in virulence in A. brassicicola (Kim and Lawrence, unpublished results). 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). 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, 2009). 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, ∆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 (2009) and 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 ∆abste12 and ∆abnik1 during plant inoculations resulted in a complete restoration of pathogenicity. These results strongly suggest, once again, 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 knocked out 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 ∆abste12 mutant exhibited a hyper anastomosis phenotype, but was non-pathogenic on plants.

The regulation of intracellular levels of reactive oxygen species (ROS) is critical for developmental differentiation and virulence of many pathogenic fungi. In a recent study, we have demonstrated that a novel peroxisomal transmembrane protein, TmpL, is necessary for regulation of intracellular ROS levels, tolerance to external ROS, and required for infection of plants by the necrotroph A. brassicicola and mammals by the opportunistic pathogen Aspergillus fumigatus. In both fungi, TmpL encodes a predicted hybrid membrane protein containing an AMP-binding domain, six putative transmembrane domains, and an experimentally-validated FAD/NAD(P)-binding domain. Localization and gene expression analyses in A. brassicicola indicated that TmpL is associated with the Woronin body, a specialized peroxisome, and strongly expressed during conidiation and initial invasive growth in planta. A. brassicicola and A. fumigatus ∆TmpL-replacement knockout mutants exhibited abnormal conidiogenesis, accelerated aging, an enhanced oxidative burst
during conidiation and hypersensitivity to oxidative stress when compared to wild-type or reconstituted strains. Moreover, *A. brassicicola* ΔTmpL mutants, although capable of initial penetration, exhibited dramatically reduced invasive growth on Brassicas and Arabidopsis. Similarly, *A. fumigatus* ΔTmpL mutants were dramatically less virulent than the wild-type and reconstituted strains in two distinct murine models of invasive Aspergillosis. Inoculation of NADPH oxidase knockout mice with *A. fumigatus* ΔTmpL did not result in susceptibility to the mutant, suggesting host-derived or extracellular ROS was not responsible for the reduction in virulence.

Constitutive expression of the Yap1 ortholog, the yeast transcription factor that positively regulates expression of antioxidant genes under stress conditions, resulted in high levels of the expression of genes associated with oxidative stress tolerance in *A. brassicicola* ΔTmpL mutants. Yap1 overexpression in the ΔTmpL background complemented the majority of observed developmental phenotypic changes and partially restored virulence on plants. Yap1-GFP fusion mutants, utilizing the native Yap1 promoter, exhibited constitutive nuclear localization in the *A. brassicicola* ΔTmpL background. Collectively, these results strongly suggest oxidative stress homeostasis, but not hypersensitivity to extracellular ROS, as the underpinning of the developmental and pathogenic defects observed in these studies.

In addition to studies centered upon TmpL, we also assisted the Cramer Lab at Montana State University in analysis of the SrbA transcription factor in *A. fumigatus* (Wilger et al, 2008). In this study, we found that SrbA was required for adaptation to hypoxia (low oxygen environment), drug resistance, oxidative stress, and virulence in a murine model of invasive Aspergillosis.

In summary, considerable progress has been made over the last several years regarding identification of *A. brassicicola* virulence genes. More than a hundred and fifty 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 plant pathogenic fungi. We have also expanded our efforts to begin the examination of orthologous genes in *A. fumigatus* to determine their role in animal pathogenesis.

**Alternaria pathogenomics and human airway disorders**

In medical mycology, Alternaria species are gaining importance as emerging human invasive pathogens, particularly in immuno-compromised 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,
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; Plaza et al., 2003; 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; Thaker et al., 1995). 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. 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 lab also suggest that this protein has phytotoxic properties and is a putative virulence factor (Lawrence, unpublished results).

In several recent studies related to chronic rhinosinusitis (CRS), antigen extracts derived from A. alternata appear to have more potent immunostimulatory activity towards CRS 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 CRS 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.
We have recently discovered, in collaboration with the Kita Laboratory, that the integrin protein CD11b is critical for human eosinophils to physically sense the glucan component in Alternaria cell walls (Yoon et al, 2008). We have also shown that Alternaria antigens possess very potent immunological activities by directing the maturation of naïve murine T cells to become Th2 type (Th2) cells that are commonly associated with allergic airway inflammation (Kobayashi et al, 2009). Lastly, we have demonstrated that Alternaria proteases are very immunogenic and induce production of the Th2 thymic stromal lymphopoietin by airway epithelial cells through a G-protein coupled receptor (PAR-2) (Kouzaki et al, 2009). Thymic stromal lymphopoietin is enhanced under asthma-like conditions in mice conditioning antigen-presenting cells in order to orientate the differentiation of T cells coming into the lungs towards a Th2 profile. We have also shown that Alternaria proteases can cause the degranulation of human eosinophils and release of eosinophil-derived neurotoxin also via interaction with PAR-2 (Matsuwaki et al, 2009). Currently we are in the process of identifying the proteins responsible for such potent immunological activities.

**Acknowledgements**

Many of the studies described were supported by the Virginia Bioinformatics Institute, Mayo Research Foundation, 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 (2008) Alternaria functional genomics, invited speaker for *Dothideomycetes Comparative Genomics Jamboree*, DOE-Joint Genome Institute, Walnut Creek, CA, November 11.


Lawrence CB (2009) Alternaria genomics: Applications in plant and human health research, keynote speaker for *Molecular Plant-Microbe Interactions, Plant Molecular Biology and Biotechnology Symposium*, Ohio State University, Columbus, OH, April 24, 2009.
Intellectual property


Peer-reviewed publications in reporting period


**Other cited references**


Iuliana M. Lazar
lazar@vbi.vt.edu
Associate Professor, Virginia Bioinformatics Institute
Associate Professor, Department of Biological Sciences, Virginia Tech
Adjunct Associate Professor, Department of Biomedical Engineering, Virginia Tech

Group contributors: Xu Yang, Milagros Perez, Jenny Armenta.

Fast Proteomic Fingerprinting in Cancerous Cells

Abstract. To demonstrate the power of mass spectrometry detection for large-scale biomarker screening in cancer research, three proteomic technologies for fast biomarker fingerprinting in complex cellular extracts have been developed. Experiments were conducted using MCF-7 breast cancer cells as a model system, and a micro liquid chromatography system interfaced to a linear trap quadrupole mass spectrometer. Fast proteomic profiling of whole cellular extracts was performed by one of the following techniques: (a) data-dependent liquid chromatography (LC)-tandem mass spectrometry of unlabeled cell extracts, (b) data-dependent liquid chromatography-tandem mass spectrometry with pulsed Q dissociation detection of isotope tags for relative and absolute quantitation (iTRAQ)-labeled samples, and (c) multiple reaction monitoring-mass spectrometry of low abundant proteins that could not be detected with data-dependent tandem mass spectrometry. The data-dependent liquid chromatography-tandem mass spectrometry analysis of unlabeled and iTRAQ-labeled MCF-7 cells enabled the identification of 796 and 389 proteins ($P<0.001$), respectively. For multiple reaction monitoring/mass spectrometry screening studies we developed a library of 9,677 peptides ($P<0.001$) representing ~1,572 proteins from human breast cancer cells. For each protein, the library provides the number and sequence of detectable peptides, charge state, spectral count, molecular weight, parameters that characterize the quality of the tandem mass spectrum, peptide retention time, and top 10 most intense product ions that correspond to a given peptide. Preliminary work demonstrates that mass spectrometry-based technologies have the potential to revolutionize biomarker research by enabling fast, sensitive and reliable detection of large panels of biomarkers.

Keywords: cancer; mass spectrometry; biomarkers.
Cancer is a life-threatening disease affecting humans worldwide. It is estimated that every year ~500,000 people will die from cancer in the United States alone. With early detection, the mortality rate is expected to be significantly reduced. Biomarker discovery/screening is an expanding field in proteomics aimed at the identification of proteins that change expression levels or post-translational modifications in response to the onset of a disease. The development of novel methods that would enable sensitive screening for large panels of biomarkers, instead of just one, represents one of the greatest needs in the field. Recent improvements in the performance of mass spectrometry in terms of sensitivity, speed of analysis and capability to unambiguously identify protein components in a sample have rapidly increased its applicability for biomarker discovery.

To demonstrate the power of liquid chromatography tandem mass spectrometry (LC-MS/MS) for rapid proteomic profiling and biomarker screening applications, three separate one-step data acquisition/processing strategies were investigated and perfected for the analysis of MCF-7 breast cancer cellular extracts (Lazar et al, unpublished data). These strategies included data-dependent LC-MS/MS analysis of unlabeled extracts for biomarker screening applications, data-dependent LC-pulse Q dissociation (PQD)-MS/MS of isotope tags for relative and absolute quantitation (iTRAQ) labeled samples for differential expression profiling, and multiple reaction monitoring (MRM)-MS for targeted analysis of low abundant proteins. Data-dependent LC-MS/MS analysis techniques aim at generating a comprehensive proteome profile of complex cellular extracts, relying on the selection of the most intense ions that co-elute during LC-MS analysis for MS fragmentation and structure elucidation. By properly selecting the timing of various MS events to coincide with the LC elution time-window of peptides, sequentially lower intensity ions are selected for fragmentation. Eventually, this sample analysis strategy enables the detection of a very large number of peptides in complex samples, making it very useful for large-scale screening applications. The method is applicable to both qualitative and quantitative investigations. PQD is a novel MS detection approach that was recently implemented on linear ion trap instruments for precursor ion activation and dissociation. The technique allows for the trapping of low m/z ions, and is particularly useful for the analysis of peptides that are labeled with isobaric tags that, upon MS/MS fragmentation, break down from the parent peptide and produce low m/z signature ions for quantitative evaluations.
The one-step data-dependent LC-MS/MS approach involving the analysis of unlabeled and iTRAQ-labeled MCF-7 cellular extracts resulted in the identification of 796 and 389 proteins, respectively. To increase the total number of identified proteins, five or six LC-MS/MS replicates of each sample were performed. The data were filtered by selecting only proteins that had $P<0.001$ and were matched by peptides that passed the Xcorr vs. charge state filter set at 1.9, 2.2 and 3.8, respectively. The rate of false positive identifications was preserved at values <5%, as is commonly required for the evaluation of large sets of proteomic data. For the purpose of these experiments, the estrogen positive MCF-7 cells were cultured in the presence of 17 $\beta$-estradiol (E2) at concentrations commensurate with the physiological levels of this hormone, i.e., 1 nM. In addition, for testing the technology for iTRAQ labeled samples, the cells were cultured in the presence of E2 (10 pM) and tamoxifen (Tam, 1 $\mu$M), a non-steroidal drug prescribed in breast cancer therapy.

A literature search for recently reported putative cancer biomarkers resulted in the generation of a list of $\sim$1100 proteins. By using our one-step data dependent LC-MS/MS analysis approach, a total of 156 markers (from the list of $\sim$1100) were identified in the MCF-7 extracts. Most of these biomarkers have been correlated with multiple types of cancer, 54 being correlated with breast cancer. The data-dependent experiment that involved the analysis of iTRAQ-labeled MCF-7 cells enabled the detection of 389 proteins ($P<0.001$) and 16 up/down regulated proteins with >2 fold change in expression level. In addition, the experiment enabled the detection of 64 proteins that, regardless of change in expression levels during the E2/Tam treatment, were found to be up- or down-regulated in other biological experiments that targeted the discovery of cancer biomarkers. These results are remarkable in demonstrating the effectiveness of the data-dependent LC-MS/MS strategy for the detection of biomarker protein panels, and clearly support the potential applicability of the method for biomarker fingerprinting in complex cellular extracts.

For the purpose of MRM experiments, a library of 1,572 proteins (all with $\geq$2 spectral counts) matched by 9,677 peptides (all with $P<0.001$) was generated. For each protein the library provides the sequence of detectable peptides, charge state, spectral count, MH$^+$, parameters that characterize the quality of the tandem mass spectrum ($P$-value, DeltaM, Xcorr, DeltaCn, Sp, number of matching $a$, $b$, $y$ ions in the spectrum), retention time, and top 10 most intense product ions that correspond to any given peptide. The spectral count of each peptide at the detectable charge state reflects the propensity for identification of that peptide. The $P$-value and the SEQUEST scores reflect the quality of the tandem mass spectrum that led to the identification of the peptide and represent
useful parameters for selecting the most relevant transitions for the identification of a protein. Up to ten MRM transitions can be set up for each peptide (parent ion) and the retention time can be used to exclude contaminant signals. To obtain a qualitative view of how well our protein pool represents the human proteome, a chart reflecting the experimental frequency of the 1,572 proteins identified as a function of molecular weight (that ranged from ~5,000 to ~1,000,000 Da) was constructed and compared to a similar chart reflecting the theoretical
protein distribution downloaded from the Swiss-Prot/ExPASy online protein knowledgebase. The experimental and theoretical distributions were fairly similar, illustrating that our dataset comprised a representative set of proteins and that our experimental protocol performed well in sampling the human proteome (see Figure).

The applicability of the peptide library/MRM analysis for the identification of trace-level biomarkers in proteomic samples was demonstrated for the detection of a set of seven putative protein markers (e.g., tumor protein D54, 14-3-3 sigma, calreticulin precursor, 4F2 cell-surface antigen, cathepsin D precursor, Ki-67 antigen and PCNA). Peptides with the largest number of spectral counts and best SEQUEST scores (most importantly with the lowest \( P \)-values) were selected for MRM. The experimental results demonstrate that many proteins undetectable by data-dependent analysis (e.g., Ki-67 and PCNA) at concentration levels <1 nM can be identified by MRM analysis.

Conferences and workshops


Intellectual property


Peer-reviewed publications in reporting period


Pedro Mendes
mendes@vbi.vt.edu
Associate Professor, Virginia Bioinformatics Institute
Adjunct Associate Professor, Department of Biochemistry, Virginia Tech
Adjunct Associate Professor, Department of Cancer Biology, Wake Forest University Medical School
Professor, School of Computer Science, University of Manchester

Group contributors: Hui Cheng, Stefan Hoops, Aejaaz Kamal, Revonda Pokrzywa.

Software, Standards and Methods for Computational Systems Biology

Abstract. Work on the development of the COPASI (Complex Pathway Simulator) software continued and included the organization of the 1st Annual COPASI Users’ Workshop and the release of four new versions of the software. The ability to allow principled annotation of models through the MIRIAM (Minimum Information Requested in the Annotation of Biochemical Models) standard and to apply optimization to sensitivity analysis are two new features included in the software. The latter feature was used to develop a new method of global sensitivity analysis to rapidly assess the importance of each parameter of a model in a wide range of values. A comparison was completed of generic kinetic functions for use in biochemical network modeling that focused on how well they can fit generic metabolic data. Finally, we continued to participate in the development of the Systems Biology Markup Language (SBML) and have designed an entirely new standard for associating systems biology data with models – Systems Biology Results Markup Language (SBRML).

Keywords: systems biology; biochemical networks; computer simulation; computational modeling; software development; community standards; global sensitivity analysis; enzyme kinetics.
Scientific Progress

Systems biology studies how the interactions between molecules in a cell contribute to the systems’ behavior. This research requires a strong computational component, not only to analyze data but also to simulate dynamics and provide a framework to integrate knowledge. Our research covers all of these aspects of computational systems biology.

Modeling and simulation software

COPASI (Complex Pathway Simulator) is an open-source software package for the modeling and simulation of biochemical networks that we have been developing for nearly a decade. COPASI is now established as one of the main tools for this type of research and there are numerous peer-reviewed publications that report simulations carried out with COPASI (nearly 100, as indexed by ISI). With funding from the National Institute of General Medical Sciences (NIGMS), we produced four new versions of COPASI, resulting in a total of 5500 downloads, during this year. The new features added in the last year have improved the modeling capabilities of the software in several areas. COPASI now supports MIRIAM (Minimum Information Requested in the Annotation of Biochemical Models)-compliant annotation and it is possible to import and export these annotations to Systems Biology Markup Language (SBML). The optimization feature became more powerful by allowing results of Metabolic Control Analysis (MCA) and Sensitivity Analysis to be optimized. COPASI can display graphical layouts of the reaction network, with the ability to overlay time course simulation results over them. Layouts can be imported through SBML files in the SBML layout extension format. COPASI simulation capabilities have been integrated with CellDesigner and Systems Biology Workbench (SBW), which are the other most frequently used systems biology software applications. Additionally, we enhanced the usability by improving the interface based on user feedback received through the online User Support Forum and tutorials. We held the first annual COPASI User Workshop in June 2008 as part of the VBI Summer Institute, which included participants coming from both the United States and Europe.

Modeling methodologies

A strong focus of our group’s research is the development of new methodologies for modeling in systems biology. In the last year we worked on two separate methods. The first involved a collaboration with Ursula Kummer and Sven Sahle from the University of Heidelberg, focusing on the problem of assessing how strongly each parameter of a model affects the model’s performance, which is known as sensitivity analysis. While standard methods exist for assessing this in a short range...
of parameter values, we were interested in the more general case of *global* sensitivity analysis that does not require the restriction of small parameter ranges. Due to the nonlinearity of biochemical models, it is hard to assess how each parameter affects the model if we allow them to change within a wide range of values. In such case a Monte Carlo sampling is usually carried out, but this computational procedure is quite expensive. Instead, we proposed a new method that uses global optimization algorithms to rapidly assess which parameters have the least effect on model outcome. This new method was only possible with some of the new features introduced in COPASI, as described above.

A second method focused on the use of approximate kinetic functions to construct models, which is important in initial stages of modeling when accurate or complete information about the biochemical network is not available. To investigate this we used an artificial network that we had created earlier, called the Claytor network, which includes metabolism, signal transduction and gene expression. Using the metabolism data from the Claytor network, we examined a case in which one is interested in modeling metabolism without knowledge of gene expression. In the literature, there are multiple instances where the metabolome is studied independently from the transcriptome and proteome. By comparing the behavior of the metabolism portion of the Claytor network to that of the full network it was possible to extrapolate how well those metabolism-focused studies captured the behavior of the complete network. We examined two scenarios: a) in an ideal case, the true kinetics of the system were assumed to be known; b) in a more realistic scenario, the kinetics were supposed to be unknown, *a priori*, and were fitted by a variety of generalized rate equations (mass action, lin-log, and convenience kinetics). We assessed the performance of these different generalized kinetic functions in terms of how well their parameters could be estimated and how well the final system reproduced the original data. This assessment revealed various strengths and weaknesses of each of the methods. Revealing, for example, that lin-log is quite hard to fit, or that, surprisingly, mass action is able to reproduce the original data quite well.

**Standards**

We continued our involvement in the creation of community standards for systems biology. SBML, an application-independent file format to specify systems biology models, is the main standard in the field and has been a focal point for the computational systems biology community and a catalyst for new developments in this area. Dr. Stefan Hoops from VBI continued serving as an SBML Editor, serving as part of a small group of researchers in charge of writing the specifications of the language. Their effort resulted in an updated specification for SBML Level 2.
While SBML has allowed systems biology models to be transferred between a large range of software applications seamlessly, we noted that there is no standard for representing systems biology data (either from experiments or simulations). Due to the growing need to represent such data in a format that can be associated with a network model, we created a draft proposal for the Systems Biology Results Markup Language (SBRML). SBRML provides a means to describe numerical data and associate the measured entities with specific elements of an SBML model. Along with the data, SBRML files contain all the information of how it was obtained, as well as how it should be interpreted in the context of a biological network. We expect this to become a standard way to communicate and store many different kinds of systems biology data.

Conferences and workshops


Peer-reviewed publications in reporting period


Biswarup Mukhopadhyay
biswarup@vbi.vt.edu
Associate Professor, Virginia Bioinformatics Institute
Adjunct Professor, Department of Biochemistry and Biology, Virginia Tech


Redox Biology of Methanogenic Archaea and Mycobacteria, Microbial Gasification of Coal and PEPCK Mechanism

Abstract. We have been investigating coenzyme F_{420} metabolism in the mycobacteria, sulfur metabolism and redox buffering systems in methanogenic archaea, the pathways for oxidation of hydrocarbons and reduction of CO_{2} to methane in coal beds, and the mechanism of phosphoenolpyruvate carboxykinase (PEPCK). Mycobacteria express a glucose-6-phosphate dehydrogenase (Fgd) that is F_{420} dependent and generates F_{420}H_{2}. F_{420}H_{2}, a deazaflavin derivative, is found primarily in the methanogenic archaea. In the bacterial domain, F_{420} is present in the Actinobacteria phylum, which includes the mycobacteria. We have found that F_{420}H_{2} chemically reduces NO_{2} to NO, which could help Mycobacterium tuberculosis to combat NO_{2} stress; this bacterium causes tuberculosis in humans. M. tuberculosis is more sensitive to NO_{2} than NO. Since an activated macrophage produces NO and converts this compound in its acidified phagosomes to NO_{2} for a more aggressive attack, M. tuberculosis would use F_{420}H_{2} to reduce NO_{2} back to NO and lower the effectiveness of the antibacterial actions of macrophages. While investigating the redox buffering systems of Methanocaldococcus jannaschii, we have characterized a NADH oxidase that appears to be a new member of the group 3 flavin-dependent protein disulfide reductases. We are also identifying the relevant biochemical pathways that are operational for hydrocarbon degradation in coal bed methane wells. The work on human liver cytosolic GTP-PEPCK, which is a gluconeogenic enzyme and plays a direct role in type 2 diabetes, has the goal of developing therapeutics that will lower the blood glucose level in a person with type 2 diabetes, without inactivating the enzyme and causing hypoglycemia. Our results show that the fully conserved Tyr235 of the human enzyme is not essential for catalysis, but influences the reaction through an anion-quadrupole interaction, which has been rarely encountered in enzyme catalysis. An interference with this interaction might provide a way to lower the activity of the enzyme.

Keywords: coenzyme F_{420}; tuberculosis; phosphoenolpyruvate carboxykinase; diabetes; archaea; redox buffer; coal gasification.
**Scientific Progress**

*Coenzyme F$_{420}$ and nitrogen oxide metabolism of the mycobacteria*

Coenzyme F$_{420}$, a 7,8-didemethyl-8-hydroxy-5-deazaflavin derivative (Figure 1), is a 2-electron or hydride transfer restricted redox catalyst ($E^\circ' = -360$ mV) similar to the nicotinamides ($E^\circ' = -320$ mV) (Eirich et al, 1978; Walsh, 1986). F$_{420}$ is found in all methanogenic and certain nonmethanogenic archaeb, where it participates in energy metabolism, NADP reduction, oxygen detoxification, and sulfite reduction (Johnson and Mukhopadhyay, 2005; Schauer et al, 1986; Seedorf et al, 2004; Yamazaki et al, 1980). In the bacterial domain, F$_{420}$ is found in certain members of the Actinobacteria phylum, such as *Mycobacterium* species (Purwantini et al, 1997). These organisms express an F$_{420}$-dependent glucose-6-phosphate dehydrogenase (Fgd, Reaction 1) (Purwantini and Daniels, 1998; Purwantini et al, 1997). The physiological fate of F$_{420}$H$_2$ produced by Fgd is unknown. An insertional inactivation of *fbiC*, an essential gene for the synthesis of the deazaflavin chromophore or catalytic unit of F$_{420}$ (Choi et al, 2002), renders *Mycobacterium tuberculosis* hypersusceptible to acidified nitrite (Darwin et al, 2003).

Glucose-6-phosphate + F$_{420}$  $\rightarrow$ 6-phosphogluconate + F$_{420}$H$_2$

(Reaction 1)

This *in vitro* treatment simulates an environment inside the phagosomes of an infected-activated macrophage, which produces nitric oxide (NO) by the action of inducible nitric oxide synthase (iNOS or NOS2) (MacMicking et al, 1997a). Upon acidification of a phagosome, nitrite, a major product of NO oxidation, is converted to nitrous acid (HNO$_2$; $pK_a$ = 3.16) (da Silva et al, 2006), which in turn dismutates to NO and NO$_2$ (Darwin et al, 2003; Stuehr and Nathan, 1989); NO$_2$ arises also from a reaction of NO with O$_2$ (Ignarro et al, 1993). These observations suggest that the pathogenic mycobacteria could use F$_{420}$H$_2$ to combat an attack of reactive nitrogen intermediates generated by the activated macrophage. We have found that reduced F$_{420}$ (F$_{420}$H$_2$) reduces NO$_2$ to NO (Figure 2) (Reaction 2) (Purwantini and Mukhopadhyay, 2009). This reaction would be useful to *M. tuberculosis* in combating the host defense system. Although both NO and
NO$_2$ exhibit antimycobacterial activities, NO$_2$ is more potent in killing *M. tuberculosis* (Yu et al, 1999).

\[
\text{F}_{420}H_2 + NO_2 \rightarrow \text{F}_{420} + NO + H_2O
\]

(Reaction 2)

Therefore, as an activated macrophage produces NO and converts this compound in its acidified phagosomes to NO$_2$ (Ignarro et al, 1993; Stuehr and Nathan, 1989) for bringing about a more aggressive attack, *M. tuberculosis* could use Fgd-generated F$_{420}H_2$ (Reaction 1) to reduce NO$_2$ back to NO (Reaction 2) and lower the effectiveness of the antibacterial action of macrophages. To test this hypothesis, we have constructed and tested the characteristics of two mutants of *Mycobacterium smegmatis*, a close relative of *M. tuberculosis* and a nonpathogen (Purwantini and Mukhopadhyay, 2009). One of these mutants could not synthesize F$_{420}$ (due to the deletion of *fbiC* gene) and the other was unable to reduce this coenzyme (due to an inactivation of the *fgd* gene). These strains were approximately fourfold more sensitive to NO$_2$ than the wild-type strain (Purwantini and Mukhopadhyay, 2009) (Figure 3). When the mutants were complemented with replicative vectors expressing the respective genes (pEP-*fbiC* and pEP-*fgd*), they were able to tolerate NO$_2$ even better than the wild-type cells (Purwantini and Mukhopadhyay, 2009) (Figure 3); the increased tolerance was likely caused by above-normal level of F$_{420}H_2$ generated by the over-expressed gene.
In summary, $\text{F}_{420}\text{H}_2$-dependent $\text{NO}_2$ reduction reaction is a defense tool for the mycobacteria against $\text{NO}_2$ stress (Purwantini and Mukhopadhyay, 2009). It should be noted that NO has often been cited as the agent that kills $\textit{M. tuberculosis}$ under both aerobic and low-oxygen or hypoxic conditions (MacMicking et al, 1997b; Nathan, 2008), although, as mentioned above, $\text{NO}_2$ has been shown to be a more potent antimycobacterial agent (Yu et al, 1999), and it is produced under aerobic conditions (Darwin et al, 2003; Ignarro et al, 1993; Stuehr and Nathan, 1989). Therefore, the $\text{F}_{420}\text{H}_2$-dependent defense against nitrosative stress would be useful to $\textit{M. tuberculosis}$ when it grows aerobically and causes active tuberculosis.

The overall phenomenon is similar to the inhibition of neoplastic transformation of human cells by $\gamma$-tocopherol (Cooney et al, 1993). It has been suggested that the nitrosation of primary amines of DNA bases by NO, which leads to cancer-causing mutations, requires oxidation of this oxide to $\text{NO}_2$ (Cooney et al, 1993). $\gamma$-Tocopherol protects cells from $\text{NO}_2$-induced carcinogenesis because it efficiently reduces $\text{NO}_2$ back to NO (Cooney et al, 1993).

---

**Figure 3. Requirement for reduced $\text{F}_{420}$ for $\text{NO}_2$ detoxification in $\textit{M. smegmatis}$.** The sensitivities of wild-type, $\text{fgd}::\text{aph}, \Delta\text{fbic}::\text{aph}, \text{fgd}::\text{aph (pEP-fgd)},$ and $\Delta\text{fbic}::\text{aph (pEP-fbiC)}$ strains to $\text{NO}_2$ were studied. pEP-fgd and pEP-fbiC, *E. coli*-mycobacteria shuttle vectors expressing $\text{fgd}$ and $\text{fbiC}$ from an unregulated promoter. In each case, the cells were exposed to $\text{NO}_2$ by incubating a corresponding diluted suspension in growth medium with spermine-NONOate under air; a control culture received water in place of spermine-NONOate solution. Spermine-NONOate generated NO slowly and the reaction of NO with oxygen provided a steady supply of $\text{NO}_2$. In each case, including the control, the number of surviving cells per milliliter in a culture is presented as a percentage of the value recorded for a sample drawn before incubation with spermine-NONOate (or water). As a result, the value even for the control culture was <100%. Each error bar was calculated from measurements in triplicate. Grey bar, control culture; black bar, $\text{NO}_2$-exposed culture. Reproduced from reference Purwantini and Mukhopadhyay, 2009.
Structural biology of GTP-PEPCK

GTP-dependent phosphoenolpyruvate (PEP) carboxykinase (GTP-PEPCK) is a gluconeogenic and glyceroneogenic enzyme (Hanson and Patel, 1994). By virtue of its gluconeogenic activity, this enzyme plays a direct role in the development of type 2 diabetes (Valera et al, 1994). It has been shown that treatment with a GTP-PEPCK inhibitor such as sodium tungstate or vanadate (Kiersztan et al, 2002; Marzban et al, 2002; Mosseri et al, 2000; Munoz et al, 2001), or silencing of hepatic cytosolic PEPCK by RNA interference (Gomez-Valades et al, 2006), lowers the serum glucose level in diabetic rats. Accordingly, the long-term goal of our research is to develop GTP-PEPCK inhibitors that can be used for the treatment of type 2 diabetes. Our particular interest has been in the compounds that lower the activity of the enzyme, but do not fully inactivate it. A full inactivation of PEPCK will cause hypoglycemia (Moller, 2001; Opherk et al, 2004). Accordingly, the work in the laboratory is focused on locating residues of PEPCK that influence, but do not determine the catalysis. These residues fall into two classes: 1. Residues that are located at or near the active site and are involved in positioning substrates (in some cases a bit of change in the relative positions of the substrates may affect, but not eliminate catalysis); 2. Residues that are not located at the active site, but are connected to catalytically critical residues at the active site via long-range interactions (an interference with such a distal site will perturb the active site geometry and result in subtle effects on catalysis). A specially designed small molecule that selectively interacts with either class of residues will lower, but not eliminate the activity of the enzyme.

In 2006, we reported that Asp84, which is located away from the active site, exerts a modulatory effect on catalysis (Case et al, 2006) and therefore belongs to Class 2 as described above. We have now found that Tyr235 of the human liver cytosolic GTP-PEPCK (cHumPEPCK) belongs to Class 1 (Dharmarajan et al, 2008). During this work we have found a rare and sought after example of the phenomenon of anion-quadrupole interaction in enzyme catalysis (Burley and Petsko, 1988; Jackson et al, 2007; Mecozzi et al, 1996; Perutz et al, 1986). We used inferences from a crystal structure (Dunten et al, 2002) which indicated that the aromatic ring of Tyr235 in human liver cytosolic GTP-PEPCK establishes an energetically favorable weak anion-quadrupole interaction with PEP carboxylate (Dharmarajan et al, 2008) (Figure 4). Therefore, it could be hypothesized that Tyr235, a fully invariant residue of GTP-PEPCKs, does not determine, but influences catalysis. From kinetic analyses of site-directed mutagenesis-derived human PEPCK variants, where position 235 was occupied by Phe, Ala and Ser, we showed that both the aromatic ring and the hydroxyl group of Tyr235 side chain help to position PEP in the

BiswaTap Mukhopadhyay  •  research report

2009 Scientific Annual Report
active site at a distance from Mn$^{2+}$ that allows efficient phosphoryl transfer and overall catalysis (Dharmarajan et al, 2008). Hence Tyr235 and distal sites that are connected to this residue and Asp84 provide avenues for lowering the activity of the human PEPCK (Case et al, 2006; Dharmarajan et al, 2008).

A redox buffering system of methanogenic archaea

In many bacteria and eukaryotes, glutathione (GSH) and GSH reductase provide a redox buffering system (Foyer and Noctor, 2005; Ritz and Beckwith, 2001). In certain bacteria and archaea, coenzyme A and coenzyme A disulfide reductase serve this role (Boylan et al, 2006; delCardayre and Davies, 1998; Hummel et al, 2005). Trypanosomes and mycobacteria fulfill this need with trypanothione and mycothiols (Bornemann et al, 1997; Fairlamb et al, 1985; Newton et al, 1996; Wang and Ballatori, 1998). These systems are absent in the methaogenic archaea. We are interested in identifying the redox buffering system of these organisms. A NADH-dependent coenzyme M (CoM) disulfide reductase (CoMDR) has been described in a methanogen (Smith and Rouviere, 1990); coenzyme M is essential for methanogenesis, which is the only avenue for energy production in a methanogen (Wolfe, 1992). The catalytic properties of CoMDR are similar to that of the group 3 flavin-dependent protein disulfide reductases (FDR) (Smith and Rouviere, 1990), that include NADH oxidases (Nox), NADH peroxidases (Npx) and CoADRs, which often have NADH oxidase activity (Argyrou and Blanchard, 2004; Claiborne et al, 1999). It could be hypothesized that CoMDR regenerates...
CoM from CoM-S-S-CoM, which could form due to oxygen exposure or normal cellular oxidative events, and therefore the enzyme is necessary for oxidative damage defense or for functioning of the cell. Also, CoM and CoMDR could constitute a redox buffering system for methanogens. We have begun testing these hypotheses with *Methanocaldococcus jannaschii*, which is a deeply rooted hyperthermophilic methanogenic archaeon that lives in deep-sea hydrothermal vents (Boone et al, 1993; Jones et al, 1983). The open reading frame MJ0649 of this anaerobe has been annotated as an NADH oxidase (Nox) homologue (Bult et al, 1996). It has the sequence signatures of group 3 flavin-dependent protein disulfide reductases, including the characteristic reactive cysteine that lies within a conserved SFXXC element and forms a Cys-sulfenic acid or Cys-SOH in the catalytic cycle (Claiborne et al, 1999) and the FAD and NADH binding sites (Case et al, 2009). From an analysis of the molecular and kinetic characteristics of a homogeneous recombinant form generated in *E.coli* we found that MJ0649 encodes an unusual NADH oxidase (Case et al, 2009). It contained one FAD per subunit, showed more similarities to CoADR than other members of the family and additionally possessed certain distinct features (Case et al, 2009). However, MJ0649 did not reduce the disulfide of coenzyme A or disulfide of coenzyme M, but oxidized NADH with molecular oxygen. These mixed properties make MJ0649 an unusual member of the FDR family.

While oxidizing NADH with $O_2$, MJ0890 produced $H_2O_2$ and $H_2O$. The latter seemed to originate from a slow peroxidase activity of the enzyme that consumed $H_2O_2$ (Case et al, 2009). The apparent $K_m$ for $O_2$ was 1.9 mM (Case et al, 2009), which is much higher than that for other Nox enzymes for which a role in oxygen detoxification has been proposed (Kengen et al, 2003; Yang and Ma, 2005). Also, within the hydrothermal vents, the natural habitat of *M. jannaschii*, the oxygen concentration cannot be above 8 µM (Case et al, 2009). The $K_m$ for NADH for MJ0649 was <3 µM (Case et al, 2009). Hence, NADH is a true substrate of the enzyme and oxygen is not likely to be a physiologically relevant reactant. Interestingly, when tested with 5,5-dithiobis(2-nitrobenzoic acid) (DTNB), a disulfide compound of no physiological relevance, MJ0649, showed significant disulfide reductase activity (Case et al, 2009). Hence, MJ0649 might reduce an as yet unidentified disulfide containing substrate and the identification of this compound could define a novel reaction and enzyme system that will perhaps lead to the redox buffering system of the methanogenic archaea (Case et al, 2009).

**Microbial conversion of coal and paraffin to natural gas**

In collaboration with Altuda Energy Corporation, San Antonio, TX, and with support from the Natural Gas industry, we
are investigating the pathways for degrading coal and paraffin in underground coal beds. The current focus is on the examination of the relevant catalytic powers of the microorganisms that inhabit these formations through enrichment, isolation and analysis of metagenomic data. The ultimate goal is to design a consortium composed of the microorganisms that inhabit the coal beds and their derivatives that will degrade coal and paraffin in the coal bed and generate methane. The technology can be applied to both the existing active or inactive coal bed methane wells as well as new wells, thereby increasing natural gas output.

Evolution of sulfite reductase methanogens - an intertwined evolutionary history of methanogenesis, sulfate reduction, and anaerobic oxidation of methane, three of the oldest respiratory energy metabolisms of earth

Sulfite is an obligatory intermediate for the reduction of sulfate to sulfide (SO$_4^{2-}$ + 2e → SO$_3^{2-}$; SO$_3^{2-}$ + 6e → S$^2$) (LeGall and Fauque, 1988) and methanogenic archaea are generally sensitive to sulfite, because this oxyanion inhibits methylcoenzyme M reductase (Mcr), which is essential for methanogenesis (4H$_2$ + CO$_2$ → CH$_4$ + 2H$_2$O) (Johnson and Mukhopadhyay, 2008). Hence sulfate reduction and methanogenesis had been thought not to occur in one organism (Johnson and Mukhopadhyay, 2005). The same is true for anaerobic oxidation of methane (AOM) (CH$_4$ → 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 AOM are transferred to a bacterium that reduces sulfate (Boetius et al, 2000). However, our calculations with the geological data have shown that in the submarine hydrothermal vents, methanogens are exposed to sulfite. The same was true for the early period of the oxygenation of earth (Johnson and Mukhopadhyay, 2005). Starting from this clue, we have shown that *Methanocaldococcus jannaschii*, a hyperthermophile that lives in the deep-sea hydrothermal vents, not only tolerates, but also uses sulfite as a sulfur source, and this ability is due to an unusual sulfite reductase (Fsr) that utilizes coenzyme F$_{420}$ as electron carrier. Fsr, a multimer of ~70-kDa subunits, is composed of two distinct domains (Johnson and Mukhopadhyay, 2005). The N-terminal half (Fsr-N) of the Fsr polypeptide is a homolog of F$_{420}$H$_2$ dehydrogenase that funnels electrons from F$_{420}$H$_2$ to a membrane-based energy transduction system in certain late evolving euryarchaea (Johnson and Mukhopadhyay, 2005). The C-terminal half of Fsr (Fsr-C) is a homolog of dissimilatory sulfite reductase subunits DsrA and DsrB (Johnson and Mukhopadhyay, 2005). Fsr homologs are present in methanogens from extremely hot and permanently cold environments (Johnson and Mukhopadhyay, 2005; Johnson and
Mukhopadhyay, 2007) and are an archaeal member of an AOM consortium (Johnson and Mukhopadhyay, 2007). Since, it has now been shown that methanogens can reduce sulfite, the concept that the development of sulfate reduction occurred in these organisms gains support. However, Fsr has a very limited distribution in the methanogens (Johnson and Mukhopadhyay, 2005; Johnson and Mukhopadhyay, 2007) and one could argue that it was acquired through horizontal transfer. It is also possible that the Fsr originated from simpler parts that originated in the methanogens. We have now found that every methanogen carries at least one open reading frame that is a homolog of Fsr-C and has the potential of encoding a small (22.4–37.2 kDa) siroheme sulfite reductase (Johnson and Mukhopadhyay, 2007); we call these open reading frames sDsr (s for small). The sDsr are diverse in their potential cofactor assembling and sulfite binding residues (Dwi Susanti and Biswarup Mukhopadhyay, manuscript under preparation). Some of the sDsr do not have iron-sulfur center assembling residues and they probably associate with ferredoxin-type proteins for obtaining electrons. Our on-going research shows that the sDsr provide a snapshot of the evolutionary processes that led not only to the Fsr, but also to the sulfite reductases of extant archaea, bacteria and eukarya.

Acknowledgements

This report contains text material and figures from our recent publications (Purwantini and Mukhopadhyay, 2009; Case et al., 2009; and Dharmarajan et al., 2008). The reported investigations have been supported by the the National Aeronautics and Space Administration (NASA), the U.S. Department of Energy, VBI-Johns Hopkins University Collaboration, and the Institute for Biomedical and Public Health Sciences (IBPHS), Undergraduate Summer Research Fellowship from the Fralin Biotechnology Center, and Graduate Fellowships from the Genetics, Bioinformatics and Computational Biology Program, and Dean’s Diversity Assistantship from The Graduate School of Virginia Polytechnic Institute and State University.

Conferences and workshops


Intellectual property


Peer-reviewed publications in reporting period


Other cited references


Genetic Design Automation

Abstract. Genetic engineering has developed around technologies enabling the targeted \textit{in vitro} recombination of DNA molecules found in living organisms. As a result, the development of new DNA molecules has been primarily focused on cloning strategies, that allow their assembly from existing DNA fragments. As chemical gene synthesis matures, the design of synthetic DNA molecules becomes the bottleneck of many biotechnology projects. It is urgent to develop representations of synthetic genetic systems that are more abstract than their DNA sequence. Abstraction makes it possible to reuse simple components to build complex systems or to break down a complex engineering problem into manageable tasks. It is particularly important to develop formal representations of the relationships between the structure of nucleic acid sequences and the biological function implemented in the gene regulatory networks that they encode. The Synthetic Biology Group at the Virginia Bioinformatics Institute is developing the foundation of a design automation framework for engineering synthetic biological systems. Improved productivity for the biotechnology industry, comparable to what has been achieved by the electronics industry over the last four decades, will result from integrating new computer languages to design DNA sequences, coupling DNA fabrication and design, and engineering a custom imaging platform to evaluate the performance of synthetic DNA molecules.

Keywords: computer-assisted design; live cell imaging; synthetic biology; genetic parts; regulatory networks; DNA fabrication; gene synthesis; cell cycle.
Scientific Progress

From Electronic Design Automation to Genetic Design Automation

Electronic Design Automation (EDA) is an engineering discipline and a category of tools for designing and producing electronic systems ranging from printed circuit boards to integrated circuits. EDA often includes computer-aided design (CAD), engineering, and manufacturing, but beyond these, the more general emphasis is on Design Automation. EDA has rapidly increased in importance in conjunction with Moore’s Law describing the scaling of semiconductor technology. Not only are modern processor and application-specific integrated circuit designs too cumbersome and complex to be designed manually, but EDA forms the backbone of a large, diverse, and multidisciplinary ecosystem that permits necessary specialization while maintaining coherent end-to-end integration.

Early integrated circuits were designed by hand, and the lithographic masks needed to produce them were laid out manually. Some of the first automation tools generated magnetic tapes of geometric data for a photoplotter — much like drafting. By the mid-1970s, automation reached beyond drafting into design, marking the start of the eventual divergence of EDA from more traditional CAD tools used in civil, architectural, and mechanical engineering. By 1980, concepts of frameworks and design languages were established in academia, appearing in commercial offerings by the mid-1980s from companies that remain dominant players today.

EDA tools are often classified by the primary engineering domains involved along the path from high-level architecture to manufacturing: architecture, logic, circuit, and layout. Each domain focuses on structural and behavioral concerns — static and dynamic perspectives — within the same abstract representation (Figure 1). Further, certain classes of tools connect representations from one or more domains. The logic of ones and zeros flowing through connected gates are related to circuits of electrical voltages and currents flowing through devices and wires. Regardless of which domains are involved, common functions form the basis of most design methodologies: for example, synthesis, simulation, validation, extraction, equivalence checking, and design rule checking. Categorizing via these abstract functions to represent the design process as a systems engineering problem exposes potential analogs with emerging tools in biology-derived domains. The Synthetic Biology group is laying the foundation of a Genetic Design Automation (GDA) discipline.
Design languages

Several fields of biology have adopted the notion of genetic parts recognizing that certain biological functions can be associated with specific DNA sequences. This concept provides a finer level of granularity than the traditional notion of a gene. However, the benefits of this paradigm shift have been hampered by the lack of formalism suitable for expressing the biological function of DNA sequences.

Attribute grammars are used in computer science to translate the text of a program source code into the computational operations it represents. It is possible to translate DNA sequences into molecular interaction network models by associating attributes with DNA parts, modifying the value of these attributes using rules describing the structure of DNA sequences, and using a multi-pass compilation process (Figure 2).

Instead of assuming that biological functions are properties of individual sequences irrespective of the context in which they are placed, attribute grammars establish the semantics of DNA sequences as a property of the genetic space generated by a parts library and the syntactic rules describing how parts can be combined. This approach has been validated by generating, translating, and evaluating the phenotype encoded in 41,472 sequences of two-gene networks derived from a small library of genetic parts. This systematic exploration of a genetic space contributes to a better understanding of how genetic parts affect the robustness and detectability of complex phenotypic traits.

Fabrication

The de novo fabrication of custom DNA molecules is a transformative technology that will significantly affect the biotechnology industry. Even before the foundations of molecular biology were established, pioneers were working on the chemical synthesis of genes. Basic genetic engineering techniques for manipulating DNA in vitro opened up incredible opportunities in the life sciences and biotechnology industry. However, genetic engineering has now moved beyond the introduction of single genes into cells to multi-gene cassettes, and is rapidly progressing towards whole-genome engineering. The synthesis of DNA molecules is increasingly the time- and cost-limiting step in genetic engineering.

Today, most multi-gene engineering projects involve ad hoc methods of DNA assembly. A variety of PCR-based methods, including SOEing (Splicing by Overlapping Extension) and SLIC (Sequence and Ligation-Independent Cloning), are in common use alongside more traditional restriction enzyme-based assembly methods. Their essential feature is the piecing together of existing DNAs that are cloned from natural sources. These techniques have several
Figure 1. Electronic Design Automation stack. A high-level representation of the relationships between the different tools integrated in design automation frameworks.

Figure 2. Architecture of a DNA compiler. The compiler translates a DNA sequence into a mathematical model of the gene network it encodes. The input for this process, the DNA sequence, is first broken down into parts by the scanner. The order of the parts in the sequence is validated by the parser according to the grammar rewriting rules. After validation by the parser, the sequence is translated by applying semantic actions attached to the rules to transform the series of parts into a set of chemical equations. The resulting equations can then be solved using existing simulation engines.
limitations. The use of restriction sites within natural sequences necessitates a labor-intensive custom cloning strategy that is difficult to automate. Molecular biologists often reach a tacit compromise between obtaining a desired sequence and the number of steps in the cloning process they are willing or able to undertake in constructing it.

Theoretically, DNA fabrication methods that are rooted in chemical synthesis could transform synthesis into a generic, predictable, and scalable process allowing for the generation of any DNA molecule. By liberating the process from the confines of preexisting sequences, the problem of composition design becomes orthogonal to the problem of physical construction. In other words, the question becomes how does one design something useful once complete creative freedom is realized? Therefore, as gene synthesis becomes a commodity, biologists will spend more time designing custom DNA molecules and characterizing their performance and less time constructing them.

Since this idealistic vision of large-scale, affordable gene synthesis has not yet materialized despite rapid progress, our group has initiated the development of a custom DNA fabrication platform that could be integrated into the GDA framework.

Performance evaluation

It is necessary to develop tools to collect phenotypic data having a dense time resolution from individual cells in order to properly evaluate the performance of genetic constructs. Combining the use of fluorescent molecular probes and time-lapse microscopy enables us to collect phenotypic data that provide insights into the inherent fluctuations of gene expression mechanisms. This setup also allows us to detect cell division events from which it is possible to derive cell lineages that may be important in analyzing epigenetic effects. In some cases, the localization of proteins within specific cellular compartments is an important phenotypic trait that can be measured in microscopic images.

Time-lapse quantitative microscopy requires developing data reduction strategies capable of extracting biologically meaningful data from the 10,000 or so images typically collected in a single experiment.

We have therefore developed algorithms embedded into a software suite automating the image processing steps necessary to quantitatively analyze data collected in time-lapse microscopy experiments. Some of the tasks integrated in this suite are cell identification, mapping of cells between frames, the assignment of cells to lineages, quantification of cellular fluorescence, analysis of localization of cell fluorescence in specific subcellular components.
**Outreach and leadership**

A systematic effort to identify potential applications for the technologies developed in the group was conducted by creating opportunities to identify potential users. GenoCAD, the design automation environment developed by the group, was presented at four international conferences (Bio-IT World, Intelligent Systems in Molecular Biology, International Conference on Systems Biology, Pacific Rim Summit on Industrial Biotechnology and Bioenergy). Six venture capital funds were given the opportunity to review a business plan describing the transfer of the GenoCAD technology into the marketplace. The feedback collected during this process will steer future development efforts in a direction attractive to potential investors and end-users. This technology transfer effort was featured on the cover of the Roanoke Times on May 21, 2008.

We have also been invited to participate in several events focused on the development of standards for the synthetic biology community. We organized a workshop on Computer Assisted Design of synthetic genetic systems in London. We provided advice to the Defense Advanced Research Projects Agency (DARPA) and the Defense Threat Reduction Agency (DTRA) to help these organizations articulate research programs focused on defense applications of synthetic biology. We also served on review panels for the National Science Foundation, the European Science Foundation, and Microsoft Research.

In January, Dr Peccoud joined the editorial board of *PLoS ONE*, an international, peer-reviewed, open-access, online publication publishing reports on primary research from any scientific discipline.

**Conferences and workshops**


Peer-reviewed publications in reporting period


Gene Interaction Networks for Functional Analysis of Complex Biological Processes

Abstract. Gene interaction networks are being developed in two model plants, *Arabidopsis* and rice, to dissect complex biological processes such as abiotic stress response and resistance. Plants have complex mechanisms of acclimation and adaptation to respond and survive environmental stresses. Analysis of these protective mechanisms will provide insights into plant stress response and resistance. The complex responses to environmental stress need to be considered at a global level to study the multiple interactive components. To construct genome-scale functional gene interaction networks, functional genomics data was probabilistically integrated, bringing together diverse interaction data types including gene coexpression, protein-protein and protein-DNA interactions, co-evolutionary functional protein linkages and small-RNA expression networks. A drought response subnetwork was then derived, composed of genes linked by functional interactions to drought responsive genes. Drought responsive genes from *Arabidopsis* and rice were compared using ortholog predictions, revealing a surprisingly high congruence that supports evolutionary conservation in gene functions. We screened knockout mutants of 200 drought responsive regulatory genes for altered drought response phenotypes. Those knockout mutants with altered drought response phenotypes were used as genetic perturbations to probe the gene network and their altered drought transcriptome was analyzed to reveal a network of regulated genes that are a subset of the drought response regulon. Iterative integration of the perturbation experimental data into the drought network will validate the network and improve its predictive value.

Keywords: gene network; *Arabidopsis*; rice; maize; switchgrass; drought stress; microarray; phosphoproteome.
Scientific Progress

The integration of functional genomics data using gene interaction networks can provide a systems-level view of complex biological processes. To validate such networks, genetic perturbations can be used as a reductionist discrete experimental approach to probe the system using variants of gene function. The iterative integration of the results from genetics experiments to a global view can provide a validated systems view of biological processes that have predictive value. We use this approach, which is shown in the Figure, to study a variety of important biological processes. Since the reproductive stage of cereals is very drought sensitive and has the most significant effect on yield losses, we have developed an iterative predictive-experimental model system to compare rice and Arabidopsis and analyze this biological process, which is described below. Studies on these plant models underpin research in other crops (maize, soybean, switchgrass), as shown in the Figure.

Comparative drought response transcriptome analysis

To form a basis for understanding the function of common drought responses in plants, we compared the two model plants Arabidopsis and rice (the sequenced genotypes Colombia and Nipponbare, respectively) to obtain a genome-wide view across dicots and monocots. We used controlled soil water deficit by progressive drought treatment to simulate field drought conditions for vegetative and reproductive stages in the two plants. RNA from the drought stressed (DS) and well-watered (WW) plants were used for transcriptome studies using Affymetrix arrays. This information was statistically analyzed and \( q \)-values <0.01 were used to select differentially expressed (DE) genes. The DE genes of rice and Arabidopsis at the vegetative and reproductive stages were compared using orthologous genes (InParanoid) and showed a high correspondence between the tissue stages in each species, as well as between. Around 10% of up-regulated rice genes in vegetative tissue and 20% in reproductive tissue had Arabidopsis orthologs that were similarly regulated. Interestingly, there were 101 common up-regulated and 60 down-regulated orthologs between the two species in both tissue drought treatments. Analysis of the cis-regulatory elements and transcription factors of the drought responsive genes will reveal the regulatory program of drought response (Krishnan and Pereira, 2008).

Arabidopsis gene interaction network

Gene interaction networks have been found useful in understanding functions in a number of organisms. In Arabidopsis, we curated, evaluated
and selected high-confidence interactions from a diverse set of evidence and datasets, including coexpression, protein-protein interactions, protein-DNA interactions and predicted co-evolutionary linkages. Quantitative data in the form of interaction scores or confidence values were retained, and when not present, calculated based on topological parameters. The individual weighted networks were evaluated against a “gold standard” set of interactions between genes participating in the same biological process (as defined by “specific” Gene Ontology biological process categories). From each set of network data, the top interactions that, together, accounted for a reasonable precision were filtered out and these high-confidence interactions from the individual network types were integrated to produce the first genome-scale gene interaction network in *Arabidopsis*, covering 20,138 genes (∼80% of the genome) connected by 104,678 edges. This network forms the basis of our systems level analysis of *Arabidopsis* biological processes.

**Comparative analysis using a rice gene interaction network**

To explore the global context of the genes regulated by drought stress in the model crop, a similar network was assembled. First, all available evidence of interactions in rice (experimental or predicted) in datasets from the Biomolecular...
Interaction Network Database, IntAct, the Rice Kinase Database, AT Reactome, the Generation Challenge Programme and InteroPORC were assembled to build a network of 3,377 proteins connected by 14,382 interactions. Second, high confidence interaction networks in *Arabidopsis* (total of 118,502 edges connecting 10,904 genes/proteins) were transferred to rice through orthologous genes (interlogs). A union of these two datasets gave a network model for rice covering 11,761 genes (∼25% of genome). The network was evaluated for containing functional interactions by gauging the number of interactions between genes identified as being responsive to drought (5,990) and known resistance genes (37), out of which 2,589 were present in the network. A statistically significant number of edges (5,499) connect these 2,589 genes among themselves ($P$-value<1E-05; it is extremely hard to observe ∼5,500 edges between an arbitrary set of ∼2,600 genes). Hence the network favorably connects functionally related genes, meaning neighbors of these ∼2,600 genes in the network are probably important for drought response/resistance. A total of 6,531 such neighbors were recovered from the network, giving a total of 12,546 genes. The network inference in rice, which is based on *Arabidopsis* data, will be iteratively refined by identifying condition (drought)-specific functional *Arabidopsis*-rice orthologs using active network analysis (discussed below), which will be used to transfer interactions from *Arabidopsis* to rice.

**Genetic perturbation of networks**

Functional genomics tools in the form of knockouts and other mutant resources have been developed in *Arabidopsis* (Arabidopsis Biological Resource Center) and rice through international collaborations (Krishnan et al, 2009). To probe the drought gene interaction networks experimentally using genetic perturbations, we selected drought responsive *Arabidopsis* regulatory genes, including transcription factors, protein kinases and phosphatases. Arabidopsis knockout mutants of these genes were tested in controlled soil water deficit drought assays for their drought response phenotypes. Drought response of mutants and wild-types was measured as either growth under drought or increase in biomass (dry-weight) under drought as compared to well-watered control (see area of Figure labeled Genetics). Those mutants that show a significant reduction under drought (normalized to growth under well-watered conditions) compared to that of the wild-type are regarded as drought sensitive. Likewise, mutants that show a significant increase (normalized) compared to a wild-type control are regarded as resistant. Around 10% of the mutants show an altered drought response using these criteria, which was further validated using other morpho-physiological measurements.
A set of knockout mutants with altered drought response phenotypes were analyzed for their altered transcriptome and revealed a mutant-drought DE gene subnetwork, which is a subset of the drought network. Integration of a number of drought subnetworks that are responsive to different genetic perturbations will provide an experimentally validated network that will have more predictive value. In addition, parallel studies of the drought phosphoproteome and metabolome (see Figure) will provide new data and an independent validation of gene and pathway functions in the networks. The comparative analysis of the Arabidopsis orthologs of rice genes involved in reproductive stress responses is a working model to dissect this critical biological process.

Conferences and workshops


**Peer-reviewed publications in reporting period**


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 a rapid pace thanks to new DNA sequencing technologies. Comparative genomics is one of the main beneficiaries of the increase 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, algorithms to help infer bacterial genome evolution, web-based infrastructure for genome annotation and analysis, and phage display sequence analysis. Highlights of the year were the publications of results concerning the evolution of multi-chromosome genomes in bacteria based on *Agrobacterium* data, a comparative genomics paper on *Brucella*, the *Pseudomonas syringae* ptoT1 paper, and a paper describing the Genome Reverse Compiler.

Keywords: bacterial genomics; bioinformatics; genome analysis; *Agrobacterium*; *Azotobacter*; *Pseudomonas*; *Brucella*.
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 in 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. During the reporting period, the *Agrobacterium* consortium published its main findings (Slater et al, 2009). We were able to propose a generalized mechanism for secondary chromosome evolution in bacteria thanks to a detailed analysis of the C58, S4, and K84 genomes and several others in the *Rhizobiales* order. A technique pioneered by our group, multiple genome alignments based on ortholog groupings, was key to the discovery of this mechanism.

*Azotobacter vinelandii* is a γ-proteobacterium free-living aerobe species that has nitrogen-fixation capabilities and is a widely used model for biochemistry studies. The goal of the project is to obtain a detailed view of *A. vinelandii*’s genome and couple that with the extensive knowledge of its biochemical properties. At the time of this writing, a paper describing the genome and the genes responsible for the protection of oxygen-sensitive processes has been submitted.

*Pseudomonas syringae*, a plant pathogen with many different strains, is also in the γ-proteobacteria group. Each strain has its own set of plant 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. Such studies can throw light on the interplay between pathogen evolution and the early stages of human agriculture, an idea advanced by Boris Vinatzer. In the reporting period, a paper describing the first genome in the batch, strain ptoT1, was published (Almeida et al, 2008). Our studies revealed that ptoT1 has a repertoire of type III secretion system genes that differs significantly from that of *P. syringae* DC3000, the preeminent model for type III secretion system studies.
**Genome analysis tools**

Andrew Warren, a Ph.D. student, finished the development of the Genome Reverse Compiler (GRC). The GRC is a stand-alone, open-source, efficient annotation tool for prokaryotic genomes. In the reporting period, a description of GRC was published (Warren and Setubal, 2009) and GRC was used to annotate the *P. syringae* ptoT1 genome. Kuan Yang, a Ph.D. student, is developing computational tools for the reconstruction of ancestral bacterial genomes. Such tools will allow a more detailed analysis of bacterial evolution, possibly leading to insights about present-day organisms. Nalvo Almeida, a visiting scholar from the Federal University of Mato Grosso do Sul in Brazil, did extensive work on tools for genome analysis, concentrating on gene content comparison among related strains. Almeida’s tools were a key factor in the *Pseudomonas* project mentioned above and in the *Brucella* work mentioned below.

**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. During the reporting period the highlight was the publication of a paper comparing ten *Brucella* genomes (Wattam et al, 2009), revealing that horizontal gene transfer has happened in these organisms, despite their intracellular lifestyle. A key component of this work was the concept of “shared anomalous regions”, which are genomic regions that have compositions that are statistically anomalous and that are shared by several members of the *Brucella* genus. This “shared anomalous region” concept allowed us to overcome the lack of specificity in a genome analysis tool that finds such regions looking at only one genome at a time (see Figure).

Joseph Gillespie, in the PATRIC team, led work on the analysis of type IV secretion system genes in members of the genus *Rickettsia* (Gillespie et al, 2009). This work represents a detailed and careful comparison of type IV genes in several *Rickettsial* genomes, uncovering differences that may play a key role in the pathogenicity of these organisms.

**Other projects**

During the reporting period, the first reports of the Plant-Associated Microbe Gene Ontology project (pamgo.vbi.vt.edu) were published in a special issue of *BMC Microbiology*. Our
contribution was a review of bacterial secretion systems and their coverage by newly developed Gene Ontology terms (Tseng et al, 2009).

In addition to ongoing collaborations described in last year’s report, we started joint work with Hamza El-Dorry, from the American University in Cairo. He is leading the Metagenomics Red Sea project. This project has collected biological samples from several locations and depths in the Red Sea, with a focus on the brine pools, regions 2,500 m deep with high salinity and temperatures of about 70°C. Metagenomics projects also require sophisticated computational tools to extract information from the sampled DNA sequences, and a major concern is the understanding of the phylogenetic diversity present in the samples. The expectation is that interesting discoveries about the oceanic microbial biodiversity will be made. Our group will play a major role in this project.

During the reporting period our collaboration with the Arap/Pasqualini lab at the M.D.
Anderson Cancer Center in Houston, Texas, made progress, with Dr. Nalvo Almeida developing a web-based system to house phage display data and facilitate data analysis. This web-based system was necessitated in part because of the use of next-generation DNA sequencing technologies in phage display, which has created a huge amount of data. Our group and Emmanuel Dias-Neto from the Arap/Pasqualini group are currently preparing a manuscript describing the first results.

**Conferences and workshops**

Setubal J (2008) Comparative genomics and evolution of *Brucella*, *Rickettsia* and *Coxiella*, *General Meeting of the American Society for Microbiology*, Boston, MA, June 1-5.


**Peer-reviewed publications in reporting period**


Abstract. A current focus of the Biochemical Profiling Group at the Virginia Bioinformatics Institute is the development of a high-throughput metabolomics platform for systems biology that can be used for the discovery of metabolic biomarkers and gene function elucidation. Our platform is built on a combination of untargeted metabolite profiling, metabolic fingerprinting and targeted analysis. Analytical techniques based on mass spectrometry provide sample analysis with high sensitivity and coverage of a wide range of metabolites. This platform is being successfully used to elucidate early metabolic responses to abiotic stress in plants, identify unique metabolic signatures associated with the progression of malignancy in human breast epithelium cells, and identify unique metabolic signatures associated with response to various drugs in the malaria parasite *Plasmodium falciparum*.
Scientific Progress

Malaria metabolomics

We have been using metabolomics to study malaria and identify the mode of action of antimalarial drugs in collaboration with Dr. David Sullivan’s group at the W. Harry Feinstone Department of Molecular Microbiology and Immunology at the Johns Hopkins Bloomberg School of Public Health.

Four of the more than 100 known species of Plasmodium are known to cause malaria in humans. Among these, Plasmodium falciparum infections are the most virulent, causing severe anemia or cerebral malaria which can result in death. The disease symptoms are most severe in young children and pregnant women. Malaria continues to be a major threat in the developing world, with more than 1 million clinical episodes and 3000 deaths every day. Ninety percent of the disease-associated mortality occurs in sub-Saharan Africa, despite the fact that malaria is indigenous in 40% of the world population. Malaria control depends on mosquito interventions and diagnosis and treatment of symptomatic cases. A licensed vaccine for malaria has not yet become a reality and antimalarial drugs are the only available method of treatment. While chloroquine, the first synthetically developed antimalarial, proved to be an almost magical cure for more than thirty years, the development and spread of chloroquine-resistant parasites has made it virtually ineffective in most parts of the world. Currently, artemisinin, a plant-derived antimalarial, is the only available drug that is globally effective against the parasite.

The malaria parasite in the human (also mosquito) host has a distinct morphology and biochemistry at different stages of its life cycle. Infection in a human host is initiated by the bite of an infected mosquito. The sporozoite stage first invades and multiplies inside hepatocytes. Once released from hepatocytes as merozoites, they infect erythrocytes and initiate a rapid and cyclic developmental pattern, which leads to the signs and symptoms commonly associated with malaria. Upon invading an erythrocyte, a P. falciparum merozoite quickly transforms into a ring stage parasite requiring about 24 hours to morphologically progress to a trophozoite stage, which, by definition, has visible heme crystals. DNA synthesis occurs during the late trophozoite stage, before nuclear division, which morphologically marks the schizont stage. Completion of schizogony produces 8–32 merozoites, which begin the 48-hour cycle again. In less than 1% of the asexual cycles, all the merozoites from a single infected erythrocyte will differentiate into the sexual gametocyte stage, which is infective for the mosquito.
Control interventions, whether by drug or a vaccine, are generally stage-specific. For example, most of the currently available drugs (except artemisinin) are not effective against gametocytes, thus allowing unabated parasite transmission. The short-lived merozoite stage is also mostly resistant to drugs, although experimental evidence is limited. The ring stage is very transcriptionally active, and relatively resistant to malaria drugs. Late ring to trophozoite stage is associated with more than 70% of host hemoglobin ingestion and catabolism to amino acids. The quinolines are most active during this hemoglobin degradation stage, with evidence pointing to interference with heme crystallization. Antifolates are active against late trophozoite and early schizont stages associated with DNA synthesis. Atovaquone collapses mitochondrial membrane potential by binding cytochrome $b_6$ complex and has broad stage activity including non-erythrocyte stages. Artemisinins have a broader window of stage activity, although early rings and late schizonts are not as susceptible to artemisinins and most other drugs. These observations indicate a stage-specific metabolism as targets for individual drugs.

The parasite food vacuole, nucleus, apicoplast, and mitochondria have been identified as important, uniquely susceptible organelles for current malaria drugs. Each has a uniquely Plasmodium-specific metabolism. The food vacuole efficiently degrades hemoglobin, making crystals and exporting short peptides into the cytosol for amino acid requirements. How quinolines target food vacuole function is still debated. The nucleus produces up to 32 progeny in five twofold divisions termed schizogony and DNA synthesis is targeted by the antifolates. The apicoplast has unique fatty acid synthesis, which is targeted by the antibacterials like clindamycin and tetracycline. The single acristate mitochondria lack a functioning Krebs cycle, but make essential heme, iron sulfur clusters and maintain membrane functions inhibited at cytochrome $b_{6}$ complex by atovoquone. After activation by heme or iron to form a radical, artemisinin probably works in the cytosol either by specific inhibition of a Ca$^{2+}$-ATPase 6 or diverse “innocent bystander” molecules in proximity to radicals formed.

Global transcriptomics or proteomics approaches have not elucidated specific mechanisms of action or immediate, altered biochemical pathways for the quinolines or artemisinins. Unlike yeast and bacteria, which have a robust transcriptional response to environmental changes including drugs, Plasmodium has a minimal (or no) transcriptional response to external stimuli or drugs. Summarizing excellent transcriptional data from many malaria labs, the intraerythrocytic P. falciparum has a streamlined, programmed transcriptional pattern where tight linkage of transcription and translation
produces proteins just in time for stage-specific functions. The pattern is also highly conserved amongst genotypic diverse isolates, with the most conservation in functional pathways like glycolysis, proteasome, plastid biosynthesis and cytoplasmic transcription and translation.

Metabolomics is the study of cells by measuring profiles of all, or a large number, of their metabolites. Metabolomics was originally proposed as a method of functional genomics, but its utility extends well beyond that – it is useful whenever an assessment of changes in metabolite levels is important. Examples exist already for applications in assessing responses to environmental stress, toxicology, drug discovery, nutrition, cancer, diabetes, and natural product discovery. This demonstrates the applicability of metabolomics in general biomedical research. Metabolite profiling, whether targeted at specific metabolite classes or untargeted, can also be applied as a tool in systems biology, where metabolite snapshots are used to study cellular dynamics through mathematical models. There are three major approaches used in metabolomics studies: (a) targeted analysis; (b) metabolite profiling; and (c) metabolic fingerprinting.

Metabolite profiling is the measurement of the levels of a set of metabolites in a sample. Metabolite profiling is a well-established activity in biomedical sciences, used routinely on biological fluids as a form of diagnosis which aids in characterizing the health state of a patient. These profiles were formerly composed of a small number of metabolites (and perhaps proteins), but this is now being extended to much larger numbers. Recently, researchers have been paying more attention to metabolite profiling as an extension to functional genomics. This application of metabolite profiling in functional genomics is similar to transcript and protein profiling, and, like these, it would be useful to establish the complete composition of the cell in terms of metabolites – the metabolome.

Metabolic fingerprinting does not attempt to identify all the metabolites in the sample. It rather considers a total profile, or fingerprint, as a unique pattern characterizing a snapshot of the metabolism in a particular cell line or tissue. Pattern recognition tools are used to classify the fingerprints and identify the specific features of the profile that are characteristic for each pattern. Metabolic fingerprinting has been widely used to discover specific metabolic patterns of diseases. Metabolic fingerprinting is particularly useful with pattern recognition and discriminant analysis techniques.

Evolutionarily, a parasite’s selection of erythrocyte as a target cell for development offers a significant metabolic challenge. Unlike other cell types in a human body, erythrocytes are terminally differentiated, lack a nucleus, mitochondria and ribosomal machinery and have hemoglobin as a primary constituent,
representing 90% of the total protein content. While this selection likely offers parasites some hitherto unknown biological advantage, it also necessitates an extensive metabolic self-engineering effort by the parasite. Fortunately for *Plasmodium* metabolomics, the somewhat limited metabolism of the uninfected erythrocyte provides less of a host background and has also been extensively characterized. The mature erythrocyte has four classical pathways: glycolysis, pentose-phosphate shunt, adenosine nucleotide metabolism, and the 2,3-diphosphoglycerate shunt which involve 39 metabolites and 32 internal metabolic reactions in addition to 19 primary and currency exchange fluxes. Unlike the host cell, we have little knowledge of the global metabolic changes that occur in the parasite and host erythrocyte during its development. The full metabolic repertoire may be more complex in the sense that a single gene leads to different transcripts, single transcripts result in proteins with different posttranslational modifications, and single proteins may accept different substrates and have different products. However, the advances in different extractions, mass spectroscopy technology and bioinformatic assembly of data make both metabolic profiles and fingerprinting possible. Metabolic footprinting has been used to classify unknown phenotypes of yeast genetic mutants by characterization of the extracellular metabolites. Use of a sulfur auxotroph in *Mycobacterium tuberculosis* identified novel sulfated metabolites and their pathways. Metabolic changes may be one of the first responses to drug action with protein changes and transcriptional changes close behind. In the study of *Plasmodium* biology, analysis of metabolite changes may be even more important because of the minimal response in transcription. Metabolomics, therefore, may allow identification of the most likely metabolic consequences of various antimalarial drugs and the discovery of potential metabolic targets for novel antimalarial drugs.

Currently the most developed technology for rapid profiling of large number of metabolites is gas chromatography coupled to electron impact (EI) quadropole mass spectrometry (GC-MS). We can simultaneously profile several hundred chemically diverse compounds including organic acids, amino acids, sugars, sugar alcohols, aromatic amines and fatty acids using this approach. These compounds can be separated and quantitated by GC-MS either directly or following a simple chemical derivatization procedure.

GC-MS profiling was performed using a ThermoFinnigan single quadropole TRACE DSQ GC-MS system consisting of ThermoFinnigan TRACE GC dual split-splitless injector-dual column capillary gas chromatograph with microplate CTC PAL headspace and liquid robotic autosampler and DSQ Electron Impact Ionization and Chemical
Ionization-capable mass selective detector, and additional ThermoFinnigan TRACE GC dual split-splitless injector-dual column capillary gas chromatograph with flame ionization (FID) and thermal conductivity (TCD) detectors and autosampler.

GC-MS methodology includes solvent extraction of metabolites from cell sample fractionation into polar and non-polar fractions, two-step derivatization and GC-MS separation. We have tried several methods for metabolite extraction and analysis from uninfected and \textit{P. falciparum}-infected human red blood cells. For non-targeted GC-MS profiling, the best results were achieved using a Covaris E100 High Performance Extraction and Dissolution System (Covaris, Inc.) that utilizes ultra-high frequency sound waves for the methanol extraction. During extraction, ribitol is added as internal standard. After liquid-liquid partitioning with methanol/water and chloroform, the polar phase (methanol/water) is dried and derivatized with methoxyamine hydrochloride in pyridine and \textit{n}-methyl-\textit{n}-trimethylsilyltrifluoracetamide. An aliquot of a retention time standard is added prior to trimethylsilylation. Replicate control samples, consisting of all of the reagents minus cells, are also processed and run through the GC-MS.
A GC-MS analysis using the Trace DSQ (ThermoFinnigan) GC-MS system results in the metabolic profile for a derivatized uninfected red blood cell (A), ring (B) and trophozoite (C) samples (see Figure). Ring stage *P. falciparum* is transcriptionally active but metabolically less active while the trophozoite stage is more metabolically active. The metabolomes of ring and trophozoite stages of the parasite were analyzed and compared by GC-MS.

Ring and trophozoite preparations were normalized by quantified hemozoin, metabolites extracted, derivatized, and analyzed by GC-MS. Metabolites were identified based upon retention time and mass spectral matching with a custom-made, proprietary library of standards. Peak areas were integrated, and response ratios of metabolite to the internal standard ribitol were calculated. Trophozoites had more and higher amounts of metabolites than rings. Many of the citric acid cycle metabolites such as succinate, malate, and fumarate were significantly higher in trophozoites than in rings, despite the absence of an intact citric acid cycle in the parasite. Many sugars were also higher in trophozoites than in rings (glucose, fructose, galactose). Thin layer chromatography analysis of radio-labeled, saponin-purified trophozoite lipid extracts indicates that glucose is effectively taken up by the parasite and converted into triacylglycerides and phospholipids which are needed for membrane growth and parasite replication. This information on the metabolic state of different developmental stages of *P. falciparum* will help to identify specific metabolic changes associated with drug treatment.

Our group also recently published results from a global lipidomic profiling study. LC-MS and MS/MS analysis of lipids extracted from purified trophozoites stage compared to lipids extracted from a subcellular fraction associated with hemozoin confirmed the neutral lipids responsible for heme crystallization. Our metabolomics analysis has shown that antimalarial drugs have different metabolic response patterns and we propose to use metabolic fingerprinting to classify different antimalarial drugs based on their metabolic response pattern.

Typically, biomarker identification is accomplished by applying computational methods to identify a small collection of metabolic markers able to classify samples based on mass spectrometry measurements. While a number of such markers have been described, few have proven to be of substantial benefit over conventional variables. We are using a new mathematical method developed by the Laubenbacher group at VBI, called “persistent homology”, which was originally described for the analysis of point clouds in high dimensional spaces, and not previously used to analyze molecular data, to assign a global signature to each drug-treated *P. falciparum* sample that does not rely on individual markers, such as...
expression patterns of individual ions as markers. Instead, it takes into account the geometry of the data points obtained from, for example, combined chromatography/mass spectrometry analysis. This signature can be used to compare signatures of different samples. While this approach has been used successfully in several application areas, such as image classification and the classification of diabetes patients based on phenotypic variables, it has not been used on high-throughput data in molecular biology. We will use this new technology on our metabolomics data obtained from drug-treated and untreated *P. falciparum*-infected red blood cells to classify known and novel candidate antimalarial drugs.

**Conferences and workshops**


**Peer-reviewed publications in reporting period**


Abstract. Considerable progress has been made through expansion of the infrastructure of the Cyberinfrastructure Group (CIG), which has been deployed through large-scale projects such as the PathoSystems Resource Integration Center (PATRIC), one of eight national Bioinformatics Resource Centers (BRC), and smaller-scale, yet highly innovative projects such as the Pathogen Interaction Gateway. CIG has extensively developed and implemented web-service-based components that can be configured and redeployed across projects as needs evolve without extensive re-writing. CIG’s infrastructure, while currently applied to the domain of infectious diseases, can be quickly re-deployed into other domains using the same data types and analysis routines. Furthermore, CIG continues to use its model of extensive collaboration with diverse scientific communities across the globe to jointly discover new biological knowledge, as well as focusing internally on biological research that CIG’s infrastructure, data and analysis are uniquely positioned to develop. Some select examples are shown for key host-associated bacteria.

Keywords: Sinorhizobium meliloti; symbiosis; two-component regulatory system; secretomes; 2-D-hydroxyacid dehydrogenase; glyoxylate reductase; hydroxypyruvate reductase; Rickettsia; Coxiella burnetii; pseudogene; host response; pathosystem; cyberinfrastructure; infectious disease; genomics; gene expression; proteomics; web services; interoperability; protein-protein interactions.
Scientific Progress

*S. meliloti* 1021 *loss-of-function deletion mutation in chvI and its phenotypic characteristics*

Bacteria have developed signaling systems for eliciting a variety of adaptive responses to their environments. Two-component regulatory systems (TCSs) are one of the most important mechanisms for these responses. They control many functions including regulating proteins important for pathogenesis and symbiosis. In *Sinorhizobium meliloti*, the TCS ExoS/ChvI regulates succinoglycan biosynthesis and flagellar genes. This system is not well characterized although it plays a crucial role in establishing the symbiosis between *S. meliloti* and its host plant. We constructed a *chvI* mutant by completely deleting the open reading frame encoding this gene (Wang et al, 2008; see “Conferences and workshops” section). Our results demonstrate that ChvI is intimately involved in regulatory networks involving the cell envelope and metabolism. However, its precise role within the regulatory network remains to be determined (Wang et al, 2009).

*Comparative modeling of the cell cycle – A homolog of the GcrA cell cycle regulator is present and essential in S. meliloti Rm1021*

*Caulobacter crescentus* has been used as a model for examining cell cycle regulation in bacteria. Several regulators that play essential roles in cell cycle progression have been identified. For example, GcrA is essential and is a global regulator of multiple cell cycle functions in *C.

The genes involved in and controlling bacteroid differentiation have not been well characterized, though changes in the bacterial cell cycle are obvious. We have demonstrated that *gcrA* is essential in *S. meliloti* using both integrated and knockout approaches. The promoter region of *gcrA* has been identified using rapid amplification of 5’ complementary DNA ends. We have isolated by density centrifugation synchronized populations of *S. meliloti* cells and analyzed their progression through the cell cycle. *GcrA* transcriptional fusions have been constructed, and a more detailed analysis of *gcrA* function is currently underway.

**Analysis of ten Brucella genomes reveals evidence for horizontal gene transfer despite a preferred intracellular lifestyle**

*Brucella* spp. infect a wide range of warm-blooded land and marine vertebrates and may cause brucellosis. The recent availability of ten genomes representing six of the species enabled a detailed comparison amongst themselves as well as to relatives in *Rhizobiales*. Phylogenomic analysis of ortholog families shows limited divergence but distinct radiations producing four clades: *B. abortus-melitensis, B. suis-canis, B. ovis*, and *B. ceti*. *Brucella* phylogeny does not appear to reflect the phylogeny of their preferred hosts. Only *B. suis* 1330 appears to have an intact β-ketoadipate pathway responsible for utilization of plant-derived compounds. In contrast, this pathway in the other species is highly pseudogenized, which is consistent with the “domino theory” of gene death. Distinct shared anomalous regions (SARs) are found in both chromosomes as the result of horizontal gene transfer unique to *Brucella* and not shared with its closest relative *Ochrobactrum*, which suggests their acquisition occurred in spite of a predominantly intracellular life style. In particular, SAR 2-5 appears to have been acquired by *Brucella* after it became intracellular. The SARs contain many genes, including those involved in O-side chain synthesis and type IV secretion, which if mutated or absent significantly affect the ability of *Brucella* to survive intracellularly in the infected host (Wattam et al, 2009).

**An anomalous T4SS in Rickettsia is evolutionarily conserved**

Thirteen *Rickettsia* genomes have been sequenced. *Rickettsia* are obligate intracellular Alphaproteobacteria responsible for various human diseases. Laboratory efforts have identified several genes involved in host cell invasion, yet no bona fide virulence factors correlating infection with reduced host fitness have been detected (Azad et al, 2008; Gillespie et al, 2009). Genomic studies (reviewed in Gillespie et al, 2008b) revealed a type IV secretion system (T4SS) similar (yet reduced in component genes) to the vir archetype of the phytopathogen *Agrobacterium*, which suggests...
Figure. Characteristics of the *Rickettsia* type IV secretion system (T4SS). (A) Comparison of the vir T4SS of *Agrobacterium tumefaciens* with the *rvh* (*Rickettsiales* vir homolog) T4SS of *Rickettsia*. Boxes depict T4SS genes: grey = conserved, white = missed by automated annotation, black = probable pseudogenes. X = deleted gene in *rvh* relative to *vir*. (B) Structure of the *Rickettsia rvh* archipelago across 13 genomes. Rvh genes are grouped into five islets as follows: box = islet A (rvhB3, rvhB4a, rvhB6a-e), open circle = islet B (rvhB2), triangle = islet C (rvhB9a, rvhB8a, rvhB7, rvhB8b, rvhB9b, rvhB10, rvhB11, rvhD4), D = islet D (rvhB1), and filled circle = islet E (rvhB4b). The conserved genomic locale of the rvh genes is shown at left, with minor deviations in the genomes of *R. bellii* str. RML 369-C, *R. canadensis* and *R. felis* shown at right. (C) Phylogeny estimation based on 18 Rvh sequences. Single most parsimonious tree of 12716 steps (6858 parsimonious characters of 28554 total characters). Branch support (all branches 100%) is from 1 million bootstrap replications.
that this transporter functions in eukaryotic cell pathogenesis. Our recent bioinformatic analysis (Gillespie et al, 2009) identifies several additional vir-like genes and suggests the Rickettsia T4SS is capable of substrate transport (see Figure). Atypical T4SS characteristics, such as duplication and genomic dispersion of scaffold genes, are conserved across the Rickettsiales. Affinities to the Legionella vir homolog (lvh) T4SS and other non-Alphaproteobacteria T4SSs hint at a lateral acquisition of the Rickettsiales T4SS, an event possibly facilitated by contact during co-infection within common protozoan or animal host cells. Thus, our study identifies distantly related bacterial species as candidates for a genetic model for further characterization of this secretion system. Given recent findings of plasmids and pronounced lateral gene transfer in Rickettsia, this anomalous T4SS, herein renamed rvh (Rickettsiales vir homolog), warrants consideration as a facilitator of genome plasticity aside from its speculative role in virulence.

Comparative genomics reveal extensive transposon-mediated genomic plasticity and diversity among potential effector proteins in Coxiella

Genetically distinct isolates of Coxiella burnetii, the causative agent of human Q fever, display different phenotypes with respect to in vitro infectivity/cytopathology and pathogenicity for laboratory animals. Correlations between C. burnetii genomic groups and human disease presentation (acute versus chronic) have been described, suggesting that isolates have distinct virulence characteristics. We deciphered the whole-genome sequences of the K (Q154) and G (Q212) human chronic endocarditis isolates and the naturally attenuated Dugway (5J108-111) rodent isolate to provide a more complete understanding of C. burnetii’s genetic diversity, evolution, and pathogenic potential. Cross-genome comparisons that included the previously sequenced Nine Mile (NM) reference isolate (RSA493) revealed both novel gene content and disparate collections of pseudogenes that may contribute to isolate virulence and other phenotypes. While C. burnetii genomes are highly syntenic, recombination between abundant insertion sequence (IS) elements has resulted in chromosomal rearrangement of syntetic blocks and DNA insertions/deletions. The numerous IS elements, genomic rearrangements, and pseudogenes of C. burnetii isolates are consistent with genome structures of other bacterial pathogens that have recently emerged from nonpathogens with expanded niches. The observation that the attenuated Dugway isolate has the largest genome with the fewest pseudogenes and IS elements suggests that the lineage of this isolate is at an earlier stage of pathoadaptation than the NM, K, and G lineages (Beare et al, 2008).
Dynamic state measurements for discovery of targets for countermeasures against microbial and viral pathogens

While progress has been made toward developing countermeasures for some biothreat pathogens, the “one bug-one drug” approach is difficult to extend to all emerging and re-emerging biological threats. New broad-spectrum strategies focus on developing countermeasures that act against multiple agents. Dynamic response measurements obtained in each of the pathosystems can be applied to support this new broad-spectrum strategy. In collaboration with various researchers, we have generated and/or analyzed human and mouse gene expression responses to over 20 different pathogens. For example, our recent collaborative work has determined that the interferon-regulated antiviral endoribonuclease, RNase-L, serves an essential and previously unrecognized function in the innate immune response to gram positive and gram negative bacteria (Li et al, 2008b). The differentially regulated genes in response to hemorrhagic fever (Djavani et al, 2009; Djavani et al, 2007, see “Other cited references” section) in a monkey model have yielded diagnostic and mechanistic markers associated with early host response.

Enhanced CIG infrastructure

We have developed a simple but effective literature retrieval system that quickly identifies publications relevant to an organism, genome, or gene of interest using PubMed and Entrez Programming Utilities (eUtils) from the National Center for Biotechnology Information. A user can access relevant publications from organism, genome, and gene level pages on the PATRIC website. Our system automatically derives search terms using genome metadata and/or functional annotation and database identifiers of a gene/protein. Multiple search terms are combined to form a search string and then queried in real-time against PubMed using eUtils. Results are parsed and presented on the website with summary and direct links to abstracts and full text articles. In addition, results can be filtered by area of interest (i.e. countermeasures, diagnosis, disease, epidemiology, or gene expression) and/or time period (i.e. past week, past month, or past year) using pre-computed links. Results are always based on the latest information available in PubMed since our system always queries PubMed in real-time.

As with PathPort, the PATRIC website now supports creation of multiple gene sets/groups. Users can now save genes of interest and results from various searches as a new group or add them to one of the previously created groups. The user can create new groups, delete/edit
existing groups, or download groups and their members to the local machine. We have also integrated a group manipulation tool called Gene Set Explorer (GSE), developed by CIG.

Protein-protein interactions play a key role in initiating infection for many pathogens. The Pathogen Interaction Gateway (PIG, http://molvis.vbi.vt.edu/pig/) (Driscoll, 2009) is a resource containing integrated host-pathogen protein-protein interactions from several public sources. PIG provides an easy to use web interface for accessing and using data that would otherwise require the navigation of several websites. It is also a platform upon which various tools can be developed for identifying potential drug targets for therapeutics. The PIG resource currently focuses on known human-pathogen interactions (viral and bacterial) and contains 20,113 host-pathogen protein-protein interactions for 206 different pathogen strains.

Conferences and workshops


Online bioinformatics training tutorial

An online bioinformatics training tutorial was developed as an education and outreach tool. The course presents the available bioinformatics resources and tools, using Francisella tularensis as the example organism, and contains exercises to give the user some hands-on experience. The tutorial was developed as part of the Middle-Atlantic Regional Center of Excellence for Biodefense and Emerging Infectious Diseases (MARCE) bioinformatics training program and is available from the MARCE website (http://www.marcebiodefense.org/index.php/marce/resources/tools/marce_bioinformatics_online_tutorial).


Gillespie JJ (2009) Phylogenomics as a tool for studying the evolution of vertebrate pathogenesis: an example from Rickettsia, Department of Entomology, North Carolina State University, Raleigh, North Carolina, USA, April 20.


**Peer-reviewed publications in reporting period**


Other cited references

Brett Tyler
bmyler@vt.edu
Professor, Virginia Bioinformatics Institute
Professor, Department of Plant Pathology, Physiology and Weed Science, Virginia Tech


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 produced improved sequences of the Arabidopsis pathogen Hyaloperonospora arabidopsidis and the soybean pathogen Phytophthora sojae, sequenced three more strains of Phytophthora sojae using Roche GS-FLX™ Titanium technology, and begun to sequence the genome of the fish pathogen Saprolegnia parasitica. Functional genomics analysis of these genomes identified virulence genes that encode the proteins that can enter plant and humans by binding to membrane lipids. Once inside plant cells, the proteins suppress plant defense reactions such as programmed cell death. We have completed a project focused on the dynamics of how genes from soybean and its pathogen Phytophthora sojae are expressed during infection. To infer genetic regulatory networks from such data, we are developing mathematics and computer science-based methods for inferring and modeling biological processes, including a method for analyzing data at a whole pathway level and for minimizing uncertainty in the inferred models.

Keywords: Phytophthora; Saprolegnia; oomycete; soybean; genome sequences; microarray analysis; regulatory networks; systems genetics; effector proteins; virulence.
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 produced an improved version of the draft genome sequence of the obligate parasite of Arabidopsis, Hyaloperonospora parasitica, by incorporating Illumina DNA and RNA sequences. A key finding from this study was the discovery that highly repetitive sequences make up more than 25% of the genome. We have also been continuing a project to finish the genome sequence of Phytophthora sojae, taking it from draft quality to near-perfect quality.

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). In 2008 we conducted a summer training workshop for the scientific community in the use of the terms and also published eight articles on our work in a special issue of BMC Microbiology.

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

<table>
<thead>
<tr>
<th>Collaborators</th>
<th>Research Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jim Beynon</td>
<td>Warwick University, United Kingdom</td>
</tr>
<tr>
<td>Jeffrey Boore</td>
<td>Genome Project Solutions, Inc.</td>
</tr>
<tr>
<td>Sandra Clifton</td>
<td>Washington University Genome Sequencing Center</td>
</tr>
<tr>
<td>Alan Collmer</td>
<td>Cornell University</td>
</tr>
<tr>
<td>Candace Collmer</td>
<td>Wells College</td>
</tr>
<tr>
<td>Ralph Dean, David Bird</td>
<td>North Carolina State University</td>
</tr>
<tr>
<td>Igor Grigoriev</td>
<td>United States Department of Energy Joint Genome Institute</td>
</tr>
<tr>
<td>John McDowell</td>
<td>Plant Pathology, Physiology and Weed Science at Virginia Tech</td>
</tr>
<tr>
<td>Richard Meyers, Jeremy Schmutz</td>
<td>Stanford University</td>
</tr>
<tr>
<td>TM Murali</td>
<td>Computer Science at Virginia Tech</td>
</tr>
<tr>
<td>Nicole Perna</td>
<td>University Wisconsin, Madison</td>
</tr>
<tr>
<td>Jane Rogers</td>
<td>Sanger Sequencing Center, Cambridge, United Kingdom</td>
</tr>
<tr>
<td>Carsten Russ, Chad Nusbaum</td>
<td>The Broad Institute, Massachusetts Institute of Technology</td>
</tr>
<tr>
<td>Pieter van West</td>
<td>University of Aberdeen, Scotland</td>
</tr>
<tr>
<td>Owen White</td>
<td>The Institute for Genomic Research (TIGR), J Craig Venter Institute</td>
</tr>
</tbody>
</table>
Figure 1. Entry of oomycete and fungal effector proteins into plants by binding phosphoinositides.

Figure 2. Expression of early effectors prevents plant responses to late effectors.
Biochemistry and 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. *P. sojae* and *Phytophthora ramorum* contain nearly 400 predicted effector genes and *H. parasitica* contains around 130. Many fungal proteins also contain effector genes and most oomycete proteins share two motifs at the N-terminus called RXLR and dEER. We showed that these two motifs, together with surrounding sequences, enable the proteins to enter plant and human cells. We also showed that effector proteins from fungal pathogens and malaria parasites also contain RXLR-like motifs at their N-terminus that enable entry into plant and human cells. We discovered a small molecule that can block effector protein entry, paving the way for the identification of a new class of therapeutic agents that can protect plants, animal and humans from oomycete, fungal and parasite infection.

With assistance from our collaborators we also showed that more than half of the effector proteins from *P. sojae* and several from *H. arabidopsidis* suppress a key plant defense process called programmed cell death, appearing to do so by attacking several different components of the host cell signaling pathways. We showed that *P. sojae* produces at least two waves of effector proteins. The second wave, which peaks around 12 hours after infection starts, can suppress the host plants’ ability to recognize and respond to the presence of characteristic molecules from the pathogen called PAMPs (Pathogen-Associated Molecular Patterns). The first wave of effectors is already expressed at the moment that infection

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

<table>
<thead>
<tr>
<th>Collaborators</th>
<th>Research Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Madan Bhattacharyya</td>
<td>Iowa State University</td>
</tr>
<tr>
<td>Daniel Capelluto</td>
<td>Biological Sciences at Virginia Tech</td>
</tr>
<tr>
<td>Mark Gijzen</td>
<td>Agriculture Canada, London, Ontario</td>
</tr>
<tr>
<td>Rays Jiang</td>
<td>Broad Institute</td>
</tr>
<tr>
<td>Jonathon Jones</td>
<td>Sainsbury Laboratories, Norwich, United Kingdom</td>
</tr>
<tr>
<td>Chris Lawrence</td>
<td>Virginia Bioinformatics Institute</td>
</tr>
<tr>
<td>John McDowell</td>
<td>Plant Pathology, Physiology and Weed Science at Virginia Tech</td>
</tr>
<tr>
<td>Weixing Shan</td>
<td>Northwest Agricultural and Forestry University, China</td>
</tr>
<tr>
<td>Chris Taylor</td>
<td>Donald Danforth Plant Science Center, St. Louis</td>
</tr>
<tr>
<td>Yuanchao Wang</td>
<td>Nanjing Agricultural University, China</td>
</tr>
</tbody>
</table>
begins and can suppress the ability of plants to recognize and respond to effectors that form part of the second wave.

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 soybean during *P. sojae* infection of cultivars that have different levels of quantitative resistance against the pathogen. This year we largely completed the analysis of transcriptional changes in a set of 300 recombinant inbred lines that segregate for quantitative resistance in order to map the soybean genetic loci responsible for resistance and any associated transcriptional changes. We found that more than 98% of soybean gene transcripts change significantly in their levels during infection, indicating that the entire physiology of the plant is remodeled during this process. To sift through the very large number of genes that might be responsible for resistance, we used a procedure called causal analysis to identify genes with a high probability of controlling resistance. We discovered one gene that is strongly predicted to control resistance, and are currently studying the mechanism by which this happens. To understand how the plant implements resistance, we developed a new method for jointly studying transcriptional data at a whole pathway level. The analysis revealed that resistant soybean plants have elevated levels of mRNAs associated with terpenoid, phenylpropanoid and suberin metabolism even before the pathogen infection begins.

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

<table>
<thead>
<tr>
<th>Collaborators</th>
<th>Research Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anne Dorrance and Steven St. Martin</td>
<td>Ohio State University</td>
</tr>
<tr>
<td>Ina Hoeschele</td>
<td>Virginia Bioinformatics Institute</td>
</tr>
<tr>
<td>Saghai Maroof</td>
<td>Crop, Soil, and Environmental Sciences at Virginia Tech</td>
</tr>
<tr>
<td>TM Murali</td>
<td>Computer Science at Virginia Tech</td>
</tr>
</tbody>
</table>

Table 4. Collaborators for genetic network inference project.

<table>
<thead>
<tr>
<th>Collaborators</th>
<th>Research Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ina Hoeschele, Reinhard Laubenbacher, and Henning Mortveit</td>
<td>Virginia Bioinformatics Institute</td>
</tr>
<tr>
<td>Pedro Mendes</td>
<td>Manchester Centre for Integrative Systems Biology, United Kingdom and Virginia Bioinformatics Institute</td>
</tr>
</tbody>
</table>
Genetic network inference

A major problem that arises when analyzing systems biology data in order to infer genetic regulatory networks is the sparsity and, often, the uneven sampling of the data. We have developed an algorithm to quantitate the uncertainty that arises in such experiments based on 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) RXLR-mediated entry of Avr1b into plant cells does not require pathogen encoded machinery, Oomycete Molecular Genetics Conference, Birnam, Scotland, May 6-8.

Tyler BM (2008) Coevolutionary battle between oomycete pathogens and hosts, American Society for Microbiology (ASM), Boston, MA, June 2-4.


**Peer-reviewed publications in reporting period**


2009 Faculty Fellow
Research Report
John Tyson

Keywords: cyclin-dependent kinases; molecular noise; transcription-translation coupling; feed-forward loops; stochastic simulation algorithms.
Scientific Progress

Stochastic models of cell cycle regulation in budding yeast

Dr. Tyson’s research group has embarked on an ambitious project to convert its well-established deterministic models (nonlinear ordinary differential equations) into detailed molecular mechanisms suitable for stochastic simulation. The work is supported by the National Institute of General Medical Sciences. To this end, postdoctoral research associate Sandip Kar has built a ‘toy’ model of the reactions governing the abundance of cyclin B-dependent kinase (Clb2:Cdk1) in yeast cells, and simulated random fluctuations intrinsic to these reactions by Gillespie’s stochastic simulation algorithm (SSA). His simulations also account for ‘extrinsic’ fluctuations in the mechanics of cell division (not exactly 50:50) and in the random distribution of proteins and mRNAs to the daughter cells. His results recently appeared in the Proceedings of the National Academy of Sciences as an invited article in a special feature on Complex Systems: From Chemistry to Systems Biology (Kar et al, 2009). Dr. Kar’s simulations suggest that the variability observed among dividing yeast cells cannot be squared with reported abundances of cell-cycle mRNAs (~1 mRNA molecule per cell per gene) and reported mRNA half-lives (15–25 min). Our simulations are consistent with abundances of 5–10 mRNAs per cell per gene and half-lives of 3–5 min. Recent measurements of specific mRNAs in yeast cells (Zenklusen et al, 2008) are closer to our estimates than to the originally reported values. It may be that yeast cells employ specific mechanisms to suppress the noise in protein levels that would be expected from low abundances of long-lived mRNA molecules. We are exploring some likely mechanisms by which cyclins might be buffered from excessive fluctuations.

In a second project, postdoctoral research associate Debashis Barik explored methods to speed up SSA simulations of enzyme-catalyzed reactions with large disparities in time scales, without sacrificing accuracy. He found that the ‘total quasi-steady state approximation’ fits this description and published his results in the Biophysical Journal (Barik et al, 2008). Lately he has been pursuing an alternative formulation of the cell cycle control system, based on the fact the cyclin-dependent kinases (Cdks) always phosphorylate their substrates on multiple sites (6–12 sites, typically). For technical reasons, this simple observation solves many problems associated with stochastic modeling of complex reaction networks in which kinases and phosphatases work on each other. This approach is revolutionizing the way we think about cell cycle control and the ways we model it both deterministically and stochastically.
Figure. Cdk signal transduction by feed-forward loops. Feed-forward loops, involving a Cdk substrate executor protein (EP) and its transcription factor (TF), are proposed to function as transducers between cell cycle regulatory signals (periodic fluctuations in Cdk activity) and cell cycle responses, such as initiation of DNA synthesis or cell division. Adapted from Csikasz-Nagy et al, 2009, Molecular Systems Biology 5: 236.

In a third project, graduate student Ted Ahn is carrying out approximate stochastic simulations of Kathy Chen’s model of a particularly interesting mutant of budding yeast in which the Clb2 protein (a B-type cyclin) is overexpressed and the Clb5 protein is removed. This mutant has the curious property that it is inviable when grown on glucose (a rich carbon source) but partially viable when grown on raffinose (a poor carbon source). The stochastic model predicts the division properties of wild-type yeast cells and the mutant strain on glucose and raffinose, and these predictions are being compared with experimental observations being made in Jean Peccoud’s laboratory at the Virginia Bioinformatics Institute.

Signal transduction by feed-forward loops in budding yeast

Successful progression through the cell cycle requires precise temporal coordination of the activities of hundreds of ‘executor’ proteins (EPs) involved in cell growth, DNA synthesis, mitosis, and cell division. In eukaryotes, Cdks, in combination with cyclins A and B, play central roles in regulating the production, activation, inactivation, and destruction of these EPs (see Figure). From genome-scale data sets of budding yeast, we have identified 126 EPs that are regulated by Cdk through direct phosphorylation of the EP itself and through phosphorylation of the transcription factors (TFs) that control expression of the EP. Hence,
each of these EPs is regulated by a feed-forward loop (FFL) from Cdk. By mathematical modeling, we have shown that such FFLs can activate EPs at different phases of the cell cycle, depending on the effective signs (+ or −) of the regulatory steps (1 2 / 3) of the FFL. The steps are as follows: 1, phosphorylation of TF; 2, transcription of the gene encoding EP; and 3, direct phosphorylation of EP. Feed-forward loops are of two types: coherent (sign1 * sign2 * sign3 = +) and incoherent (sign1 * sign2 * sign3 = −). Coherent FFLs of type (+ + / +) and (− − / +) are expected to drive the production of EPs that are active in S/G2/M phases of the cell cycle, when Cdk activity is high; whereas FFLs of type (+ − / −) and (− + / −) should drive the production of EPs that are active in G1 phase of the cell cycle, when Cdk activity is low. Incoherent FFLs of type (− + / +) and (+ − / +) are expected to drive the production of EPs that are active at the G1-to-S phase transition, when Cdk activity is abruptly increasing; whereas FFLs of type (+ + / −) and (− − / −) should drive the production of EPs that are active at the M-to-G1 phase transition, when Cdk activity is abruptly decreasing. These expectations are fulfilled for several EPs for which we know their roles and the signs of steps 1, 2, and 3. Our theory makes many predictions about regulatory triads (Cdk–TF–EP) for which we do not yet have complete information. If future experiments confirm these predictions, then we will have good reason to believe that the signal transduction properties of FFLs explain how one (or a few) Cdk signal(s) can drive a host of cell cycle responses in correct temporal sequence. This work, which appeared recently in *Molecular Systems Biology*, resulted from a collaboration of scientists in Hungary, Italy, Denmark, England, and the Virginia Bioinformatics Institute (Csikasz-Nagy et al, 2009).

**Conferences and workshops**


Tyson J (2008-2009) Invited speaker at Rice University, Purdue University, Columbia University Medical School, Albert Einstein College of Medicine, University of Chicago.


Peer-reviewed publications in reporting period


Other cited references
