Virginia Bioinformatics Institute at Virginia Tech

Scientific Annual Report 2006

Virginia Bioinformatics Institute
Washington St. (0477)
Blacksburg, VA 24061
ph: 540.231.2100
Fax: 540.231.2606
email: info@vbi.vt.edu
web: www.vbi.vt.edu

Table of Contents

Introduction

Research reports — VBI Faculty

Research reports — VBI Fellows

Introduction

Dear friends,

The Virginia Bioinformatics Institute (VBI) has made considerable progress over the past year. It is therefore gratifying to see the diligent work of faculty, staff and collaborators represented in this our second scientific annual report. At VBI, one of our primary objectives is to channel scientific innovation into transdisciplinary research projects that emphasize the importance of team-based science. One of the goals of the scientific annual report is to highlight the very latest research accomplishments of the Institute that have arisen from this collaborative approach to science.

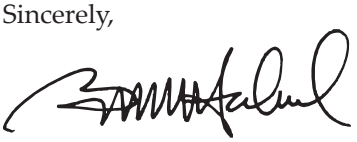
Work at VBI spans many disciplines but has at its foundation the sciences of biology, mathematics and information technologies. Faculty at VBI use a wide variety of methods and tools that range, for example, from simulation and modeling of biological networks, statistical genetics and large-scale comparative genomics to computational systems biology, microfluidic mass spectrometry and different cyberinfrastructure-based resources. These tools and methods are being used to investigate plant and animal pathogens, make advances in infectious disease research and to better understand a wide range of biological, information, social and technological systems.

In this report, readers will not only be able to see the diversity of research projects underway at VBI but also the many synergies that exist between different projects and collaborators. At VBI, we benefit greatly from collaborations with other departments on the Virginia Tech campus as well as other leading national and international universities and institutes. The projects described here show how these relationships are flourishing and bringing new perspectives to our scientific undertakings. Reports from two of VBI's faculty fellows – Virginia Tech faculty who work closely with VBI researchers – specifically describe projects involving computational systems biology as well as the simulation and analysis of molecular regulatory systems in cell biology.

The work described in this report reflects our willingness to undertake team-based scientific initiatives directed at some of the key research challenges facing society today. It is my sincere hope that the research described here will serve as a catalyst for future work and further groundbreaking collaborations.

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

Sincerely,



Bruno Sobral
Executive and Scientific Director
Virginia Bioinformatics Institute

2006 Research Reports

from the Faculty at the Virginia Bioinformatics Institute

Network Dynamics and Simulation Science Laboratory

Christopher L. Barrett, cbarrett@vt.edu
Professor, Virginia Bioinformatics Institute & Computer Science, Virginia Tech

The Network Dynamics and Simulation Science Laboratory is pursuing an advanced research and development program for interaction-based modeling, simulation, and the associated analysis, experimental design, and decision support tools for understanding large biological, information, social, and technological systems. Extremely detailed, multi-scale computer simulations allow formal and experimental investigation of these systems. The need for such simulations is derived from questions posed by scientists, policy makers, and planners involved with very large complex systems. The theoretical foundations of our work are rooted in the concept of interaction-based computing and discrete dynamical systems. We are currently pursuing projects in the following programmatic areas: integrated high-performance simulation and data service architectures; human population dynamics and associated social networks in urban environments and at the national scale; epidemiology and the spread of infectious diseases; computational and behavioral economics and commodity markets; next generation computing and telecommunication systems; and computational systems biology.

Keywords: interaction-based computing, simulation, and modeling; infectious diseases and public health; grid computing and service-oriented architectures; activity-based models of urban regions; population dynamics; social networks; complex systems; discrete dynamical systems; computational complexity; algorithmic semantics.

Scientific and technical progress

Significant progress has been made in achieving important programmatic goals during the last twelve months. Here we highlight some of the achievements. In the reporting period, we have developed version 1.0 of Simfrastructure, a service- and grid computing-oriented modeling tool for socio-technical, biological, and information systems. We have also developed version 1.0 of Simdemics, a scalable high-performance computing-based service environment for general reaction diffusion systems. Other milestones include the successful development of scalable algorithms for simulating epidemics and other reaction diffusion systems. We have now generated a synthetic population consisting of 250 million individuals endowed with daily activity patterns where the activities are performed at real locations. In addition, the EpiSims (Epidemiological Simulation System) project has progressed to represent various disease outbreak interventions. The tool is being used by the National Institutes of Health (NIH) Models of Infectious Disease Agent Study to support pandemic preparedness.

For applications in the fields of commodities and energy markets, we have developed version 1.0 of Sigma (Simulation of Generic Markets), a service-oriented modeling architecture for analyzing large commodity and energy markets. We have also enhanced TRANSIMS (Transportation Analysis Simulation Systems) to support the U.S. Department of Transportation sponsored transport planning study for

Contributors:
Karla Atkins, Richard Beckman, Keith Bisset, Jiangzhuo Chen, Stephen Eubank, V. S. Anil Kumar, Achla Marathe, Madhav V. Marathe, Henning Mortveit, Julia Paul, Paula Stretz

Student contributors: Kevin Allen, Deepti Chafekar, Abhijit Deodhar, Aseem Deshpande, Subodh Lele, Bryan Lewis, Farid Merchant, Kashmira Phalak, Sameer Tupe

the Washington DC region. In addition, we participated in a team effort to implement a TRANSIMS-based traffic simulation on reconfigurable computing hardware.

Our work has also comprised advanced computer and mathematical research in the area of high-performance computing, discrete dynamical systems, and communication networks. Notable achievements in those areas include the following:

- First provable multi-criteria approximation algorithms for scheduling on unrelated parallel machines
- First algorithmic results for approximately computing the capacity of arbitrary wireless networks
- Extension of computational and mathematical theory of sequential dynamical systems

Programmatic progress

The work of the Network Dynamics and Simulation Science Laboratory (NDSSL) has resulted in progress in several key program areas. First, we have established two epidemiological simulations for projects for situation assessment and response analysis at the Defense Threat Reduction Agency. Second, our group has participated as a partner institution and co-principal investigator in the Centers for Disease Control Center of Excellence in Medical Informatics, which is led by the University of Utah Medical School. Third, we have established a project to develop a comprehensive national incidence characterization and management system. We have also continued to contribute as a principal institution in the NIH MIDAS (Models of Infectious Disease Agent Study) project.

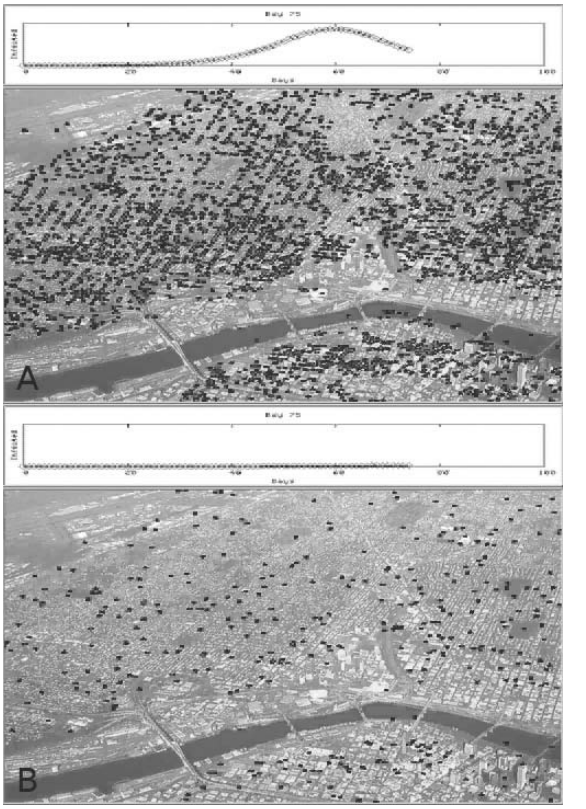


Fig.1. Example of Simfrastructure coupled with Episims simulating two outcomes of an avian flu outbreak in Portland, OR, using synthetic individuals. (A) The number of people by location infected on a given day when no interventions were enacted. In this base case, approximately 630 000 individuals, or 39% of the population, were infected by day 100. (B) The effect of intervention by sequestration of 75% of the population. In this case, 47 610 individuals, or 3% of the population, were infected by day 100.

National and international leadership

The work of the NDSSL has also involved extensive interactions and leadership roles at the national and international levels. We have been advising an NIH steering committee for developing NIH modeling methods for contagious diseases. We have also been advising the U.S. Department of Health and Human Services and the White House on approaches for pandemic planning. Our group has also served on the U.S. Northern Command coordinating committee for integration of models among NIH, U.S. Department of Defense, and U.S. Department of Homeland Security for crisis planning and management. It has also served on a European Union program review and planning committee on complexity science.

In the reporting period, I have represented the NDSSL group as an invited speaker at several conferences, including the Dagstuhl Workshop on Algorithmic Aspects of Large and Complex

Fig. 2. Fuel tank study, Washington DC area.

Networks, the Society for Industrial and Applied Mathematics Data Mining Workshop on Spatial Data Mining: Consolidation and Renewed Bearing, the Massachusetts Institute of Technology seminar series on social networks and epidemics and the Distinguished speaker series in Crisis Management and Public Health at the University of Utah Medical School. In addition, I was the plenary speaker at the Canadian National Research Council, Canadian Congress on Computing, Social Sciences and Humanities. In the reporting period, I also served on the External Advisory Board for the National Center for Advanced Secure Systems at the National Center for Supercomputing Applications and represented the efforts of the NIH/National Institute of General Medical Science in social behavior modeling to the United States Congress.

Outreach and education

Two new Virginia Tech courses were introduced by NDSSL members. They include Mathematics of Simulation, an undergraduate course in the Mathematics Department at Virginia Tech, and Randomized Algorithms, a graduate course in the Computer Science Department. We have also organized and lead an invited two-week course in complexity science (Complexity in Real World Systems and their Simulations) sponsored by the European Union EXYSTENCE complex systems program and the ERC. The venue for the course was the Institute for Scientific Interchange Foundation, Torino, Italy.

In the last twelve months, two new seminar series were started: an interdisciplinary seminar series on Critical Infrastructures, organized jointly with the Department of Electrical and Computer Engineering at Virginia Tech, and The Science of Complex Networks.

Publications

Arciniegas I, Marathe A (2005) Important variables in explaining real-time peak price in the independent power market of Ontario, *Utilities Policy* 3(1), pp. 27-39.

Barrett C, Anil Kumar VS, Marathe M, Thite S, Istrate G (2006) Strong edge coloring for channel assignment in wireless radio networks. In P*roceedings of the First I.E.E.E. International Workshop on Foundations and Algorithms for Wireless Networking (FAWN'06)*. March 13, pp 106-110. IEEE Computer Society Press.

Barrett C, Eidenbenz S, Kroc L, Marathe M, Smith J (2005) Probabilistic multi-path vs. deterministic single-path protocols for dynamic ad-hoc network scenarios. In *Proceedings of the 2005 ACM Symposium on Applied Computing*. pp. 1166-1173. New York, NY: ACM Press.

Barrett C, Eidenbenz S, Kroc L, Marathe M, Smith J (2005) Parametric probabilistic routing in sensor networks. *ACM/Baltzer J. Mobile Networks and Applications (MONET)* **10(4)**: 529-544.

Bisset K, Eubank S, Marathe M, Mortveit M (2005) The design and implementation of Simdemics. *NDSSL Technical Report*. **05-017**.

Bisset K, Atkins K, Barrett C, Beckman R, Eubank S, Marathe M, Marathe A, Mortveit H, Stretz P, Anil Kumar VS (2006) Synthetic data products for societal infrastructures and proto-populations: Data Set 1.0. *NDSSL Technical Report*. **06-006**: January 24.

Chen J, Liu JX, Noubir G, Sundaram R (2005) Minimum energy accumulative routing in wireless networks. *Proc. IEEE INFOCOM* **3**: 1875-1886. doi: 10.1109/INFCOM.2005.1498466.

Hunt H III, Marathe M, Stearns R, Rosenkrantz D (2005) Towards a predictive complexity theory of periodically specified problems: A survey. In *Computational Complexity and Statistical Physics*, Moore C, Istrate G, Percus A (eds) pp 285-318. Oxford: Oxford University Press.

Hansson A, Mortveit H, Tripp J, Gokhale M (2005) Urban traffic simulation modeling for reconfigurable hardware. *Industrial Simulation Conference (ISC05)* pp 291-298. Berlin.

Hansson A, Mortveit A, Reidys C (2005) On asynchronous cellular automata. *Advances in Complex Systems* **8**: 521-538.

Istrate G, Hansson A, Marathe M, Thulasidasan S, Barrett C (2006) Semantic compression of TCP traces. In *Proceedings of the IFIP NETWORKING Conference*, Boavida F, Plagemann T, Stiller B, Westphal C, Monteiro E (eds) pp 123-135. Lecture Notes in Computer Science, vol. 3976, Berlin/Heidelberg: Springer Verlag.

Anil Kumar VS, Marathe M, Parthasarathy S, Srinivasan A (2005) Algorithmic aspects of capacity in wireless networks. In *Proceedings of the 2005 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems* **33 (1)**:133-144.

Anil Kumar VS, Marathe M, Parthasarathy S, Srinivasan A (2005) Approximation algorithms for scheduling on multiple machines. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05)* pp 254-263, October.

Anil Kumar VS, Marathe M, Parthasarathy S, Srinivasan A, Zust S (2005) Provable algorithms for parallel sweep scheduling for unstructured meshes. In *19th International Parallel and Distributed Processing Symposium (IEEE IPDPS)* p 26.

Anil Kumar VS, Marathe M, Parthasarathy S, Srinivasan A (2005) Scheduling on unrelated machines under tree-like precedence constraints. *8th International Workshop on Approximation Algorithms for Combinatorial Optimization Problems* pp 146-157, 22-24 August 2005, UC Berkeley.

Lloyd E, Liu R, Marathe M, Ramanathan R, Ravi S (2005) Algorithmic aspects of topology control problems for ad hoc networks, *ACM/Baltzer J. Mobile Networks and Applications (MONET)* **10**: 19-34.

Tripp J, Hansson A, Mortveit H, Gokhale M (2005) Metropolitan road traffic simulation on FPGAs. 13th *IEEE Symposium on Field-Programmable Custom Computing, Machines* pp 117-126.

Toroczkai Z, Eubank S (2005) Agent-based modeling as a decision making tool: How to halt a smallpox epidemic. *The Bridge, Eleventh Annual Symposium on Frontiers of Engineering, organized by the National Academy of Engineering*. September 22-24, 2005, Niskayuna, NY, **35(4)**: 22-27.

Tripp J, Hansson A, Gokhale M, Mortveit H (2005) Partitioning hardware and software for reconfigurable supercomputing applications: A case study. In *Proceedings of the 2005 ACM/IEEE Conference on Supercomputing*. November 12-18, p 27, Seattle, Washington: IEEE Computer Society.

Bioinformatics inspired by a tree

Allan W. Dickerman, dickerman@vt.edu
Assistant Professor,Virginia Bioinformatics Institute

The concept that an evolutionary history of common ancestry unites all living organisms was extremely productive for our understanding of biology in the pre-genomic era. Now we are in a position to extend and amplify that lesson at the levels of genes, chromosomes and whole genomes for the many organisms that are fully known at the sequence level. Our group is pursuing a large-scale analysis method to describe genome-wide assemblies of phylogenetic models for individual sequence regions, focusing first on protein-coding genes. This "GeneTrees" database provides a basis for collaboration with external workers with interests in various taxonomic groups. Opportunities to apply phylogenetic principles to genomic and biological problems have brought termite endosymbionts, plant pathogens, and other diverse systems into the group portfolio. In another area, our work on bioinformatics, transcriptomics and transgenic analysis of essential gene functions in the plant *Arabidopsis thaliana*, under the "SeedGenes" project, is approaching completion. Integration of these diverse themes in a common phylogenomic perspective remains the long-term goal.

Keywords: phylogenomics; evolutionary biology; *Arabidopsis* embryo development; plant pathogens; diagnostic microarray.

GeneTrees phylogenomics infrastructure

A primary interest of the group is to develop the analysis infrastructure needed to describe the history of common ancestry for all components of all genomes in a particular sphere of interest. The focus is to explore and leverage those cases where explicit phylogenetic modeling provides crucial information not available at the level of simple pairwise comparisons typical of much comparative genomics. Our current activities consist of developing the infrastructure and populating the "GeneTrees" database with the raw material for further investigations. Several specific applications of the database are discussed here. Development of the GeneTrees database has been done with extensive input from Yuying Tian of the VBI Cyberinfrastructure Group.

Our current database and analysis system focuses on phylogenetic models of alignable protein sequences. Using protein sequences confers simplicity by forcing all sequences into a coherent orientation, which can be missing in DNA alignments, where repeats, inversions, and other patterns lead to very confusing tangles. Protein coding regions tend to be conserved enough to be recognizable at the moderate-to-deep evolutionary distances that characterize most comparisons among model species. A set of automated scripts has been developed that will take all predicted protein sequences for a given set of genomes and infer a set of homology groups consisting of protein regions that participate in reasonable multiple sequence alignments. Parameterized criteria determine the stringency for including or excluding proteins from homology sets. This process starts with all-versus-all pairwise alignment summaries and mines these connections for dense clusters using heuristic methods. These dense clusters form seed sequences for an initial alignment which is then used to build a hidden Markov model (HMM). The HMM is used to evaluate all candidate members based on pairwise similarity. This is a rough description of the process for finding coherent multiple sequence homology groups that we have been

Contributors:
Johanna C. Craig, Elena Shulaeva, Eric K. Nordberg, Marc L. Fisher, Ruth Howe

Table 1. Number of entities in the three largest components of the GeneTrees database.

Database name	Genomes	Sequences	Homology groups
prokaryote2	325	890 081	20 985
viral3	1606	45 749	4752
eukaryotes	21	375 204	18 353

evolving over the last few years. In the past year, we have started using the program MrBayes (Ronquist et al, 2003) almost exclusively for phylogenetic inference. This procedure has the advantage that it uses a Markov chain Monte Carlo (MCMC) method to successively refine estimates of parameters of the likelihood model so that they need not be specified *a priori*.

We currently have three well-developed databases focusing on prokaryotic genomes, viral genomes, and a set of fully sequenced eukaryotes. The size of the databases is summarized in Table 1. We have recently completed an incremental update function that allows us to add new genomes to an existing database. Applying this update in March 2006 nearly doubled the size of the prokaryote2 database from 184 to 325 genomes.

A recent innovation to GeneTrees is a feature that allows the user to "pull out" coherent subtrees from large trees. This produces trees with fewer deep gene duplications: each branch of a major duplication is pulled out as a separate gene phylogeny. This tidies up large, unwieldy trees to focus on more strictly orthologous subtrees. Including in the database trees that are subsets of larger homology groups poses the challenge of exposing to the user the nature of relationships among homology groups, such as how many sequences are shared and whether one is a proper subset of the other. Presenting these patterns in a manner useful for browsing is an area of future development. We are moving to put the web interface to this database online for external access.

Alpha-proteobacteria multi-locus phylogeny

The occasion of the Virginia Bioinformatics Institute (VBI) symposium on the alpha-proteobacteria on April 26-28 prompted a research effort to build a strongly supported phylogeny for the fully sequenced genomes. This idea was proposed by Dr. Kelly Williams of Dr. Sobral's CyberInfrastructure Group, who found that there were 68 genomes in the group completed or near completion. He proposed applying a multi-locus phylogenetic analysis to obtain a tree representing the largest number of species studied to date in this way for the group. The results would help in the analysis of the distribution of a peculiar trait of the histidine amino-acyl synthetase discovered by Dr. Williams and the subject of his presentation to the alpha-proteobacteria symposium. We mined the GeneTrees database for genes that looked to be single-copy in the 14 species of alpha-proteobacteria contained in the database at the time, resulting in over 300 candidate genes. We then identified the homologs of each in the full set of 68 genomes. After filtering genes without suitable representation, we obtained a set of 115 genes found in all or nearly all 68 alpha-proteobacterial genomes plus 10 outgroup species. By selecting only the strongly aligning regions of these 115 multiple-sequence alignments (omitting gapped and highly variable regions), we obtained a high-quality dataset of over 33 000 aligned amino acid positions. Tree-building on this dataset with MrBayes (Ronquist et al, 2003) yielded a single topology, which is unusual as the MCMC method usually samples a large number of topologies within the region of high likelihood. This appears to be a property of the very large number of characters in our data set and suggests we have attained sampling that is extensive enough to converge on a very well-supported model.

Arabidopsis SeedGenes project

The SeedGenes project, of which Dr. David Meinke of Oklahoma State University is the principal investigator, concluded its funding in the spring of 2006. The website describing all proven early developmental lethal mutations in *Arabidopsis* is active and will remain so for the

immediate future. We are awaiting a decision on the renewal proposal.

Dr. Johanna Craig is finishing up the data analysis on the transcriptomics experiments we performed on early developmental seeds (globular to heart stage). A crucial aspect of our experimental design was to compare wild-type seeds to a mutant characterized by having no visible embryo. Genes under-expressed in the mutant relative to the wild-type are presumptively involved in embryo-specific functions. This helps dissect embryo biology from the remaining tissues such as seed coat and endosperm that predominate at the earliest embryonic stages. Comparison of the 16 Affymetrix GeneChips® from our experiments to the large body of expression data we obtained from the Nottingham Arabidopsis Stock Centre (www.nasc.org) has revealed a very obvious pattern of genes that are expressed strongly in seed development, beginning after flowering and ending before or at dessication of the silique. Of this set, genes with higher expression in wild-type versus the embryo-minus mutant are of particular interest.

Ms. Elena Shulaeva has been very productive bringing the final experimental effort begun under the SeedGenes project to completion. This is a promoter-reporter transformation project targeting some of the early seed-specific genes. Regions of *Arabidopsis* genes covering approximately 1500 base pairs upstream of the start of translation were amplified by PCR and cloned in front of a green fluorescent protein (GFP) in a transformation vector. Transformation into plants will occur over the summer of 2006. This project has served as a training project for Ms. Ruth Howe, who graduated from Blacksburg High School as co-valedictorian in June, 2006. In the fall, Ruth will attend Washington University in St. Louis. She intends to study molecular engineering.

A microarray for plant pathogen identification

We received news in early 2006 that our proposal to the United States Department of Agriculture (USDA) to design an Affymetrix microarray to detect and identify pathogens of plants had been approved. Dr. Chris Lawrence of VBI is co-principal investigator and Dr. Stephen Goodwin of the USDA and Purdue University is a funded collaborator. The project will begin with extensive analysis of all available fungal and bacterial ribosomal DNA (rDNA) sequences to predict suitable combinations of 25-mer probes. The plan is to design thousands of probes to span the phylogenetic diversity of known plant pathogens in the bacteria and fungal groups, with some representation of viral sequences. The design will balance representation of probes to relatively conserved as well as highly variable regions of the rRNA genes. This should result in a chip that can identify novel species, or at least position them roughly in a phylogenetic context, as well as recognize known pathogens with robust support. This should provide a chip that can be used in diverse applications. Work on developing protocols for labeling total RNA, including bacterial and fungal ribosomal RNA (rRNA) will be performed by Elena Shulaeva. Tissue from infected plants will be provided by the laboratories of Drs. Chris Lawrence and Stephen Goodwin. Testing will initially weed out probes with non-specific binding (false-positives). Later in the three-year project we will carefully analyze the sensitivity and specificity for particular serious pathogens.

Termite endosymbiotic bacteria

Marc Fisher received his Ph.D. in May for work on the endosymbiotic bacteria of termites. Drs. Dini Miller and Carlyle Brewster of the Department of Entomology at Virginia Tech were his co-advisors. A significant part of Dr. Fisher's research was carried out at VBI under my supervision. Fisher extracted total RNA from the gut lumen of the termite *Reticulitermes flavipes*. With help from Elena Shulaeva, Fisher used previously published conserved, prokaryote-specific, primers to amplify an approximately 1400-bp region of the small-subunit ribosomal RNA gene by polymerase chain reaction (PCR). This PCR product consisted of a diverse set of RNA products from the various prokaryotes in the termite gut. This diverse primary PCR product was cloned into a simple cloning vector and transformed into *Escherichia coli*. From there, we pursued two complementary approaches to studying the diversity and identity of the organisms represented. The first was by sequencing 42 rDNA inserts by

the VBI Core Laboratory Facility and at the GeneLab/BioMed facility within the Louisiana State University School of Veterinary Medicine (Baton Rouge, Louisiana). The other was fingerprint analysis of restriction fragments, or ARDRA (amplified rDNA restriction analysis). A total of 512 clones was analyzed from which 261 different ARDRA profiles were found. Representatives were found from six major bacterial phyla - Proteobacteria, Spirochaetes, Bacteroidetes, Firmicutes, Actinobacteria, and the recently proposed Endomicrobia. Fisher applied ecological diversity models to estimate that the expected number of different ribotypes existing in this bacterial ecosystem is between 600 and 1300.

Computational approaches to inferring gene functions

Regular meetings were held throughout the period with Drs. T. M. Murali of the Department of Computer Science at Virginia Tech, Brett Tyler of VBI, and Ph.D. student Eric Nordberg. The subject was how to utilize high-throughput biological data to best infer gene functions. A key focus was Dr. Murali's graphical approach embodied in his program 'GAIN'. We believe our work has pushed the theoretical basis of these algorithms to new levels, applying Bayesian principles and biological insight. These sessions also led to work on the phylogenetic study of the genomic pedigree of the organismal group *Phytophthora* (of the chromalveolate group) in collaboration with Dr. Brett Tyler's group at VBI. In particular, we addressed whether one can observe remains of a photosynthetic genome in this obligate heterotroph, as would be expected from the widespread presence of autotrophy amongst other chromalveolates. This work found mention, briefly, in the publication of the *Phytophthora sojae* and *P. ramorum* genomes (Tyler et al, in press).

References

Fisher ML (2006) Comparison of subterranean termite (Rhinotermitidae: Reticulitermes) gut bacterial diversity within and between colonies and to other termite species using molecular techniques (ARDRA and 16S rRNA gene sequencing). Dissertation, Virginia Polytechnic Institute and State University.

Karaoz U, Murali TM, Letovsky S, Zheng Y, Ding C, Cantor CR, Kasif S (2004) Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc. Natl. Acad. Sci. USA*. **101**(9): 2888-2893.

Ronquist F, Huelsenbeck JP (2003) MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**:1572-1574.

Publications

Tyler BM, Tripathy S, Zhang X, Dehal P, Jiang RHY, Aerts A, Arredondo F, Baxter L, Bensasson D, Beynon JL, Chapman J, Damasceno CMB, Dickerman A, Dorrance AE, Dou D, Dubchak I, Garbelotto M, Gijzen M, Gordon S, Govers F, Grunwald N, Huang W, Ivors K, Jones RW, Kamoun S, Krampis K, Lamour K, Lee MK, McDonald WH, Medina M, Meijer HJG, Nordberg E, Maclean DJ, Ospina-Giraldo MD, Morris P, Phuntumart V, Putnam N, Rash S, Rose JKC, Sakihama Y, Salamov A, Savidor A, Scheuring C, Smith B, Sobral BWS, Terry A, Torto-Alalibo T, Win J, Xu Z, Zhang H, Grigoriev I, Rokhsar D, Boore J (2006) *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science*, (in press)

Zhao C, Craig JC, Petzold HE, Dickerman AW, Beers EP (2005) The xylem and phloem transcriptomes from secondary tissues of the *Arabidopsis* root-hypocotyl. *Plant Physiol.* **138**(2): 803-818.

Genetic architecture of quantitative traits

Ina Hoeschele, inah@vt.edu
Professor, Virginia Bioinformatics Institute & Statistics, Virginia Tech

The common theme of our research is the desire to understand how the joint action and interaction of multiple genes determines complex phenotypes of interest. We have analyzed a factorial microarray experiment focused on evaluating soybean cultivars differing in quantitative resistance to the *Phytophthora sojae* pathogen. Different methods were evaluated for high-dimensional mapping of expression quantitative trait loci. A structural equation model is being implemented for gene network inference based on genetical genomics experiments that provide causal inference and strong constraints on network topology. We have evaluated our new method for haplotype reconstruction in the context of joint linkage and linkage disequilibrium mapping in complex human pedigrees. A Markov chain Monte Carlo method has been investigated for the estimation of genetic parameters and for marker-assisted selection of radiological health traits of the limbs of riding horses.

Keywords: statistical genetics; quantitative trait locus mapping; genetical genomics; statistical design and analysis of microarray expression experiments; gene network inference.

Introduction

Statistical genetics includes the search for and characterization of genes affecting human health and economic traits of plants and animals, the evolution of genes in natural populations, the evolution of genomes and species, the analysis of DNA, RNA and protein sequence and structure, as well as statistical design and analysis of genomic, transcriptomic, proteomic, and metabolomic experiments. In the field of statistical genetics, our group contributes to (i) the statistical design and analysis of 'omics' experiments, in particular experiments with multi-factorial treatment and covariance structures, (ii) the identification of genes and gene (interaction) networks influencing quantitative traits of economic importance in animals and plants or complex diseases in animal models and humans, and (iii) the development and application of statistical and computational methods for genetic analyses of high-dimensional phenotypes in genetical genomics and systems genetics.

Contributors:
Chiranjeet Chetia, Alberto de la Fuente, Guimin Gao, Bing Liu, Yongcai Mao, Kathrin Friederike Stock

Analysis of factorial microarray experiment

We have analyzed a factorial Affymetrix GeneChip® microarray experiment performed in Dr. Brett Tyler's laboratory at the Virginia Bioinformatics Institute (VBI). Eight soybean cultivars differing in quantitative disease resistance were expression profiled using a design with two time points for post-infection, mock and pathogen-inoculated plants, and four overall experimental replicates, with a total of 128 GeneChips. After the elimination of genes determined as not expressed in the factorial experiment, background correction as implemented in the GC robust multiarray average (found to be superior to other methods), and quantile normalization, we performed gene- and probe-level linear mixed model analyses (LMMA). Using a False Discovery Rate controlling procedure, we determined which genes had significant effects of cultivar, infection, time and their interactions. While for most genes the results from gene- and probe-level LMMA were in good agreement, there were differences for some genes with many probe outlier values. Including probe-by-cultivar and probe-by-cultivar-by-time interactions in the model improved agreement and increased the power of the probe-level analysis.

High-dimensional expression quantitative locus mapping

This work is motivated by a large genetical genomics data set being created in Dr. Brett Tyler's laboratory at VBI in collaboration with Dr. Saghai Maroof (Crop & Soil Environmental Sciences, Virginia Tech) and Ohio State University researchers. A population of 300 recombinant inbred lines is being phenotyped for quantitative disease resistance, expression profiled and genotyped for DNA markers on a genome-wide scale. In preparation for the analysis of this data set, we are implementing, developing, and comparing methods for the identification of expression Quantitative Trait Loci (eQTL). QTL mapping can be performed on the expression profiles (etraits) to identify QTL regions influencing the expression of certain genes. An eQTL that influences the expression of a gene or genes not located in its QTL region is termed a *trans*-eQTL, while a *cis*-eQTL influences the expression of at least one gene located in its region. The current approach of genome-wide eQTL analysis of each etrait separately using standard QTL mapping methods for individual etraits does not have optimal power and precision. Our approach is to explore the use of dimension reduction, the search for *cis*-linkages only at the genomic location of each gene, and the search for *trans*-QTL by simultaneously fitting a regulatory gene candidate and its closest marker with simulated data and a yeast segregant population (Liu, de la Fuente and Hoeschele, unpublished results). *Cis*- and *trans*-mapping and Principal Component Analysis (PCA) mapping considerably increase the power and resolution of eQTL analysis over standard single etrait eQTL mapping.

Gene network reconstruction via structural equation modeling

Gene network reconstruction is based on genetical genomics experiments that provide causal inference and strong constraints on network topology. If gene A is located in QTL region B that influences the expression of gene C, then A is considered a direct or indirect candidate regulator of target gene C. Based on the QTL mapping results consisting of a list of candidate and a list of target genes for each eQTL, we have provided a statistical approach for evaluating candidate regulators that improves on our previous method (Bing & Hoeschele, 2005), and we have constructed an encompassing (partially) directed network (EDN) including both gene and QTL nodes. The EDN contains directed links from each *trans*-eQTL and its candidate regulator(s) to the targets, a directed link from each *cis*-eQTL to its *cis*-linked gene(s), and directed links from its *cis*-linked gene(s) to its targets. This EDN contains only those regulatory relationships where the regulator is polymorphic. Some non-polymorphic regulators can also be included based on partial correlation analysis. The EDN is sparsified via Structural Equation Modeling (SEM) being implemented in Maximum Likelihood and Bayesian frameworks (Liu, de la Fuente and Hoeschele, unpublished results). While Bayesian network analysis is a popular tool for gene network inference, SEM does not require discretization of data and accommodates cycles or feedback loops. The method is currently being tested on artificial genetical genomics experiments and a yeast data set.

Joint linkage and linkage disequilibrium mapping in pedigrees

Identity-by-descent (IBD) matrix calculation is an important step in QTL analysis using variance component models. To calculate IBD matrices efficiently for large pedigrees with large numbers of loci, an approximation method based on the reconstruction of haplotype configurations for the pedigrees was proposed (Gao & Hoeschele, 2005). The new method was compared with a Markov chain Monte Carlo (MCMC) method (Loki) in terms of QTL mapping performance on simulated pedigrees. Both methods yield almost identical results for the estimation of QTL positions and variance parameters, while the new method is much more computationally efficient than the MCMC approach for large pedigrees and large numbers of loci. The proposed method uses a subset of as little as 50 haplotype configurations, but it produces QTL mapping results that are essentially identical to those obtained by computing the IBD matrix using the MCMC algorithm in Loki with 100 000 iterations. The proposed method was also compared with an exact method (Merlin) in small simulated pedigrees, where both methods produce nearly

identical estimates of position-specific kinship coefficients. The new method can be used for fine-mapping with joint linkage disequilibrium and linkage analysis and has been implemented in a C++ program available for academic research.

Genetic analyses of radiological health traits in riding horses

Because strength and soundness of the equine locomotory system are of great importance in all sectors of the horse industry, inclusion of radiological health traits of the limbs in current breeding schemes of the Warmblood riding horse has been suggested. We have therefore evaluated a Bayesian analysis of a multivariate, generalized linear mixed model including discrete health traits and a correlated continuous trait, incorporating DNA marker data, and implemented via a MCMC algorithm. Evaluation was based on accuracy of genetic parameter estimation dependent on population and data structure, and choice of priors in the Bayesian analysis. Efficiency of marker-assisted selection for radiological health traits was also evaluated (Stock, Distl and Hoeschele, unpublished results).

Current and future directions

Over the next year, a major focus will be the analysis of the soybean genetical genomics experiment ongoing in Dr. Brett Tyler's laboratory at VBI. Over the next five years, another focus will be the continued implementation and evaluation of SEM (and possibly other methods capable of modeling feedback control) for gene network inference based on genetical genomics experiments, and its application to the soybean genetical genomics experiment with incorporation of causal relations among gene expressions and disease phenotypes and host-pathogen interactions. This generalizes to the joint identification of causal relationships among DNA variants, transcripts and component traits of complex diseases. Such an approach should significantly strengthen our ability to investigate the complex interactions among DNA variants, disease and environment. A further generalization is the integration of proteomics and metabolomics data.

Conditional on the outcome of current funding applications, we expect to initiate work on two collaborative projects focusing on a mouse model for lung cancer developed at Wake Forest University Medical School (Winston-Salem, NC). The first project uses genetic linkage analysis of an Advanced Intercross (AI) to identify candidate genes responsible for the observed phenotypic variance in lung tumor incidence of adult mice following *in utero* exposure to chemical carcinogens. Currently, AI populations are not analyzed correctly by the existing software packages for inbred line crosses and their use leads to incorrect significance results and power calculations. To account for the unequal correlation structure of the AI, it is necessary to extend current packages to mixed models, while maintaining computational efficiency. The second project is a microarray gene expression experiment comparing two different tumor types corresponding to early and late stages of tumorigenesis to identify differentially expressed genes and pathways that are activated in the conversion from early to late stage. We will apply model-based and probabilistic graph-based methods for pathway inference to infer a common network and test for differences between the two groups.

Publications

Betthauser JM, Pfister-Genskow M, Xu H, Gouleke PJ, Lacson JC, Koopang RW, Liu B, Hoeschele I, Eilertsen KJ, Leno GH (2006) Nucleoplasmin facilitates reprogramming and in vivo development of bovine nuclear transfer embryos. *Molecular Reproduction and Development* advanced online publication; doi 10.1002/mrd.20493

Bing N, Hoeschele I (2005) Genetical genomics analysis of a yeast segregant population for transcription network inference. *Genetics* **170**: 533-542.

Gao G, Hoeschele I (2005) Approximating identity-by-descent matrices using multiple haplotype configurations on pedigrees. *Genetics* **171**: 365-376.

Modeling and simulation of biochemical networks

Reinhard Laubenbacher, reinhard@vbi.vt.edu
Professor, Virginia Bioinformatics Institute & Mathematics, Virginia Tech

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

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

Introduction

"All processes in organisms, from the interaction of molecules to the complex functions of the brain and other whole organs, strictly obey [these] physical laws. Where organisms differ from inanimate matter is in the organization of their systems and especially in the possession of coded information" (Mayr, 1988). It is the task of systems biology to elucidate those differences. At the molecular level, part of this task focuses on the discovery of biochemical networks, from gene regulatory networks to protein and metabolic networks. The goal is to use suitably designed system-level experiments, such as the collection of time courses of DNA microarray measurements, to make mathematical models of the dynamic networks that underlie the measurements, without biasing the discovery process through the use of prior assumptions about the network. This methodology is becoming known as "top-down" modeling. The Applied Discrete Mathematics Group at the Virginia Bioinformatics Institute (VBI) is focused in part on the development and application of tools from discrete mathematics for this purpose.

The philosophy underlying our approach has been characterized by R. Karp as follows: "It appears that the transcriptional control of a gene can be described by a discrete-valued function of several discrete-valued variables. […] A regulatory network, consisting of many interacting genes and transcription factors, can be described as a collection of interrelated discrete functions and depicted by a "wiring diagram" similar to the diagram of a digital logic circuit" (Karp, 2002). We are developing computational methods to reverse-engineer these discrete functions from state transition measurements of all network variables obtained by perturbing the network experimentally. The principal experimental driver of the project is the study of the oxidative stress response network in *Saccharomyces cerevisiae*. In collaboration with the groups of Drs. Mendes and Shulaev at VBI, and supported by National Institutes of Health (NIH) grant Nr. RO1 GM068947-01, we have worked to integrate the modeling and data generation processes. The experimental part of this yeast systems biology project is described in more detail in the report of the Shulaev laboratory (see pp. 66-71); the Mendes report contains further information on the joint

Contributors:
Applied Discrete Mathematics Group: Miguel Colon-Velez, Edgar Delgado-Eckert, Elena Dimitrova, Abdul Salam Jarrah, John McGee, Brandilyn Stigler, Alan Veliz-Cuba, Paola Vera- Licona, Dedra Wright

modeling project (see pp. 35-40). Another application of our mathematical tools involves the analysis of agent-based models of immune response to Epstein-Barr virus infection, partially supported by NIH grant Nr. RO1 AI062989-01. This project is not described here in detail due to space limitations. The major focus of the project is on the approximation of stochastic interaction-based models by deterministic state space models. The mathematical underpinning of our modeling methods is being developed in part with support from National Science Foundation (NSF) grant Nr. DMS-0511441.

Results

In the course of the last year, our group has made significant progress on the theoretical aspects of our method, as well as on the software design front. The ultimate goal of this project is to make available a comprehensive, user-friendly software package for the reverse-engineering of biochemical networks that provides default settings, as well as customization capabilities for advanced users. The package will include components 1-4 below.

1. Data discretization. We have refined our method from the previous year in response to feedback from applications and other users. In particular, we have obtained a better understanding of the potential problems arising in data discretization in general and in using our method in particular. For instance, the presence of different time scales in the data can possibly result in the introduction of periodicities in the discretized data. Care must therefore be taken in validating the discretized data by comparison to the original data. The method is described in detail in a manuscript that has been submitted for publication (Dimitrova, McGee and Laubenbacher, 2006, unpublished results). In order to make the method more usable, we are planning to implement it in the statistics package R rather than just stand-alone C++ code.

2. Description of the model space and model selection. One of the distinct advantages of our methodology is that we are able to describe the entire space of dynamic models that are consistent with a given data set of time course measurements in a way that does not rely on model enumeration. The enumeration method entails a technical choice, however, a total ordering of the variables in the network. In the absence of biological information that could inform an appropriate choice of ordering, the random choice of an ordering can influence subsequent model selection. This year we have succeeded in developing a method that finds all possible (static) minimal wiring diagrams of models that does not depend on any random choices. We have developed a statistical measure for the selection of most likely wiring diagrams. The results, which have been obtained in collaboration with Michael Stillman from Cornell University, have been submitted for publication (Jarrah, Laubenbacher, Stigler and Stillman, 2006, unpublished results).

We have also succeeded in developing a criterion that measures how much the original method is dependent on the choice of variable ordering, in the sense of measuring the number of different outcomes of model selection. This criterion is the basis of an algorithm that determines which additional measurements should be performed in order to make the modeling output independent of order choice (Dimitrova, Jarrah and Laubenbacher, 2006, unpublished results).

3. Optimization algorithm for robust models. One of the ingredients in Step 2 is an exact interpolation algorithm, which is very sensitive to the presence of noise in the discretized data. While data discretization does absorb some noise present in experimental data, we can expect that for microarray data, for instance, approximately 5% noise can be expected. In order to avoid overfitting of data, we have developed an evolutionary algorithm that optimizes between model complexity and data fit. The algorithm is complex, and most of this last year was taken up with coding, testing, and determination of optimal algorithm parameters.

4. Relationship to other discrete modeling methods. Ever since the introduction of Boolean networks as models for gene regulatory networks by S. Kaufman (Kaufman, 1969), discrete models have provided an alternative and complementary point of view for differential equations models. Our modeling paradigm of polynomial dynamical systems over finite number fields encompasses Boolean

networks and has the advantage of having a rich mathematical foundation. The two other discrete modeling methods in systems biology that share this feature are the so-called logical models of Snoussi and Thomas (Snoussi & Thomas, 1993) and Petri nets (Pinney et al, 2003). We have developed and implemented an algorithm this year that shows the equivalence of polynomial dynamical systems and logical models, in the sense that one can be translated into the other without loss of information. The advantage for logical models is that our reverse-engineering technology becomes available to the logical model framework. On the other hand, logical models are more intuitive than polynomial models, which gain thereby a more easily interpretable framework (details are included in a manuscript under preparation: Colon-Velez, Jarrah and Laubenbacher, 2006, unpublished results). We are currently working on a similar equivalence between polynomial dynamical systems and a certain class of Petri nets.

5. A unified software environment. We have chosen to develop the comprehensive software package described earlier within Macaulay2 (Grayson and Stillman: Macaulay2, a software system for research in algebraic geometry, see www.math.uiuc.edu/Macaulay2), an open-source, symbolic computation system that is specialized for computations in algebra and algebraic geometry, and is optimized for the type of computations relevant to our work, as opposed to Mathematica, Maple, or Matlab, which are less well-suited. Macaulay2 is one of the most commonly used customized packages around the world. This last year saw an intense collaboration with Michael Stillman, one of the co-developers of Macaulay2. With his help, we have optimized several of our algorithms and are developing a Macaulay2 library for polynomial dynamical systems that can be distributed with the main Macaulay2 package, making it widely available.

6. A mathematical foundation for poly-nomial dynamical systems. While discrete models tend to be more intuitive than continuous models, they suffer from an important drawback, namely the absence of an extensive collection of mathematical tools for the analysis of the relationship between model structure and resulting dynamics, which is in contrast to the powerful methods available for continuous models. During this last year, we have made significant progress in studying this relationship for nonlinear polynomial dynamical systems. Part of this work has been done in collaboration with the Barrett group at VBI, a study of stochastic, sequentially updated polynomial systems. In particular, we have developed computationally feasible methods to predict key features of model dynamics for special, but large, families of systems (Colon-Reyes, Jarrah, Laubenbacher and Sturmfels, 2006, unpublished results; Jarrah, Laubenbacher and Vera-Licona, 2006, unpublished results; Laubenbacher & Pareigis 2006).

7. Method validation. We have devoted part of this last year to the validation of our top-down modeling technology, partly in collaboration with the Mendes group. We are using real experimental data sets for model validation, including data from a published yeast cell cycle data set (Spellman et al, 1998), a data set from *Caenorhabditis elegans* embryonal development (Baugh et al, 2005), and functional magnetic resonance imaging (fMRI) data generated by collaborators at Rutgers University Medical School. The advantage of real data is that they present challenges that are difficult to replicate with simulated data. But two features make their use difficult at this time, namely the fact that real data sets are typically quite small compared to the number of network nodes, and typically information about cellular networks is incomplete, at best. Consequently, we have primarily used simulated data sets, generated by the Mendes group using Gepasi.

References

Baugh LR, Hill AA, Claggett JM, Hill-Harfe K, Wen JC, Slonim DK, Brown EL, Hunter CP (2005) The homeodomain protein PAL-1 specifies a lineage-specific regulatory network in the *C. elegans* embryo. *Development* **132**: 1843-1854.

Karp RM (2002) Mathematical challenges from genomics and molecular biology. *Notices Amer. Math. Soc.* **49** (5): 544-553.

Kauffman SA (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.* **22**(3): 437-467.

Laubenbacher R, Pareigis B (2006) Update schedules of sequential dynamical systems, *Discr. Appl. Math.* **54**: 980-994.

Mayr E (1988) *Toward a new philosophy of biology.* Harvard University Press, Cambridge, MA, p.2.

Pinney JW, Westhead DR, McConkey GA (2003) Petri net representations in systems biology. *Biochem. Soc. Trans.* **31**: 1513-1515.

Snoussi E, Thomas R (1993) Logical identification of all steady states: the concept of feedback characteristic states. *Bull. Math. Biol.* **55**: 973-991.

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol.Cell* **9**: 3273-3297.

Publications

Babson E, Barcelo H, Delongueville M, Laubenbacher R (2006) Homotopy theory of graphs, *J. Algebraic Combinatorics* **24**: 31-44.

Garcia L, Jarrah AS, Laubenbacher R (2006) Sequential dynamical systems over words, *Appl. Math.Comp.* **174**: 500-510.

Laubenbacher R (2005) System identification of biochemical networks using discrete models. In *Computation of Biochemical Pathways and Networks*, Kummer U (ed), pp. 87-94. Berlin: Petronius Verlag.

Laubenbacher R (2005) Algebraic models in systems biology. In *Algebraic Biology 2005*, Anai H, Horimoto K (eds) pp. 33-40. Tokyo: Universal Academy Press, Inc.

Laubenbacher R, Mendes P (2005) A discrete approach to top-down modeling of biochemical networks. In *Computational Systems Biology*, Eils R, Kriete A (eds) pp. 229-247. Burlington, MA: Elsevier.

Laubenbacher R, Pareigis B (2006) Update schedules of sequential dynamical systems, *Discr. Appl. Math.* **54**: 980-994.

Alternaria pathogenomics

Christopher B. Lawrence, lawrence@vbi.vt.edu
Associate Professor, Virginia Bioinformatics Institute & Biology, Virginia Tech

The focus of our research group is the study of the interactions of fungi with plants and humans. We are using different experimental approaches to characterize the defense responses of resistant and susceptible plant hosts when they are attacked by destructive necrotrophic fungi. For this purpose, we are looking in detail at the way the necrotrophic fungus *Alternaria brassicicola* interacts with the model flowering plant *Arabidopsis* as well as closely related *Brassica* crops. Our group uses a functional genomics approach to the molecular dissection of infection, identifying genes involved in pathogenicity from the fungus and, concomitantly, the genes involved in both disease susceptibility and resistance in the plant. The data generated from these studies may provide insight into the design and implementation of new strategies for controlling necrotrophic pathogens in economically important crops. Humans are constantly exposed to *Alternaria* and other airborne fungi. Mounting evidence suggests that *Alternaria* play a critical role in the development of chronic airway diseases. Our group is therefore also studying the role of fungi in chronic airway diseases such as asthma, allergy, and chronic rhinosinusitis. In this report, we provide an update on the progress we have made in the following projects: the *Alternaria brassicicola*-Brassicaceae pathosystem, the *A. brassicicola* genome sequencing project, the study of *Alternaria*-human interactions (allergy, asthma and chronic sinusitis), and our other research interests in the area of fungal biotechnologies.

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

Introduction

The so-called "rots" that are caused by necrotrophic fungi are among the most destructive of plant diseases. They inflict substantial tissue damage on their hosts in advance of and during hyphal colonization. Thus, necrotrophs obtain the vast majority of the nutrients required for lifecycle completion from dying or dead tissue. Necrotrophs represent the largest class of fungal plant pathogens; hitherto, our understanding of host-parasite interactions involving this class of pathogens is overall poorly understood. These fungi are tremendously important economically: although they represent just 4% of fungal diversity they cause ~80% of foliar losses due to fungal diseases in some parts of the world (Richard Oliver, Director, The Australian Centre for Necrotrophic Fungal Pathogens, personal communication; Rotem, 1994).

Although they are sometimes considered somewhat primitive in comparison to the more sophisticated biotrophs which depend on a living host to acquire nutrients and complete their life cycle, necrotrophic pathogenic fungi must also be highly specialized in order to successfully avoid, or suppress, host resistance responses. In general, necrotrophic fungi employ a variety of mechanisms to circumvent the host plant defense response by either interfering with the activation of the response or negating its effect. It has been shown in some instances that one form of host defense suppression is due to the action of secreted toxic molecules that cause programmed cell death reminiscent of apoptosis in mammals. Some important genera, e.g. *Alternaria*, accomplish this by producing low-molecular-weight, host-specific/selective, phytotoxic secondary metabolites. Moreover, there are specific examples of species

Contributors:
Mihaela Babiceanu, Yangrae Cho, Kwang-Hyung Kim, Carlos Mauricio La Rota, Violetta Macioszek, Graciella Santopietro

within these genera that produce both host and non-host-specific toxins. There is another group of pathogens that possess a critical necrotrophic stage in their life cycle that do not produce host-specific toxins — although some produce other, non-host-specific phytotoxins — and the molecular basis for pathogenicity in these organisms remains largely unknown. This group contains many important plant-pathogenic fungal genera including *Botrytis*, *Colletotrichum*, *Fusarium*, *Leptosphaeria*, *Magnaporthe*, *Mycosphaerella*, *Sclerotinia*, and *Stagnospora*. However, the recent completion of the *Magnaporthe grisea* and *Fusarium graminearum* whole-genome sequencing and associated high-throughput functional genomics projects via international collaborations has provided more insight into virulence mechanisms employed by these fungi for rice and wheat infections, respectively.

As mentioned, toxins produced by necrotrophs can be of a "host-specific" or "non-host-specific" nature, are diverse in chemical structure and include secondary metabolites, cyclic peptides, and even proteins such as the host-specific Ptr toxins produced by the wheat pathogen, *Pyrenophora tritici-repentis* (Lichter et al, 2002). In some plant-pathogen systems, these toxins have been shown to be the primary determinant of pathogenicity. In other cases, these toxins clearly serve to increase virulence. In most scenarios, host plant resistance mechanisms in response to true necrotrophic fungi are complex and not well understood, but appear to at least partially function by interfering with the ability of the pathogen to suppress defenses and/or initiate host programmed cell death via toxins. Our research is primarily focused on plant interactions with *Alternaria brassicicola*.

Alternaria-plant interactions: the *Alternaria brassicicola*-Brassicaceae pathosystem

Brassicaceae, the crucifer plant family, consists of approximately 3500 species in 350 distinct genera. However, the most important crop species from an economic perspective are found within the single genus, *Brassica*. These crop species include *Brassica oleracea* (vegetables), *Brassica rapa* (vegetables, oilseeds, and forages), *Brassica juncea* (vegetables and seed mustard), and *Brassica napus* (oilseeds and root vegetables). *A. brassicicola* causes black spot disease (also called dark leaf spot) on virtually every important *Brassica* spp. and is of worldwide economic importance (Sigareva & Earle, 1999; Westman et al, 1999). High levels of resistance/immunity to this fungus have been reported in weedy cruciferous plants such as *Arabidopsis thaliana*, *Camelina sativa* and *Capsella bursa-pastoris*, but no satisfactory source of resistance has been identified among cultivated *Brassica* species (Conn et al, 1988; King, 1994; Sigareva & Earle, 1999; Westman et al, 1999; Otani et al, 2001). Of the very few *Brassica* species or breeding lines that have been reported to possess some limited level of resistance, the genetic basis appears to be somewhat complex and involves additive and dominant gene action (King, 1994). Additionally, due to polyploidization within the Brassicaceae plant family with different species containing diverse genomes and number of chromosomes, numerous breeding efforts employing hybridization (traditional breeding approaches and somatic hybridization) between highly resistant wild species and cultivated Brassicas have proven time and time again unsuccessful due to interspecies incompatibility (Conn et al, 1988; King, 1994; Sigareva & Earle, 1998; Westman et al, 1999).

Despite our limited understanding of *A. brassicicola* pathogenesis mechanisms, a substantial amount of work has been done to characterize resistance mechanisms to *A. brassicicola* using the model plant *Arabidopsis thaliana*. Natural variation in susceptibility and resistance to *A. brassicicola* has been shown to exist in *Arabidopsis* ecotypes and several mutants have been identified that confer increased susceptibility to this fungus (Otani et al, 2001; Kagan & Hammerschmidt, 2002; reviewed by Thomma, 2003; Lawrence et al, unpublished results). Further, the enormous genomic resources that are available for *Arabidopsis* make it an ideal system to identify which signaling pathways are important for resistance to necrotrophic fungal pathogens such as *A. brassicicola*. For example, Penninckx and coworkers were able to show that jasmonic acid levels increased dramatically when *Arabidopsis* plants were challenged with *A. brassicicola*, which resulted in the expression of *PDF1.2*, a gene that encodes for an antifungal, defensin-like peptide (Penninckx et al, 1996).

Two other genes were activated coordinately with *PDF1.2* upon *A. brassicicola* challenge including *PR-3*, a basic chitinase, and *PR-4*, a hevein-like protein (Thomma et al, 1998). Thomma and colleagues demonstrated that the methyl jasmonate insensitive mutant, *coi1-1*, is more susceptible to *A. brassicicola* than wild-type Columbia (Col-0) (Thomma et al, 1998). *PAD3*, a gene encoding a cytochrome *P*450 monooxygenase that is essential for synthesis of the *Arabidopsis* phytoalexin, camalexin, is also more susceptible to mutation (Thomma et al, 1999; Zhou et al, 1999). These studies clearly suggest that jasmonic acid signaling and camalexin synthesis are required for resistance to *A. brassicicola*. Loss of camalexin cannot be the reason for enhanced susceptibility of *coi1*, as *Alternaria*-induced camalexin levels are wild-type in this mutant (Thomma et al, 1999; van Wees & Glazebrook, 2003). Thomma and coworkers also found that *NahG* plants, which have markedly reduced salicylic acid levels, remain resistant to *A. brassicicola*, indicating that salicylic acid is not required for resistance (Thomma et al, 1998). Further substantiation for these observations came from the finding that the *Arabidopsis* mutation *esa1* enhances susceptibility to *A. brassicicola* and exhibits a severe reduction in both camalexin production and jasmonate-dependent gene induction (Tierens et al, 2002).

Several large-scale gene expression studies have been undertaken to dissect *Arabidopsis* resistance to *A. brassicicola*. Schenk and coworkers used microarray analysis to identify 168 genes upregulated during an interaction between the *Arabidopsis* ecotype Col-0 and *A. brassicicola*, but the roles of these genes in resistance have yet to be determined (Schenk et al, 2000). Schenk and colleagues have also examined gene expression in distal uninoculated tissues during *A. brassicicola* infection, finding 35 genes with altered expression (Schenk et al, 2003). Van Wees and Glazebrook identified 645 genes induced by *A. brassicicola* infection in wild-type Columbia (Col-0) and pad3 plants, indicating that Pad3 does not have a major effect on the early stages of defense signaling (Van Wees & Glazebrook, 2003). Interestingly, 265 of the 645 *A. brassicicola* induced genes identified by Van Wees and Glazebrook required COI1 for full expression, suggesting a major role of jasmonic acid signaling in responses to *Alternaria* infection (Van Wees & Glazebrook 2003). Thus, based on the research to date, it seems clear that jasmonic acid and camalexin both play critical roles in resistance to *A. brassicicola*. These results collectively indicate that the *A. brassicicola* – *Arabidopsis* interaction has already become a very useful model pathosystem to study necrotrophic fungal pathogenesis, defense signaling pathways, and the genetic basis for host resistance. In contrast, mechanisms of *A. brassicicola* pathogenicity are very poorly understood, with no known virulence factors identified to date.

In our laboratory, we are taking a functional genomics approach to elucidate both molecular aspects of fungal pathogenicity and host plant response to *A. brassicicola* infection in both *Arabidopsis* and cultivated Brassicas. Our research thus far can be divided into three major areas: 1) Large-scale generation and analysis of expressed sequence tags (ESTs) derived from various *A. brassicicola*-host interactions; 2) functional analysis of fungal pathogenicity; and 3) genome-wide analysis of gene expression in cultivated Brassicas and *Arabidopsis* during fungal infection.

Over 17000 ESTs from various cDNA libraries derived from *Alternaria*-infected plant tissues have been generated. We have performed bioinformatic analysis of ESTs using Blast algorithms (BlastX, BlastN) for sequence comparisons using primarily the Genbank NR and fungal genome databases available at the Broad Institute (www.broad.mit.edu/). We have also performed Interpro analysis (www.ebi.ac.uk/interpro/). InterPro is a database of protein families, domains and functional sites in which identifiable features found in known proteins can be applied to unknown protein sequences. The results of these analyses and a review of the current literature have identified potential fungal pathogenicity factors and host genes involved in susceptibility. Here we will primarily describe research related to fungal genes, as analysis of host genes is only just beginning. However, we are particularly interested in host genes involved in programmed cell death since this process is thought to be required for infection by necrotrophic pathogens. A summary of this initial work has been recently published (Cramer et al, 2006).

From our analyzed EST data, we have selected an initial set of 50 fungal genes for functional analysis of phenotypic changes in virulence. These genes encode proteins putatively involved in cell wall/cuticle degradation, Mitogen Activated Protein (MAP) kinase signaling, and secondary metabolite biosynthesis. We have also targeted genes predicted to encode phytotoxic proteins. In this regard we have concomitantly found it necessary to develop a high throughput method for obtaining fungal gene knockouts. Gene knockout plays a critical role in identification of fungal pathogenicity/virulence factors. Unlike RNA interference (RNAi)-based mutational approaches, targeted gene knockout primarily depends on homologous recombination between a disruption construct and a nascent gene. In most filamentous fungi, mutant generation has been the most rate-limiting step for the functional analysis of individual genes due to low efficiencies of both transformation and targeted integration. We have recently developed an extremely efficient approach for generating targeted gene disruption mutants in *A. brassicicola* (Cho et al, 2006).

Using our knockout technology, we have identified several interesting genes thus far with a role in pathogenicity. For example, disruption of a MAP kinase gene resulted in a non-pathogenic phenotype (Cho et al, unpublished results). Moreover, disruption of this MAP kinase revealed a role in both the positive and negative regulation of fungal hydrolytic enzyme genes depending upon type and level of nutrient sources present in the immediate environment. Disruption of a specific non-ribosomal peptide synthase gene (*AbNRPS2*) revealed a dramatic effect on pathogenicity, sporulation, spore cell wall structure, and spore viability. To our knowledge, this is the first example of a secondary metabolite-associated gene being critical for formation of the fungal spore cell wall (Kim et al, unpublished results). Another class of secondary metabolite-related genes we are investigating are the polyketide synthases (PKS). Disruption of two PKS genes in *A. brassicicola* has revealed a putative role in pathogenicity and these genes are thought to be responsible for the production of phytotoxins.

In summary, we have identified thousands of fungal genes expressed during plant infection and have selected candidates using bioinformatics for functional analysis. Moreover, we now have developed a reliable, rapid method for targeted gene disruption. The *Alternaria brassicicola* genome sequencing project has recently been funded by the 2005 National Science Foundation-United States Department of Agriculture (NSF-USDA) Interagency Microbial Genome Sequencing Program (Principal Investigator, Christopher Lawrence). We anticipate that our high throughput functional analysis pipeline will be extremely useful for researchers worldwide interested in analyzing genes involved in necrotrophic fungal pathogenicity.

The *Alternaria brassicicola* genome sequencing project

As mentioned above, *A. brassicicola* is a necrotrophic fungus that is an economically important pathogen of Brassicas. Moreover, it has been used as a model necrotroph for studies with *Arabidopsis* for over a decade. From a human health perspective, the genus *Alternaria* is clearly associated with, if not the causal agent of, chronic respiratory disorders such as allergy, asthma, and chronic rhinosinusitis. The genome of the haploid fungus *A. brassicicola* (strain ATCC 96866) is estimated to be nearly 30 Mb in size, distributed among 9 to 10 chromosomes (Akamatsu et al, 1999). A consortium of several research labs, led by our group at the Virginia Bioinformatics Institute and including researchers at VBI (Dr. Brett Tyler), Washington University Genome Sequencing Center (Dr. Sandra Clifton), North Carolina State University (Dr. Tom Mitchell) and Colorado State University (Drs. Dennis Knudson and Susan Brown), is determining and annotating the genome sequence. The strategy for sequencing used a combination of traditional BAC fingerprinting, with clone by clone sequencing from BAC and Fosmid ends and a separate whole genome shotgun (WGS) from sheared libraries having 6-fold coverage. This approach integrates the results from fingerprinting data and sequence assembly data providing better continuity to the WGS assembly. A preliminary assembly of nearly 300 000 WGS reads and the BAC and Fosmid end sequence data revealed a genome size of approximately 30 Mbp based on 838 supercontiguous sequences/scaffolds

with 83% (~25.5 Mbp) of the 30-Mbp sequence contained in 11 supercontiguous sequences. Identification of genes, non-coding RNAs and other features will be performed with the aid of *ab initio* gene-finding programs as well as incorporating existing ESTs. A preliminary survey of the version 1.0 assembly, using gene prediction programs such as FgeneSH utilizing an *A. brassicicola*-specific training set, estimates approximately 11 000 protein-coding genes. The project is also using experimental procedures such as massively parallel signature sequencing (MPSS) (Brenner et al, 2000) as novel annotation tools for gene prediction. A subset of ESTs derived using 454 sequencing technology will also be incorporated into the next version of the assembly. Finally, the project includes the development of a community annotation bioinformatics platform. The machine annotated sequence is expected to be released in late 2006. As described in the next section, the sequence information has already proven extremely valuable for identifying target fungal genes involved in plant pathogenesis and human respiratory diseases.

Alternaria-human interactions: allergy, asthma, and chronic sinusitis

Sensitivity to the fungus *Alternaria alternata* and most likely other species within the genus is believed to be a common cause of asthma. Epidemiological studies from a variety of locations worldwide indicate that *Alternaria* sensitivity is closely linked with the development of asthma (Gergen & Turkeltaub, 1992; Halonen et al, 1997). In addition, up to 70 % of mold-allergic patients have skin test reactivity to *Alternaria* (Schonwald, 1938). *Alternaria* sensitivity has been shown to not only be a risk factor for asthma, but can also directly lead to the development of severe and potentially fatal asthma (Gergen & Turkeltaub, 1992; Halonen et al, 1997; O'Hollaren et al, 1992). Additionally, *Alternaria* sensitization has been determined to be one of the most important factors in the onset of childhood asthma in the southwest deserts of the United States and other arid regions (Halonen et al, 1997; Peat et al, 1993). *Alternaria* spores are routinely found in atmospheric surveys in the United States and in other countries (Hoffman, 1984). Moreover, *Alternaria* spores are the most frequently encountered of any fungus in these surveys highlighting the ubiquitous nature of this genus. Fungal exposure differs from pollen exposure in quantity (airborne spore counts are often 1000-fold greater than pollen counts) and duration (*Alternaria* exposure occurs for months, whereas ragweed pollen exposure occurs less frequently). This type of concentrated, lengthy exposure is similar to that of other asthma-associated allergens such as those found in cat dander and dust mites and may be at least partially responsible for both the chronic and severe nature of asthma in *Alternaria*-sensitive individuals.

Although some research has been performed on the physiological and molecular identification of *Alternaria* allergens, only three major and five minor allergenic proteins have been described from one highly ubiquitous species, *A. alternata* (Sanchez & Busch, 2001). Our laboratory was the first to identify the major allergen homolog Alta1 in *A. brassicicola*, a species other than *A. alternata* (Cramer & Lawrence, 2003). Moreover, we have recently determined in collaboration with Dr. Barry Pryor at the University of Arizona that over 52 *Alternaria* species and related taxa possess Alta1 homologs, suggesting that all species are potentially allergenic (Hong et al, 2005). The biological role of these allergens in the development of allergy and asthma is poorly understood. Other than a few studies demonstrating binding of these allergens to IgG/IgE-specific antibodies in human sera from patients diagnosed as being *Alternaria* sensitive, virtually nothing is known about how these highly immunoreactive proteins interact with the host.

In one current project in our laboratory, the secretomes of three species of *Alternaria* are being surveyed for the presence of IgG/IgE-reactive proteins using a proteomics approach. Subsequently, a collection of recombinant antigenic proteins will be produced, applied to lung epithelial cells, and various host immune responses will be profiled, such as the production of antimicrobial proteins, chemokines, cytokines, and the expression patterns of Toll-like receptor genes. In addition, we will investigate gross ultrastructural changes in treated cells. We believe the interaction of secreted *Alternaria* allergens/antigens with lung epithelial cells represents a unique, highly physiologically-relevant model *in vitro* system for studying the

primary host-pathogen interface. For example, not only have ungerminated *A. alternata* spores been shown to contain the major allergen Alta1, secretion of this protein has been reported to dramatically increase during germination (Mitakakis et al, 2001). Thus, it is highly conceivable that airway epithelial cells would be the primary cell type to be exposed to both fungal proteins constitutively present in the spore cell wall and secreted during the germination process following attachment to airway epithelium. The physiological and molecular basis of host-pathogen signaling at this interface could undoubtedly be critical in the predisposition to and onset of asthma, allergy, and chronic sinusitis. There is clearly a need to elucidate the role of *Alternaria* immunoreactive proteins in the development of respiratory disorders from both diagnostic and immunotherapeutic perspectives.

We have established a consortium with Dr. Hirohito Kita at the Mayo Medical School (Rochester, MN) to investigate the role of *Alternaria* in the pathogenesis of chronic rhinosinusitis. Earlier data generated by Dr. Kita have revealed that secreted *Alternaria* antigens (and not antigen preparations from other ubiquitous fungi such as *Aspergillus*) specifically cause immune cells to secrete proinflammatory cytokines only in CRS patients and not in normal individuals. This observation highlights the importance of *Alternaria* in the etiology of this disease. Using a combination of immunology, proteomics employing the *A. brassicicola* genome sequence, and fungal biotechnology, several secreted *Alternaria* proteins have been identified that trigger eosinophilic inflammation, activation of eosinophils once recruited, and cause profound effects on inflammatory cytokine production in a mouse model, dendritic cells, and *in vitro* airway epithelial cell systems. These fungal proteins may also be considered drug targets for future development of therapeutic strategies.

Dr. Rathore at the Virginia Bioinformatics Institute has also been contributing to this project by aiding us in the expression of recombinant fungal proteins in an *E. coli* expression system.

References

Akamatsu H, Taga M, Kodama M, Johnson R, Otani H, Kohmoto K (1999) Molecular karyotypes for Alternaria plant pathogens known to produce host-specific toxins. *Curr. Genet.* **35**: 647-656.

Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, Roth R, George D, Eletr S, Albrecht G, Vermaas E, Williams SR, Moon K, Burcham T, Pallas M, DuBridge RB, Kirchner J, Fearon K, Mao J, Corcoran K (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature Biotechnol*. **18**: 630-634.

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

Conn KL, Tewari JP, Dahiya JS (1988) Resistance to *Alternaria brassicae* and phytoalexin-elicitation in rapeseed and other crucifers. *Plant Sci.* **56**: 21-25.

Cramer R, Lawrence CB (2003) Cloning of a gene encoding an Alt a 1 isoallergen differentially expressed in the phytopathogenic fungus, *Alternaria brassicicola* during Arabidopsis infection. *Appl. Environ. Microbiol.* **69**: 2361-2364.

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

Gergen PJ, Turkeltaub PC (1992) The association of individual allergen reactivity with respiratory disease in a national sample: data from the Second National Health and Nutrition Examination Survey, 1976-80 (NHANES II). *J. Allergy Clin. Immunol.* **90**: 579-588.

Halonen M, Stern DA, Wright AL,Taussing LM, Martinez FD (1997) Alternaria as a major allergen for asthma in children raised in a desert environment. *Am. J. Respir. Crit. Care Med.* **155**:1356-1361.

Hoffman DR (1984) Mould allergens. In *Mould allergy* AL-Doory Y, Domson J (eds) pp. 104-116. Philadelphia: Lea & Febiger.

Hong SG, Cramer RC, Lawrence CB, Pryor BM (2005) Alta1 allergen homologs from *Alternaria* and related taxa: analysis of phylogenetic content and secondary structure *Fungal Genet. Biol.* **42**:119-129.

Kagan IA, Hammerschmidt R (2002) Arabidopsis ecotype variability in camalexin production and reaction to infection by *Alternaria brascicicola*. *J. Chem. Ecol*. **28**: 2121-2140.

King SR (1994) Screening, selection, and genetics of resistance to *Alternaria* diseases in *Brassica oleracea*. *Ph.D Thesis, Cornell University, Ithaca, New York*. Diss. Abst. Int. 55/0 B:2471.

Lichter A, Gaventa JM, Ciuffetti LM (2002) Chromosome-based molecular characterization of pathogenic and non-pathogenic wheat isolates of *Pyrenophora tritici repentis*. *Fungal Gen. Biol*. **37**: 180-189.

Mitakakis TZ, Barnes C, Tovey ER (2001) Spore germination increases allergen release from Alternaria. *J. Allergy Clin. Immunol*. **107**: 388-390.

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

Otani H, Kohnobe A, Narita M, Shiomi H, Kodama M, Kohmoto K (2001) A new type of host-selective toxin, a protein from *Alternaria brassicicola*. In *Delivery and Perception of Pathogen Signals in Plants*, Keen NT, Mayama S, Leach JE, Tsuyumu S (eds) pp 68-76. St. Paul, MN: APS Press.

Peat JK, Tovey CM, Mellis CM, Leedre SR, Woolcock AJ (1993) Importance of house dust mite and Alternaria allergens in childhood asthma: an epidemiological study in two climatic regions of Australia. *Clin. Exp. Allergy* **23**:812-820.

Peninckx IAMA, Eggermont K, Terras FRG, Thomma BPHJ, De Samblanx GW, Buchala A, Metraux JP, Manners JM, Broekaert WF (1996) Pathogen induced systemic activation of a plant defensin gene in *Arabidopsis* follows a salicylic acid-independent pathway. *Plant Cell* **8**: 2309-2323.

Rotem J (1994) *The Genus Alternaria. Biology, Epidemiology, and Pathogenicity*, APS Press, St. Paul, Minnesota.

Sanchez H, Bush RK (2001) A review of *Alternaria alternata* sensitivity. *Rev. Iberoam Micol*. **18**: 56-59.

Schenk PM, Kazan K, Wilson I, Anderson JP, Richmond T, Somerville SC, Manners JM (2000) Coordinated plant defense responses in *Arabidopsis* revealed by microarray analysis. *Proc. Natl. Acad. Sci. USA* **97**: 11655-11660.

Schenk PM, Kazan K, Manners JM, Anderson JP, Simpson RS, Wilson IW, Somerville SC, Maclean DJ (2003) Systemic gene expression in Arabidopsis during an incompatible with *Alternaria brassicicola*. *Plant Physiol*. **132**: 999-1010.

Schonwald P (1938) Allergenic molds in the Pacific Northwest. *J. Allergy* **9**: 175-179.

Sigareva MA, Earle ED (1999) Camalexin induction in intertribal somatic hybrids between *Camelina sativa* and rapid cycling *Brassica oleracea*. *Theor. Appl. Genet*. **98**: 164-170.

Thomma BPHJ, Eggermont K, Penninckx IAMA, Mauch-Mani B, Vogelsang R, Cammue BPA, Broekaert WF (1998) Separate jasmonate-dependent and salicylate-dependent defense-response pathways in Arabidopsis are essential for resistance to distinct microbial pathogens. *Proc. Natl. Acad. Sci. USA* **95**: 15107-15111.

Thomma BP, Nelissen I, Eggermont K, Broekaert WF (1999) Deficiency in phytoalexin production causes enhanced susceptibility of *Arabidopsis thaliana* to the fungus *Alternaria brassicicola*. *Plant J.* **19**: 163-171.

Thomma BPHJ (2003) *Alternaria* spp. from general saprophyte to specific parasite. *Mol. Plant Pathol*. **4**: 225-236.

Tierens KFMJ, Thomma BPHJ, Bari RP, Garmier M, Eggermont K, Brouwer M, Penninckx IAMA, Broekaert WF, Cammue BPA (2002) Esa1, an *Arabidopsis* mutant with enhanced susceptibility to a range of necrotrophic fungal pathogens, shows a distorted induction of defense responses by reactive oxygen generating compounds. *Plant J.* **29**: 131-140.

van Wees SC, Glazebrook J (2003) Loss of non-host resistance of *Arabidopsis* NahG to *Pseudomonas syringae* pv. *phaseolicola* is due to degradation products of salicylic acid. *Plant J*. **33**: 733-742.

Westman AL, Kresovich S, Dickson MH (1999) Regional variation in *Brassica nigra* and other weedy crucifers for disease reaction to *Alternaria brassicicola* and *Xanthomonas campestris* pv. *campestris*. *Euphytica* **106**: 253-259.

Zhou N, Tootle TL, Glazebrook J (1999) Arabidopsis *PAD3*, a gene required for camalexin biosynthesis, encodes a putative cytochrome P450 monooxygenase. *Plant Cell* **11**: 2419-2428.

Publications

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

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

Funnell DL, Lawrence CB, Pedersen JF, Schardl CL (2005) Expression of the tobacco β-1,3-glucanase gene, PR-2d, following induction of SAR with *Peronospora tabacina*. *Physiol. Mol. Plant Pathol*. **65**: 285-296.

Hong SG, Cramer RC, Lawrence CB, Pryor BM (2005) Alta1 allergen homologs from Alternaria and related taxa: analysis of phylogenetic content and secondary structure *Fungal Genet. Biol*. **42**:119-129.

Li Q-S, Lawrence CB, Xing H, Davies M, Everett NP (2006) Increased pathogen resistance and yield in transgenic plants expressing combinations of the modified antimicrobial peptides based on indolicidin and magainin. *Planta* **223**: 1024-1032.

Microfluidic mass spectrometry for proteomic and biomarker applications

Iuliana M. Lazar, lazar@vbi.vt.edu
Assistant Professor, Virginia Bioinformatics Institute & Biology, Virginia Tech

The long-term objective of our research is to develop a microfluidic platform that integrates mass spectrometry detection for high-throughput screening and/or discovery of biomarkers in cancer cells and tissues. Microfluidic architectures present distinguishing capabilities that facilitate process integration, multiplexing and high-throughput processing of minute amounts of sample. The development of novel technologies that will enable fast and cost-effective discovery of prognostic/diagnostic markers that can be used collectively with increased specificity and sensitivity, will significantly enhance our capacity to intervene in disease detection, prevention and therapy. To date, we have developed a bioanalytical protocol that enabled the identification of more than 2000 proteins (~1900 with $P<0.001$) in the MCF7 breast cancer cell line, by using conventional liquid chromatography (LC) and ion trap–electrospray ionization mass spectrometry (ESI-MS) detection. The rate of false positive protein identifications was as low as 0.4 %. The typical reproducibility in detecting overlapping proteins over replicate runs was in excess of 90% for proteins matched by ≥2 unique peptides. According to their biological function, approximately 220 proteins were involved in cancer-relevant cellular processes, and over 25 proteins were previously described in the literature as putative cancer biomarkers (i.e. differentially expressed between normal and cancerous cell states). Among these, biomarkers such as proliferating cell nuclear antigen (PCNA), cathepsin D, E-cadherin, 14-3-3-sigma, antigen Ki-67, TP53RK and calreticulin were identified. The microfluidic LC-MS analysis of a selected, protein-rich MCF7 sample sub-fraction, enabled the confident identification of 39 proteins with $P<0.001$ and of five cancer-specific biomarkers. These findings demonstrate the potential applicability of these chips for high-throughput biomarker screening applications.

Keywords: proteomics; cancer; mass spectrometry; microfluidics; biomarker discovery and screening.

Introduction

The capability to provide an integrated view of the dynamic molecular cell profile, at the proteomic and metabolic levels, represents a major landmark for biological research. Relevant questions that must be addressed relate to the identity and expression level of protein components in a cell, the nature, site and number of post-translational modifications, and the specific functions associated with these proteins. The generated information is critical for differentiating normal versus diseased cell states. The development of novel technologies that will enable high-throughput explorations is essential to provide a timely solution to proteomic investigations. Our research is focused on two main themes: (1) the development of fully integrated, stand-alone microfluidic devices that integrate mass spectrometry detection for high-throughput proteomics investigations, and (2) the development of bioanalytical strategies for global proteomic profiling of cancer cells and tissues (qualitative profiling, differential protein expression analysis, and characterization of post-translational modifications). The discovery of novel and specific molecular markers for early disease detection is extremely valuable for cancer patients who need urgent intervention to increase survivability rates. In spite of worldwide research that has resulted in the discovery of a range of potential biomarkers, very few

Contributors:
Hetal Sarvaiya, Phichet Trisiripisal

are used in the clinical practice. The efficient merger of novel microfluidic technologies with powerful MS detection strategies will enable the development of effective microfluidic–MS platforms for cancer biomarker discovery and screening.

Additional smaller projects focus on (1) the development of a mass spectrometric bioanalytical platform for the analysis of cancer cells and tissues (collaboration with Dr.Yong Chen, Wake Forest University Cancer Center), (2) the analysis of flavonoid enzyme interactions in *Arabidopsis thaliana* by mass spectrometry (collaboration with Dr. Brenda Winkel, Department of Biological Sciences, Virginia Tech), and (3) the isolation of class A penicillin-binding proteins (PBPA) complexes from *Bacillus subtilis* (collaboration with Dr. David Popham, Department of Biological Sciences, Virginia Tech).

Methods

Microfluidic devices were fabricated from glass substrates (1.6 mm thick) precoated with chrome and positive photoresist. Photomasks for transferring the desired microfluidic channel pattern onto the glass substrate were prepared with AUTOCAD software and then outsourced for preparation to designated manufacturers. The fabrication of the microchips included the following steps: proper alignment of the substrate and photomask, exposure to ultraviolet light, developing and removal of the irradiated photoresist, selective removal of the underlying chrome layer, wet chemical etching of the channel pattern in the glass substrate (etched depth of 1.5–50 μm), removal of the remaining chrome layer, drilling of access holes in the cover plate, cleaning, and thermal bonding of the substrate to a cover plate by gradual heating to 550°C. Fluidic propulsion was accomplished using electrically- and pressure-driven mechanisms. Fluidic manipulations were optimized with a Nikon epi-fluorescent microscope. Capillary separation columns were prepared from reversed phase C18 packing using 3 μm and 5 μm particles. Mass spectrometric detection was accomplished using electrospray ionization with an ion trap LTQ system (Thermo Electron). MCF7 cancer cells were cultured to 70 % confluence, harvested, lysed, and the soluble fraction was digested with trypsin and fractionated using strong cation exchange (SCX) separation columns. The SCX fractions were analyzed using an Agilent micro liquid chromatography (LC) system and microfluidic chips.

Results and Discussion

Qualitative evaluation of the MCF7 breast cancer cell line. Proteomic protocols are being developed for confident identification of a large number of proteins, differential protein expression analysis and biomarker discovery. The MCF7 breast cancer cell line was chosen as a model system as it is an estrogen receptor-positive cell line that retains several characteristics of the differentiated mammary epithelium. The soluble fraction of the cellular extract is typically processed by using a shotgun protocol–whole tryptic digestion of the entire cellular extract, SCX fractionation and LC-MS/MS analysis (Wolters et al, 2001).

Several series of optimization strategies were performed to improve the experimental set-up and the data acquisition and database search protocols. A data filtering strategy was developed to enable confident identification of a large number of proteins and potential biomarkers. Over 2000 proteins were identified using multiple data filtering parameters and *P*-value sorting. From the total number of proteins, approximately 90% were identified with $P<0.001$, and approximately 1000 proteins were identified by ≥2 unique peptides (of which >99 % had $P<0.001$). The false positive identification rates and the effectiveness of the data selection criteria were evaluated by searching a composite database that contained the forward and reverse directions of the protein entries from the National Center for Biotechnology Information (NCBI) database. For proteins with $P<0.001$, the false positive identification rates were ~0.1 % at the peptide level and ~0.4 % at the protein level, i.e. much lower than previously reported (Peng et al, 2002). The reproducibility in detecting overlapping proteins over replicate runs exceeded 88–90% for proteins matched by ≥2 unique peptides.

The identified proteins were categorized based on their cellular location and biological

function. Approximately 220 cancer-relevant proteins were found. These were involved in cellular processes such as: cell differentiation (17), cell growth and proliferation (62), cell cycle regulation (61), cell adhesion (19), apoptosis (42) and DNA repair (17). Among these, over 25 proteins were previously described in the literature as putative cancer biomarkers, including those found to be up or down regulated in cancerous cell states. Biomarkers such as PCNA, cathepsin D, E-cadherin, 14-3-3-sigma, Ki-67, TP53RK, calreticulin and keratins 8, 18, 19, etc., were identified. A separation of a relevant SCX fraction and the GO (gene ontology) categorization of the identified proteins are shown in Fig. 1.

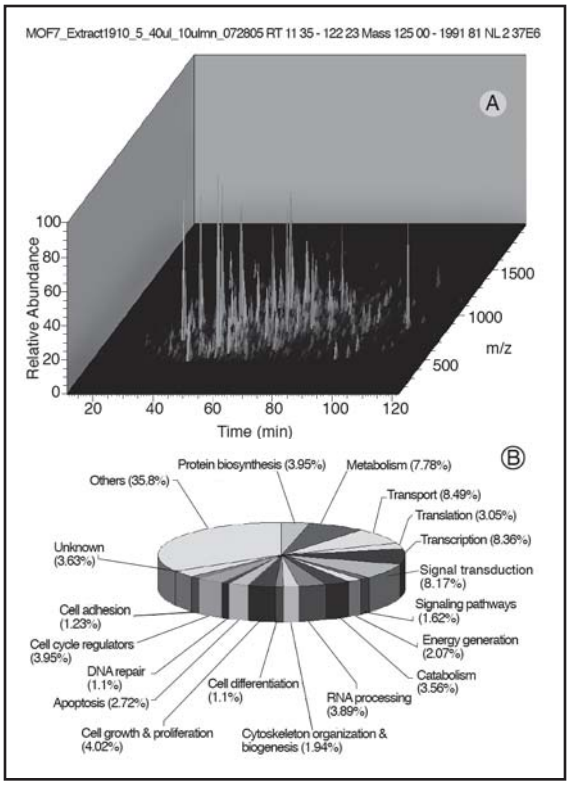


Fig. 1. (A) Data-dependent LC-MS/MS separation of a protein digest SCX sub-fraction from the MCF7 breast cancer cellular extract using conventional instrumentation. (B) Biological process categorization of the identified proteins in the MCF7 breast cancer cell line.

Microfluidic LC device for biomarker screening applications. A microfluidic LC chip that integrates all the necessary components for stand-alone operation of a miniaturized LC platform, i.e. pump, valve, separation column and electrospray interface, was developed. Two fully functional separation systems can be incorporated within a 3" by 1" glass microchip. A multichannel architecture that uses electroosmotic flow (EOF) pumping principles enables eluent propulsion and sample valving (Lazar & Karger, 2002). The configuration of the pumping device consists of hundreds of parallel micro/nano channels that generate EOF. The mechanism of operation relies on electroosmotic pumping principles; however, the pump enables stable flow and pressure generation in electrical field-free regions on the chip. The choice for an EOF pumping system to run the microfluidic LC was dictated by several reasons: (1) EOF pumps can generate high pressures (hundreds of bars), (2) manufacturing is extremely simple and reliable, (3) the same structure can be effectively utilized for sample loading and valving, and (4) the design enables stand-alone operation, multiplexing and high-throughput analysis.

The multiple open channel configuration has a much larger hydraulic resistance than any of the other functional elements on the chip. As such, it acts as a valve that is open to material transport through an electrically-driven mechanism, but is closed to material transport through a pressure-driven mechanism. The same multichannel structure can be used as an EOF eluent pump, and as an EOF sample valve. For very shallow microchannels, the hydraulic resistance is so large that one set of microchannels can be used for pumping, and several other sets for valving. A sample can be manipulated within the chip only when a potential differential is applied between adequate access ports on the chip.

A protein/biomarker-rich fraction (SCX fraction 7 of the MCF7 cellular extract) was subjected to microfluidic LC-MS analysis. A base peak chromatogram is shown in Fig. 2. Sample volumes loaded on the chip for analysis were estimated to be ~1 μl. Peak widths at half height were 15–30 s, the separation efficiency was in the 45 000–180 000/channel range, and peak capacity was estimated to be around 80–100. Two EOF pumps each containing 200 pumping nanochannels and an EOF valve with 100 nanochannels on each arm were used to operate this microfluidic LC platform. The variability between the electrical currents measured in the

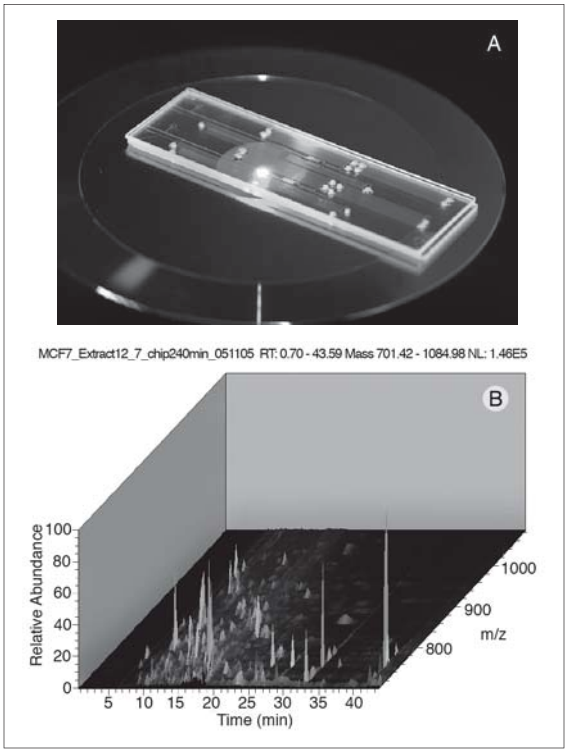


Fig. 2. (A) Microfluidic LC-MS chip with ESI-MS detection. (B) Data-dependent microfluidic LC-MS/MS of an SCX fraction of the MCF7 cancer cell extract.

two pumps on a chip was in the 5–15 % range, eluent flow rates were typically in the 50-80 nL/min range, and relative standard deviation (RSD) values of flow rates measured within one experiment were <11%.

The performance of the microchip LC-MS system was comparable to that obtained with conventional instrumentation, providing an overlap of 75% in proteins that were identified by more than 2 unique peptides. The microfluidic LC analysis of the protein-rich SCX fraction enabled the confident identification of 77 proteins by using conventional data filtering parameters (X_{corr} = 1.9, 2.2 and 3.8 for z =1, 2 and 3, respectively) and of 39 proteins with P<0.001. Five putative biomarker proteins were identified with the microfluidic LC-MS chip: PCNA, cathepsin D, and keratins 8, 18 and 19, thus demonstrating the potential applicability of this chip for future high-throughput biomarker screening applications.

Conclusions

Following a series of optimization studies, we were able to develop a strategy for sensitive and confident identification of a large number of proteins in the MCF7 breast cancer cell line. This strategy has enabled identification of approximately 2000 proteins (P<0.001) and ~25 well-established cancer biomarker components. A microfluidic LC chip that enabled the simultaneous identification of a panel of five cancer-specific biomarkers was developed. The microfluidic environment enables the implementation of a high-throughput, contamination-free analysis strategy. It is envisioned that the advance of low-cost, disposable microfluidic–MS platforms that have the capability to identify protein co-expression patterns in one single run, will enable the biomedical research community in the near future to perform (1) high-throughput proteomic investigations, (2) screening and discovery of prognostic/diagnostic biomarkers, and (3) large scale population screening.

Conference presentations

In the period covered by this report, the Lazar group gave oral and poster presentations at the following events or locations: the *28th International Symposium on Capillary Chromatography and Electrophoresis*, Las Vegas, NV, USA, May 22-25, 2005 (invited presentation); Predicant Biosciences, May 25, 2005; the *53rd Conference on Mass Spectrometry and Allied Topics*, San Antonio, TX, USA, June 5-9, 2005 (poster presentation); the National Cancer Institute (NCI) sponsored conference on *Moving Biosensors to Point-of-Care Cancer Diagnostics*, Rockville, MD, USA, June 8-9, 2005 (poster presentation); the HUPO 2nd Annual Conference, Boston, MA, USA, March 11-15, 2006 (poster presentation).

Acknowledgment

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

References

Wolters DA, Washburn MP, Yates JR III (2001) An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* **73**: 5683-5690.

Peng J, Elias JE, Thoreen CC, Licklider LJ, Gygi SP (2002) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* **2**: 43-50.

Lazar IM, Karger BL (2002) Multiple open-channel electroosmotic pumping system for microfluidic sample handling. *Anal. Chem.* **74**(24): 6259-6268.

Publications

Lazar IM, Grym J, Foret F (2005) Microfabricated devices: a new sample introduction approach to mass spectrometry. *Mass Spectrom. Rev.* advance online publication 28 February 2006; doi: 10.1002/mas.20081

Lazar IM, Trisiripisal P, Sarvaiya H (2006) Microfluidic liquid chromatography system for proteomic applications and biomarker screening. *Anal. Chem.* advance online publication 4 July 2006; doi: 10.1021/ac060434y

Software, models, and experiments for systems biology

Pedro Mendes, mendes@vbi.vt.edu
Associate Professor, Virginia Bioinformatics Institute &
Adjunct Professor of Biochemistry, Virginia Tech

Modeling and simulation of biochemical networks are essential activities to aid in the understanding of cellular behavior and to facilitate quantitative interpretation of modern experiments. Systems biology combines modeling, simulation and quantitative experiments in a cooperative way, where the results of one of these activities are used to define subsequent iterations of the others. There are two main routes for creating models: one based on in vitro kinetic properties of the participating enzymes which are used to "synthesize" a model of the entire pathway (bottom-up modeling); the other is based on reverse-engineering the pathway dynamics using measurements of the response of the system to perturbations (top-down modeling). For bottom-up modeling, we are developing a method to characterize every single isoenzyme of the yeast pentose-phosphate pathway and build a comprehensive model of this important metabolic pathway. For top-down modeling, we have continued to develop a database system to store large-scale systems biology data, applied several algorithms for data analysis and data reduction, and continued developing a method for inference of biochemical models from time course data. Another important computational task in systems biology is to organize and analyze data from large-scale technologies, such as microarrays and metabolomics. To that extent we have continued the development of our DOME system. Finally, we have been applying machine learning algorithms to metabolomics data for biomarker discovery and characterization of time courses.

Keywords: systems biology; metabolomics; biochemical dynamics; data fusion; enzyme kinetics.

Modeling and simulation of biochemical networks

COPASI. We have been engaged in a collaboration with the Kummer group at EML Research to construct a new biochemical modeling and simulation software COPASI, a successor to Gepasi (Mendes, 1993; Mendes, 1997). In the past 12 months, the collaboration has resulted in a series of improvements that were released in four new test versions (www.copasi.org). Despite their beta version status, these software releases were downloaded in large numbers (about 500 per version) and we are aware of several groups that already use COPASI for research and education. The new features added were a command line version that can be used to run in batch mode or controlled by other programs, an arbitrary function optimizer, parameter estimation capabilities, a faster method of stoichiometric analysis (Vallabhajosyula et al, 2006), and the ESRS optimization algorithm (evolution strategy with stochastic ranking; Runarsson and Yao, 2000).

COPASI can now run in a command line shell and be controlled from other programs, which opens up the possibility that the software be used as a simulation engine for other software tools. We are planning to do exactly that with a system to create synthetic models of gene networks (Mendes et al, 2003) and to automate our top-down modeling method (see

Contributors:
Diogo Camacho, Hui Cheng, Autumn Clapp, Kimberly Heard, Stefan Hoops, Aejaaz Kamal, Xing Jing Li, Ana M. Martins, Bharat Mehrotra, Saroj Mohapatra, Revonda Pokrzywa, Wei Sha

next section). Other groups have also inquired about the possibility of having COPASI as a simulation engine.

Top-down modeling. Top-down modeling refers to the inference of dynamical models based solely on the observation of the dynamics of the system in response to environmental or genetic perturbations. Our efforts in top-down modeling are based on fitting a linear

$$\frac{dx_i}{dt} = \gamma_i \left(\sum_{j=1}^{N} a_{ij} x_i + \sum_{k=1}^{p} \beta_{ik} y_k + C_i \right)$$

ordinary differential equation (ODE) model to the observed dynamics. Each observable corresponds to a differential equation in the model in the form of Equation 1:

Where x_i are the variables, a_{ij} the strength of the action of x_j on x_i, y_k the external perturbations, β_{ik} the strength of perturbation y_k on x_i, and C_i represents the effects not explained by the linearization. γ_i is always equal to unity except where x_i has been knocked-out of the system (e.g. by genetic mutation), when it becomes zero. The matrix composed of the a_{ij} coefficients quantifies the interactions between the variables. This approach is similar to one developed by the Laubenbacher group at VBI where they discretize the data and then use algebraic methods to reach similar representations. Our two groups are investigating the extent to which the two methods can be used in a synergistic way as they require the same input data and arrive at similar interpretations, but use different methods to achieve this.

The model of Eq. 1 is formally similar to models that use the Jacobian of the system (the partial derivatives of the ODEs) (e.g. Yeung et al, 2002). However, since we fit the model parameters to many trajectories that are not close to a steady state, a_{ij} represents something different than the Jacobian elements. The Jacobian matrix is not constant and changes from state to state; to estimate the elements of the matrix requires trajectories that are close to a single steady state. In our method, the a_{ij} coefficients are constant and quantify the overall action of one variable onto another. Unlike the Jacobian methods, our approach does not require that each variable be perturbed directly. This is an advantage because those variables are intracellular metabolites, transcripts and proteins, and most cannot be perturbed directly because they do not cross cellular membranes. On the other hand, molecules that can enter cells, such as substrates, hormones, or toxicants, are easy to perturb – our approach is directed to this type of experimental perturbations.

While the approach is sound in theoretical terms, in practical terms it is very hard to fit a system of equations (Eq. 1) to highly nonlinear trajectory data. We are developing an iterative method that first solves Eq. 1 with low accuracy; from this solution a decision is made on which variable to knock-out to obtain further trajectories that will be used to improve the solution of Eq. 1, iterating this process as needed. This cyclic process has the virtue of prioritizing the mutations that would provide maximal information content for building the model.

Bottom-up modeling. The traditional way to construct biochemical pathway models is to use the *in vitro* enzyme kinetic properties of each enzyme. We have started modeling the yeast pentose-phosphate pathway (PPP) based on our own kinetics assays of its enzymes. The PPP is an important metabolic pathway that serves two functions: it provides precursors for biosynthesis of macromolecules (e.g. pentoses for nucleic acids), and maintenance of the NADPH/NADP+ ratio. The latter plays a major role in the response to various stresses, particularly oxidative stress. A feature of the yeast PPP is that several of its enzymes are encoded by two different genes and these are expressed under different conditions. Our hypothesis is that the different enzymes have different regulatory properties, "tuned" to serve one of the two main functions of the pathway (reductive power for stress protection or precursors for biosynthesis). To construct a model of this pathway to test this hypothesis requires that we isolate each and every one of those genes, characterize their protein products *in vitro* and use these data to construct a model of the pathway.

Purification of all of these isoenzymes is through the use of the yeast MORF mutants created in the laboratory of Mike Snyder (Gelperin et al, 2005). These mutants allow for overexpression of each single yeast gene under galactose induction. The overexpressed

genes have C-terminal tags, allowing them to be purified in a straightforward way. However, at the end of the purification not all tags are removed from the C-terminus, which means that their kinetic properties may be different from the native yeast proteins. We have established that for glucose-6-phosphate dehydrogenase, the kinetic parameters of the MORF protein are very similar to the native enzyme. Thus, we have started to purify the other PPP enzymes from MORF mutants. This will lead to a first, approximate model of the PPP in yeast. Estimation of enzyme kinetic parameters and mechanisms is being carried out with COPASI.

MIRIAM – standards for model annotation. We have been strong supporters of the establishment of standards for systems biology, which started with the creation of the Systems Biology Markup Language (SBML; Hucka et al, 2003). More recently, we participated in the establishment of standards for model annotation (MIRIAM; Le Novere et al, 2005). This addresses the problem with a large number of published biochemical models which are partially lost because they are not properly described and their source is not available electronically. The MIRIAM standard calls for models to be made available in SBML and its components specified unequivocally. MIRIAM requires that metadata about those molecular entities be kept in the annotation fields. MIRIAM applies only to published models and also requires that the encoded version of the model contain full reference to the original publication (even if this is some kind of electronic publication, such as a web page). Finally, MIRIAM requires that the SBML file of the model corresponds one-to-one to the description of the model in the publication, and that all figures of the publication can be reproduced from the SBML model encoding.

Integration of diverse large-scale data sets

An important aspect of systems biology is that it uses diverse large-scale data sets. However, the use of such data sets is not trivial. It requires that multidimensional data sets be reduced to a manageable size to create dynamical models, it requires sophisticated methods for managing the data sets, and requires algorithms to combine the data of diverse sources.

DOME: management and integration of systems biology data. We have been developing a client-server system, DOME, to manage and integrate transcriptomic, proteomic, and metabolomic data. DOME is composed of a relational database, a web-based front-end, and a set of analysis and visualization methods. In the past 12 months, DOME was expanded to include front-end functions for project managers to edit information about the biological system and the way in which the experiments were performed (metadata). The DOME schema was also extended to be able to represent features found in our data sets that required changes. In particular, we have identified that the Affymetrix *Vitis vinifera* (grape) GeneChips® contain many redundant probe sets that target the same RNA sequences but return quite different levels of fluorescence. Since no good solution currently exists to summarize these redundant probes, it is imperative that all data be passed on to the user who will have to make a decision on which numbers to use. In proteomics, another problem arises when a certain spot on a 2D-gel contains two or more proteins; in this case, we have a single numerical value for several entities (in the former we had several numerical values for each entity). The schema and front-end were changed to reflect these issues and pass all data to the user. DOME solves the problem of data integration by providing an easy way of cross-querying data from transcriptomics, proteomics, and metabolomics and displaying them in a single numerical table.

Combining data from different sources. The problem of data integration has been a major issue in bioinformatics and several solutions have been proposed for it. In DOME, we have followed a data warehouse approach, where the various data sets are stored in a single database. The positive aspect of DOME is that it has allowed us to use the three sources of data in an integrated way, combining them all in a single table. A deeper problem of integrating these data sets has largely been ignored. This relates to the different numerical characteristics of the data obtained from different sources, such as the different dynamic range of the sensors used to measure them, their signal-to-noise ratio, or the different pre-processing steps applied to them. These differences make the data sets hard to combine with each other to produce valid results and require special means of post-

processing to allow for true data integration. This numerical data integration problem is similar to what goes on in remote sensing applications, where data from a diversity of sources need to be combined. The method is usually known as "data fusion". In data fusion, one seeks to merge data sets that reflect different variables of the same system, such as an infrared satellite image with a road map of the same area; in our case, it would be, for example, measurements of gas chromatography-mass spectrometry (GC-MS) metabolite levels with microarray transcript levels. A related problem is when two data sets need to be compared that include the same variables but were measured under different conditions. This is common in clinical trials where sometimes it is needed to merge data from two drug trials carried out in different ways (different population sizes, different drug doses, etc.). This problem may be addressed by meta-analysis, a statistical methodology. In our systems biology and functional genomics experiments, both of these problems are present and thus the methods used in data fusion and in meta-analysis are relevant and being investigated.

Analysis of metabolomic data

Supervised learning. A fundamental goal of systems biology research is to understand the ability of cells to adapt to diverse environmental stimuli. Metabolomics is a means of probing these mechanisms in an unbiased way (Bino et al, 2004; Sumner et al, 2003). Unfortunately, most of the metabolites detected in such studies are yet to be identified; they can either be entirely novel metabolites or perhaps known ones that have not yet been properly annotated. It is thus important to develop strategies to prioritize which of these uncharacterized metabolites to focus on. We propose to base this on classifications of groups of metabolomics samples. The unknown metabolites that are used in the classification must then be the most important ones in the context of the differences of the samples. We have applied this rationale to a number of metabolomic studies to illustrate the potential of the method.

The supervised learning method we use was described by Jarvis and Goodacre (Jarvis and Goodacre, 2005) and is based on a genetic algorithm for variable selection applied onto discriminant function analysis (known as GA-DFA). The variable selection step carried out by the genetic algorithm is required because the discriminant analysis can only accept data sets with a number of variables between the number of classes and the number of samples. In metabolomics, the number of variables usually exceeds the number of samples, therefore the genetic algorithm selects a subset of variables that are used to discriminate the sample classes. However, because of the large indetermination in the data set, there are many subsets of variables that are able to discriminate the samples equally well. The solution is to carry out a Monte Carlo approach, where the GA-DFA algorithm is run many times and then the variables are ranked by how often they were used (Jarvis and Goodacre, 2005). The method was applied to two different data sets described below.

We were interested in characterizing several mutants of the *myo*-inositol oxygenase (MIOX) gene family of *Arabidopsis*. Several replicate, wild-type MIOX mutant plants, both knockouts and overexpressors, were grown in standard conditions and their root and leaf tissue were analyzed by non-targeted metabolomics (liquid chromatography-mass spectrometry or LC-MS, carried out by the Shulaev group at VBI). The data were reduced to fingerprints by adding all of the mass spectral scans of the whole chromatogram, resulting in a single mass spectrum. The GA-DFA algorithm was applied to these spectra. The second application was to a metabolomics study of malignant and non-malignant human cell lines, carried out in collaboration with the Shulaev group and with colleagues at the Wake Forest University Cancer Biology Department.

Classification of temporal data. The procedure described above only works when there are several metabolomics samples from each class. However, metabolomic studies are often carried out as time series. In order to be able to use such longitudinal data, it is necessary to first group several samples, i.e. find stages in the time series that are fairly similar to each other, then group all of the samples from each stage as one class, and finally apply the GA-DFA algorithm. We have used a time series of metabolite profiles from *Medicago truncatula* cell cultures elicited with the plant hormone

methyl-jasmonate. To classify the various stages of the time series, hierarchical and Bayesian clustering analyses were used; both methods indicated that there were two clear stages, with the break between them happening between the 21 h and 24 h samples. Samples were then classified into early response and late response and control samples were classified as a third class. After running the GA-DFA algorithm several thousand times, the ions of nominal mass 107, 108, 110, 111, 627, 630, 640 and 644 were chosen more than 50% of the time and are clearly related with the differences between the three sample classes. Inspection of the chromatograms defined by these masses reveals several unknown metabolites that are now prime candidates for characterization.

Acknowledgements

This work was supported by the National Science Foundation (grants DBI-0109732 and DBI-0217653), the National Institutes of Health (grant GM068947, Principal Investigator Dr. R. Laubenbacher), and the Virginia Bioinformatics Institute. The following people have contributed to very productive and enjoyable collaborations: Ursula Kummer, Rick Dixon, Lloyd Sumner, Tim Smith, Greg May, Grant Cramer, Craig Nessler, and Boris Chevone. I would like to express special thanks to Vladimir Shulaev and Reinhard Laubenbacher, and the members of their research groups at VBI. We thank Jim Walke for his invaluable administrative support to our group.

References

Bino RJ, Hall RD, Fiehn O, Kopka J, Saito K, Draper J, Nikolau BJ, Mendes P, Roessner-Tunali U, Beale MH, Trethewey RN, Lange BM, Wurtele ES, Sumner LW (2004) Potential of metabolomics as a functional genomics tool, *Trends Plant Sci.* **9**: 418-425.

Gelperin DM, White MA, Wilkinson ML, Kon Y, Kung LA, Wise KJ, Lopez-Hoyo N, Jiang L, Piccirillo S, Yu H, Gerstein M, Dumont ME, Phizicky EM, Snyder M, Grayhack EJ (2005) Biochemical and genetic analysis of the yeast proteome with a movable ORF collection, *Genes Dev.* **19**: 2816-2826.

Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A, Cuellar AA, Dronov S, Gilles ED, Ginkel M, Gor V, Goryanin II, Hedley WJ, Hodgman TC, Hofmeyr JH, Hunter PJ, Juty NS, Kasberger JL, Kremling A, Kummer U, Le Novere N, Loew LM, Lucio D, Mendes P, Minch E, Mjolsness ED, Nakayama Y, Nelson MR, Nielsen PF, Sakurada T, Schaff JC, Shapiro BE, Shimizu TS, Spence HD, Stelling J, Takahashi K, Tomita M, Wagner J, Wang J (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**: 524-531.

Jarvis RM, Goodacre R (2005) Genetic algorithm optimization for pre-processing and variable selection of spectroscopic data. *Bioinformatics* **21**: 860-868.

Le Novere N, Finney A, Hucka M, Bhalla US, Campagne F, Collado-Vides J, Crampin EJ, Halstead M, Klipp E, Mendes P, Nielsen P, Sauro H, Shapiro B, Snoep JL, Spence HD, Wanner BL (2005) Minimum information requested in the annotation of biochemical models (MIRIAM), *Nature Biotechnol.* **23**: 1509-1515.

Mendes P (1993) GEPASI: a software package for modelling the dynamics, steady states and control of biochemical and other systems. *Comput. Appl. Biosci.* **9**: 563-571.

Mendes P (1997) Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. *Trends Biochem. Sci.* **22**: 361-363.

Mendes P, Sha W, Ye K (2003) Artificial gene networks for objective comparison of analysis algorithms, *Bioinformatics* **19**: ii122-ii129.

Runarsson T, Yao X (2000) Stochastic ranking for constrained evolutionary optimization. IEEE *Trans. Evolut. Comput.* **4**: 284-294.

Sumner LW, Mendes P, Dixon RA (2003) Plant metabolomics: large-scale phytochemistry in the functional genomics era, *Phytochemistry* **62**: 817-836.

Vallabhajosyula RR, Chickarmane V, Sauro HM (2006) Conservation analysis of large biochemical networks. *Bioinformatics* **22**: 346-353.

Yeung MK, Tegner J, Collins JJ (2002) Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl. Acad. Sci. USA* **99**: 6163-6168.

Publications

Cramer GR, Cushman JC, Schooley DA, Quilici D, Vincent D, Bohlman MC, Ergul A, Tattersall EAR, Tillett R, Evans J, Delacruz R, Schlauch K, Mendes P (2005) Progress in bioinformatics – the challenge of integrating transcriptomic, proteomic and metabolomic information. In *Proceedings of the VII International Symposium on Grapevine Physiology and Biotechnology*, International Society for Horticultural Science, *Acta Horticulturae* vol. 689 Williams LE (ed) pp. 417-425.

Laubenbacher R, Mendes P (2005) A discrete approach to top-down modeling of biochemical networks. In *Computational Systems Biology*, Eils R, Kriete A (eds) pp. 229-247. Burlington, MA: Elsevier.

Lei Z, Elmer AM, Watson BS, Dixon RA, Mendes P, Sumner LW (2005) A two-dimensional electrophoresis proteomic reference map and systematic identification of 1367 proteins from a cell suspension culture of the model legume *Medicago truncatula*. *Mol. Cell. Proteomics* **4**: 1812-1825.

Le Novere N, Finney A, Hucka M, Bhalla US, Campagne F, Collado-Vides J, Crampin EJ, Halstead M, Klipp E, Mendes P, Nielsen P, Sauro H, Shapiro B, Snoep JL, Spence HD, Wanner BL (2005) Minimum information requested in the annotation of biochemical models (MIRIAM). *Nature Biotechnol.* **23**: 1509-1515.

Mendes P, Camacho D, de la Fuente A. (2005) Modelling and simulation for metabolomics data analysis. *Biochem. Soc. Trans.* **33**: 1427-1429.

Rodriguez-Fernandez M, Mendes P, Banga JR (2005) A hybrid approach for efficient and robust parameter estimation in biochemical pathways. *Biosystems* **83**: 248-265.

Methanogenic archaea, tuberculosis, coalbed methane, and type 2 diabetes

Biswarup Mukhopadhyay, biswarup@vt.edu
Assistant Professor, Virginia Bioinformatics Institute &
Adjunct Professor of Biochemistry and Biology, Virginia Tech

We have discovered that certain deeply rooted methanogenic archaea possess a new type of sulfite reductase (Fsr) that uses F_{420}, a deazaflavin coenzyme, as the electron carrier; previously described sulfite reductases use nicotinamides or cytochromes. The N-terminal half of the 70-kDa Fsr polypeptide is a homolog of the electron funneling unit of a membrane-based energy transduction system (a homolog of the mitochondrial Complex I) of the late-evolving archaea and the C-terminal half is a homolog of bacterial and archaeal dissimilatory sulfite reductases. Methanogens that contain Fsr tolerate sulfite and use sulfite as a sole sulfur source; other methanogens are killed by this oxyanion. These observations raise the possibility that methanogenesis and sulfate reduction, which are two of the most ancient metabolisms on earth and considered mutually exclusive, existed once in one organism, which probably performed sulfate-dependent anaerobic oxidation of methane. Our laboratory has also been developing microbial enrichments for converting coal to methane and isolating methanotrophs for mitigating methane-induced mine explosion. In collaboration with the Johns Hopkins University, Rotinsulu Pulmonary Hospital (Bandung, Indonesia), and Institut Teknolgi Bandung, we are working on the development of diagnostics, vaccines and therapeutics for tuberculosis. Studies of human phosphoenolpyruvate carboxykinase and an archaeal type phosphoenolpyruvate carboxylase have also allowed us to develop new avenues for designing drugs for treating type 2 diabetes and *Clostridium perfringens* infection.

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

Remnants of ancient metabolism in archaea

Our model system is the metabolisms of archaeal organisms found in submarine hydrothermal vents. In this context, we study the ecophysiology of *Methanocaldococcus jannaschii*, a deeply rooted hyperthermophilic archaeon (optimal growth temperature 85°C) and an inhabitant of the vents (Jones et al, 1983). It is a strict anaerobe and a chemolithoautotroph. The only energy source *M. jannaschii* can use is hydrogen. It performs anaerobic oxidation of hydrogen with CO_2 as the electron acceptor and in the process generates methane. For this reason, it is called a methanogen (Jones et al, 1983).

Methanogenesis

$$CO_2 + 4H_2 \longrightarrow CH_4 + 2H_2O$$

This archaeon has a considerably small genome (1.66 Mbp; Bult et al, 1996), yet it synthesizes a complete cell from H_2, CO_2, H_2S and minerals. Hence, it might represent a minimum requirement for a life form to exist independently. It is thought that life emerged on earth at a hydrothermal vent-type site as an autotroph (Wachtershauser, 2000) and a deeply rooted organism such as *M. jannaschii* probably

Contributors:

Christopher L. Case, Deanna M. Colton, Lakshmi Dharmarajan, Ashley M. Hoffman, Eric F. Johnson, Jessica L. Kraszewski, Endang Purwantini (collaborator), Jennifer P. Stieber, Dwi Susanti, Francisca Tanoerahardjo (collaborator)

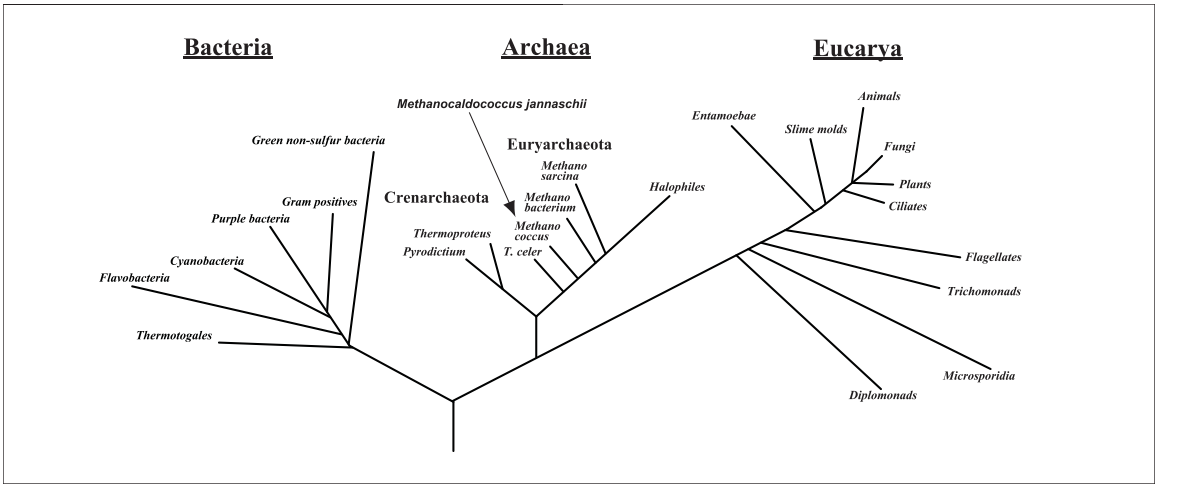
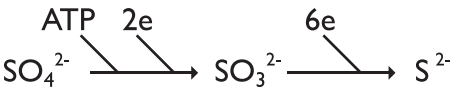


Fig. 1. 16S rRNA sequence based tree of life (Wheelis et al, 1992).

harbored remnants of the last common ancestor of all extant life. The DNA replication and transcription apparatus of archaea such as *M. jannaschii* are similar to those of the eukaryotes (Bell et al, 2001; Kelman & Kelman, 2003). Therefore, studies of *M. jannaschii* may reveal how a eukaryotic cell originated and evolved.

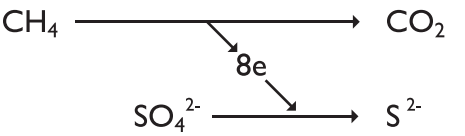
Our most recent study has focused on whether dissimilatory sulfate reduction and methanogenesis from CO_2, two of the oldest types of energy-conserving respiratory systems on earth that developed at least 3.7 and 2.7-3.2 billion years ago, respectively (Dhillon et al, 2003; Leigh, 2002), at one time existed in one organism.

Sulfate reduction



Certain bacteria and archaea carry the dissimilatory sulfate reduction pathway, whereas biological methanogenesis is restricted to the methanogenic archaea (Leigh, 2002). Anaerobic oxidation of methane is also an ancient system (Leigh, 2002). Here, methane oxidation is carried out by a consortium of anaerobic sulfate-reducing bacteria and methane-oxidizing archaea; the latter process employs a reversed archaeal methanogenesis pathway (Leigh, 2002).

Anaerobic methane oxidation



We want to know whether at one time anaerobic oxidation of methane was carried out by one organism. Such an organism would have possessed both the methane oxidation and sulfate reduction pathways and given rise to a non-sulfate-reducing methanogen as well as a non-methanogenic sulfate reducer. To answer these questions, we have looked into the possibility of methanogenesis or methane oxidation pathways co-existing with a sulfate reduction system. It is certain that the complete biological sulfate reduction pathway developed only after an ancient cell learned to deal with sulfite, because sulfite is an obligatory intermediate in the reduction of sulfate to sulfide and is toxic to cells of all types (Wedzicha, 1992). This safeguard is even more important for a methanogen with sulfate reduction ability, because sulfite reacts with and inhibits methylcoenzyme M methyl reductase, an essential enzyme for methanogenesis and the only source of energy for the methanogens (Becker & Ragsdale, 1998; Balderston & Payne, 1976). The process of anaerobic oxidation of methane faces a similar problem, because the first step for anaerobic methane oxidation likely involves methylcoenzyme M methyl reductase

(Boetius et al, 2000). Therefore, we have focused our search on looking for a remnant of an ancient sulfite reductase or sulfite detoxification system in a methanogen. We believed that such a unit could exist in methanogens in a deep-sea hydrothermal vent for the following reasons. A mixing of cold seawater that permeates through the chimney wall with sea water in the vent brings the temperature of the nutrient rich vent fluid down from 350°C to a level where life can exist and thrive (Jannasch & Mottl, 1985; Corliss et al, 1979) (Fig.2). Since the atmosphere of the earth became oxygenic, this mixing has also been bringing oxygen into the vents. This oxygen is removed though a reaction with sulfide that is present in the vent fluid and this process helps to maintain the anaerobic conditions that are required for the growth of a methanogen (Jannasch & Mottl, 1985; Boone et al, 1993), but has the potential of producing sulfite. Therefore, a vent methanogen had to deal with this harmful consequence of the appearance of oxygen on earth. It seems plausible that the prior development of a robust sulfite reduction system poised them for this task. Accordingly, *M. jannaschii* is the right organism for our investigations.

We found that *M. jannaschii* not only tolerates sulfite, but also uses it as a sole sulfur source (Johnson & Mukhopadhyay, 2005). Two hydrogen-utilizing methanogens, *Methanothermococcus thermolithotrophicus* and *Methanothermobacter thermautotrophicus*, have similar ability (Daniels et al, 1986). However, the genomes of *M. thermautotrophicus* and *M. jannaschii* do not possess a clear homolog of a sulfite reductase (Bult et al, 1996; Smith et al, 1997); the genome sequence of *M. thermautotrophicus* has yet to be determined. Starting with the hypothesis that the observed sulfite tolerance and reduction ability of these archaea were due to a relic that we have been looking for, we have studied sulfite metabolism of *M. jannaschii*. This work has led to the discovery of a new type of sulfite reductase composed of a 70-kDa polypeptide (Johnson & Mukhopadhyay, 2005). The enzyme uses coenzyme F_{420} (Fig. 3), a naturally occurring deazaflavin, as an electron carrier and reduces toxic sulfite to sulfide, an essential nutrient. We call this enzyme Fsr (F_{420}-dependent sulfite reductase) (Fig.4). The previously described sulfite reductases use nicotinamides and cytochromes as electron carriers (Fig. 6) (Johnson & Mukhopadhyay, 2005). We found that the genomes of *Methanopyrus kandleri*, which is also a vent methanogen that can grow at a temperature as high as 100°C, and *Methanothermobacter thermautotrophicus*, which grows optimally at 65°C, possess Fsr homologs (Johnson & Mukhopadhyay, 2005). Structural analyses showed that Fsr carries the likely ancestor of two electron transfer

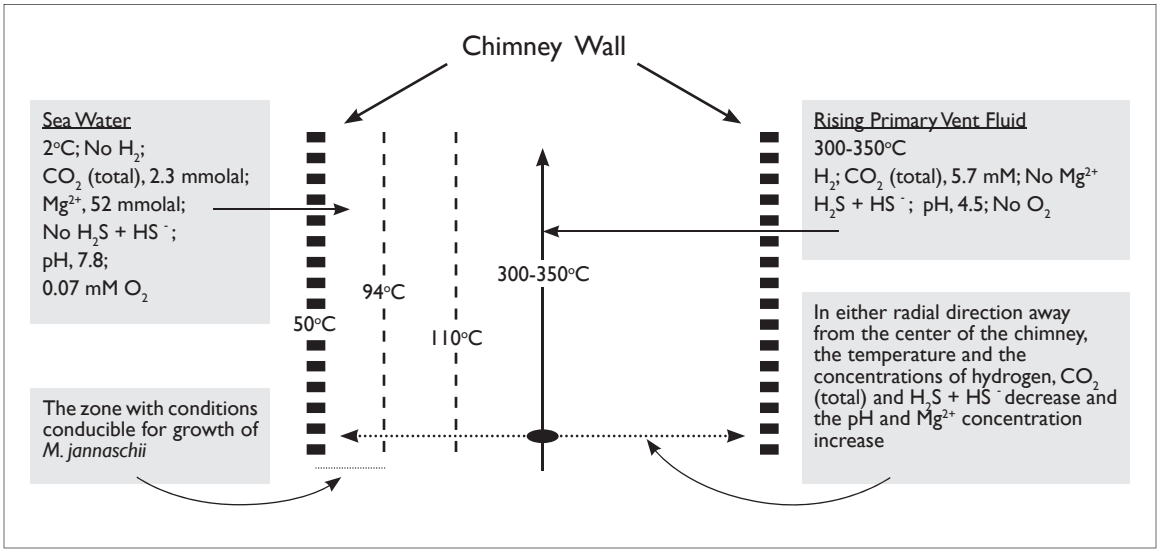


Fig.2. The habitat of *M. janaschii* (adapted from McCollom and Shock, 1997).

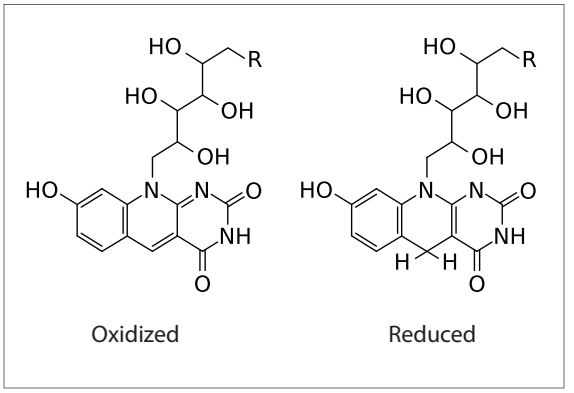


Fig. 3. Coenzyme F_{420}, a 5-deazaflavin derivative.

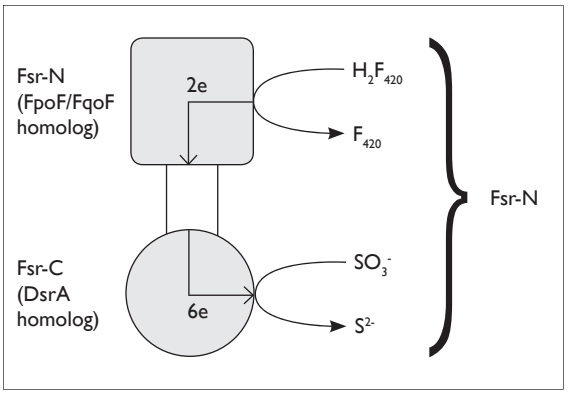


Fig. 4. Functional units of coenzyme F_{420}-dependent sulfite reductase of *M. jannaschii*. Fsr-N and Fsr-C, N- and C-terminal halves of Fsr. Fsr-N, a H_2F_{420} dehydrogenase unit; FqoF/FpoF, electron funneling units of the H_2F_{420} dehydrogenases of *Methanosarcina* and *Archaeoglobus*, respectively, two late-evolving archaeal genera (see Fig. 5) (Bruggemann et al, 2000; Baumer et al, 2000).

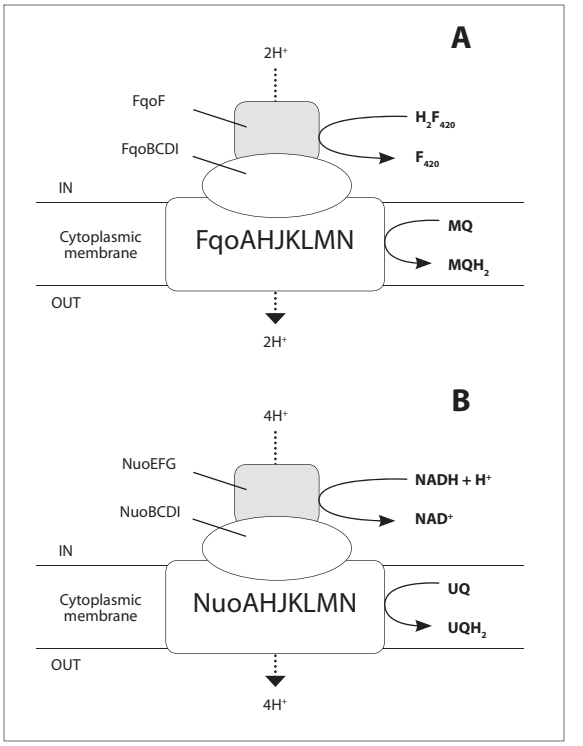


Fig. 5. Archaeal H_2F_{420} dehydrogenase and bacterial/mitochondrial NADH dehydrogenase complex (Complex I). H_2F_{420} dehydrogenase complex (A) (Bruggemann et al, 2000; Baumer et al, 2000) is homologous to Respiratory Complex I of *E. coli* and mitochondria (B) (Baumer et al, 2000). FqoF is a Fsr-N homolog (see Fig. 4). MQ, menaquinone; UQ, ubiquinone. FqoF uses protein-associated flavin for transferring electrons from H_2F_{420} to [4Fe-4S] centers. NuoEFG transfers electrons from NADH.

proteins that perform independent roles in the bacteria and late-evolving archaea (Figs 5 and 6). The N-terminal half of Fsr is an H_2F_{420} dehydrogenase (Fig. 5) and a possible ancestor for the electron input unit for a membrane-based energy transduction system of certain late-evolving archaea (Fig. 6). The latter is homologous to bacterial NADH dehydrogenase and mitochondrial Complex I. The C-terminal half of Fsr represents a dissimilatory siroheme sulfite reductase that is found in *Archaeoglobus* and dissimilatory sulfate reducing bacteria (Deppenmeier, 2004). The specific activity of Fsr is much higher than would be expected for a role in anabolic sulfide production. This highly active catabolic type of activity fits a detoxification role, where Fsr will convert sulfite present in the hydrothermal vent water into sulfide rapidly; it thus spares the organism from inhibition of its sole energy-producing pathway. One could imagine that *fsr* provided a selective advantage to an ancestral methanogen for surviving sulfite exposure when oxygen appeared on earth. Recent reports suggest that the development of a fully oxic atmosphere followed a protracted oxygenation period (Poulton et al, 2004; Kah et al, 2004; Shen et al, 2003), where a small supply of oxygen was quickly and fully sequestered in a process that could have generated sulfite, an incomplete oxidation product of sulfide. At a later time, *fsr* allowed the development of the sulfate reduction pathway within a methanogen. We investigated whether the genome of *M. jannaschii* carries the remnants of the rest of an ancient sulfate reduction pathway where as yet unidentified proteins reduced sulfate to

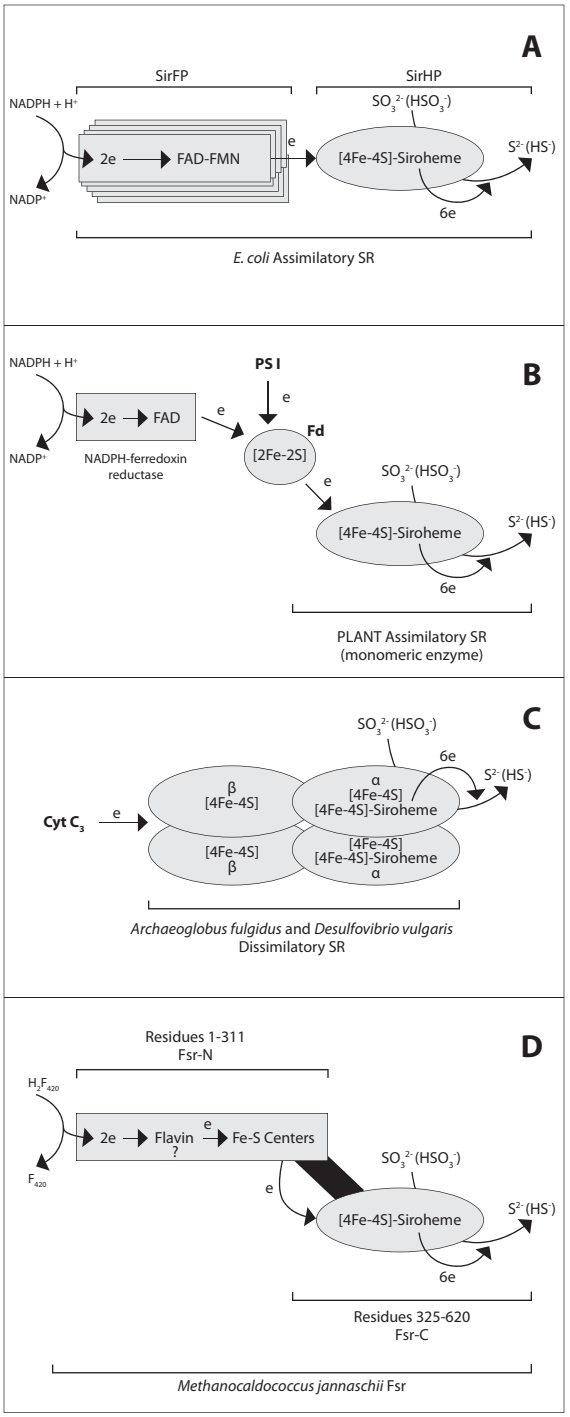


Fig. 6. Assimilatory and dissimilatory sulfite reductases. SR, or Sir, sulfite reductase; HP, heme-containing protein; FP, flavoprotein; Fd, ferredoxin; PS I, photosystem I; Cyt, cytochrome. Most dissimilatory sulfite reductases are $\alpha_2\beta_2$ proteins. The style of this figure has been adopted in part from Nakayama et al, 2000. The quaternary structure for *M. jannaschii* Fsr is not known. "?" in D indicates that it is not known whether Fsr contains bound flavin.

sulfite and Fsr reduced sulfite to sulfide. The *M. jannaschii* genome indeed bears signs of this possibility. The open-reading frames MJ0066 and MJ0973 show some sequence similarities to a 3′-phosphoadenosine-5′-phosphosulfate reductase and sulfate adenylyltransferase or ATP sulfurylase (Bult et al, 1996), which participate in the reduction of sulfate to sulfite in certain bacteria and archaea (Leustek et al, 2000). *M. jannaschii* and Fsr therefore provide an opportunity to examine how sulfate reduction and methanogenesis, two of the oldest energy-producing systems in the living world, led to certain parts of the metabolic pathways in extant organisms. Our current work focuses on the following: i) distribution of *fsr* in hydrothermal vent microorganisms; ii) structural and functional relationships of Fsr with the homologs of Fsr-N and Fsr-C (Figs 4–6); search for the remnants of the enzymes that reduced sulfate into sulfite in the sulfate-reducing methanogen or methanotroph; and search for a sulfate-reducing methanotroph.

Tuberculosis

We are using bioinformatics-based genome analysis, transposon mutagenesis and natural product sensitivity screening to identify the targets and compounds of potential value in an effort to develop therapeutics for the treatment of tuberculosis. About 10% of the worldwide cases of *Mycobacterium tuberculosis* infections are found in Indonesia (Corbett et al, 2003). We have established collaborative sites at Institut Teknologi Bandung (ITB) and Rotinsulu Pulmonary Hospital (Bandung, Indonesia) for our work on diagnostics and vaccines for tuberculosis. The team has isolated 30 clinically unique strains of *M. tuberculosis* and is currently identifying their genomic and antigenic differences. We have generated five *M. tuberculosis* antigens in purified recombinant forms and these proteins are currently being tested for their potential as vaccine or diagnostic reagents; the work involves (1) screening of blood samples collected from healthy and tuberculosis-infected patient volunteers for the presence of antibodies against the test antigens and (2) the ability of these antigens to stimulate interferon-γ production by the blood cells.

Coal bioconversion to methane and mitigation of methane-induced mine explosion

By 2025, natural gas consumption is expected to increase to almost 35 trillion cubic feet (Tcf), or 26% of United States delivered energy consumption, according to the U.S. Energy Information Administration. This represents an increase of about 52% from the 2003 level. In contrast, domestic gas production is expected to increase from 19.5 Tcf in 2001 to 26.4 Tcf in 2025. Imports, particularly of liquefied natural gas, are expected to cover the gap between consumption and production.

One source of natural gas production in the United States is coal bed methane. In 2002, coal bed methane accounted for about 9.89% of the total natural gas reserve and 8.34% of the total natural gas production in the United States. A development that will increase coal bed methane production will have a major beneficial impact on energy security in the United States. Isotope data have indicated that some of the methane in certain coal bed methane reservoirs is the product of recent microbiological activities. One way to achieve an improvement in coal bed methane production would be to enhance these microbiological activities and, in the process, convert part of the coal to methane. In collaboration with Altuda Energy Corporation (San Antonio, TX), we are therefore developing microbial enrichments with coalbed samples and testing them for coal grading activities. As a part of this collaboration, we are also isolating microorganisms that can consume methane at a low concentration of oxygen. Such organisms could be used for removing methane gas from coalmines.

Phosphoenolpyruvate carboxylase and phosphoenolpyruvate carboxykinase

In 2004, we discovered a new type of phosphoenolpyruvate carboxylase that is widespread in the archaea, absent in the eukaryotes and present in only three bacteria, one of which is *Clostridium perfringes*, a human and animal pathogen (Patel et al, 2004). Our data suggest that this new enzyme could be essential in the organisms that have it.

We are studying the structure-function relationships of the *C. perfringes* enzyme with the goal of developing therapeutics. In humans, phosphoenolpyruvate carboxykinase catalyzes the first committed step for glucose synthesis. This activity supplies glucose between meals. On the other hand, an increased and unregulated activity of this enzyme leads to type 2 diabetes. Our structural biology investigation of human phosphoenolpyruvate carboxykinase is focused on identifying sites on the enzyme for the development of a synthetic allosteric agent that will slow down but not fully inhibit the enzyme, and thereby alleviate type 2 diabetes.

Acknowledgements

The work in this laboratory has been funded by three Phase I grants and two Phase II grants from the U.S. Department of Energy, a grant from the U.S. National Aeronautics and Space Administration, a start-up fund from the Virginia Bioinformatics Institute, and a grant from the Institute for Biomedical and Public Health Sciences to Biswarup Mukhopadhyay, a Johns Hopkins University–Virginia Bioinformatics Institute collaboration grant to Endang Purwantini and Biswarup Mukhopadhyay, and an Indonesian International Joint Research Program III (RUTI III) grant from the Ministry of Research and Technology of the Republic of Indonesia to Endang Purwantini.

References

Balderston WL, Payne WJ (1976) Inhibition of methanogenesis in salt marsh sediments and whole-cell suspensions of methanogenic bacteria by nitrogen oxides. *Appl. Environ. Microbiol.* **32**: 264-269.

Baumer S, Ide T, Jacobi C, Johann A, Gottschalk G, Deppenmeier U (2000) The $F_{420}H_2$ dehydrogenase from *Methanosarcina mazei* is a redox-driven proton pump closely related to NADH dehydrogenases. *J. Biol. Chem.* **275**: 17968-17973.

Becker DF, Ragsdale SW (1998) Activation of methyl-SCoM reductase to high specific activity after treatment of whole cells with sodium sulfide. *Biochemistry* **37:** 2639-2647.

Bell SD, Magill CP, Jackson SP (2001) Basal and regulated transcription in Archaea. *Biochem. Soc. Trans.* **29:** 392-395.

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

Boone DR, Whitman WB, Rouviére P (1993) Microbiology. Diversity and taxonomy of methanogens. In *Methanogenesis: ecology, physiology, biochemistry and genetics*, Ferry JG (ed), pp. 35-80, New York: Chapman and Hall.

Bruggemann H, Falinski F, Deppenmeier U (2000) Structure of the $F_{420}H_2$:quinone oxidoreductase of *Archaeoglobus fulgidus* identification and overproduction of the $F_{420}H_2$-oxidizing subunit. *Eur. J. Biochem.* **267:** 5810-5814.

Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD, Kerlavage AR, Dougherty BA, Tomb JF, Adams MD, Reich CI, Overbeek R, Kirkness EF, Weinstock KG, Merrick JM, Glodek A, Scott JL, Geoghagen NS, Venter JC (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273:** 1058-1073.

Corbett EL, Watt CJ, Walker N, Maher D, Williams BG, Raviglione MC, Dye C (2003) The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. *Arch. Intern. Med.* **163:** 1009-1021.

Corliss JB, Dymond J, Gordon LI, Edmond JM, von Herzen RP, Ballard RD, Green K, Williams D, Bainbridge A, Crane K, van Andel TH (1979) Submarine thermal springs on Galápagos Rift. *Science* **203:** 1073-1083.

Daniels L, Belay N, Rajagopal BS (1986) Assimilatory reduction of sulfate and sulfite by methanogenic bacteria. *Appl. Environ. Microbiol.* **51:** 703-709.

Deppenmeier U (2004) The membrane-bound electron transport system of *Methanosarcina* species. *J. Bioenerg. Biomembr.* **36:** 55-64.

Dhillon A, Teske A, Dillon J, Stahl DA, Sogin ML (2003) Molecular characterization of sulfate-reducing bacteria in the Guaymas Basin. *Appl. Environ. Microbiol.* **69:** 2765-2772.

Jannasch HW, Mottl MJ (1985) Geomicrobiology of deep-sea hydrothermal vents. *Science* **229:** 717-725.

Johnson EF, Mukhopadhyay B (2005) A new type of sulfite reductase, a novel coenzyme F_{420}-dependent enzyme, from the methanarchaeon *Methanocaldococcus jannaschii*. *J. Biol. Chem.* **280:** 38776-38786.

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

Kah LC, Lyons TW, Frank TD (2004) Low marine sulphate and protracted oxygenation of the Proterozoic biosphere. *Nature* **431:** 834-838.

Kelman LM, Kelman Z (2003) Archaea: an archetype for replication initiation studies? *Mol. Microbiol.* **48:** 605-615.

Leigh JA (2002) Evolution of energy metabolism. In *Biodiversity of Microbial Life: Foundation of Earth Biosphere*, Staley JT & Reysenbach AL (eds) pp. 103-120, New York: John Wiley & Sons.

Leustek T, Martin MN, Bick JA, Davies JP (2000) Pathways and regulation of sulfur metabolism revealed through molecular and genetic studies. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **51:** 141-165.

McCollon TM, Shock EL (1997) Geochemical constraints on chemolithoautotrophic metabolism by microorganisms in seafloor hydrothermal systems. *Geochim. Cosmochim. Acta* **61:** 4375-4391.

Nakayama M, Akashi T, Hase T (2000) Plant sulfite reductase: molecular structure, catalytic function and interaction with ferredoxin. *J. Inorg. Biochem.* **82:** 27-32.

Patel HM, Kraszewski JL, Mukhopadhyay B (2004) The phosphoenolpyruvate carboxylase from *Methanothermobacter thermautotrophicus* has a novel structure. *J. Bacteriol.* **186:** 5129-5137.

Poulton SW, Fralick PW, Canfield DE (2004) The transition to a sulphidic ocean approximately 1.84 billion years ago. *Nature* **431:** 173-177.

Shen Y, Knoll AH, Walter MR (2003) Evidence for low sulphate and anoxia in a mid-Proterozoic marine basin. *Nature* **423:** 632-635.

Smith DR, Doucette-Stamm LA, Deloughery C, Lee H, Dubois J, Aldredge T, Bashirzadeh R, Blakely D, Cook R, Gilbert K, Harrison D, Hoang L, Keagle P, Lumm W, Pothier B, Qiu D, Spadafora R, Vicaire R, Wang Y, Wierzbowski J, Gibson R, Jiwani N, Caruso A, Bush D, Safer H, Patwell D, Prabhakar S, McDougall S, Shimer G, Goyal A, Pietrokovski S, Church GM, Daniels CJ, Mao J-I, Rice P, Lling J, Reeve JN (1997) Complete genome sequence of *Methanobacterium thermoautotrophicum* Delta H: functional analysis and comparative genomics. *J. Bacteriol.* **179:** 7135-7155.

Wachtershauser G (2000) Origin of life. Life as we don't know it. *Science* **289:** 1307-1308.

Wedzicha BL (1992) Chemistry of sulphiting agents in food. *Food Addit. Contam.* **9:** 449-459.

Wheelis ML, Kandler O, Woese CR (1992) On the nature of global classification. *Proc. Natl. Acad. Sci. USA* **89:** 2930-2934.

Publications

Johnson EF, Mukhopadhyay B (2005) A new type of sulfite reductase, a novel coenzyme F_{420}-dependent enzyme, from the methanarchaeon *Methanocaldococcus jannaschii*. *J. Biol. Chem.* **280:** 38776-38786.

Strategies for malaria control

Dharmendar Rathore, rathore@vbi.vt.edu
Assistant Professor, Virginia Bioinformatics Institute

Malaria, a devastating disease caused by *Plasmodium* parasites, is responsible for 10% of all the disease-associated mortality in children under the age of 5. The majority of these fatalities are caused by infections with *Plasmodium falciparum*, the most lethal form of the human malaria parasite. Furthermore, the incidence of malaria is on the rise globally, primarily due to reasons associated with the emergence and spread of drug-resistant parasites and declining environmental conditions in developing countries. In such a deteriorating situation, any effective intervention strategy to control this disease will save the lives of millions of children worldwide. We have actively pursued a genomics approach towards the identification of parasite factors that lead to the successful onset and sustenance of malaria infection in its human host. This has been pursued by selectively expressing a selected set of hypothetical proteins that were predicted to be present on the surface of the parasite and could most likely be involved in pathogenesis. Our aim is to identify parasite components that may be developed either as a vaccine or a drug target. As part of this endeavor, we identified a fasciclin domain-encoding malarial protein that plays multiple roles in the lifecycle of the parasite and could potentially be developed as a vaccine and a drug target. This report summarizes our research efforts towards exploitation of this protein as a novel therapeutic target.

Keywords: malaria; *Plasmodium falciparum*; vaccine development; drug development; infectious disease.

Antimalarial drug development

While malaria infection begins with the invasion of hepatocytes by *Plasmodium* sporozoites inoculated by an infected mosquito, clinical symptoms of malaria, which include high fever, chills and anemia, are due to the subsequent infection and rapid multiplication of the parasite inside red blood cells (RBCs). To sustain its rapid pace of development, the parasite cannibalizes hemoglobin (Hb), which represents 90% of the total protein present inside an RBC; approximately 75% of this hemoglobin is degraded during the erythrocytic stage of development (Francis et al, 1997). The degradation of hemoglobin releases heme, which is extremely toxic to the parasite. To protect itself, the parasite rapidly detoxifies heme into a crystalline product called hemozoin (Fig. 1). Several ancient antimalarials known to traditional healers as well as some of the most potent drugs developed during the pre-

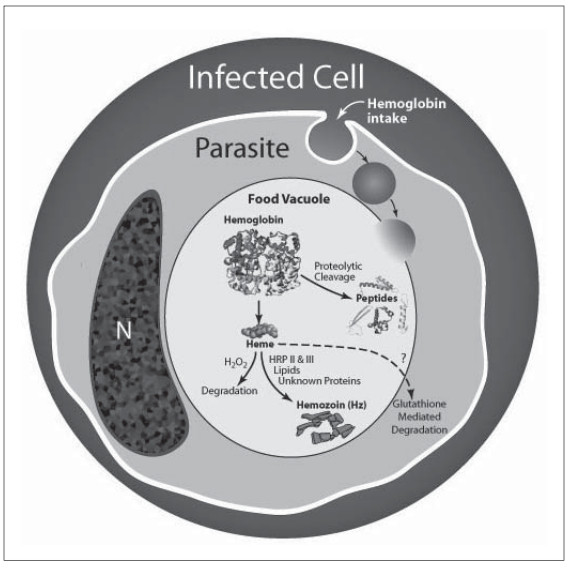


Fig. 1. Schematic representation of hemoglobin processing by the malaria parasite. A cytostome-mediated intake and transport delivers hemoglobin to the food vacuole where it undergoes proteolytic processing. Heme is released as part of this process which is converted into hemozoin.

Contributors:
Dewal Jani, Rana Nagarkatti, Kristal Cooper

genomic era, exert their antimalarial activity by blocking heme detoxification. Chloroquine, the most widely used antimalarial in history, binds to heme (Chou et al, 1980; Sullivan et al, 1996b) with a 1:2 stoichiometry (Leed et al, 2002) and prevents its detoxification, leading to heme toxicity and death of the parasite. Similarly, artemisinin, which is currently the most effective drug for the treatment of malaria, has been shown to form heme adducts that cannot be detoxified (Kannan et al, 2002). As advancements in our understanding of the detoxification process have been made, new leads are continuously being explored for the identification of novel compounds that can block this pathway and heme detoxification continues to be an important target in antimalarial drug discovery. However, the way in which heme is converted to hemozoin is poorly understood (Egan et al, 2001; Fitch et al, 1999; Sullivan et al, 1996a).

We have demonstrated that the fasciclin domain-containing protein is responsible for this conversion. The protein detoxifies heme in a dose-dependent manner and can convert up to 50% of free heme into hemozoin. All of the currently available inhibitors of hemozoin synthesis act by binding to the substrate (heme). The fasciclin domain-containing protein is therefore a completely novel target for malaria drug development. Due to its potent detoxification activity, we have named this protein Heme Detoxification Protein or HDP. We also found that HDP is present in all the members of the *Plasmodium* genus and is functionally conserved.

In collaboration with the Broad Institute of MIT and Harvard University at Cambridge, MA, we will start high throughput screening of compound libraries to identify potential compounds that can inhibit HDP activity. At least 100 000 compounds will be screened. Efforts are already underway towards assay optimization for the high throughput screening and development of reagents for this assay. Target molecules that are identified by the initial screening will be evaluated for their inhibitory activity in an *in vitro* hemozoin formation assay, parasite growth assay as well as in a rodent malaria model.

HDP as a malaria vaccine candidate

The invasion of liver cells by *Plasmodium* parasites leads to the onset of malaria in the host. Parasite proteins that interact with the host liver cells are therefore obvious targets for designing intervention strategies for malaria. We recently found that HDP is involved in the adhesion of *Plasmodium* sporozoites to liver cells and antibodies raised against this protein can prevent >94% of *P. falciparum* sporozoites from invading liver cells. This invasion inhibitory activity is comparable to the activity observed with antibodies raised against circumsporozoite protein, the only other major parasite protein involved in cell invasion and a vaccine candidate that is currently undergoing phase III clinical trials. We have subsequently mapped the epitope to a 25-amino-acid region, which is highly conserved across different species of the parasite. This suggests that this fragment is playing an important role in the invasion process and could become part of a multi-epitope vaccine against malaria. We have also identified the ortholog of this protein in *Plasmodium yoelii* (mouse malaria) parasites and shown that its protein also participates in the receptor ligand interactions. This suggests that HDP plays an important role in the onset of infection across the *Plasmodium* genus.

Cell-mediated immune responses directed against the *Plasmodium* parasite residing inside the liver cells play an important role in conferring protection against malaria (Ballou, 2005; Levine, 2004). Therefore, antigens that are expressed during the hepatocytic development of the parasite are increasingly being tested as vaccine candidates for malaria (Prieur et al, 2004). In collaboration with Dr. John Sacci at the University of Maryland, we found that HDP is expressed by the parasite during the exo-erythrocytic stages of development inside the liver cells (Fig. 2), thus suggesting that the protein not only plays an important role during the liver stages of development, but it has the potential of being processed and presented by the major histocompatibility complex (MHC) class I molecules.

Furthermore, anti-HDP antibodies were found in 60% of the malaria-infected subjects residing in a malaria-endemic area of Mali, which suggests that the protein is recognized

by the host immune responses. In collaboration with Dr. Kirsten Lyke at the University of Maryland, we will investigate cell-mediated immune response against HDP in peripheral blood mononuclear cells obtained from children infected with malaria. Using its ortholog, we will also investigate its potential to serve as a vaccine in a *P. yoelii*-based mouse malaria model. These experiments will provide an important impetus towards developing HDP as a malaria vaccine candidate for humans.

Developing an attenuated malaria parasite vaccine

Historically, vaccines based on attenuated pathogens have been the most successful. While a licensed malaria vaccine is currently not available, in an experimental setup, exposure of human subjects to radiation-attenuated parasites protects individual when they are subsequently challenged with virulent parasites (Hoffman et al, 2002). An attenuated, pathogen-based vaccine works by allowing the pathogen a limited amount of development in its host. This permits the host immune system to recognize and memorize its unique signatures (proteins) before eliminating the infection. Subsequently, when the host is attacked by the virulent strain, the immune system is able to recognize the invader and quickly mounts a response that eliminates the infection. Consequently, development of a malaria parasite with an impaired growth is a prerequisite for developing an attenuated vaccine. However, rapid multiplication of the malaria parasite in its human host is one of its hallmarks and, therefore, the major objective of this investigation is to develop genetically attenuated *P. falciparum* malaria parasites with a defective cell-division machinery that can be exploited as an attenuated vaccine for malaria. This is a collaborative effort with the laboratory of Dr. Sanjai Kumar at the Food and Drug Administration in Rockville, Maryland.

Centrin, a protein involved in cell division (Salisbury et al, 2002), is being developed as the target for this project. In *Leishmania*, a parasite responsible for leishmaniasis in humans, the targeted deletion of the centrin gene leads to attenuation and the development of the parasites in human macrophages is severely compromised (Selvapandiyan et al, 2004). The

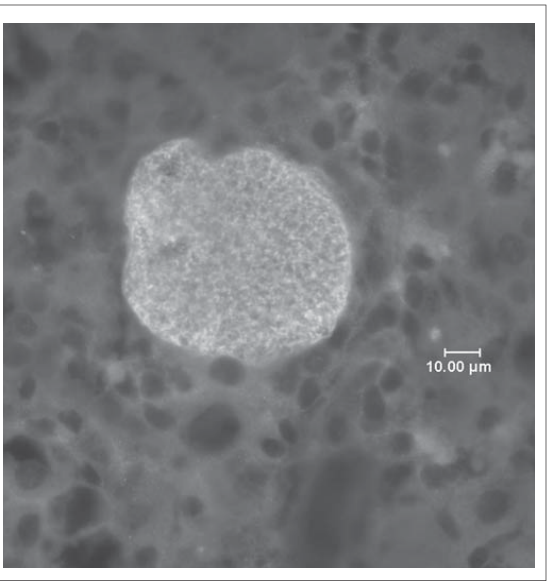


Fig. 2. Expression of Heme Detoxification Protein (HDP) during the development of *P. falciparum* parasites inside the liver. A chimeric mouse system where foci of human hepatocytes have been developed in the mouse liver was utilized for this assay (Sacci et al, 2006). Animals were infected with *P. falciparum* sporozoites and sacrificed on day seven after the onset of infection. Sections from liver tissue were prepared and the expression of HDP was detected by immunofluorescence using anti-HDP antibodies. HDP was highly expressed during intra-hepatocytic stages of the lifecycle of the parasite.

P. falciparum genome encodes four isoforms of centrin (Gardner et al, 2002). We have developed knock-out vectors for centrin and have initiated efforts towards developing centrin gene-knockouts in *P. falciparum* parasites. Centrin gene knock-out parasites will be tested for their virulence in culture and in *Aotus nancymai* monkeys, a non-primate model for *P. falciparum* malaria. This will be the first investigation of its kind in malaria, where the targeted knock-out of a gene involved in cell division will be exploited towards development of a whole organism vaccine.

Other projects

Immunological parameters of* P. vivax *malaria in endemic populations. *Plasmodium vivax* is the second most prevalent species of malaria and is responsible for almost half of the

total malaria cases in the world. Unfortunately, most of the vaccine development efforts have been focused towards developing a vaccine that will only protect against *P. falciparum* malaria. Furthermore, only a limited amount of immunological characterization of *P. vivax* infections in endemic regions has been performed. We are supporting the efforts of the Malaria Program of the Naval Medical Research Center in Silver Spring, Maryland, for immunological characterization of *P. vivax* malaria infections in the malaria endemic regions of Indonesia and Peru. This project is funded by a contract from the US Department of Defense. We have produced recombinant *P. vivax* antigens that are now being utilized for evaluating host immune responses in subjects infected with *P. vivax* parasites. These studies will provide a deeper understanding of the host immune responses to *P. vivax* malaria, which will be utilized for designing vaccines against this parasite.

Fungal antigens in chronic rhinosinusitis. Members of the *Alternaria* genus have recently been shown to cause chronic rhinosinusitis. In a collaboration with Drs. Chris Lawrence at the Virginia Bioinformatics Institute and Hirohito Kita at the Mayo Clinic, we have recombinantly produced and purified fungal antigens that have been shown to be capable of inciting an immune response in the host. This immune response ultimately leads to pathogenesis associated with this disease. This work will provide vital clues about the molecular triggers expressed by the fungi, which could be exploited for developing therapies in the future.

Acknowledgements

This work has been funded through an ASPIRES grant program, Department of Defense contract, Johns Hopkins University-Virginia Bioinformatics Institute Infectious disease research grant and funds from the Virginia Bioinformatics Institute. We thank our collaborators Drs. Sanjai Kumar, John Sacci, Kirsten Lyke and Wandy Beatty for their help. We also thank June Mullins for her help in graphic design.

References

Ballou WR (2005) Malaria vaccines in development. *Expert Opin. Emerg. Drugs* **10**: 489-503.

Chou AC, Chevli R, Fitch CD (1980) Ferriprotoporphyrin IX fulfills the criteria for identification as the chloroquine receptor of malaria parasites. *Biochemistry* **19**: 1543-1549.

Egan TJ, Mavuso WW, Ncokazi KK (2001) The mechanism of beta-hematin formation in acetate solution. Parallels between hemozoin formation and biomineralization processes. *Biochemistry* **40**: 204-213.

Fitch CD, Cai GZ, Chen YF, Shoemaker JD (1999) Involvement of lipids in ferriprotoporphyrin IX polymerization in malaria. *Biochim. Biophys. Acta* **1454**: 31-37.

Francis SE, Sullivan DJ Jr, Goldberg DE (1997) Hemoglobin metabolism in the malaria parasite *Plasmodium falciparum*. *Annu. Rev. Microbiol.* **51**: 97-123.

Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan MS, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DM, Fairlamb AH, Fraunholz MJ, Roos DS, Ralph SA, McFadden GI, Cummings LM, Subramanian GM, Mungall C, Venter JC, Carucci DJ, Hoffman SL, Newbold C, Davis RW, Fraser CM , Barrell B (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**: 498-511.

Hoffman SL, Goh LM, Luke TC, Schneider I, Le TP, Doolan DL, Sacci J, de la Vega, P, Dowler M, Paul C, Gordon DM, Stoute JA, Church LW, Sedegah M, Heppner DG, Ballou WR, Richie TL (2002) Protection of humans against malaria by immunization with radiation-attenuated *Plasmodium falciparum* sporozoites. *J. Infect. Dis.* **185**: 1155-1164.

Kannan R, Sahal D, Chauhan VS (2002) Heme-artemisinin adducts are crucial mediators of the ability of artemisinin to inhibit heme polymerization. *Chem. Biol.* **9**: 321-332.

Leed A, DuBay K, Ursos LM, Sears D, De Dios AC, Roepe PD (2002) Solution structures of antimalarial drug-heme complexes. *Biochemistry* **41**: 10245-10255.

Levine MM (2004) *New generation vaccines*. Marcel Dekker, Taylor & Francis, New York London.

Prieur E, Gilbert SC, Schneider J, Moore AC, Sheu EG, Goonetilleke N, Robson KJ, Hill, AV (2004) A *Plasmodium falciparum* candidate vaccine based on a six-antigen polyprotein encoded by recombinant poxviruses. *Proc. Natl. Acad. Sci. USA* **101**: 290-295.

Sacci JB Jr, Alam U, Douglas D, Lewis J, Tyrrell DL, Azad AF, Kneteman NM (2006) *Plasmodium falciparum* infection and exoerythrocytic development in mice with chimeric human livers. *Int. J. Parasitol.* **36**: 353-360.

Salisbury JL, Suino KM, Busby R, Springett M (2002) Centrin-2 is required for centriole duplication in mammalian cells. *Curr. Biol.* **12**: 1287-1292.

Selvapandiyan A, Debrabant A, Duncan R, Muller J, Salotra P, Sreenivas G, Salisbury JL, Nakhasi HL (2004) Centrin gene disruption impairs stage-specific basal body duplication and cell cycle progression in Leishmania. *J. Biol. Chem.* **279**: 25703-25710.

Sullivan DJ Jr, Gluzman IY, Goldberg DE (1996a) Plasmodium hemozoin formation mediated by histidine-rich proteins. *Science* **271**: 219-222.

Sullivan DJ Jr, Gluzman IY, Russell DG, Goldberg DE (1996b) On the molecular mechanism of chloroquine's antimalarial action. *Proc. Natl. Acad. Sci. USA* **93**: 11865-11870.

Publications

Rathore D, Jani D, Nagarkatti R, Kumar S (2006) Heme detoxification and antimalarial drugs: known mechanisms and future prospects. *Drug Discov. Today: Therapeutic Strategies* advance online publication 11 July 2006; doi: 10.1016/j.ddstr.2006.06.003

Rathore D, McCutchan TF, Sullivan M, Kumar S (2005) Antimalarial drugs: current status and new developments. *Expert Opin. Investig. Drugs* **14**(7): 871-883.

Computational biology for mitochondrial medicine

David C. Samuels, dsamuels@vbi.vt.edu
Assistant Professor, Virginia Bioinformatics Institute

Our research focuses on the application of computational and mathematical methods to biomedical research, mainly in the area of mitochondrial medicine. Our major project this year is on the mitochondrial toxicity of the nucleoside analog drugs used as antiviral agents in the treatment of HIV/AIDS. This toxicity involves the metabolism of DNA precursors within mitochondria. These metabolic pathways are also of interest to us since a number of serious genetic diseases are caused by mutations in genes for enzymes and transporters involved in this metabolism. For both of these projects we have developed an initial computational model of the biochemistry of these interacting metabolic pathways. A second area of research involves the susceptibility of the mitochondrial DNA molecule to mutation and the relationship of these mutations to the aging process. This research has overlaps with the nucleotide metabolism project, since alterations in nucleotide metabolism are one cause of the mitochondrial DNA mutations that arise as part of the normal aging process. Finally, we discuss a small project on the intracellular life-cycle of the influenza virus. We end this report with a discussion of possible future projects for our group. One new area of interest to us concerns the interactions of intracellular pathogens with the mitochondria of the host cell.

Keywords: DNA; nucleotide metabolism; antiviral drugs; mitochondrial toxicity; aging; mutation mechanisms.

The toxicity of antiviral nucleoside analog drugs

The first successful treatment for human immunodeficiency virus (HIV) infection was AZT (azidothymidine; zidovudine), an analog of the nucleoside thymidine. AZT is a pro-drug, which must be metabolized within the cell to its active form, AZT triphosphate (AZTTP). This metabolism uses the enzymes for the transport and phosphorylation of thymidine (Samuels, 2006a). Like thymidine triphosphate (dTTP), AZTTP can be incorporated into a replicating DNA strand. However, AZT lacks the chemical structure needed to add the next nucleotide to the DNA strand. The incorporation of AZTTP therefore stops DNA replication, a process called "chain termination". This is the antiviral activity of the drug against viruses such as HIV, which must produce a DNA molecule as part of their life-cycle.

AZT is just one example of a class of drugs known as nucleoside analogs. The current AIDS therapy is Highly Active Anti-Retroviral Therapy (HAART). This therapy typically uses two nucleoside analogs in combination with one other drug, usually a protease inhibitor. This therapy has been very successful at reducing the level of the HIV virus, but it does not remove the virus completely, so it must be continued as a life-long therapy. The long-term use of this therapy is a challenge. Multiple drugs are used in HAART as an attempt to prevent, or at least delay, adaptation of the virus to the drugs. However, adaptation of the virus still occurs and it is normal to have to cycle the patients through different combinations of drugs over the course of their treatment.

Nucleoside analogs can interfere with the replication of any DNA molecule, not just viral DNA. The analogs used as antiviral agents have their strongest interaction with the viral DNA polymerases, and, in general,

Contributors:
Patrick C. Bradshaw, Harsha Rajasimha

have very little interaction with the host DNA polymerases. The analogs that are currently used in HAART do have some interaction with the mitochondrial DNA polymerase (Pol-γ) and these drugs do have toxic effects, which are due to mitochondrial dysfunction (Nolan & Mallal, 2004). A study in 2000 showed that over a treatment period of 45 weeks, 21% of patients discontinued HAART due to toxic side-effects, while only 5% discontinued due to failure of the therapy to control the virus (Monforte et al, 2000). Mitochondrial toxicity of these drugs is a very serious issue in the care of these patients. The need to cycle patients through different combinations of the drugs to avoid virus adaptation means that the use of nucleoside analogs with appreciable mitochondrial toxicity cannot be avoided.

The fundamental mechanism for the toxicity of nucleoside analogs is not clear. Incorporation of the analog into mitochondrial DNA (mtDNA) is only one of many proposed mechanisms (Samuels, 2006a). This metabolism involves four parallel pathways for the import and phosphorylation of the four natural nucleosides — thymidine, deoxyadenosine, deoxycytidine, and deoxyguanosine — within the mitochondrion. These four pathways share enzymes at various points, coupling the pathways together through competitive inhibitions. The direct raw materials for the production of new DNA are the four triphosphate nucleotides, dATP, dGTP, dCTP and dTTP. The relative concentrations of these nucleotides within the mitochondrion have a strong effect on the fidelity of replication of mtDNA. The addition of one or more nucleoside analogs to this metabolism could cause mitochondrial toxicity through competitive inhibition at any point in this metabolism, if that inhibition eventually results in altered triphosphate nucleotide concentrations. Furthermore, the probability of incorporation of the analog into the mtDNA strand depends on the ratio of the analog and natural nucleotide triphosphate concentrations, such as with AZTTP/dTTP for example. Alterations in dTTP concentrations caused by inhibitions at any point in this metabolism could affect AZT incorporation into mtDNA and cause toxicity. A systems biology approach to this problem is needed to understand the interactions between these coupled metabolic pathways and explore the combined effects of multiple mechanisms of toxicity.

To investigate these interacting metabolic pathways, we have constructed a computational model representing this metabolism within a single mitochondrion, including the metabolism of AZT (Bradshaw et al, 2005a). This mitochondrial metabolism is coupled to analogous pathways in the cytoplasm through the transport of nucleotide di- and triphosphates between the cytoplasm and the mitochondrion via a protein called the DeoxyNucleotide Carrier (DNC). In this simulation, we modeled this metabolism in three simulated cell types: rapidly dividing cells, slowly dividing cells, and non-dividing cells. The different cell types are represented in the computational model by different levels of nucleotides in the cell cytoplasm. The mitochondrial toxicity of nucleotide analogs has complicated tissue dependence with different tissues affected by different analogs. AZT toxicity primarily affects post-mitotic cells. In this first publication from this project (Bradshaw et al, 2005a), we calculated the probability of AZT incorporation into mtDNA in the three modeled cell types to see what assumptions would lead to the toxicity behavior observed *in vivo*. Our initial set of kinetic parameters for this model resulted in an insignificant rate of AZT incorporation in all three cell types. Analysis of the model showed that the DNC was actually playing a protective role in the model by transporting phosphorylated AZT out of the mitochondrion and into the cytoplasm. Decreasing the activity of the DNC transporter for AZT produced a large increase in the probability of AZT incorporation into mtDNA in post-mitotic cells (an indication of toxicity) with far less of an effect in slowly or rapidly dividing cells. Based on the simulation results, we concluded that further experimental investigations should focus on the DNC activity with AZT.

Future work. Further development of this model, funded by the National Institutes of Health (NIH), has just begun in April 2006. We will extend the model to include effects such as Pol-γ infidelity due to imbalances in nucleotide triphosphate pools. We also will develop a computational model of another analog, lamivudine (3TC). This deoxycytidine

analog is the one most often used in HAART in combination with AZT. We will merge the AZT and 3TC models to look for interaction effects that would not occur with either drug alone. A major part of the future work will be a new collaboration with Edward McKee, a biologist at the University of Notre Dame/ Indiana University School of Medicine. McKee is measuring AZT metabolism in isolated rat heart and liver mitochondria, and he will begin measurement of 3TC metabolism as part of this collaboration.

The metabolism of DNA precursors in mitochondria

Since mtDNA replication must occur continuously in all cell types independent of the cell cycle, mitochondria have their own set of metabolic pathways for the import and phosphorylation of DNA precursors from the cytoplasm. A number of genetic diseases are caused by mutations in the enzymes of these pathways. Our computational work is aimed first at understanding the normal function of this metabolism in healthy cells. The eventual goal is to apply that knowledge to these genetic diseases.

From our initial simulation model of this metabolism (Bradshaw and Samuels, 2005), we have proposed a list of five testable hypotheses: (1) The supply of DNA precursors within each mitochondrion is too small to support even a single mtDNA replication event; therefore replication is limited by the rate of import of nucleotides and nucleosides from the cytoplasm. (2) The time required to complete mtDNA replication in post-mitotic cells is an order of magnitude longer than the often stated time of 1 hour measured in cell culture. (3) Mitochondria in post-mitotic cells act as net sources of deoxynucleotides by importing nucleosides from the cytoplasm, phosphorylating them internally and then exporting most of this material back out into the cytoplasm. (4) Based on metabolic control analysis, the thymidine kinase 2 (TK2) and nucleoside diphosphate kinase (NDPK) enzymes have the most control over mtDNA replication rates in post-mitotic cells, but the deoxynucleotide carrier (DNC) has the most control in rapidly dividing cells. (5) Rapidly dividing cells derive most of their mtDNA precursors from the cytoplasm as phosphorylated nucleotides, not as nucleosides. After this paper was published, we found evidence from the older experimental literature (Piko et al, 1984) that supports our second hypothesis, which was the most surprising of the five hypotheses. In a related paper collaborating with the medical schools at Columbia University and at the University of Newcastle, we presented the first measurements of the amount of mtDNA depletion in patients with a mutation in the *TK2* gene, one of the enzymes in this metabolism (Durham et al, 2005).

Future work. We are developing projects with two new experimental collaborators based on this model. William Lewis, a pathologist at the Emory University School of Medicine, has a number of mouse lines that are genetically engineered to over-express some of the enzymes in this metabolism. Some of these mouse lines have cardiomyopathy phenotypes due to mtDNA damage; others do not have these phenotypes. We are applying for funding for a joint project to apply our computational model to these mouse lines in an attempt to explain the phenotype variation. If that is successful, we will then use the computational model to predict the phenotypes of crosses of these mouse lines, and then carry out the lab experiments at Emory University School of Medicine. Lewis is also an expert on cardiomyopathy due to AZT toxicity (Lewis et al, 2003), and he will be collaborating with us on that project too. The other new collaborator is Chris Mathews from Oregon State University. Mathews' laboratory is one of the few research groups with experience at measuring the concentration of nucleotides within mammalian mitochondria. We are applying for joint funding to measure the quantities of these nucleotides in tissue samples from patients with genetic defects in this metabolism for comparison with our computational models of these diseases.

We are also developing a new computational collaboration with Yang Cao, a faculty member in the Department of Computer Science, Virginia Tech. Cao is an expert in discrete stochastic biochemical simulations. Our preliminary simulation is continuous and deterministic. However, the results of this simulation show that under some conditions the number of substrate molecules in these reactions can

decrease to approximately 100 or less. Discrete stochastic models will allow us to explore the role of fluctuations under these conditions.

Physical properties of mtDNA related to the rate of aging

We calculated the binding energy of the DNA strands of the mitochondrial genomes of 76 mammalian species and found a direct relationship with the longevity of these species (Samuels, 2005). Extension of this calculation to three short-lived invertebrates and one yeast species showed that this relationship extended down a further two orders of magnitude in lifespan. We developed a physical explanation for this relationship based on the local separation of the DNA strands due to thermal fluctuations, called DNA bubbles. Since separation of the DNA strands increases the risk of damage to the DNA molecule, we made the hypothesis that lifespan would be proportional to the equilibrium constant between the protected double-stranded form of mtDNA and the vulnerable single-stranded form. Using thermodynamics, this constant can be calculated and this equation had the correct functional form to fit our data relating mtDNA binding energy and longevity. The least-squares fit parameter in these equations corresponds to the nucleation size of the DNA bubbles. Our data gave a value of 22 base pairs (plus or minus 2 bp), which is in excellent agreement with recent experiments showing that this nucleation size is approximately 20 bp (Zeng et al, 2004).

Curiously, all of the mtDNA genomes that we analyzed had average binding energies less than the value for a random sequence with equal base frequencies. We speculated that this indicated there was an evolutionary force to decrease the binding energy, perhaps to make strand separation for DNA replication and transcription easier. The implication is that physical properties of the DNA molecule can also be subject to evolution, and that this can limit the range of potential sequences that evolution can explore.

Other mitochondrial research projects

The global human population has been classified into approximately 20 haplogroups based on small variations in their mitochondrial genomes. A comparison of the mtDNA sequences of individuals from different haplogroups typically gives differences of about 20 to 50 base pairs in a genome of 16.6 kbp. A large number of papers have been published claiming that certain mtDNA haplogroups have a larger or smaller propensity for various medical conditions. Many of these studies may have been underpowered due to the small number of subjects involved and therefore some of the reports of associations may be false positives. To deal with this question, we developed a Monte-Carlo simulation of these studies with a human genetics medical statistician at the University of Edinburgh and a clinician at the University of Newcastle (Samuels et al, 2006). Using the results of these simulations, we developed a new model equation to allow researchers to calculate power curves for disease association studies involving mitochondrial haplogroups.

Influenza modeling

In a collaboration with Karen Duca (Virginia Bioinformatics Institute), Robert Gogal (Virginia College of Osteopathic Medicine), Mark Jones and Paul Plassman (The Bradley Department of Electrical and Computer Engineering) at Virginia Tech, and Larisa Gubareva (Infectious Diseases) at the University of Virginia Health System, we have been developing a preliminary model of the initial stages of influenza infection. This work has been funded by a seed grant from the Virginia Tech – University of Virginia – Carilion Health System Collaborative Research Grants program. The purpose of this project is to carry out a small number of experiments on specific strains of influenza virus to develop a quantitative data set that can then be used in computational models. A set of three computational models is being developed: an intracellular viral replication model (Samuels), a model of the host cell response to the virus (Duca), and a multicellular model of the spread of the virus between cells (Jones and Plassman).

Long term directions for our research

In recent years, increasing evidence suggests that many intracellular pathogens interact with the mitochondria of their hosts, and that this interaction is a critical part of their pathogenesis (D'Augustino et al, 2005). Often this interaction involves control of the host apoptosis signaling pathways (Everett & McFaddan, 2001), which are primarily controlled by the mitochondria. We are interested in applying systems control analysis to the apoptosis signaling network, and we have begun a collaboration on this subject with two engineers, Pushkin Kachroo and Michael Hsaio from the Bradley Department of Electrical and Computer Engineering at Virginia Tech.

Not all interactions of pathogens and mitochondria involve apoptosis control, at least not in any obvious manner. For example, recent clinical data from HIV-positive but treatment-naïve patients show that the HIV virus itself somehow causes depletion of mtDNA in the patients' blood cells (Maagaard et al, 2006). I would be very interested in developing research projects in this area. This would build on our past work on mtDNA modeling and integrate our new work on AIDS treatments. Pursuing these new research goals will depend on success in recruiting new laboratory and clinical collaborators from the infectious disease research community. Our work on antiviral drug toxicity and nucleotide metabolism is a first step into this community and we have been successful in recruiting new laboratory collaborators for those projects.

Acknowledgements

This work has been supported by the Virginia Bioinformatics Institute, the Virginia Tech – University of Virginia – Carilion Health System Collaborative Research Grants program, and by National Institutes of Health grant DK070533.

References

D'Augostino DM, Bernardi P, Chiecho-Bianchi L, Ciminale V (2005) Mitochondria as functional targets of proteins coded by human tumor viruses. *Adv. Canc. Res.* **94**: 87-142.

Everett H, McFadden G (2001) Viruses and apoptosis: meddling with mitochondria. *Virology* **288**: 1-7.

Lewis W, Day BJ, Copeland WC (2003) Mitochondrial toxicity of NRTI antiviral drugs: An integrated cellular perspective. *Nat. Rev. Drug Dis.* **2**: 812-822.

Maagaard A, Holberg-Petersen M, Kvittingen EA, Sandvik L, Bruun JN (2006) Depletion of mitochondrial DNA copies/cell in peripheral blood mononuclear cells in HIV-1-infected treatment naïve patients. *HIV Med.* **7**: 53-58.

Monforte AA, Lepri AC, Rezza G, Pezzotti P, Antinori A, Phillips AN, Angarano G, Colangeli V, De Luca A, Ippolito G, Cagesse L, Soscia F, Filice G, Gritti F, Narciso P, Tirelli U, Moroni M (2000) Insights into the reasons for discontinuation of the first highly active antiretroviral therapy (HAART) regimen in a cohort of antiretroviral naïve patients, *AIDS* **14**: 499-507.

Nolan D, Mallal S (2004) Complications associated with NRTI therapy: update on clinical features and possible pathogenic mechanisms. *Antivir. Ther.* **9**: 849-863.

Piko L, Bulpitt KJ, Meyer R (1984) Structural and replicative forms of mitochondrial DNA in tissues from adult and senescent BALB/c and Fischer 344 rats. *Mech. Ageing and Dev.* **26**: 113-131.

Zeng Y, Montrichok A, Zocchi G (2004) Bubble nucleation and cooperativity in DNA melting. *J. Mol. Biol.* **339**: 67-75.

Publications

Bradshaw PC, Li JX, Samuels DC (2005a) A computational model of mitochondrial AZT metabolism. *Biochem. J.* **392**: 363-373.

Bradshaw PC, Rathi A, Samuels DC (2005b) Mitochondrial-encoded membrane potential transcripts are pyrimidine-rich while soluble protein transcripts and ribosomal RNA are purine-rich. *BMC Genomics* **6**: Art # 136.

Bradshaw PC, Samuels DC (2005) A computational model of mitochondrial deoxynucleotide metabolism and mtDNA replication. *Am. J. Physiol. Cell Physiol.* **288**: C989-C1002.

Durham SE, Bonilla E, Samuels DC, DiMauro S, Chinnery PF (2005) Mitochondrial DNA copy number threshold in mtDNA depletion myopathy. *Neurology* **65**: 453-455.

Durham SE, Samuels DC, Chinnery PF (2006) Is selection required for the accumulation of somatic mitochondrial DNA mutations in post-mitotic cells? *Neuromuscular Disorders* advance online publication 8 May 2006; doi: doi:10.1016/j.nmd.2006.03.012

Samuels DC (2005) Life span is related to the free energy of mitochondrial DNA. *Mech. Aging and Dev.* **126**: 1123-1129.

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

Samuels DC (2006b) Computational models of mitochondrial DNA in aging. In *Handbook of Models for Human Aging*, Conn PM (ed) pp 591-600. Academic Press.

Samuels DC, Carothers AD, Horton R, Chinnery PF (2006) The power to detect disease associations with mitochondrial DNA haplogroups. *Am. J. Hum. Gen.* **78**: 713-720.

Bacterial genomics and bioinformatics

João C. Setubal, setubal@vbi.vt.edu
Associate Professor, Virginia Bioinformatics Institute & Department of Computer Science, Virginia Tech

This report gives an overview of the activities of my research group at the Virginia Bioinformatics Institute. It consists of descriptions of projects and publications for the reporting period April 2005 to March 2006. I locate most of my research in the interplay between the analysis and interpretation of biological results arising from genome sequencing projects and the development of bioinformatics tools to facilitate the interpretation and creation of such results. The Setubal research group works primarily on bioinformatics for bacterial genome annotation and sequence analysis. Currently, there are eight ongoing projects, each of which is briefly described.

Keywords: *Agrobacterium*; *Azotobacter vinelandii*; *Pseudomonas syringae*; *Xanthomonas*; genome projects; genome annotation; sequence analysis.

Bioinformatics for genome sequencing projects

Agrobacterium *biovar-type strains.* *Agrobacterium* is a well known genus of bacterial plant pathogens. The genus is divided into three biovars that are based on differences in biochemical and physiological characteristics. The genome of *Agrobacterium tumefaciens* C58, a representative of biovar 1, was published by Wood et al (2001). In this project, the goal is to sequence and analyze the genomes of a biovar 2 representative (*Agrobacterium radiobacter* K84) and a biovar 3 representative (*Agrobacterium vitis* S4). At the time that this report was written, the sequences have been completed, annotation is almost done, and general genome analyses are underway. Of particular interest is the relationship between the genomes of all three biovars and those of *Rhizobium etli* (Gonzalez et al, 2006) and *Rhizobium leguminosarum* (Young et al, 2006), which have recently been published. It is known that these genera are closely related (there is some contention that the two genera should be merged); however, *A. tumefaciens* and *A. vitis* are plant pathogens, whereas *R. etli* and *R. leguminosarum* are nitrogen-fixing plant symbionts (*A. radiobacter* is a biocontrol agent for *A. tumefaciens* C58). Therefore, the careful comparison of these five genomes should yield interesting insights into the genomic basis and evolution of bacterial plant pathogenicity and symbiosis. In this respect, a research lead that I am currently following deals with the development of computational tools to reconstruct the detailed evolutionary history of whole bacterial replicons, including events such as genome rearrangements and lateral gene transfer.

This project is being undertaken by a consortium that, besides the Virginia Bioinformatics Institute (VBI), includes the University of Washington (E. Nester), Seattle Pacific University (D. Wood), Arizona State University (S. Slater), the Monsanto Corporation (B. Goldman), Hiram College (B. Goodner), the University of Illinois at Urbana-Champaign (S. Farrand), and Cornell University (T. Burr). The contribution of the Setubal group to this project has been to provide the bioinformatics infrastructure for genome annotation and analysis (through the Genome Annotation Tool or GAT, described later in this report), and the genomic annotation and analyses themselves.

Azotobacter vinelandii. *Azotobacter vinelandii* is a free-living nitrogen-fixing bacterium, which is found in soils world-wide, and which possesses features of nitrogen and energy metabolism relevant to agriculture. This organism has been studied for about 100

Contributors:
Zheng Cai, Anjan Purkayastha, Joshua Shallom, Jian Sun, Andrew Warren

years by numerous scientists throughout the world. Before Joshua Lederberg's discovery of sexuality in *Escherichia coli*, *A. vinelandii* was the experimental organism of choice for many investigators during the emergence of biochemistry as a dominant discipline within the life sciences.

In this project, the goal is to sequence and annotate the genome of *A. vinelandii*. It is undertaken by a consortium similar to the one for *Agrobacterium*, as described previously in this report: VBI, Seattle Pacific University (D. Wood), the Monsanto Corporation (B. Goldman), Hiram College (B. Goodner), and Virginia Tech (D. Dean). At the time that this report was written, the sequence had been completed (5.3 Mbp, no plasmids). Annotation and analyses started in June 2006. The contribution of the Setubal group is similar to that described for the *Agrobacterium* project.

Xanthomonas axonopodis *pv.* aurantifolii. *Xanthomonas axonopodis* pv. *aurantifolii* is a gamma-proteobacterium that associates with citrus. Two South American strains are well known and designated by the letters B and C. Strain B causes an attenuated version of citrus canker, whereas strain C causes a hypersensitive response on certain varieties of citrus. Together with *Xanthomonas axonopodis* pv. *citri*, which is the causative agent of citrus canker (also known as cancrosis A), they form a good system for the study of citrus canker. This project aims to sequence and annotate the genomes of strains B and C, and compare them with the sequence and annotated genome of strain A, which was sequenced and published in 2002 (da Silva et al, 2002).

This multi-center project is being undertaken in Brazil and is led by a group from Universidade Estadual Paulista (Unesp) at the Campus de Jaboticabal. A draft sequence of both genomes is already completed, and annotation and analysis are underway. The data for this project are not housed at VBI, and therefore the GAT system is not being used. Our contribution is to perform various bioinformatics tasks and assist in the genome analysis.

Bioinformatics tools

Two main genome project tools are the focus of our group. One is the Genome Annotation Tool, or GAT, which is a system that includes a database, a web-based interface, and a pipeline for automated annotation. GAT was first designed and developed by myself and J. Jhaveri, and it is now being maintained and further developed by J. Sun. Currently, GAT houses the genomes of *A. tumefaciens* C58, *A. vitis* S4, and *A. radiobacter* K84. Inclusion of *Azotobacter vinelandii* will happen shortly.

The second tool is the Genome Reverse Compiler, which is an annotation tool for prokaryotic genomes. Its name and philosophy are based on analogy to a high-level programming language compiler. In this analogy, the genome is a "program" in a certain "low level" language that humans cannot understand. The goal of GRC is, given the sequence of any prokaryotic genome, to output its human-readable corresponding "high-level program" (its annotation). It is a reverse compiler because the direction from low-level to high-level is the opposite of the one in traditional computer language compilers. The name compiler is also used because our goal is to make GRC work in a completely automatic manner and make it available as a stand-alone tool, assuming standard input formats and providing equally standard outputs (such as a GenBank file). Automated annotation pipelines are common throughout laboratories that work on genomes. However, to the best of our knowledge an open source, stand-alone, easy-to-run, lab-independent annotation tool such as GRC does not exist. Another design goal for GRC is efficiency. GRC should be able to process average prokaryotic genomes in a matter of a few hours. GRC is in its early stages of development. In the first version, its features are: it only annotates protein-coding genes; it requires as separate inputs annotated genomes from organisms closely related to the input genome; it relies completely and shamelessly on sequence similarity to make gene product assignments; moreover, similarity computations make use of the National Center for Biotechnology Information (NCBI) BLAST program (therefore, this version is not strictly speaking stand-alone). GRC is the thesis project

of my PhD student Andrew Warren (Department of Computer Science, Virginia Tech).

Bacterial evolution of *Pseudomonas syringae*

Bacteria can evolve quickly relative to the time scale of evolution since they are able to acquire entire blocks of new DNA from other organisms or the environment. These blocks of DNA are called genomic islands. We have compared the three sequenced genomes of the bacterial plant pathogen *Pseudomonas syringae* to identify genomic islands. At a first approximation, all DNA regions that are not shared between all three strains are putative genomic islands that have been acquired during the microevolution of the three sequenced strains to adapt to different hosts. We have aligned the genomes using MAUVE (Darling et al, 2004) to determine these strain-specific regions. Subsequently, we have processed the results with our own scripts or manually. We are currently analyzing the identified putative islands regarding their content in mobile elements, flanking repeat or tRNA elements, GC content, codon usage and dinucleotide frequency. We also intend to leverage these results to analyze unsequenced field isolates of *P. syringae*. From a bioinformatics standpoint, this research should yield new tools for detailed whole genome analysis. This is joint work with B. Vinatzer (Department of Plant Pathology, and Weed Science, Virginia Tech), Liqing Zhang (Department of Computer Science, Virginia Tech) and Zheng Cai (Genetics, Bioinformatics and Computational Biology PhD program student).

Gene ontology terms for standardized annotation of plant-associated microbe genomes (PAMGO)

In this project, the goal is to expand the Gene Ontology (Ashburner et al, 2000) to include terms that can be used to annotate genes involved in host-microbe interactions; furthermore, the goal is to apply the terms developed for certain microbial genomes and to make them available to the community to facilitate further dissemination of the developed terms. It is a consortium project that includes VBI (B. Tyler and J. Setubal), Cornell University (A. Collmer), Wells College (C. Collmer), the University of Wisconsin-Madison (N. Perna), North Carolina State University (R. Dean and D. Bird), and The Institute for Genomic Research (O. White and M. Gwinn-Giglio). The Setubal group contributes with term development and a re-annotation of the genes involved in host-microbe interactions in the *A. tumefaciens* C58 genome, as well as their orthologs in *A. vitis* S4 and *A. radiobacter* K84. The bioinformatics research component of this project is the exploration of the limits of automatic transfer of GO terms among related organisms, a crucial component for any high-throughput annotation tool, such as the GRC described earlier in this article. Currently, J. Sun is adding capabilities for GO curation to GAT and J. Shallom is helping with term development and C58 reannotation. The current status of this project is available at http://pamgo.vbi.vt.edu.

PathoSystems Resource Integration Center

The PathoSystems Resource Integration Center, PATRIC (http://patric.vbi.vt.edu), was established at VBI to create a public, web-based, comprehensive bioinformatics resource for select priority pathogens (the bacteria *Brucella*, *Coxiella*, and *Rickettsia* and the following viruses: caliciviruses, coronaviruses, hepatitis A, hepatitis E, and rabies). The ultimate goal is to help speed up the development of vaccines, diagnostics, and therapeutics for the diseases caused by these organisms. The data types covered by the resource include genome sequence, comparative genomics, polymorphisms, gene expression, proteomics, pathways and host-pathogen interaction data.

During 2005-2006, a new automated genome annotation pipeline, a new web-based curation interface, and a user-friendly web portal were developed by the PATRIC team. PATRIC already provides previously available annotations on all its pathogens (approximately 350 genomes). At the time of writing this report, we have also updated the DNA annotations of all viral genomes (except coronaviruses), *Brucella melitensis* 16M, *Rickettsia prowazekii*, and *Coxiella burnetii* genomes.

Work on PATRIC is made possible by a contract awarded to Bruno Sobral from the

National Institute of Allergy and Infectious Disease, part of the National Institutes of Health. I am co-Principal Investigator and I supervise the work of the curation group led by A. Purkayastha. A description of preliminary results concerning *Rickettsia* appeared in Hance et al (2005).

Editorial activities

During the period covered by this report S. Verjovski-Almeida and I chaired the *Brazilian Symposium on Bioinformatics 2005*, held in São Leopoldo, Brazil, July 27-29, 2005, as part of the annual meeting of the Brazilian Computer Society. The proceedings of BSB 2005 have been published by Springer Verlag (Setubal & Verjovski-Almeida, 2005). Out of 45 full papers submitted to BSB 2005, 15 were selected by an international program committee of 40 members to appear in the proceedings and for oral presentations. An additional 16 extended abstracts were also selected and presented.

The best full papers of BSB 2005 were invited for expanded and revised submissions to the journal *Computers in Biology and Medicine* (Elsevier). A special issue of CB&M has recently been published with the resulting five papers (Setubal & Verjovski-Almeida, 2006).

Additional student projects

My PhD student L. Digiampietri, from the University of Campinas (Brazil), is doing work on webservices-based workflows for bioinformatics (co-advised by C. Medeiros). Preliminary results were published in Digiampietri et al (2005). D. Lorenzini, from the University of São Paulo, did his PhD work on a small Expressed Sequence Tag (EST) project for immune-related genes in the tarantula spider *Acanthoscurria gomesiana*, and results were published in Lorenzini et al (2006). His advisor was S. Daffre. My contribution was to help set up and fine-tune the bioinformatics pipeline created to collect and analyze the EST data. T. Venâncio, a bioinformatics PhD student from the University of São Paulo, co-advised by S. Verjovski-Almeida, is analyzing evolutionary interesting and host-related genes in the parasite *Schistosoma mansoni*, based on data contributed by the work of Verjovski-Almeida et al (2003).

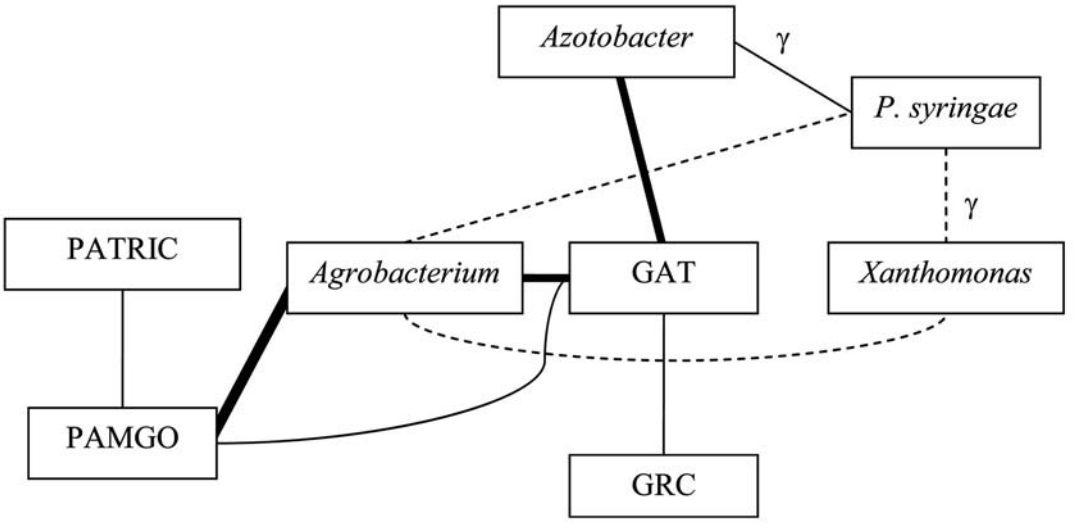


Fig. 1. A summary of the main projects of the Setubal research group. The connections between the projects are shown as follows: thick edges mean an explicit link between projects; dashed edges indicate that the organisms linked are bacterial plant pathogens; other edges represent conceptual relationships. GAT, Genome Annotation Tool; GRC, Genome Reverse Compiler; PATRIC, PathoSystems Resource Integration Center project; PAMGO, Plant-Associated Microbe Gene Ontology project; γ, γ-proteobacteria.

Other publications

During the reporting period Setubal, Moreira & da Silva (2005) published a review of the current status of genomics of bacterial plant pathogens. Setubal, Reis, Matsunaga, and Haake (2006) published their work on automated detection of lipoproteins in spirochetal genomes, which is briefly described in the 2005 VBI scientific annual report (available upon request).

References

Ashburner M et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25**: 25-29.

da Silva ACR et al (2002) Comparison of the genomes of two *Xanthomonas* pathogens with differing host specificities. *Nature* **417**(6887): 459-463.

Darling AC, Mau B, Blattner FR, Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14**(7): 1394-1403.

Digiampietri LA, Medeiros CB, Setubal JC (2005) A framework based on Web service orchestration for bioinformatics workflow management. *Genet. Mol. Res.* **4**(3): 535-542.

Gonzalez V et al (2006) The partitioned *Rhizobium etli* genome: genetic and metabolic redundancy in seven interacting replicons. *Proc. Natl Acad. Sci. USA* **103**(10): 3834-3839.

Hance ME, Czar MJ, Azad A, Purkayastha A, Snyder EE, Crasta OR, Setubal JC, Sobral BW (2005) The pathogen resource integration center: implications for rickettsial research. *Ann. N.Y. Acad. Sci.* **1063**: 459-465.

Lorenzini D, da Silva PI Jr, Soares MB, Arruda P, Setubal J, Daffre S (2006) Discovery of immune-related genes expressed on hemocytes of the tarantula spider *Acanthoscurria gomesiana*. *Dev. Comp. Immunol.* **30**(6):545-556.

Setubal JC, Verjovski-Almeida S (2005) *Advances in Bioinformatics and Computational Biology. Proceedings of the Brazilian Symposium on Bioinformatics, BSB 2005.* Lecture Notes in Bioinformatics. vol. 3594. Berlin: Springer-Verlag.

Setubal JC, Moreira LM, da Silva ACR (2005) Bacterial phytopathogens and genome science. *Curr. Opin. Microbiol.* **8**, 595-600.

Setubal JC, Verjovski-Almeida S (2006) Brazilian Symposium on Bioinformatics. *Comp. Biol. Med.* advanced online publication 17 April 2006; doi: 10.1016/j.compbiomed.2006.03.00 1.

Setubal JC, Reis M, Matsunaga J, Haake DA (2006) Lipoprotein computational prediction in spirochaetal genomes. *Microbiology* **152**:113-121.

Verjovski-Almeida S et al (2003) Transcriptome analysis of the acoelomate human parasite *Schistosoma mansoni. Nature Genet.* **35**:148-157.

Wood DW et al (2001) The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. *Science* **294**: 2317-2323.

Young JP et al (2006) The genome of *Rhizobium leguminosarum* has recognizable core and accessory components. *Genome Biol.* **7**(4): R34.

Publications

Digiampietri LA, Medeiros CB, Setubal JC (2005) A framework based on Web service orchestration for bioinformatics workflow management. *Genet. Mol. Res.* **4**(3): 535-542.

Hance ME, Czar MJ, Azad A, Purkayastha A, Snyder EE, Crasta OR, Setubal JC, Sobral BW (2005) The pathogen resource integration center: implications for rickettsial research. *Ann. N.Y. Acad. Sci.* **1063**: 459-465.

Lorenzini D, da Silva PI Jr, Soares MB, Arruda P, Setubal J, Daffre S (2006) Discovery of immune-related genes expressed on hemocytes of the tarantula spider *Acanthoscurria gomesiana*. *Dev. Comp. Immunol.* **30**(6):545-556.

Setubal JC, Verjovski-Almeida S (2005) *Advances in Bioinformatics and Computational Biology. Proceedings of the Brazilian Symposium on Bioinformatics, BSB 2005.* Lecture Notes in Bioinformatics. vol. 3594. Berlin: Springer-Verlag.

Setubal JC, Moreira LM, da Silva ACR (2005) Bacterial phytopathogens and genome science. *Curr. Opin. Microbiol*. **8**, 595-600.

Setubal JC, Verjovski-Almeida S (2006) Brazilian Symposium on Bioinformatics. *Comp. Biol. Med.* advanced online publication 17 April 2006; doi: 10.1016/j.compbiomed.2006.03.00 1.

Setubal JC, Reis M, Matsunaga J, Haake DA (2006) Lipoprotein computational prediction in spirochaetal genomes. *Microbiology* **152**:113-121.

Applications of metabolomics to functional genomics

Vladimir Shulaev, vshulaev@vbi.vt.edu
Associate Professor, Virginia Bioinformatics Institute & Horticulture, Virginia Tech

Metabolomics, the global analysis of all cellular metabolites, is rapidly moving to the forefront of genomics research, complementing the well-established areas of transcriptomics and proteomics (Shulaev, 2006). Originally proposed as a functional genomics technique (Oliver et al, 1998), metabolomics is now applied far beyond genomics studies. It is used to study mutant phenotypes (Raamsdonk et al, 2001), evaluate response to environmental stress (Rizhsky et al, 2004; Viant et al, 2003), study growth and development (Martins et al, 2004; Tweeddale et al, 1998), in drug discovery (Watkins & German, 2002), and human disease (Griffiths & Stubbs, 2003; Watkins et al, 2002) and nutrition research (German et al, 2002). Recently, metabolomics has been applied as a tool in systems biology (Mendes, 2001). There are three major approaches used in metabolomics studies: (a) targeted analysis, (b) metabolite profiling and (c) metabolic fingerprinting (Shulaev, 2006). Our research is focused on developing metabolomics technology and the application of high throughput metabolite profiling to study stress response in microorganisms, plants and animals.

Keywords: metabolomics; metabolite profiling; mass spectrometry gas chromatography-mass spectrometry; liquid chromatography-mass spectrometry; capillary electrophoresis-mass spectrometry; cancer; yeast; Arabidopsis.

Metabolomics and yeast systems biology

Oxidative stress has been associated with human diseases including cancer, autoimmune disease, and ageing. Lately, oxidative stress response in many organisms has been studied on a genomic scale using high throughput profiling methodologies, including transcriptomics, proteomics and metabolomics. In collaboration with the Laubenbacher and Mendes groups at the Virginia Bioinformatics Institute (VBI), we use a systems biology approach to study oxidative stress response in *Saccharomyces cerevisiae* (this project is funded by National Institutes of Health (NIH) Grant R01 GM068947-01 "*A new mathematical modeling approach to biochemical networks, with an application to oxidative stress in yeast*" to Reinhard Laubenbacher). This project is aimed at developing novel modeling methods for the reverse-engineering of biochemical networks. These methods will be applied to data from experiments specifically designed for modeling, focusing on the regulatory network in *S. cerevisiae* involved in oxidative stress response. Our group is generating genomics, proteomics, and metabolomics data sets for this project.

We have studied the effect of oxidative stress in *S. cerevisiae* caused by the exposure to cumene hydroperoxide (CHP) using a time course experiment. *S. cerevisiae* strain BY4743 was exposed to 0.16 mM CHP and samples were taken at different time points following peroxide treatment. In order to capture both fast and slow responses to CHP, samples were collected on a logarithmic time scale. Yeast cultures were grown in a controlled batch condition using a bioreactor control system (BioFlo 110 from New Brunswick Scientific). Three replicate cultures were exposed to CHP at the mid-log phase, while the other three cultures were not treated with the peroxide and were treated as controls. At different time intervals following CHP exposure, samples were collected from the

Contributors:
Diego Cortes, Hope Gruszewski, Holly Johnson, Teruko Oosumi, Wei Sha, Joel Shuman, Leepika Tuli

fermentors by spraying the cultures into cold, buffered methanol (-40°C) to rapidly quench the metabolism and preserve RNA, protein and metabolite species for further quantitative analysis. Cell pellets were separated from the culture medium, washed and frozen in liquid nitrogen. Samples were split into three parts for transcriptomics, proteomics and metabolomics analysis.

For gene expression profiling, RNA was extracted with the hot phenol method and profiling was performed using the Affymetrix GeneChip® system with Yeast Genome S98 arrays (Affymetrix, Santa Clara, CA). All gene expression profiling experiments were performed in the VBI Core Laboratory Facility (CLF). For data analysis, we used Robust Multichip Average (RMA) for microarray data summarization and normalization, and a two-way ANOVA gene-by-gene model to assess the significance of differences between transcripts across two different experimental conditions. Genes with similar expression pattern were discovered by *k*-means clustering using the TIGR Multiexperiment Viewer (MeV). High-throughput GoMiner was used to determine gene ontology categories that were significantly affected by CHP and heat maps were built to visualize the responses of pathways that were likewise significantly affected by it. Several pathways, including response to stress, carbohydrate metabolism, protein catabolism and signal transduction, were significantly up-regulated in strain BY4743 following CHP exposure.

We have established methods for non-targeted analysis using gas chromatography-mass spectrometry (GC-MS) and liquid chromatography-mass spectrometry (LC-MS). Non-targeted GC-MS profiling of yeast cultures is based on methods described by Roessner et al. (Roessner et al, 2000). We previously used this technique for metabolite profiling of two distinct growth phases of *S. cerevisiae* cultures (Martins et al, 2004). We have further refined this methodology by automating most of the metabolite extraction and derivatization steps. Using robotics in sample preparation significantly improved the robustness and reproducibility of the GC-MS assay. We have further optimized a new LC-MS separation protocol for yeast metabolite profiling using capillary LC with a monolithic capillary column. To achieve higher representation of the yeast metabolome we compared seven different extraction protocols. Metabolites were extracted from the aliquots of the same sample using various protocols, and separated on a monolithic column using a 90-minute acetonitrile gradient prior to MS detection. The efficiency of the extracted method was evaluated based on the number and chromatographic quality of the resolved peaks. We will use this approach for non-targeted LC-MS profiling for the non-volatile metabolites.

Currently, a combination of techniques is used in proteomics research: two-dimensional gel electrophoresis, image analysis, mass spectrometry and bioinformatics tools. The outcome of the proteomics experiment depends on 2 two key factors: (1) proper separation of complex protein mixtures (commonly by 2D gel electrophoresis), and (2) positive identification of proteins by mass spectrometry. We performed a comprehensive study of the effect of several protein extraction, separation and visualization techniques on the total amount and complexity of the proteins extracted, identified and quantitated from yeast cells. After comparing results for two extraction protocols (RIPA buffer and Yeast Protein Extraction Kit, Calbiochem), four staining methods (Coomassie, Silver, Flamingo, and Sypro Ruby stains), and two protein separation techniques (2D-PAGE and PF2D) in conjunction with protein identification by LC-MS, we have found the best protocol to study changes in protein expression under oxidative stress induced by CHP.

Transcriptomics, proteomics, and metabolomics experimental data generated by this project will be used by the Mendes and Laubenbacher groups for the reverse-engineering of biochemical networks, using modeling methods based on continuous and discrete mathematics.

Woodland strawberry (*Fragaria vesca*) as a model for fruit functional genomics

Fruit is an essential part of human diet, since it contains a large number of beneficial phytochemicals like antioxidants and anticancer

agents (Eichholzer et al, 2001; Macheix et al, 1991; Schieber et al, 2001; Selmar 1999; Swanson, 1998). We are working on identifying the genes involved in the biosynthesis and regulation of these phytochemicals and on identification of novel compounds with potential health benefits.

Many fruit species, including almonds, apples, plums, peaches, pears, raspberries, sour cherries, sweet cherries and strawberries belong to the *Rosaceae* family. Recently an effort has been made to develop a model for genomics research in *Rosaceae*. The woodland strawberry (*Fragaria vesca*) provides an attractive model species due to its small genome size, short reproductive cycle, and small plant size. It is also easy to propagate and manipulate in the greenhouse and produces a large number of seeds per plant. To facilitate genomics research in strawberries, we have developed a high throughput platform for reverse and forward genetics in *F. vesca* based on T-DNA insertional mutagenesis, gene identification, and phenotype profiling. The highly efficient transformation protocol developed by our group (Oosumi et al, 2006) makes it feasible for systematic production of a large collection of T-DNA insertional mutant lines that can be used for both forward and reverse genetic studies.

For reverse genetics research, we established a high-throughput method for amplification of T-DNA flanking genomic sequences in *F. vesca* by optimizing the thermal asymmetric interlaced (TAIL) PCR method. We obtained a success rate of over 90% with transgenic plants tested using three arbitrary degenerate (AD) primers (AD1, AD2, and AD3) previously designed for *Arabidopsis thaliana*. Secondary TAIL-PCR products were sequenced directly using a T-DNA-specific primer. A large number of the T-DNA tagged sequences analyzed had significant similarities to genes or proteins of other plants.

To verify our T-DNA tagging system as a forward genetics tool, we screened a subset of T-DNA insertional mutant lines for obvious morphological mutants. Four lines with altered leaf shapes were found (Oosumi et al, 2006). One of those lines was used for genetic and molecular characterization of the putative mutant. Segregation analysis based on Green Fluorescent Protein (GFP) expression, phenotype, and PCR analysis with the primers specific to T-DNA flanking genomic DNA indicated that plants with the putative mutant phenotype were homozygous T-DNA insertional mutants. We amplified cDNA of this putative T-DNA tagged gene by rapid amplification of cDNA ends (RACE) and determined the entire cDNA sequence of a new strawberry gene that has high similarity to unknown mRNAs of *Arabidopsis* and rice.

Metabolomics and cancer

Breast cancer is a common and all too often fatal human disease. To date, the etiology of human breast cancer remains obscure. In collaboration with the Wake Forest Comprehensive Cancer Center and several research groups at VBI, we aim to build a comprehensive systems biology understanding of cancer derived from mathematical modeling of complex biochemical and regulatory networks. This novel systems biology approach to cancer research will lead to advances in cancer diagnostics and treatment.

Although a significant role in cancer initiation and progression is attributed to changes in RNA and protein expression levels and regulation, changes in small molecules can provide important mechanistic clues on cancer development. There is a strong body of evidence supporting an important role for metabolic regulation in cancer, including breast cancer. Malignant cells undergo significant changes in metabolism including a re-distribution of metabolic networks (Boros et al, 2003). These metabolic changes result in different metabolic landscapes in cancer cells versus normal cells. Metabolomics as a global approach is especially useful in identifying overall metabolic changes associated with a particular biological process and identifying the metabolic networks most affected. Moreover, metabolomics provides an additional layer of information that can be linked with transcriptomics and proteomics data to obtain a comprehensive view of the biological system.

Most cancer metabolomics studies to date have been done using metabolic fingerprinting or profiling with NMR spectroscopy of tissue extracts or *in vivo* magnetic resonance

spectroscopy. Using spectroscopic techniques, it is possible to differentiate several tumor types in humans and animal models (Devos et al, 2004; Lukas et al, 2004; Tate et al, 1998; Tate et al, 2000). But while techniques based on magnetic resonance have the advantage of being non-invasive, they have low sensitivity and cannot detect molecules at low concentrations. Mass spectrometry methods are much more sensitive and are indeed the most appropriate for *in vitro* studies.

Our group is studying the progression of malignancy in human breast epithelial cells using a combination of mass spectrometry-based metabolomics and transcriptomics approaches, with the aim of identifying robust molecular signatures that can distinguish early stages of malignant transformation. These molecular signatures can potentially provide biomarkers for early detection of disease and targets for anticancer drugs with novel modes of action.

Application of metabolomics to study gene functions in Arabidopsis

Complete sequencing of several plant species has identified thousands of genes. However, the biological function of about half those genes remains unidentified. Metabolomics is now often used to study mutant biochemical phenotypes and identify the functions of unknown genes in plants.

Our group is part of the Plant Metabolomics Consortium, which focuses on developing metabolomics technologies for plant research. We work in close collaborations with the leading metabolomics laboratories in the United States, led by Dr. Basil J. Nikolau (Iowa State University), Dr. Oliver Fiehn (University of California, Davis), Dr. Lloyd Sumner (Noble Foundation, Oklahoma), Dr. Mark Lange (Washington State University), Dr. Ruth Welti (Kansas State University), Dr. Sue Rhee (The Arabidopsis Information Resource; TAIR), Dr. Julie Dickerson (Iowa State University), Dr. Eve Syrkin Wurtele (Iowa State University), Dr. Philip Dixon (Iowa State University), Dr. George Kraus (Iowa State University), and Dr. Nikki Pohl (Iowa State University). Last year, the Consortium was funded by the National Science Foundation (NSF) to develop a new tool to decipher the functions of plant genes using metabolomics (NSF 2010 grant "Metabolomics as a functional genomics tool for identifying functions of *Arabidopsis* genes in the context of metabolic and regulatory networks" to Dr. Basil Nikolau). Using complementary analytical platforms, all labs together can profile about 2000 "metabolite peaks", of which 900 are chemically defined metabolites.

In another collaborative project also funded by the NSF Arabidopsis 2010 Program, we, together with Drs. Eran Pichersky and Joseph Noel, use the combination of molecular genetics, enzymology, protein structure analysis, gene expression profiling and metabolite profiling to identify the function of all the methyltransferases belonging to the SABATH family. The *Arabidopsis thaliana* genome contains 24 related genes encoding enzymes that belong to the SABATH family of methyltransferases (MTs). Several enzymes of the SABATH family have been recently characterized as jasmonic acid methyltransferase (*AtJAMT1*), indole-3-acetic acid methyltransferase (*AtIAMT1*), benzoic acid methyltransferase (*AtBSMT*) and farnesoic acid methyl transferase (*AtFAMT*). We performed comprehensive metabolite profiling of a series of knock-out and overexpressor lines for several SABATH methyltransferases in order to identify their substrates.

Acknowledgements

Our work was supported by NIH Grant R01 GM068947-01 "A new mathematical modeling approach to biochemical networks, with an application to oxidative stress in yeast", NSF Arabidopsis 2010 Grant # 0312857 "Collaborative research on the functional of the SABATH family of methyltransferases", NIH Grant 2RO1AI045774-06A2 "*Plasmodium falciparum* metal metabolism", NSF Arabidopsis 2010 grant #0520140 "Metabolomics: A functional genomics tool for deciphering functions of *Arabidopsis* genes in the context of metabolic and regulatory networks", Virginia Tech ASPIRES grant "Strawberry functional genomics", Beckman Coulter Foundation, Virginia Bioinformatics Institute, and Virginia Tech Horticulture Department. We thank our collaborators Frank Van Breusegem, Eduardo Blumwald, Scott Campbell, Allan Dickerman,

Oluwatosin Ginsanrin, Scott Harrison, Reinhard Laubenbacher, Kim Lewers, Chunhong Mao, Pedro Mendes, Ron Mittler, Craig Nessler, Jerzy Nowak, Joseph Noel, David Oliver, Eran Pichersly, Dominique Rasoloson, Janet Slovin, David Sullivan, Richard Veilleux, Phillip Wadl, and Brenda Winkel. We thank Jim Walke for administrative support and editing. We also thank Susan Martino-Catt, Clive Evans and all of the staff in the Core Laboratory Facility at the Virginia Bioinformatics Institute.

References

Boros LG, Brackett DJ, Harrigan GG (2003) Metabolic biomarker and kinase drug target discovery in cancer using stable isotope-based dynamic metabolic profiling (SIDMAP). *Curr. Cancer Drug Targets* **3**: 445-453.

Devos A, Lukas L, Suykens JA, Vanhamme L, Tate AR, Howe FA, Majos C, Moreno-Torres A, van der Graaf M, Arus C, Van Huffel S (2004) Classification of brain tumours using short echo time 1H MR spectra. *J. Magn. Reson.* **170**: 164-175.

Eichholzer M, Luthy J , Gutzwiller F, Stahelin HB (2001) The role of folate, antioxidant vitamins and other constituents in fruit and vegetables in the prevention of cardiovascular disease: The epidemiological evidence. *Int. J. Vitam. Nutr. Res.* **71**: 5-17.

German JB, Roberts MA, Fay L, Watkins SM (2002) Metabolomics and individual metabolic assessment: the next great challenge for nutrition. *J.Nutr.* **132**: 2486-2487.

Griffiths JR, Stubbs M (2003) Opportunities for studying cancer by metabolomics: preliminary observations on tumors deficient in hypoxia-inducible factor 1. *Adv. Enzyme Regul*. **43**: 67-76.

Lukas L, Devos A, Suykens JA, Vanhamme L, Howe FA, Majos C, Moreno-Torres A, van der Graaf M, Tate AR, Arus C, Van Huffel S (2004) Brain tumor classification based on long echo proton MRS signals. *Artif. Intell. Med*. **31**: 73-89.

Macheix JJ, Sapis JC, Fleuriet A (1991) Phenolic compounds and polyphenoloxidase in relation to browning in grapes and wines. *Crit. Rev. Food Sci. Nutr*. **30**: 441-486.

Martins AM, Camacho D, Shuman J, Sha W, Mendes P, Shulaev V (2004) A systems biology study of two distinct growth phases of *Saccharomyces cerevisiae* cultures. *Current Genomics* **5**: 649-663.

Mendes P (2001) Modeling large scale biological systems from functional genomic data: parameter estimation. In: *Foundations of Systems Biology*, Kitano H (ed) pp 163-186. Cambridge, MA: MIT Press.

Oliver SG, Winson MK, Kell DB, Baganz F (1998) Systematic functional analysis of the yeast genome. *Trends Biotechnol.* **16**: 373-378.

Oosumi T, Gruszewski HA, Blischak LA, Baxter AJ, Wadl PA, Shuman JL, Veilleux RE, Shulaev V (2006) High-efficiency transformation of the diploid strawberry (*Fragaria vesca*) for functional genomics. *Planta* **223**: 1219-1230.

Raamsdonk LM, Teusink B, Broadhurst D, Zhang N, Hayes A, Walsh MC, Berden JA, Brindle KM, Kell DB, Rowland JJ, Westerhoff HV, van Dam K, Oliver SG (2001) A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat. Biotechnol.* **19**: 45-50.

Rizhsky L, Liang H, Shuman J, Shulaev V, Davletova S, Mittler R (2004) When defense pathways collide. The response of *Arabidopsis* to a combination of drought and heat stress. *Plant Physiol.* **134**: 1683-1696.

Roessner U, Wagner C, Kopka J, Trethewey RN, Willmitzer L (2000) Technical advance: simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *Plant J.* **23**: 131-142.

Schieber A, Stintzing FC, Carle R (2001) By-products of plant food processing as a source of functional compounds - recent developments. *Trends Food Sci. Tech.* **12**: 401-413.

Selmar D (1999) Cyanide in foods: Biology of cyanogenic glucosides and related nutritional problems. In: *Phytochemicals in human health protection, nutrition, and plant defense*, Romeo JT ed, New York: Kluwer Academic, pp. 369-392.

Shulaev V (2006) Metabolomics technology and bioinformatics. *Brief Bioinform*. advance online publication 18 May 2006; doi:10.1093/bib/bbl012

Swanson CA (1998) Vegetables, fruits, and cancer risk: the role of phytochemicals. In: *Phytochemicals. A new Paradigm*, Bidlack WR, Omaye ST, Meskin MS, Jahner D, eds, Lancaster: Technomic Publishing Co., pp. 1-12.

Tate AR, Foxall PJ, Holmes E, Moka D, Spraul M, Nicholson JK, Lindon JC (2000) Distinction between normal and renal cell carcinoma kidney cortical biopsy samples using pattern recognition of (1)H magic angle spinning (MAS) NMR spectra. *NMR Biomed*. **13**: 64-71.

Tate AR, Griffiths JR, Martinez-Perez I, Moreno A, Barba I, Cabanas ME, Watson D, Alonso J, Bartumeus F, Isamat F, Ferrer I, Vila F, Ferrer E, Capdevila A, Arus C (1998) Towards a method for automated classification of 1H MRS spectra from brain tumours. *NMR Biomed*. **11**: 177-191.

Tweeddale H, Notley-McRobb L, Ferenci T (1998) Effect of slow growth on metabolism of *Escherichia coli*, as revealed by global metabolite pool ("metabolome") analysis. *J. Bacteriol*. **180**: 5109-5116.

Viant MR, Rosenblum ES, Tieerdema RS (2003) NMR-based metabolomics: a powerful approach for characterizing the effects of environmental stressors on organism health. *Environ. Sci.Technol*. **37**: 4982-4989.

Watkins SM, German JB (2002) Metabolomics and biochemical profiling in drug discovery and development. *Curr. Opin. Mol. Ther*. **4**: 224-228.

Watkins SM, Reifsnyder PR, Pan HJ, German JB, Leiter EH (2002) Lipid metabolome-wide effects of the PPARγ agonist rosiglitazone. *J. Lipid Res.* **43**: 1809-1817.

Publications

Gadjev I, Vanderauwera S, Gechev TS, Laloi C, Minkov IN, Shulaev V, Apel K, Inze D, Mittler R, Van Breusegem F (2006) Transcriptomic footprints disclose specificity of reactive oxygen species signaling in *Arabidopsis*. *Plant Physiol*. **141**: 436-445.

Shimada T, Nakano R, Shulaev V, Sadka A, Blumwald E (2006) Vacuolar citrate/H(+) symporter of citrus juice cells. *Planta* advanced online publication 27 January 2006; doi:10.1007/s00425-006-0223-2

Shulaev V (2006) Metabolomics technology and bioinformatics. *Brief Bioinform* advance online publication 18 May 2006; doi:10.1093/bib/bbl012

Suzuki N, Rizhsky L, Liang H, Shuman J, Shulaev V, Mittler R (2005) Enhanced tolerance to environmental stress in transgenic plants expressing the transcriptional coactivator multiprotein bridging factor 1c. *Plant Physiol.* **139**: 1313-1322.

Yang Y, Varbanova M, Ross J, Wang G, Cortes D, Fridman F, Shulaev V, Noel JP, Pichersky E (2006) Methylation and demethylation of plant signal molecules. In: *Integrative plant biochemistry - Recent Advances in Phytochemistry*, Romeo J (ed), vol. 40, Elsevier.

Oosumi T, Gruszewski HA, Blischak LA, Baxter AJ, Wadl PA, Shuman JL, Veilleux RE, Shulaev V. (2006) High-efficiency transformation of the diploid strawberry (*Fragaria vesca*) for functional genomics. *Planta* **223**: 1219-1230.

Shulaev V, Oliver DJ (2006) Metabolic and proteomic markers for oxidative stress. New tools for reactive oxygen species research. *Plant Physiol*. **141**: 367-372.

Living inside your host: lessons from the α-proteobacteria

Bruno WS Sobral, sobral@vbi.vt.edu;
Executive & Scientific Director and Professor, Virginia Bioinformatics Institute
Professor, Plant Pathology, Physiology and Weed Science, Virginia Tech

We report on the progress made in the collective efforts of our Cyberinfrastructure and PathoSystems Biology groups over the past year. These groups are engaged in a wide range of research activities that focus on using the α-proteobacteria as a starting point for increasing our knowledge of infectious diseases in different biological systems. By investigating the interactions between hosts, pathogens and their environment - the so-called "disease triangle" - we hope to bring a new level of understanding to infectious disease research in plants, humans and other animals. Our approach is a transdisciplinary undertaking that combines and permeates diverse disciplines to address some of key challenges in the biomedical, environmental and plant sciences. In addition, we are also looking at innovative ways to train and educate the next generation of scientists. A key objective remains to develop genomic, proteomic and bioinformatic tools and data that can be readily applied to the study of infectious diseases as well as to the discovery of new vaccine, drug and diagnostic targets.

Keywords: cyberinfrastructure; infectious disease; symbiosis; proteomics; genomics.

Contributors:

PathoSystems Biology Group :
Timothy Driscoll, Matt Dyer, Raymie Equi, Jocelyn Kemp, Shenghua Li, Chunhong Mao, Erica Mason, Endang Purwantini, Jing Qiu, Xiaoyan Sheng, Chunxia Wang

Cyberinfrastructure Group:
George Abramoch-kin, Susan Baker, Cory Byrd, Stephen Cammer, Oswald Crasta, Mike Czar, Chitti Dharmanolla, Allan Dickerman, Darius Dziuda, Zhangjun Fei, Herman Formadi, Joseph Gillespie, Mark Hance, Ranjan Jha, Nithiwat Kampanya, Ron Kenyon, Chaitanya Kommidi, Christine Lee, Jian Li, Jian Lu, Lucas Mackasmiel, Srinivasrao Mane, Eric Nordberg, Anjan Purkayastha, Daphne Rainey, Harsha Rajasimha, Graciela Santopietro, Mark Scott, Patricia Seeley, João Setubal, Joshua Shallom, Shamira Shallom, Bruce Sharp, Maulik Shukla, Eric Snyder, Jeetendra Soneja, Dan Sullivan, Wei Sun, Yuying Tian, Nirali Vaghela, Nishantsinh Vaghela, Nataraj Vishnubhat, Rebecca Wattam, Rebecca Will, Kelly Williams, Tian Xue, Hyunseung Yoo, Boyu Yang, GongXin Yu, Qiang Yu, Chengdong Zhang, Fengkai Zhang, Yan Zhang

Introduction

The α-proteobacteria form a monophyletic group of organisms. They represent a diverse group of gram-negative bacteria that is part of the proteobacteria, the largest known phylum within the prokaryotes (Kersters et al, 2003; Kainth & Gupta, 2005). The breadth of microbes encompassed by the α-subdivision of the proteobacteria is truly astounding. Members have broad differences in lifestyle, habitat, ecology, and functionality, and contain species that broadly impact agriculture, medicine, industry, and the environment. This group of organisms is also diverse with respect to fundamental biological properties, including their morphology (having spiral, rod, and stalked shapes), metabolism, and physiology (encompassing phototrophs, heterotrophs, and chemoautolithotrophs) (Kersters et al, 2003). Members of the α-proteobacteria include plant and animal pathogens, symbionts (fixing atmospheric dinitrogen), organisms with small or large genomes (e.g. *Bradyrhizobium*, 9 Mb; *Rickettsia*, 1 Mb), and the biotechnologically important bacterium *Agrobacterium tumefaciens* (Tsolis, 2002). Other important pathogens, symbionts, and free-living organisms also

included in this subdivision include: *Bartonella*, *Brucella*, *Sinorhizobium*, *Mesorhizobium*, and the insect endosymbiont *Wolbachia* (Kersters et al, 2003).

According to the National Center for Biotechnology Information, 45 α-proteobacterial genomes have been completed and 68 are in progress (www.ncbi.nlm.nih.gov/genomes/lproks.cgi accessed 27 May 2006). There has also been recent interest in the α-proteobacteria since some genes found in eukaryotic cells are derived directly from α-proteobacteria, especially those related to mitochondria (Gray et al, 1999; Andersson et al, 1998; Gupta 2000).

The majority of the α-proteobacteria whose genomes have recently been sequenced, or are currently being sequenced, have successfully explored life within diverse eukaryotic cells (host associated). These host cells range from mammals to plants and insects. Why and how have these bacteria been so successful at exploring an intracellular lifestyle? Such questions are also germane to more recently evolving intracellular relationships. For example, since mitochondria are likely to have originated from an ancestral α-proteobacterium (Emelyanov, 2001a,b), one could also imagine rhizobia eventually becoming integrated into plants via the N_2-fixing "symbiosome."

Many of the studied α-proteobacteria act as infectious disease agents of specific hosts. Some are listed in the National Institute of Allergy and Infectious Diseases' (NIAID) and the Centers for Disease Control and Prevention's select-agent lists (www.cdc.gov/od/sap/). Since infectious diseases result from the interaction between hosts, pathogens, and their environments, there is a need for a fundamental understanding of the biological and developmental processes that these pathogenic systems undergo. Moreover, this need can be coupled more effectively with solving public health and biodefense challenges, at least in the context of vaccines, diagnostics, and therapeutics.

The *Sinorhizobium meliloti–Medicago truncatula* symbiosis

The primary focus of the PathoSystems Biology group is to comprehensively look at the α-proteobacterial genomes from the bias of the Rhizobiales, seeking to compare the biological systems across different species, mostly within the group. Key developments in the last year have focused on establishing enabling technological platforms for such studies – both from the perspective of lab technologies and resources, such as custom gene expression arrays, proteomics and other such approaches, as well as informatics resources such as NodMutDB (nodmutdb.vbi.vt.edu), the Rhizobiales Bioinformatics Resource Center (RhizoBRC, rhizobia.vbi.vt.edu), and data analysis pipelines.

***Proteomics and polyhydroxybutyrate synthesis in* S. meliloti.** Proteome maps were generated by the ProteomeLab PF 2D system, which separated proteins according to pI in a first-dimension chromatofocusing step and according to hydrophobicity in a reversed-phase step in the second dimension. We have identified two major proteins associated with polyhydroxybutyrate (PHB) granules, PhaP1 encoded by *SMc00777* (*phaP1* gene), and PhaP2 encoded by *SMc02111* (*phaP2* gene), in *S. meliloti* Rm1021 (Wang C, unpublished results). These two bands were selected for Matrix-Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry (MALDI-TOF MS) analysis. We have shown that these two genes are expressed, and that the proteins are synthesized when the cells accumulate PHB. To understand the functions of *phaP1* and *phaP2*, several mutants were generated and the effects on PHB formation and accumulation were studied. We also investigated the effect of mutation of these genes on nodulation and nitrogen fixation. In collaboration with VBI's Core Laboratory Facility (https://www.vbi.vt.edu/article/articleview/87/1/13), this work has established proteomics technologies in the group.

Integrated bioinformatics resources for rhizobia-legume symbiosis research. The three projects that we have been working on this year are integral components of a comprehensive bioinformatics resource for rhizobia-legume symbiosis research. First, we continued to develop the RhizoBRC to provide a comprehensive and accurate web-based resource for genomic and associated information on the sequence. We are making an effort to improve and update the

genome annotation for *Sinorhizobium meliloti* 1021. Currently, we have completed annotation of pSymA of *S. meliloti* at the nucleotide level. Second, we analyzed the conserved intergenic regions of Rhizobiales and searched for novel small RNAs (sRNAs) in *S. meliloti*. We have identified 77% of *S. meliloti* small RNAs (sRNAs) previously predicted by Rfam and published in the literature, and 14 potentially novel sRNAs. Some of these sRNA candidates will be further verified experimentally. Third, we updated the NodMutDB (Nodulation Mutant Database), the database that we have developed for depositing, organizing and retrieving information on symbiosis-related genes, mutants and published literature. In addition to *S. meliloti* and its host *Medicago truncatula*, which we have previously curated, we incorporated two more species in the database – *Bradyrhizobium japonicum* and its host, *Glycine max*. NodMutDB brings together new studies and existing mutant-based literature to facilitate our understanding of how genes function in symbiotic processes in both Rhizobia and their host plants.

RhizobialesBRC uses the information technology infrastructure developed by the PATRIC project (PathoSystems Resource Integration Center, patric.vbi.vt.edu, see later in this article for more on PATRIC). The genomic data are managed using Oracle9i/10 and the "Genomics Unified Schema (GUS) and Application Framework" (www.gusdb.org; Bahl et al, 2003; Manduchi et al, 2004). The database can be queried and the genomes can be browsed via our web interface. The sequence data analysis, such as the similarity search using BLAST and genome comparison using Mummer, can be done via the web services. Using the curation tool provided by PATRIC, we re-annotated pSymA of *S. meliloti* at the genome sequence level. The result is summarized in Table 1. We predicted a total of 1502 coding sequences (CDSs) including 223 new CDSs compared with what was published in RefSeq. Comparing with the RefSeq data, our annotation agreed with 1026 of RefSeq CDSs at the sequence level. We shortened 169 and extended 84 of the RefSeq CDSs. The comparison of our annotation and the RefSeq data is shown in Fig.1. We plan to continue annotating pSymB and the chromosome of *S. meliloti*. Once the sequence level annotation is completed, the proteins will be curated.

Table 1. Annotation summary for pSymA. CDSs, coding sequences.

Total CDSs	1502
Same as RefSeq	1026
RefSeq 5′ end shortened	169
RefSeq 5′ end extended	84
Gene added	223

Small non-coding RNAs (sRNAs) are an emerging class of gene expression regulators. sRNAs regulate gene expression at different levels: transcription, RNA modification or stability, and translation. The size of small RNAs ranges from 40 to 400 nucleotides. sRNAs can be grouped into two major classes: *cis* regulators, e.g. riboswitches, and *trans* regulators, e.g. Hfq-binding RNAs (Lai, 2003; Wassarman et al, 2002; Majdalani et al, 2005 for review). The genome projects often focus on annotating protein-encoding genes, tRNAs and rRNAs. sRNAs are often missed in the initial genome annotation because: 1) only recently, people began to recognize the abundance and importance of the regulatory small RNAs; 2) it is still difficult to identify sRNAs by sequence analysis.

We have developed a pipeline to identify and analyze the conserved intergenic regions

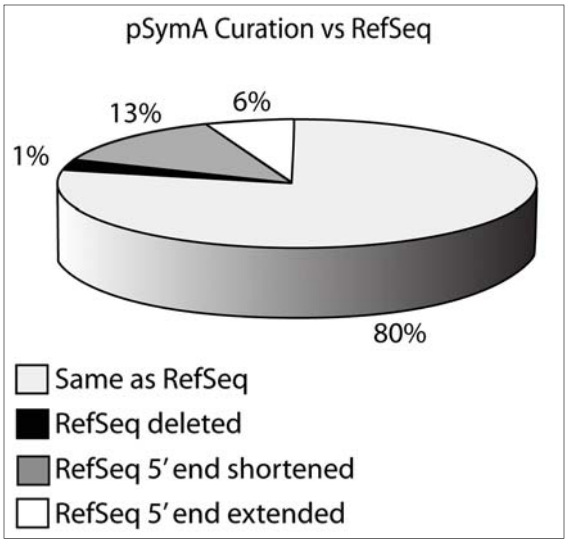


Fig.1. Comparison of RhizobialesBRC pSymA annotation with RefSeq.

(IGs) of Rhizobiales and search for novel sRNAs in *S. meliloti*.

The first four steps are automated and the last two steps are mostly manual. The pipeline result is summarized in Table 2. Out of over 5000 IGs of *S. meliloti*, 13% are conserved IGs as judged by the BLAST result. One hundred and nineteen are mapped to known RNAs and 360 are mapped to repeats. We identified 27 novel sRNA candidates by sequence alignment. 14 of these candidates are expressed according to the microarray data. One hundred percent of rRNAs and tmRNA and 98% of tRNAs were recovered (Table 3). tRNA-sec was because this tRNA is not conserved among the Rhizobiales genomes. The pipeline cut down 87% of IGs through conservation analysis; 77% of known sRNAs were still recovered. The majority of conserved IGs were mapped to repeats (consult Fig. 2 for conserved IGs analysis). For future work, it is planned to further verify the predicted sRNAs experimentally by detecting the transcripts and determining 5′ and 3′ ends of transcripts. We will also try to find possible targets of putative sRNAs and search for novel sRNAs in other Rhizobiales genomes.

***Comparative modeling of the cell cycle: from* Caulobacter crescentus *to* S. meliloti.** At some point in their life cycle, all α-proteobacteria that have an intracellular phase go through some form of differentiation with respect to their free-living stage (if one is known). The control of the bacterial cell cycle becomes important to enable successful exploitation of the intracellular environment and -- in rhizobia-plant symbioses -- must be fine-tuned so that neither partner is at risk.

Dr. John Tyson, a VBI fellow, is developing a cell cycle model for the free-living α-proteobacterium *Caulobacter crescentus*. We are collaborating on the hypothesis that a similar control over the cell cycle permits rhizobia to produce the bacteroid (intracellular) cells. Different lines of evidence have shown that at least some key regulatory genes are conserved from *C. crescentus* to *S. meliloti* (Fig.3). New experiments are needed to fully test the circuitry and develop the mathematical model that explains rhizobial bacteroid development. Comparison of the *C. crescentus*–*S. meliloti* systems might also provide insight into the *Brucella*–macrophage system, where we have already obtained transcriptional profiles from macrophages being infected by virulent and avirulent *Brucella* strains (He et al, in press).

Table 2. CIGpipe result.

Total *S. meliloti* IGs	5131
Conserved IGs from Blast result	688
Conserved IGs are known RNAs	119
Conserved IGs mapped to repeats	360
Novel sRNA candidates by alignment	27
Candidates sRNAs with expression	14

Table 3. Recovery rate from the CIPpipe.

	Recovery rate (%)
rRNA	100
tRNA	98
tmRNA	100
sRNAs from Rfam and recent publications	77

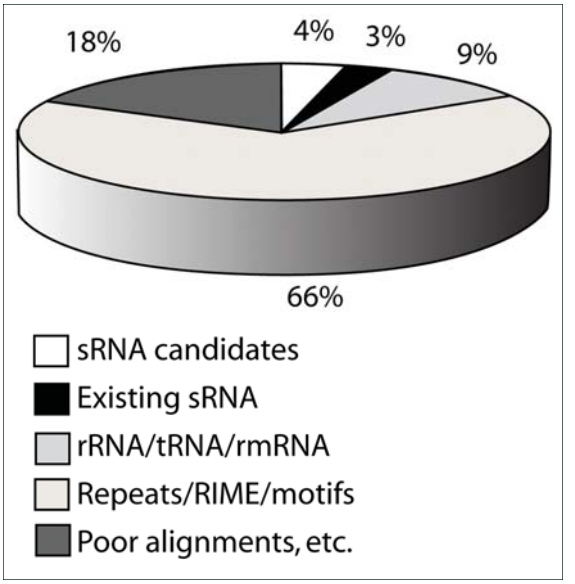


Fig. 2. *S. meliloti* conserved IGs analysis.

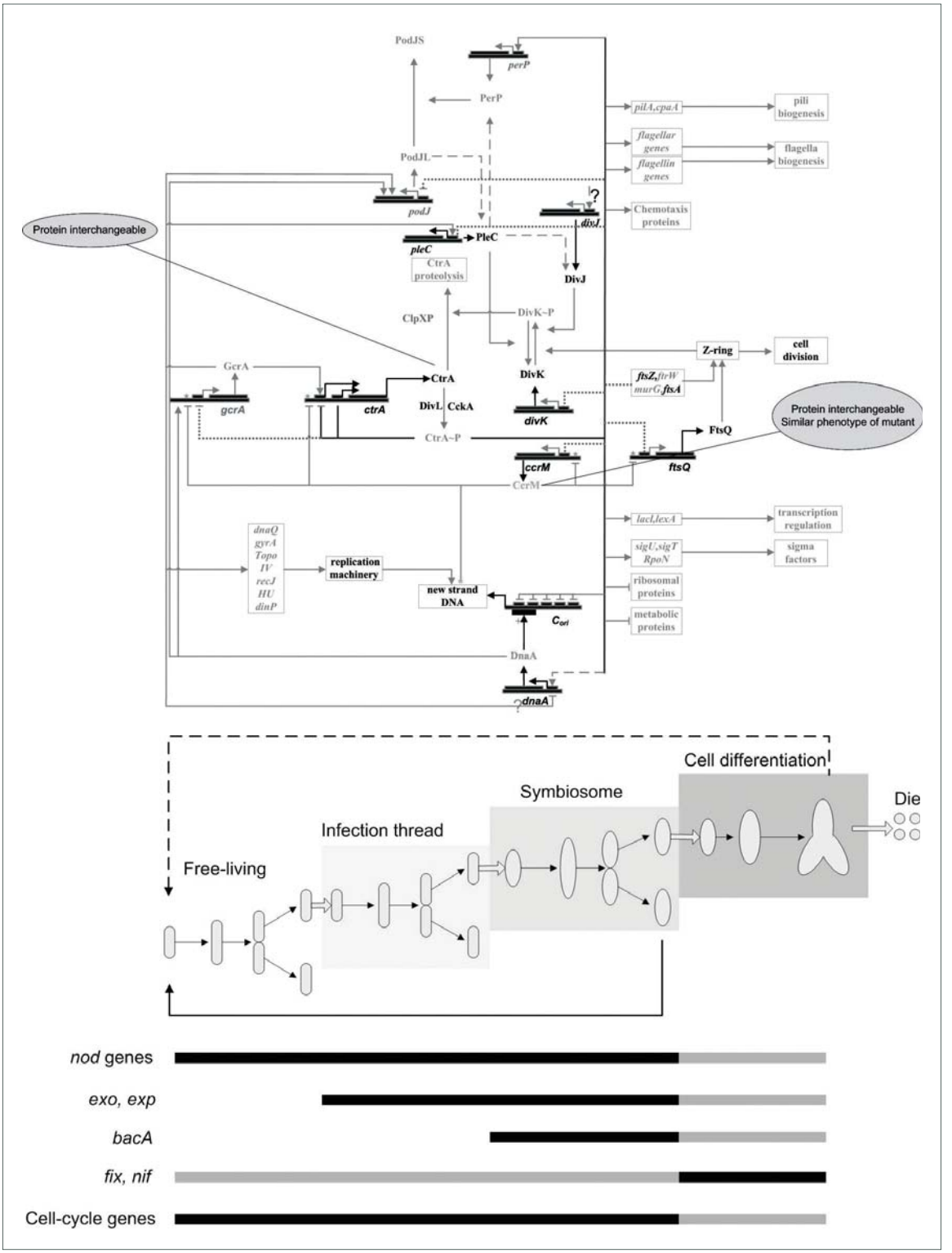


Fig. 3. The circuitry controlling cell cycle division in *S. meliloti* compared to *C. crescentus*. Gene names in bold font represent biological evidence or literature showing that these *C. crescentus* genes exist in *S. meliloti*. *perP*, *podJ*, *gcrA*, and *clpXP* represent cases for which bioinformatic analyses show that these homologs exist in *S. meliloti*. Solid lines represent biological evidence or literature supporting that these controls exist among these genes; the dashed lines (- - -) were drawn based on bioinformatic analyses showing that similar *ctrA* regulons exist among the promoters of these genes.

Cyberinfrastructure for emerging and re-emerging infectious diseases

Providing effective and efficient platforms that empower scientists and engineers to conduct transdisciplinary research (Gibbons et al, 2004) is a central part of cyberinfrastructure (CI) (Atkins, 2003). The CI Group at the Virginia Bioinformatics Institute (VBI) has established CI systems in the areas of bioinformatics and computational biology that focus on infectious diseases. The bioinformatics resources developed by the CI Group include tools for the curation of the genomes and pathosystems of a wide range of infectious organisms, database systems for acquiring, storing, and disseminating high-throughput data generated from the study of pathosystems biology (Sobral, 2002; Sobral et al, 2002; Eckart & Sobral, 2003; Lathigra et al, 2005), and software systems for analysis and visualization of the data. Education and outreach activities include the training of current and future generations of scientists and collaborative research activities with a primary goal of generating knowledge through analysis, integration and ensuring the interoperability of diverse data sets.

Summary of CI Group resources. The following summarizes the resources offered by the CI Group:

- Genome Curation Infrastructure (GCI), consisting of analytical services and web pages, has been developed for nucleotide and protein level annotation of microbial genomes. Nucleotide and protein level annotation of eight pathosystems have been completed (patric.vbi.vt.edu).

- 43 PathInfo and 30 MINet documents, which present information on 40 pathogen species, have been developed to integrate and disseminate published information on pathogens in a machine-readable format (pathport.vbi.vt.edu)

- The database system and web visualization tools have been developed to integrate and disseminate experimental data on microarrays and several proteomics data types

- Several bioinformatics tools and web services have been developed/enhanced with particular focus on microarrays (pathport.vbi.vt.edu)

- The outreach activities include training current scientists in 'omic data analysis, future scientists in disseminating the concepts of cyberinfrastructure and collaborative research services with a goal of mining the high-throughput data and discovering targets for countermeasures against emerging and reemerging infectious diseases

Genome curation infrastructure. The GCI consists of three major components: the data repository, the web site, and analytical services (Fig. 4). The underlying database, an implementation of the Genomics Unified Schema (Davidson et al, 2001) and associated tools, stores genome sequences and all data produced during the annotation and curation process. The analytical services component consists of sequence annotation pipelines and others tools used to populate the database with information for the manual curation process. For automated curation of genomes, the CI Group has developed a Java-based, domain-independent Generic Analysis Pipeline application. Specific pipeline instances are defined using GAPML (Generic Analysis Pipeline Markup Language; Newcomer, 2002) to meet requirements such as different organism types (e.g., viral versus bacterial).

The GCI web site serves as both the public face of the project and the interface through which curators view and edit annotation. Specialized web pages have been developed to review pipeline-generated annotation and enter curated data. The Gene Edit Page (GEP) provides the curators with a means to review DNA-level features and create curated features based on their interpretation. Provenance is used to record the evidence utilized by the curator to make the final decision for that particular feature. Like the GEP, the Protein Edit Page (PEP) allows the curator to review and interpret evidence generated by the Protein Analysis Pipeline (PAP). PEP is used to curate protein attributes such as function, localization and group membership, i.e., properties characteristic

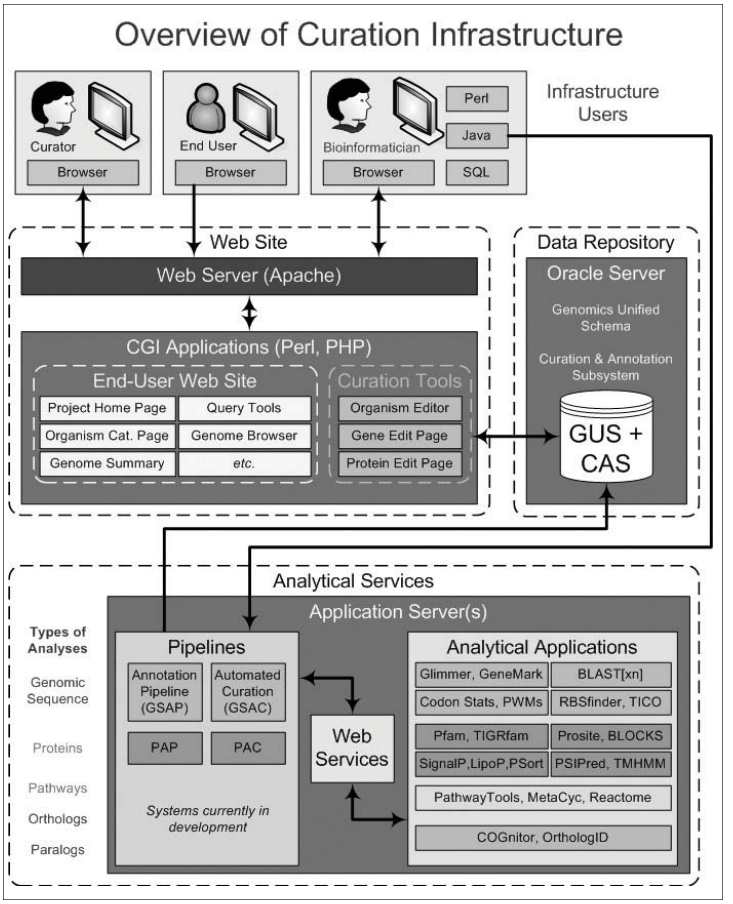


Fig. 4. Overview of the genome curation infrastructure.

of the protein as a whole. BLAST (Altschul et al, 1990) searches against the DNA and protein sequences and whole genome alignments using MUMmer (Delcher et al, 2002) are also provided. Top priorities for future development include integration of comparative genomics, pathways and experimental data from proteomics and transcriptomics.

Literature curation. The CI Group has created two XML-structured information resources to integrate and disseminate published information on pathogens in a machine-readable format. The CI Group's PathInfo documents (pathport.vbi.vt.edu/pathinfo/) (He et al, 2004) specify the taxonomy, organism characteristics, epidemiology and diagnosis of the pathogen, as well as provide information on possible hosts. The CI Group's MINet documents (pathport.vbi.vt.edu/minet/) detail the known interactions that occur between a pathogen and its host. Forty-three PathInfo and 30 MINet documents present information on 40 pathogen species (combining some closely related species). MINet documents are presented as a pathway diagram in which objects and interactions are clickable to access curated information. An illustration of the PathInfo documents is given in Fig. 5.

Microarray and proteomics database and web interfaces. The CI Group is creating infrastructure for storing the data and visualizing the results from microarray and proteomics experiments. This should allow scientists to study the dynamic responses of pathosystems with respect to specific stimuli. The microarray system has been developed based on community standards (Brazma et al, 2001). The database system is based on the Microarray GeneExpression – Object Model, MAGE-OM (Spellman et al, 2002). The web application is modeled after ArrayExpress (Brazma et al, 2003) developed at the European Bioinformatics Institute (EBI). The storage of analyzed results occurs in flat files on our application server. The CI Group's Proteomics Database hosts several datasets and data types from several organisms. The web-based system allows users to navigate individual experimental data by organism or data type. Experiment datasets are downloadable through web-based Structured Query Language (SQL) query tools and web services. The database system adheres to community standards including MIAME and MIAPE (Orchard et al, 2004), and PSI-MI (Orchard et al, 2005). Data analysis and visualization is enabled through several web-based tools for different data types such as two-dimensional gel-coupled mass spectrometry, three-dimensional image and protein-protein interactions. Future development will include integration of various data types. The web interface for visualization of the results can be found at proteinbank.vbi.vt.edu/ProteinBank.

A system has been developed for analysis and management of microarray data. The

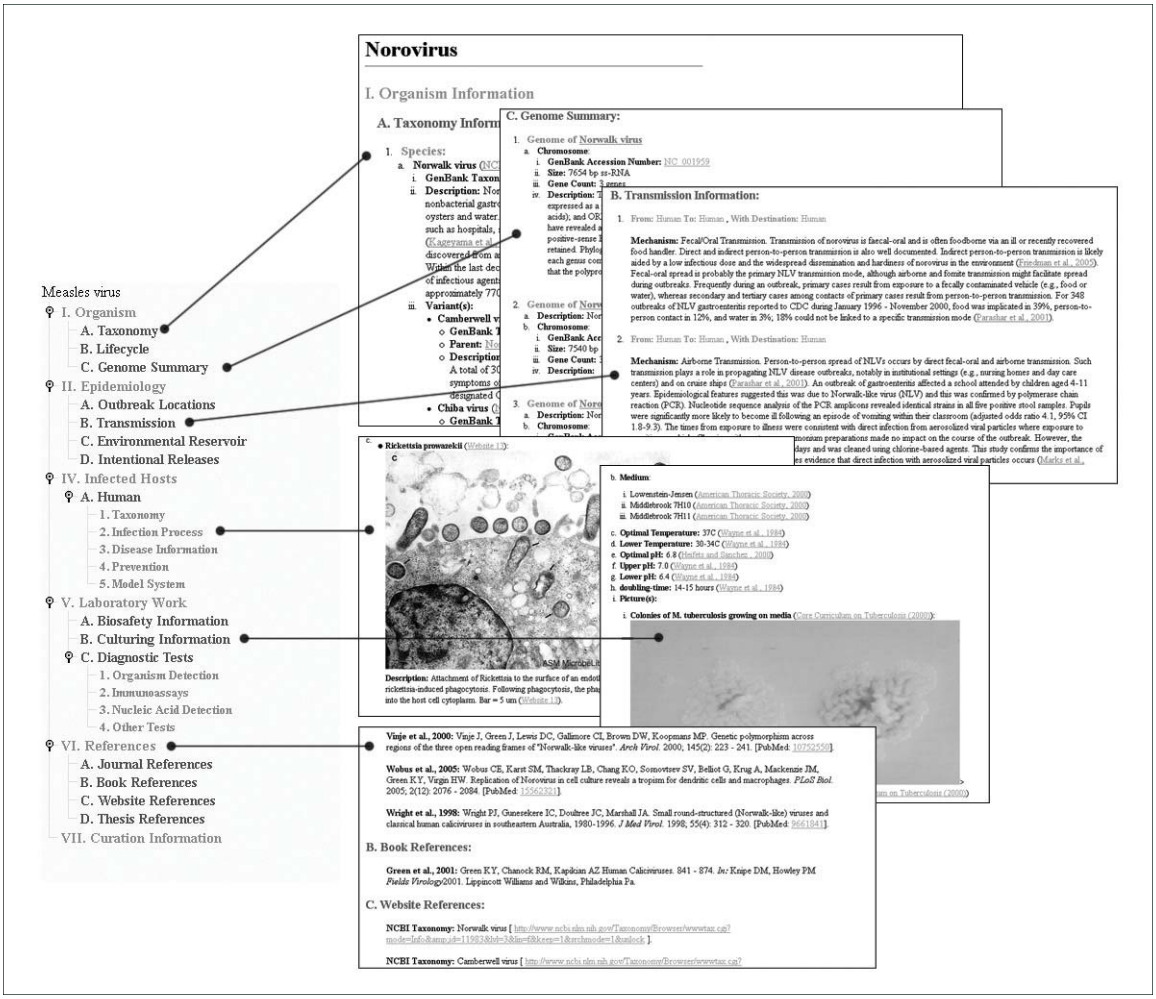


Fig. 5. A PathInfo document structure.

microarray analysis system in ToolBus/PathPort has been enhanced through rigorous requirement collection through collaboration with Dr. Brett Tyler's group at VBI. The CI Group has also developed the first version of the web-based system comprised of VBIExpressDB microarray data integrated with a pipeline for the data preprocessing and analysis in collaboration with GraphLogic (www.graphlogic.com). The system allows users to perform: 1) pre-processing of raw data, 2) basic exploratory analysis, 3) taxonomy-related analysis, and 4) facilitation of biological interpretation of the results with Gene Ontology (GO) term and pathway analysis. The system will be available at pathport.vbi.vt.edu.

Non-Human Primate Information Management System (NHP-IMS). Typically scientists developing countermeasures to infectious agents of humans must go through animal models for testing their countermeasures. In some cases, these animal models include non-human primates. Funded by the MARCE consortium, we have developed an information management system to support the needs of the Primate Research Laboratory at the University of Pittsburgh School of Medicine. The Oracle database has five custom-designed schemas, and the browser-agnostic user interface for the system is implemented in PHP, with embedded JavaScript to enhance the system interface functionality and usability. The system collects information needed to support pre-clinical animal studies and makes it available to internal and external authorized users with role-based and individual access (Fig. 6). The system currently has six functional

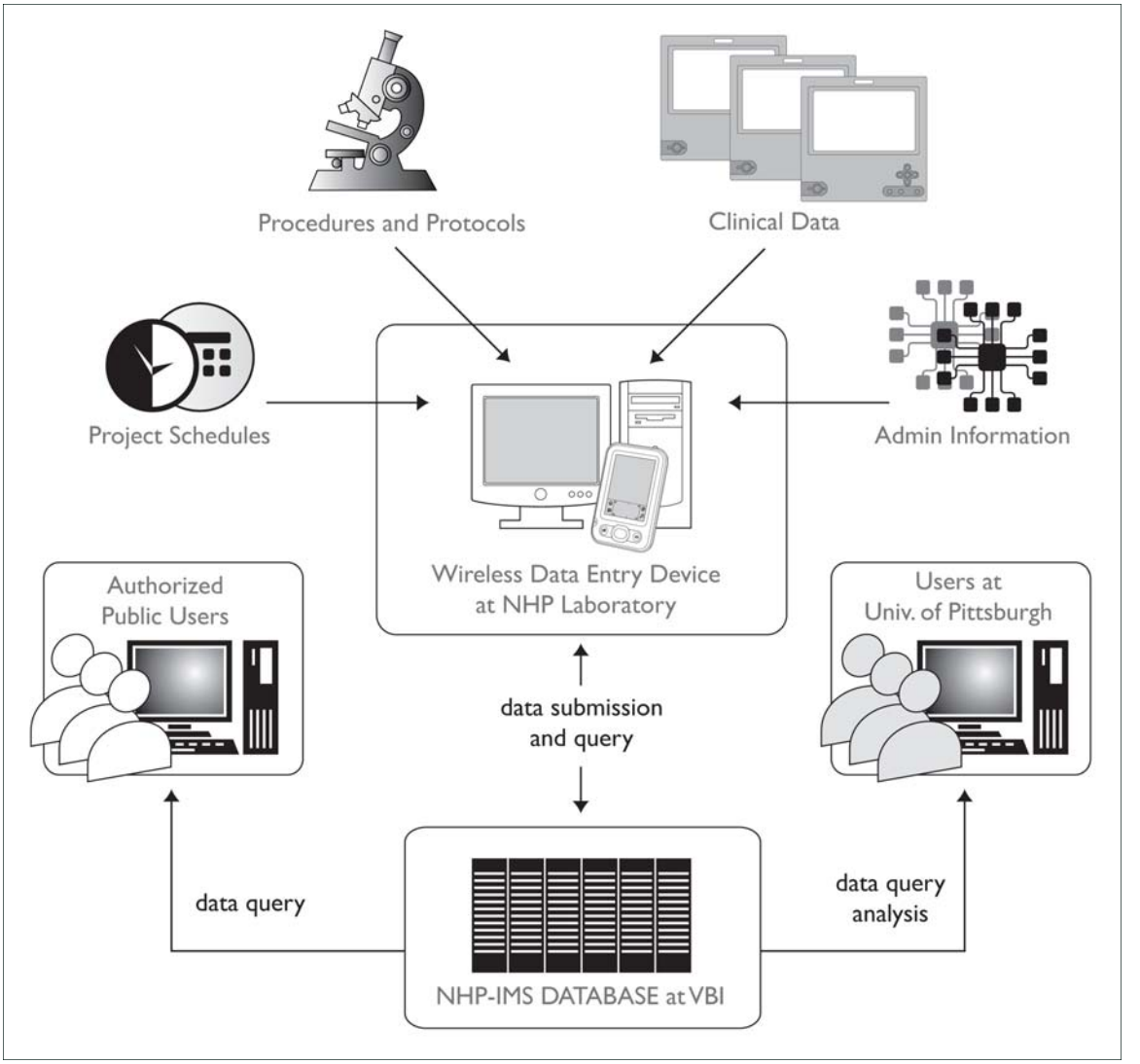


Fig. 6. Non-Human Primate Information Management System (NHP-IMS).

modules – Biographical, Protocols, Studies and Scheduling, Veterinary Procedures, Reports, and Administration.

Scientific collaborations and education and outreach

Development and deployment of CI Group resources are most effective if they are coupled with scientific collaborations with scientists trying to develop countermeasures. Through such collaborations, there is an increased understanding of the requirements of specialized information systems by the users, as well as increased understanding from the users of what is (and is not) possible. Such resource development is further enhanced and use is more likely if resource development is also tightly coupled to education and training outreach activities.

Bioinformatics training. Through the Middle Atlantic Regional Center of Excellence (MARCE) project, the CI Group has delivered bioinformatics training courses to infectious disease researchers to promote the use and understanding of bioinformatics tools and algorithms and establish dialogue with researchers. The training course lecture section reviews the fundamental applications and underlying theory of commonly used analysis

algorithms followed by a hands-on section with the ToolBus/PathPort system. We have also developed customized sessions using the data from researchers and user cases. To date, we have trained more than 70 participants representing seven MARCE institutions.

CI training for the future bioinformatics workforce. Through the leadership of Drs. Crasta, Cammer and Faulkner, the CI Group is developing a course for the students of Bluefield State College in Bluefield, West Virginia, and Galileo Magnet High School in Danville, Virginia, in partnership with the faculty of the respective institutions. The main objective is to educate current and future generations of scientists, engineers, and educators in the incorporation of CI in their teaching, training, and learning. This is being accomplished through the development and implementation of a transdisciplinary course in bioinformatics with a project-centric teaching paradigm.

Collaborative research. The CI Group's external collaborations synergize with our CI user community for the analysis and interpretation of post-sequence data, such as microarray and proteomics data, to rapidly and efficiently identify biological pathways, targets and biomarkers associated with infectious disease phenotypes of interest. Table 4 provides summaries of our current scientific collaborations. Several collaborative projects were initiated by the CI Group and are described below:

***Brucella* comparative genomics:** We have developed a genome-context-based procedure for identifying split genes due to frame-shifts and in-frame mutations, building "a complete gene set" for given organisms. Protein insertions and deletions (indels) are identified and analyzed in the context of closely related organisms to identify genes that are specific to these genome properties (Yu et al, unpublished results). Hypotheses generated from these data are being experimentally validated at Dr. Stephen Boyle's lab at Virginia Tech.

Identification and characterization of membrane-associated proteins in *Rickettsia*: The CI Group has designed an *in silico* pipeline to identify putative vaccine targets with the following properties: surface-exposed; lifestyle-critical; conserved across all sequenced *Rickettsia* species. In the first step, the *Rickettsia conorii* proteome (1374) was queried for the presence of putatively secreted proteins using the program SignalP (Nielsen et al, 1997) and 54 such proteins were identified. Using BOMP (Berven et al, 2004), members of this smaller set were then queried for the presence of a beta-barrel signature structure (usually found in surface-associated proteins and in autotransporters) yielding four proteins. All four proteins have domains that are characteristic of autotransporters and outer membrane proteins, and are conserved across the eight sequenced *Rickettsia* species. These *in silico* results will be experimentally validated at Dr. Abdu Azad's laboratory at the University of Maryland.

Characterization and whole genome sequencing of the attenuated strain S19 of *Brucella abortus*: *B. abortus* strain S19 is a spontaneously attenuated strain obtained from the virulent strain *B. abortus* 2308. Live attenuated *B. abortus* S19 has been used for many years as an effective vaccine to prevent brucellosis in cattle. *B. abortus* S19 is expected to have lost some essential but yet unknown mechanism of virulence. Comparisons of these two genomes might provide some insight into future vaccine development strategies. We have applied (pyro)sequencing technology (Margulies et al, 2005) (www.454.com) to rapidly and comprehensively determine more than 99.5% of the genome sequence of S19. The genome sequence was completed at VBI's Core Laboratory Facility (www.vbi.vt.edu/core/) through additional targeted sequencing of gaps identified by comparative analysis of the S19 contiguous sequences against the only published (when assembly was done) whole genome sequence of *B. abortus* 9-941. Comparison of 9-941 with S19 has already identified more than 200 single nucleotide polymorphisms (SNPs) with high confidence (unpublished results). One hundred and fifty-six of these SNPs are from gene coding regions and 107 SNPs affect the translation of the corresponding proteins. In addition, several insertion/deletions were identified in S19 compared to 9-941. These findings are being studied by additional sequencing.

Table 4. Examples of CI Group research collaborations.

Collaborator	Project
Dr. Maria Salvato (University of Maryland, Baltimore)	Analysis of VBI-generated expression profiling data of early response to hemorrhagic fever (Ebola virus) in *Rhesus macaque*.
Dr. James Kaper (University of Maryland, Baltimore)	Epithelial cell response to enterohemorrhagic *E. coli* O157: H7 infection.
Dr. Abdu Azad (University of Maryland, Baltimore)	Comparative genomics training, identification and characterization of unique genes among the nine available *Rickettsia* genomes in GenBank.
Dr. Brett Tyler (Virginia Bioinformatics Institute)	Creation of microarray tools supporting host and pathogen data on the same microarray.
Dr. William Petri (University of Virginia)	Expression profiling of *E. histolytica* during mouse infection stage.
Dr. Tomoyoshi Nozaki (Japan)	Comparison of transcriptomes among different *E. histolytica* strains.
Dr. George Dimopolous (John Hopkins University)	Development of microarray database for storage and presentation.
Dr. Carole Goble (MyGrid and BioMoby)	Development of interoperable Web-services and integration with Taverna.
Inpharmatica	Analyzed *Brucella* spp. proteome for functional analysis of the genes as well as the characterization of the drug targets by comparing to the proprietary database they have on the known drug targets.

Loss of a universal tRNA identity element: Rather than adding an extra 5′ guanylate post-transcriptionally as eukaryotes do, the CI Group has found, in a collaborative effort with the Dr. Sobral's PathoSystems Biology Group, that *Sinorhizobium meliloti* simply lacks any extra nucleotide on tRNA-His. This loss correlates with changes throughout the clade at the 3′ end sequence of tRNA-His and at many sites in histidyl-tRNA synthetase that might be expected to affect tRNA-His recognition, and may even affect global tRNA identity rules in the clade. This work demonstrates the value of careful attention to RNAs in genomics, and will affect the way this gene is annotated in pathogen genomes (Williams et al, unpublished results).

Functional genomic analysis of fruit flavor and nutrition pathways: Plant breeders have traditionally ignored flavor due to its genetic complexity and the tomato has suffered from this neglect. In collaboration with researchers at the University of Florida and at Cornell University, we are exploiting the diverse and well-characterized tomato germplasm base with high throughput metabolite screens to identify lines altered in flavor and nutrient composition and then using an integrated genetic, genomic and bioinformatic approach to identify genes involved in pathways for synthesis of metabolites essential to the attributes of fruit flavor and nutrient. During the past year, our collaborators have generated significant amounts of gene expression and metabolite profiling data. In order to store, manage, and assist the exploitation of these datasets, we continue to develop the two databases: tomato metabolite database and tomato expression database (Fei et al, 2006). In these databases, tools to mine, visualize, and analyze the gene expression and metabolite profiling data have been developed. In addition, tools to integrate gene expression and metabolite profiling data along with metabolic pathway information are also being implemented in the databases.

In summary, the PathoSystems Biology group has sought to further understand bacterial life inside eukaryotic host cells through α-proteobacterial models, with special emphasis on *S. meliloti*, *Brucella* spp., and rickettsia. From molecular and genomic, transcriptomic and proteomic approaches, data and technologies have been developed and applied to these systems in a comparative framework. Meanwhile, the CI Group has designed and deployed critical cyberinfrastructure for pathosystems biology, largely building from the genome through to the dynamic responses of the genome to the environment (transcriptome and proteome initially). Furthermore, new paradigms of transdisciplinary, team-oriented, problem-solving training, outreach and educational activities have been developed and deployed in concert with the evolving scientific and infrastructural goals. We feel this is a novel and effective way to develop, build and deploy cyberinfrastructure for pathosystems biology.

Acknowledgements

The work described herein is partially funded by the following awards to Dr. Bruno Sobral: Department of Defense (Contract #W911SR-04-C-0045) for work on PathPort, subcontract from Social and Scientific Systems, Inc. (DMID1-VBI) through the National Institute of Allergy and Infectious Disease (Contract #HHSN266200400061C) "Administrative Center for Biodefense Proteomic Research Programs", National Institute of Allergy and Infectious Disease Contract #HHSN266200400035C (PATRIC), subcontract from the University of Maryland (PO# S01776) through the National Institute of Allergy and Infectious Disease (Cooperative Agreement #U54 AI57168) for the "Mid-Atlantic Regional Center of Excellence (MARCE)". Dr. Zhangjun Fei was partially supported by a subcontract from the University of Florida (UF-IFAS 00057866) through the National Science Foundation (Grant #DBI-0501778); Dr. Oswald Crasta and Dr. Stephen Cammer were partially supported by awards from the National Science Foundation (Grant #OCI-0537461) to the Cyberinfrastructure Group.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**(3): 403-410.

Andersson SG, Kurland CG (1998) Reductive evolution of resident genomes. *Trends Microbiol.* **6**(7):263-268.

Atkins D, Droegemeier K, Feldman SI, Garcia-Molina H, Klein ML, Messerschmitt DG, Messina P, Ostriker JP, Wright MH (2003) Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure, National Science Foundation.

Bahl A, Brunk B, Crabtree J, Fraunholz MJ, Gajria B, Grant GR, Ginsburg H,Gupta D, Kissinger JC, Labo P, Li L, Mailman MD, Milgram AJ, Pearson DS, Roos DS, Schug J, Stoeckert CJ Jr, Whetzel P (2003) PlasmoDB: the *Plasmodium* genome resource. A database integrating experimental and computational data. *Nucleic Acids Res.* **31**: 212-215.

Berven FS, Flikka K, Jensen HB, Eidhammer I (2004) BOMP: a program to predict integral beta-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria. *Nucleic Acids Res.* **32** (Web Server issue): W394-399.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* **29**(4): 365-371.

Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, Oezcimen A, Rocca-Serra P, Sansone SA (2003) ArrayExpress – a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* **31**(1): 68-71.

Davidson SB, Crabtree J, Brunk B, Schug J, Tannen V, Overton C, Stoeckert C (2001) K2/Kleissli and GUS: Experiments in integrated access to genomic data sources. *IBM Systems Journal* **40**(2): 1-20.

Delcher AL, Phillippy A, Carlton J, Salzberg SL (2002) Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* **30**(11): 2478-2483.

Eckart JD, Sobral BW (2003) A life scientist's gateway to distributed data management and computing: the PathPort/ToolBus framework. *Omics* **7**(1): 79-88.

Emelyanov VV (2001a) Rickettsiaceae, rickettsia-like endosymbionts, and the origin of mitochondria. *Biosci Rep.* **21**(1):1-17.

Emelyanov VV (2001b) Evolutionary relationship of Rickettsiae and mitochondria. *FEBS Lett.* **501**(1):11-18.

Gibbons M, Limoges C, Nowotny N, Simon Schwartzman, Scott P, Trow M (1994) *The new production of knowledge: The dynamics of science and research in contemporary societies*, Sage Publications.

Gray MW, Burger G, Franz Lang B (1999) Mitochondrial evolution. *Science*: **283**(5407), 1476-1481.

Gupta RS (2000) The phylogeny of proteobacteria: relationships to other eubacterial phyla and eukaryotes. *FEMS Microbiol Rev.* **24**(4): 367-402.

He Y, Vines RR, Wattam AR, Abramochkin GV, Dickerman AW, Eckart JD, Sobral BW (2005) PIML: the Pathogen Information Markup Language. *Bioinformatics* **21**(1): 116-121.

Kainth P, Gupta RS (2005) Signature proteins that are distinctive of alpha proteobacteria. *BMC Genomics*. **6**: 94.

Kersters K, De Vos P, Gillis M, Swings J, Vandamme P, Stackebrandt E (2003) Introduction to the Proteobacteria. In *The Prokaryotes: An evolving electronic resource for the microbiological community*. Dworkin M (ed), 3rd edition, Release 3.12. New York: Springer-Verlag.

Lai EC (2003) RNA sensors and riboswitches: self-regulating messages. *Curr. Biol.* **13**: 285–291.

Lathigra R, He Y, Vines RR, Nordberg EK, Sobral BWS (2003) A biologist's view of systems integration for system biology - the Pathogen Portal Project. *Conference Proceedings for Stadler Symposium.* Columbia, MO: March 31-April 2. pp. 1-12.

Luyten E, Vanderleyden J (2000) Survey of genes identified in *Sinorhizobium meliloti* spp., necessary for the development of an efficient symbiosis. *Eur. J. Soil Biol.* **36**:1-26.

Majdalani N, Vanderpool CK, Gottesman S (2005) Bacterial small RNA regulators. *Crit. Rev. Biochem. Mol. Biol.* **40**: 93-113.

Manduchi E, Grant GR, He H, Liu J, Mailman MD, Pizarro AD, Whetzel PL, Stoeckert CJ Jr (2004) RAD and the RAD Study-Annotator: an approach to collection, organization and exchange of all relevant information for high-throughput gene expression studies. *Bioinformatics* **20**: 452-459.

Margulies M et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**(7057): 376-380.

Newcomer E (2002) *Understanding Web Services: XML, WSDL, SOAP, and UDDI*. Boston: Addison-Wesley.

Nielsen H, Engelbrecht J, Brunak S, von Heijne G (1997) A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int. J. Neural Syst.* **8**(5-6): 581-599.

Orchard S, Hermjakob H, Julian RK Jr, Runte K, Sherman D, Wojcik J, Zhu W, Apweiler R (2004) Common interchange standards for proteomics data: Public availability of tools and schema. *Proteomics* **4**(2): 490-491.

Orchard S, Hermjakob H, Taylor CF, Potthast F, Jones P, Zhu W, Julian RK Jr, Apweiler R (2005) Further steps in standardisation. Report of the second annual Proteomics Standards Initiative Spring Workshop (Siena, Italy 17-20th April 2005). *Proteomics* **5**(14): 3552-5.

Sobral B (2002) Senate testimony on fighting bioterrorism: Using America's scientists and entrepreneurs to find solutions. Subcommittee on Science, and Space of the Senate Commerce, Science, and Transportation Committee.

Sobral B, Eckart D, Laubenbacher R, Mendes P (2002) The role of bioinformatics in toxicogenomics and proteomics. *NATO Advanced Workshop on Toxicogenomics and Proteomics*, Prague, Czech Republic.

Spellman PT, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, Bernhart D, Sherlock G, Ball C, Lepage M, Swiatek M, Marks WL, Goncalves J, Markel S, Iordan D, Shojatalab M, Pizarro A, White J, Hubley R, Deutsch E, Senger M, Aronow BJ, Robinson A, Bassett D, Stoeckert CJ Jr, Brazma A (2002) Design and implementation of microarray gene expression markup language (MAGE-ML) *Genome Biol.* **3**(9): RESEARCH0046.

Tsolis RM (2002) Comparative genome analysis of the alpha-proteobacteria: relationships between plant and animal pathogens and host specificity. *Proc. Natl Acad. Sci. USA* **99**(20):12503-12505.

Wassarman KW (2002) Small RNAs in bacteria: diverse regulators of gene expression in response to environmental changes. *Cell* **109**: 141–144.

Publications

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

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

Hance ME, MJ Czar, Azad A, Purkayastha A, Snyder EE, Crasta OR, Setubal JC, Sobral BW (2005) The Pathogen Resource Integration Center: Implications for rickettsial research. *Ann. N.Y. Acad. Sci.* **1063**: 459–465.

He Y, Reichow S, Ramamoorthy S, Ding X, Lathigra R, Craig JC, Sobral BWS, Schurig GG, Sriranganathan N, Boyle SM *Brucella melitensis* triggers time-dependent modulation of apoptosis and down-regulation of mitochondria-associated gene expression in mouse macrophages. *Infect. Immun.*, in press.

Yang B, Xue T, Zhao J, Kommidi C, Soneja J, Li J, Will R, Sharp B, Kenyon R, Crasta O, Sobral BW (2006) Bioinformatics Web Services Provided by VBI. Regular Research Paper Paper ID #: BIC4190. BIOCOMP'06: June 26-29, 2006, Las Vegas, USA.

Zhao J, Xue T, Yang B, Williams K, Wattam R, Will R, Sharp B, Kenyon R, Crasta O, Sobral BW (2006) VBI Genome Annotation and Comparison System. Regular Research Paper Paper ID #: BIC4189. BIOCOMP'06: June 26-29, 2006, Las Vegas, USA.

Genomic and bioinformatic analysis of oomycete-host interactions

Brett Tyler, bmtyler@vt.edu
Professor, Virginia Bioinformatics Institute &
Plant Pathology, Physiology and Weed Science, Virginia Tech

Oomycete plant pathogens such as *Phytophthora* species and downy mildews cause destructive diseases of an enormous variety of crop plant species as well as forests and native ecosystems. These organisms are most closely related to algae in the kingdom Stramenopiles, and hence have evolved plant pathogenicity independently of other plant pathogens such as fungi. We have used structural genomics (genome sequence comparisons) and functional genomics (transcriptional profiling) to identify plant and pathogen genes that may be key players in the interaction between the soybean pathogen *Phytophthora sojae* and its host. In *P. sojae*, we have identified many rapidly diversifying gene families that encode potential pathogenicity factors including protein toxins, and a class of proteins (avirulence or effector proteins) that appear to have the ability to penetrate plant cells. Transcriptomic analysis of quantitative or multigenic resistance against *P. sojae* in soybean has revealed that there are widespread adjustments in host gene expression in response to infection, and that some responses are unique to particular resistant cultivars. These observations lay the foundation for dissecting the interplay between pathogen and host genes during infection at a whole-genome level.

Keywords: oomycetes; *Phytophthora*; soybean; genome sequences; transcriptomics.

Introduction

Plant pathogens from the genus *Phytophthora* cause destructive diseases of an enormous variety of crop plant species as well as forests and native ecosystems (Erwin & Ribiero, 1996). The potato pathogen *Phytophthora infestans* was responsible for the Irish potato famine, and is still a destructive pathogen of concern for biosecurity. The newly emerged *Phytophthora* species, *Phytophthora ramorum*, is attacking trees and shrubs of coastal oak forests in California, including the keystone live oak species (sudden oak death disease) (Rizzo et al, 2002). The soybean pathogen *Phytophthora sojae* is also causing serious losses to the United States soybean crop, in the order of US$ 1–2 million annual damage (Erwin & Ribiero, 1996). This species has been used for many basic studies of *Phytophthora* because it is easy to genetically manipulate. Taxonomically, *Phytophthora* pathogens are oomycetes, which are organisms that resemble fungi but belong to a kingdom of life called Stramenopiles that are most closely related to algae such as kelp and diatoms. Hence conventional fungal control measures often fail against these pathogens.

Oomycetes include many other destructive plant pathogens in addition to *Phytophthora*, in particular the downy mildews and more than 100 species of the genus *Pythium*. These organisms at least double the losses due to *Phytophthora*. The downy mildews of maize, caused by *Peronosclerospora philippinensis* and *Sclerophthora rayssiae*, are considered major biosecurity threats to the United States. The downy mildew of *Arabidopsis thaliana*, caused by *Hyaloperonospora parasitica*, is an important model system for understanding plant responses to infection.

Contributors:
Felipe Arredondo, Nathan Bruce, Marcus Chibucos, Yinghui Dan, Daolong Dou, Bing Fang, Adriana Fereira, Nikolaus Galloway, Dianjing Guo, Regina Hanlon, Rays Jiang, Andrew Kincaid, Angela Ko, Konstantinos Krampis, Bing Liu, Varun Pandey, Rajat Singhania, Brian Smith, Ken Tian, Trudy Torto-Alalibo, Sucheta Tripathy, Lachelle Waller, Xuemin Zhang, Lecong Zhou

The overall thrust of our research is to identify and characterize the genes in oomycete species that enable these pathogens to recognize their plant hosts and overcome the defenses of the plant host, and to determine the mechanisms by which they do so. The approaches we are using are centered around genome-wide technologies, namely using a combination of high-throughput experimental and bioinformatic methods to identify pathogen and host genes that participate in the interaction, and to predict the functional interactions among the products of those genes that determine the outcome of infection.

Sequencing of oomycete genomes

This project involves collaborations with teams led by Dr. Jeffrey Boore and Dr. Dan Rokhsar (DOE Joint Genome Institute), Dr. Sandra Clifton (Washington University Genome Sequencing Center), Dr. Jim Beynon (Warwick University, United Kingdom), Dr. Jane Rogers (Sanger Sequencing Centre, United Kingdom) and Dr. John McDowell (Department of Plant Pathology, Physiology and Weed Science, Virginia Tech).

The complete genome sequence of an organism is an excellent starting point for identifying genes involved in pathogenicity, or any other process of interest. The genome sequence can also aid substantially in developing genetic tools for detecting and tracking the pathogen. In collaboration with the DOE Joint Genome Institute (JGI), we have completed draft genome sequences of *P. sojae* and *P. ramorum*. The 95-Mb genome of *P. sojae* was sequenced to a depth of 9x while the 65-Mb genome of *P. ramorum* was sequenced to a depth of 7x, both by whole-genome shotgun sequencing. Furthermore, we have obtained a preliminary sequence of the 75-Mb genome of *Hyaloperonospora parasitica* at a depth of 8x. We have completed the identification and annotation of genes in the genomes of *P. sojae* and *P. ramorum*, and this process is just beginning for *H. parasitica*.

Automated gene calling software has identified 19 027 predicted genes (called gene models) in the genome of *P. sojae* and 15 743 gene models in the genome of *P. ramorum*. Of these gene models, 9768 are similar enough in sequence between *P. sojae* and *P. ramorum* that

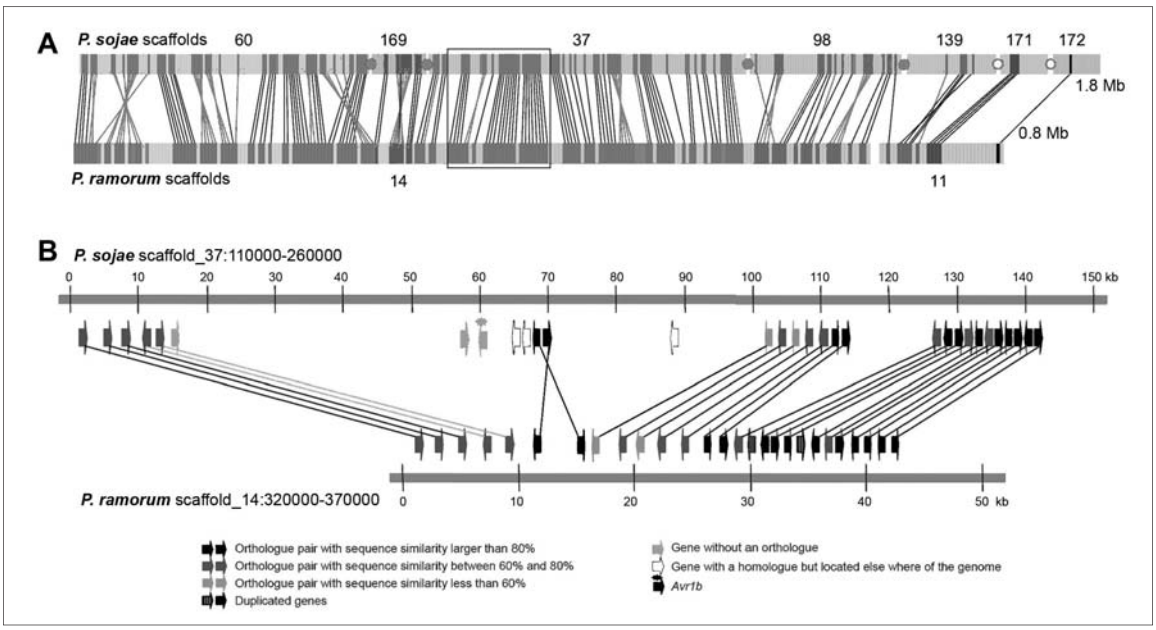


Fig. 1. Long range gene colinearity between the genomes of *P. sojae* and *P. ramorum*. In **(A)** and **(B)**, black and gray lines link orthologs of like and reversed orientation, respectively. In **(A)**, black bars indicate orthologs located in different *P. sojae* sequence scaffolds. Pale bars indicate genes without orthologs. Filled circles indicate scaffolds linked by a single end-sequenced BAC, and open circles indicate scaffolds linked by end-sequenced BAC contiguous sequences. The boxed area in **(A)** is enlarged in **(B)**.

they are predicted to have the same function. These highly similar genes, sometimes called orthologs, are arranged in very similar orders along the chromosomes (i.e. exhibit synteny). By comparing common runs of genes in sequence scaffolds from the *P. sojae* and *P. ramorum* genome sequence assemblies, we can predict which sequence scaffolds lie side by side along the chromosomes of the two species. This is illustrated in Fig.1A. Here seven scaffolds spanning a 1.8-Mbp region of the *P. sojae* genome are matched to two scaffolds of the *P. ramorum* genome sequence by the presence of common genes. Preliminary analysis of the *H. parasitica* genome sequence has revealed that gene order is also well conserved in that species compared to the two *Phytophthora* species. Despite the conservation of gene order, large segments of genome sequence in the *P. sojae* genome appear to be absent from *P. ramorum*, which partially accounts for the larger genome size of *P. sojae* (Fig.1B).

Although 9768 genes appear to be orthologous between *P. sojae* and *P. ramorum*, a very large number (49% in *P. sojae*; 38% in *P. ramorum*) have diverged so much that they have no matching gene in the other species. Of these divergent genes, 1755 *P. sojae* genes are unique to that species and 624 appear unique to *P. ramorum*. The remainder of the genes represent gene families that are present in both species but whose members do not match one another precisely. Proteins that are secreted by the pathogen during infection, as well as being divergent between *P. sojae* and *P. ramorum*, are excellent candidates for proteins that function to promote the infection process. An example is the explosive expansion and divergence of one secreted protein family called NPP1, which encodes a phytotoxin (Fig.2). Although 1-4 *NPP1* genes are present in some fungal and bacterial species, *P. sojae* and *P. ramorum* have 29 and 40 copies, respectively. *H. parasitica* also has multiple *NPP1* genes.

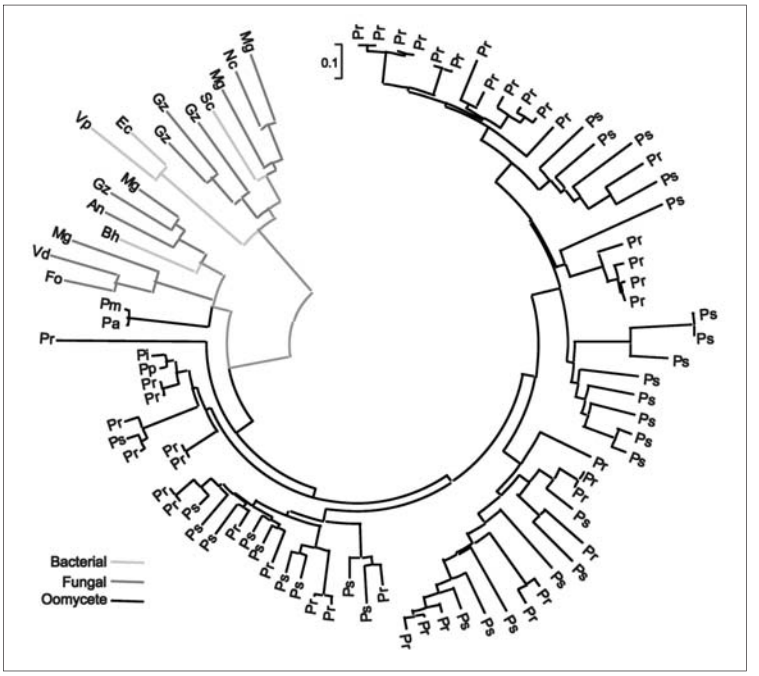


Fig. 2. Sequence divergence of the NPP1 or Nep1-like protein toxin family. This phylogram was constructed from 40 *P. ramorum* and 29 *P. sojae* sequences, plus 20 from GenBank. Protein sequences were aligned using ClustalW, and the un-rooted phylogram was made using the neighbor-joining method (MEGA 3.1). The scale bar represents 10% weighted sequence divergence. Species-of-origin are abbreviated as follows: An, *Aspergillus nidulans*; Bh, *Bacillus halodurans*; Ec, *Erwinia caratovora*; Fo, *Fusarium oxysporum*; Gz, *Giberella zeae*; Mg, *Magnaporthe grisea*; Nc, *Neurospora crassa*; Pa, *Pythium aphanidermatum*; Pi, *Phytophthora infestans*; Pm, *Pythium monospermum*; Pp, *Phytophthora parasitica*; Ps, *Phytophthora sojae*; Pr, *Phytophthora ramorum*; Sc, *Streptomyces coelicolor*; Vd, *Verticillium dahlia*; Vp, *Vibrio pommerensis*.

A major finding from the genome sequencing was the discovery of a very large, very diverse superfamily of genes with similarity to the oomycete avirulence genes, *Avr1b*-1 gene of *P. sojae* (Shan et al, 2004), *Atr*1 (Rehmany et al, 2005) and *Atr*13 (Allen et al, 2004) of *H. parasitica* and *Avr3a* of *P. infestans* (Armstrong et al, 2005). Members of this superfamily share two conserved motifs, called RXLR and dEER, a short distance from the end of the secretory leader. These genes likely encode pathogenicity proteins that can enter the host plant cell by means of the RXLR motif in order to reprogram the host cell to make it more susceptible to infection. These genes are described in more detail in the following section on avirulence genes.

A further inference from the oomycete genome sequences is that oomycetes evolved from a photosynthetic ancestor, in common with photosynthetic members of the Stramenopile kingdom such as diatoms. Many examples of oomycete genes with close affinity to plant and algal genes were found.

Population structure of* P. ramorum *inferred from single nucleotide polymorphisms. *P. ramorum*, like other oomycetes, is diploid, so that two haplotypes were recovered from the genome sequence. Comparison of the two haplotypes identified approximately 200 000 single nucleotide polymorphisms (SNPs). We designed an Affymetrix SNP GeneChip® containing probes for 880 SNPs, and used it to genetically type *P. ramorum* isolates from the United States and Europe. The results show that the two most common clones of *P. ramorum*, found in the United States and Europe, respectively, originated from a single, sexually-reproducing population, but that a new genotype of *P. ramorum* has recently entered the United States from a different origin.

VBI Microbial Database and Genome Community Annotation Tool. Analysis and visualization of gene predictions and annotations of the *P. sojae* and *P. ramorum* genomes are available at the VBI Microbial Database (VMD) (phytophthora.vbi.vt.edu) as well as from JGI Genome Portals (www.jgi.doe.gov/genomes). The database schema for VMD is based on the Genomics Unified Schema (GUS) developed at the University of Pennsylvania (www.gusdb.org/). VMD stores not only genome sequence data, but also Expressed Sequence Tag (EST) and microarray data. We built a new genome browser for viewing the genome sequence and EST data stored in VMD and also a Genome Community Annotation Tool (GCAT) for community annotation of the sequences. We have also collaborated with the VBI Cyberinfrastructure Group led by Dr. Bruno Sobral in the creation of new viewing and analysis tools for the microarray data stored in VMD.

Gene ontology terms for plant-associated microbes. The Gene Ontology (GO; www.geneontology.org) is a set of standardized terms for describing biological processes, molecular functions and cellular components of living organisms (The Gene Ontology Consortium, 2000). The use of these terms to annotate the roles of genes greatly aids in identifying similarities among organisms. In collaboration with Dr. João Setubal at VBI and researchers at Cornell University (Dr. Alan Collmer), the University of Wisconsin (Dr. Nicole Perna and Dr. Jeremy Glasner), Wells College (Dr. Candace Collmer), North Carolina State University (Dr. Ralph Dean and Dr. David Bird) and the Institute for Genome Research (Dr. Michelle Gwinn-Giglio and Owen White), we have developed new terms for annotating the contributions of genes to interactions between microbes and hosts to assist in identifying similarities among plant-associated microbes (see pamgo.vbi.vt.edu).

Function of avirulence genes during *Phytophthora* infection

Major plant disease resistance genes (R genes) that protect plants against oomycetes and many other pathogens encode intracellular receptors that detect the presence of pathogen proteins in the plant cytoplasm. These pathogen proteins, called avirulence or effector proteins, likely enter the plant cell to increase its susceptibility to infection (Chang et al, 2004). The mechanism by which fungal and oomycete proteins enter plant cells is not known. At least 10 avirulence genes have been identified genetically in *P. sojae* (Tyler, 2002) and we previously cloned one of them, *Avr1b*-1, and showed that it encoded a small secreted protein. As described above, comparison of *Avr1b*-1 with other oomycete avirulence genes and with over 350 similar genes each in the *P. sojae* and *P. ramorum* genomes identified two conserved motifs called RXLR and dEER a short distance from the end of the secretory leader. Since the four oomycete avirulence proteins interact with intracellular plant disease resistance proteins, and so are inferred to have the ability to enter the plant cell, we hypothesized that the RXLR motif, with or without the dEER motif, is involved in the ability of these proteins to enter the plant cell. We have now obtained experimental support for this hypothesis. Mutations in the RXLR motif of *Avr1b*-1 block the ability of the protein to interact with the plant when it is secreted into the extracellular space, but not when it is synthesized inside the plant cell. Proteins secreted by the malaria parasite *Plasmodium* that

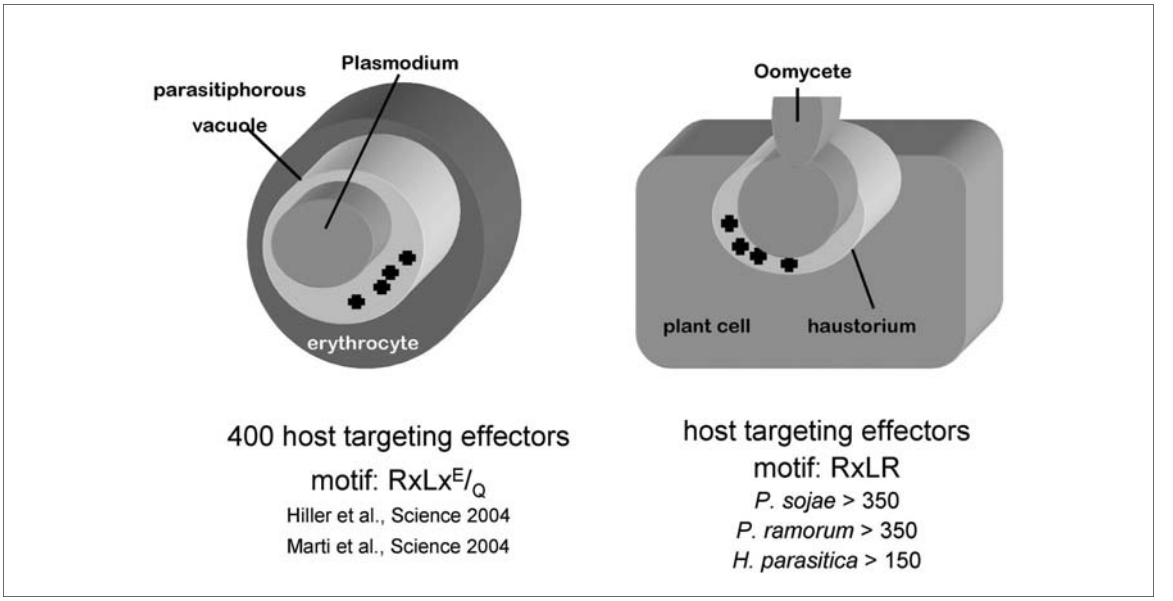


Fig. 3. Comparison of effector mechanisms between malaria and *Phytophthora*.

have the ability to cross the membrane enclosing the parasite (the parasitiphorous vacuolar membrane) into the lumen of the red blood cell contain a motif near the secretory leader, called Pexel (Marti et al, 2004) or VTF (Hiller et al, 2004), which closely resembles the RXLR motif (Fig. 3), suggesting that both pathogens use a similar mechanism to introduce virulence proteins into the cytoplasm of their host cells. We have shown that *P. sojae* transformants that overexpress the Avr1b protein have increased virulence on some soybean cultivars, supporting the hypothesis that oomycete avirulence proteins contribute positively to virulence.

Counter-play of plant and pathogen genes during *Phytophthora* infection of soybean

This project is a collaboration with Dr. Saghai Maroof (Department of Crop Soil and Environmental Science, Virginia Tech), Dr. Ina Hoeschele (Virginia Bioinformatics Institute), and Dr. Anne Dorrance and Dr. Steven St. Martin (Ohio State University).

Plant pathogenic microbes have evolved special mechanisms to defeat their hosts' defenses. To protect themselves, plants must evolve additional counter-measures against which the pathogens must in turn evolve new mechanisms of virulence. As a result of this evolutionary "arms race", large numbers of plant genes contribute to natural resistance (i.e., quantitative or multigenic resistance), and large numbers of pathogen genes contribute to virulence.

Crop breeders have found that improving multigenic resistance gives much longer protection to crops than single resistance genes, which are quickly overcome by new strains of pathogens. However, this kind of resistance is much harder to improve by conventional breeding because multigenic resistance is created by many genes making small contributions (Young, 1996). It is also much harder to study the molecular mechanisms by which the genes act.

This project is focused on characterizing mechanisms of quantitative resistance in soybean against *P. sojae* using a combination of transcriptional profiling and quantitative trait locus (QTL) mapping. We have completed an analysis of the transcriptional profiles of eight cultivars with varying levels of quantitative resistance, following inoculation with *P. sojae*. Based on pilot experiments, we assayed infected roots three and five days after infection and compared them to mock inoculated roots. The transcriptional profiles were measured using an Affymetrix GeneChip® that contains probes for

38 000 soybean genes and 15 800 *P. sojae* genes. The major findings of this analysis were:

1. A very high proportion of soybean genes detectable by the arrays (>80%) showed significant changes in expression following pathogen inoculation.

2. Each of the four most resistant cultivars showed expression changes that were unique to that cultivar, suggesting that each cultivar had some unique mechanisms of resistance.

3. A set of genes could be identified that showed expression changes common to all four resistant cultivars, suggesting some common mechanisms of resistance.

4. There were few significant differences in expression between three and five days of infection.

More details of these findings are summarized in Table 1. The identification of potentially different mechanisms of resistance in the four resistant cultivars was supported by the finding that cultivars Conrad and V71-370 contained Quantitative Trait Loci (QTLs) for resistance that mapped to different loci. These observations suggest strategies for improving quantitative resistance by hybridizing different resistant cultivars.

To further dissect mechanisms of quantitative resistance, we have recently completed the transcriptional profiling of two selected resistant and two selected susceptible cultivars over a detailed time course of infection. The final objective will be to carry out transcriptional profiling of 300 recombinant inbred progeny lines derived from a cross of a resistant and a susceptible cultivar in order to identify and characterize the soybean genetic loci responsible for quantitative resistance.

Table 1. Contrast analysis of genes up-regulated in specific subsets of cultivars. R, most resistant; M, moderately resistant; S, susceptible. Only genes showing significant responses to cultivar x infection interactions were included in the contrast analysis.

Significantly up-regulated		**Number of genes**
In cultivar(s):	**But not in cultivar(s):**	
Individual comparisons to susceptible cultivar		
Athow (R)	Sloan (S)	4622
General (R)	Sloan (S)	4097
V71-370 (R)	Sloan (S)	3430
Conrad (R)	Sloan (S)	2201
Williams (M)	Sloan (S)	4301
PI291327 (M)	Sloan (S)	1126
Each resistant cultivar compared to the other resistant cultivars		
V71-370	Conrad, General, Athow	312
Conrad	V71-370, General or Athow	241
General	Conrad, V71-370, Athow	461
Athow	Conrad, General, V71-370	857
Most resistant cultivars compared as a group to less resistant cultivars		
Resistant: Conrad, V71-370 Athow & General	Susceptible: OX20-8, PI291327 or Williams	152

Acknowledgements

We wish to thank all members of the Tyler group, all our collaborators for their contributions to this work, and the many members of the oomycete molecular genetics community for their contributions to the annotation of the *Phytophthora* genome sequences. This work was supported by grants from the National Research Initiative of the United States Department of Agriculture (USDA) Cooperative State Research, Education and Extension Service, grant numbers 00-52100-9684, 2001-35319-14251, 2002-35600-12747, 2004-35600-15055, 2005-35604-15525, 05-JV-11272138-056 and 2005-35600-16370, and from the US National Science Foundation, grant numbers MCB-0242131, DBI-0211863, EF-0412213 and EF-0523736, and by funds from the US Department of Energy Joint Genome Institute and the Virginia Bioinformatics Institute.

References

Allen RL, Bittner-Eddy PD, Grenville-Briggs LJ, Meitz JC, Rehmany AP, Rose LE, Beynon, JL (2004) Host-parasite coevolutionary conflict between *Arabidopsis* and downy mildew. *Science*. **306**: 1957-1960.

Armstrong MR, Whisson SC, Pritchard L, Bos JI, Venter E, Avrova AO, Rehmany AP, Bohme U, Brooks K, Cherevach I, Hamlin N, White B, Fraser A, Lord A, Quail MA, Churcher C, Hall N, Berriman M, Huang S, Kamoun S, Beynon JL, Birch PR (2005) An ancestral oomycete locus contains late blight avirulence gene Avr3a, encoding a protein that is recognized in the host cytoplasm. *Proc. Natl. Acad. Sci. USA* **102**: 7766-7771.

Chang JH, Goel AK, Grant SR, Dangl JL (2004) Wake of the flood: ascribing functions to the wave of type III effector proteins of phytopathogenic bacteria. *Curr. Opin. Microbiol*. **7**: 11-18.

Erwin DC, Ribiero OK (1996) *Phytophthora Diseases Worldwide*. APS Press, St. Paul, Minnesota.

Hiller NL, Bhattacharjee S, van-Ooij C, Liolios K, Harrison T, Lopez-Estraño C, Haldar K (2004) A host-targeting signal in virulence proteins reveals a secretome in malarial infection. *Science* **306**: 1934-1937.

Marti M, Good RT, Rug M, Knuepfer E, Cowman AF (2004) Targeting malaria virulence and remodeling proteins to the host erythrocyte. *Science* **306**: 1930-1933.

Rehmany AP, Gordon A, Rose LE, Allen RL, Armstrong MR, Whisson SC, Kamoun S, Tyler BM, Birch PRJ, Beynon JL (2005) Differential recognition of highly divergent downy mildew avirulence gene alleles by *RPP1* genes from two *Arabidopsis* lines. *The Plant Cell*, **17**(6): 1839-1850.

Rizzo DM, Garbelotto M, Davidson JM, Slaughter GW, Koike ST (2002) *Phytophthora ramorum* as the cause of extensive mortality of *Quercus* spp. and *Lithocarpus densiflorus* in California. *Plant Dis*. **86**: 205-214.

Shan W, Cao M, Leung D, Tyler BM (2004) The *Avr1b* locus of *Phytophthora sojae* encodes an elicitor and a regulator required for avirulence on soybean plants carrying resistance gene *Rps1b*. *Mol. Plant Microbe Interact*. **17**: 394-403.

The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nat. Genet*. **25**: 25-29.

Tyler BM (2002) Molecular basis of recognition between *Phytophthora* species and their hosts. *Ann. Rev. Phytopathol*. **40**: 137-167.

Young ND (1996) QTL mapping and quantitative disease resistance in plants. *Ann. Rev. Phytopathol*. **34**: 479-501.

Publications

Jiang RHY, Tyler BM, Whisson SC, Hardham AR, Govers F (2006) Ancient origin of elicitin gene clusters in *Phytophthora* genomes. *Mol. Biol. Evol*. **23**(2): 338-351.

Tripathy S, Pandey VN, Fang B, Salas F, Tyler BM (2006) VMD: A community annotation database for microbial genomes. *Nucl. Acids Res*. **34**: D379–D381.

Tyler BM (2006) Genomics of fungal plant pathogens. In *Encyclopedia of Plant and Crop Science*, Goodman RM (ed), Marcel Dekker, New York, USA (in press).

Tyler BM (2006) *Phytophthora sojae*: root rot pathogen of soybean and model oomycete. *Mol. Plant Pathol*. (in press).

Tyler BM, Tripathy S, Zhang X, Dehal P, Jiang RHY, Aerts A, Arredondo F, Baxter L, Bensasson D, Beynon JL, Chapman J, Damasceno CMB, Dickerman A, Dorrance AE, Dou D, Dubchak I, Garbelotto M, Gijzen M, Gordon S, Govers F, Grunwald N, Huang W, Ivors K, Jones RW, Kamoun S, Krampis K, Lamour K, Lee MK, McDonald WH, Medina M, Meijer HJG, Nordberg E, Maclean DJ, Ospina-Giraldo MD, Morris P, Phuntumart V, Putnam N, Rash S, Rose JKC, Sakihama Y, Salamov A, Savidor A, Scheuring C, Smith B, Sobral BWS, Terry A, Torto-Alalibo T, Win J, Xu Z, Zhang H, Grigoriev I, Rokhsar D, Boore J (2006) *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science* (in press).

2006 Research Reports

from the Virginia Bioinformatics Institute's Faculty Fellows

Data-driven computational systems biology

T. M. Murali, murali@cs.vt.edu
Assistant Professor, Computer Science, Virginia Tech

Our research program tackles two fundamental challenges in the field of systems biology: (i) computing the building blocks of cellular networks and investigating how they relate to each other and (ii) predicting the biological roles of those genes in sequenced genomes that have unknown or poorly understood functions. Data-driven methods underlie our solutions to both problems. We have developed methods to unearth networks of interactions activated in response to a particular stimulus (ActiveNetworks), place the active network for one stimulus in the context of response networks for other stimuli (NetworkLego), compute modules of protein interactions conserved across different organisms (GraphHopper), study combinatorial transcriptional control by integrating occurrences of known transcription factor-binding motifs with gene expression data (XcisClique), and build biologically-interpretable classifiers for diseases (xMotif). Our GAIN system (Gene Annotation using Integrated Networks) predicts gene functions by integrating gene expression data with molecular interaction networks. The VIRGO server allows biologists to upload gene expression data and use GAIN to obtain gene function predictions tuned to this data. We have compared predictions made by GAIN for *Saccharomyces cerevisiae* and *Homo sapiens* based on Gene Ontology (GO) annotations in March 2005 with annotations in March 2006 and determined that more than 50% of predictions from GAIN are correct and well-correlated with confidence values that GAIN estimates for each prediction. We have incorporated a new method for generating negative examples into GAIN, making it the first known algorithm to provably make predictions that are consistent with the true path rule of GO.

Keywords: comparative systems biology; gene function prediction; functional linkage networks; network legos; GraphHopper.

Introduction

Rapid advances in high-throughput and large-scale experiments are providing us with breathtaking new insights into cellular machinery and processes. The public availability of multiple types of data about genes, proteins, and other molecules is inspiring the study of properties of groups of molecules that act in concert (Hartwell et al, 1999) and laying the basis for advances in systems biology. Our research program tackles two fundamental challenges in systems biology: (i) computing the building blocks of cellular networks by integrating multiple types of data and discovering patterns of coordinated activity contained in these data sets and (ii) predicting the biological roles of those genes in sequenced genomes that have unknown or poorly understood functions. Data-driven methods based on graph theory, discrete algorithms, data mining, and machine learning underlie our solutions to both problems.

Building blocks of cellular networks. We have developed several techniques that integrate different types of data to find building blocks of biological circuits. The ActiveNetworks approach operates by overlaying gene expression data for a given stimulus on molecular interaction networks to obtain networks of interactions activated in response to that stimulus. The NetworkLego method compares and contrasts these active networks for different stimuli in order to find out similarities and differences between the response of the cell

Contributors:
David Badger, Gregory Grothaus, Naveed Massjouni, Adeel Mufti, Corban Rivera

to these stimuli. The GraphHopper algorithm compares protein-protein interactions for different organisms to find protein interaction modules that are conserved across multiple species. XcisClique (Pati et al, 2006) is a system to study combinatorial transcriptional control by integrating occurrences of known transcription factor-binding motifs with gene expression data. Finally, xMotif is a method that utilises biclusters in gene expression data to build an accurate nearest-neighbor classifier for identifying tissues and diseases. xMotif's classifiers are biologically interpretable by design.

Automatic prediction of gene function. Our GAIN system predicts gene functions by integrating gene expression data with molecular interaction networks. In the last 12 months, we have extended GAIN in three principal directions. First, the VIRGO server (Massjouni et al, 2006) allows a biologist to upload gene expression data and use GAIN to obtain gene function predictions tuned to this data. Second, we have comprehensively compared predictions made by GAIN for *Saccharomyces cerevisiae* and *Homo sapiens* based on Gene Ontology (GO) (Ashburner et al, 2000) annotations in March 2005 with annotations in March 2006. Our analysis demonstrates that more than 50% of predictions from GAIN are correct and well-correlated with confidence values that GAIN estimates for each prediction. Third, we have identified a flaw in the way in which function prediction algorithms generate negative examples. Our solutions make GAIN the first known algorithm to provably make predictions that are consistent with the true path rule of GO.

Building blocks of cellular networks

ActiveNetworks and NetworkLego. Publicly-available data sets provide detailed and large-scale information on diverse types of molecular interaction networks in a number of model organisms (Bork et al, 2004). These multi-modal universal networks capture a static view of cellular state. Consider a biologist who wants to study a specific system of interest, e.g., oxidative stress. This biologist may want to determine which subset of interactions in these universal networks is activated in response to oxidative stress. Further, the biologist may desire to understand which parts of the response network are also activated upon other stimuli and which parts are unique to oxidative stress, and thus warrant further study.

We have developed two new approaches called ActiveNetworks (Murali et al, Analysing Stimulus Specific Responses using ActiveNetworks, unpublished results) and NetworkLego (Rivera & Murali, Automatically Assembling the Building Blocks of Cellular Circuitry, unpublished results) that provide a comprehensive solution to these questions. The ActiveNetworks approach overlays gene expression data for a given stimulus on molecular interaction networks and computes dense subgraphs in the resulting weighted graph to obtain networks of interactions activated in response to that stimulus. The NetworkLego system compares and contrasts all computed active networks using a generalisation of hierarchical clustering to graphs. Each lego computed is a sub-network of coherently activated molecular interactions that corresponds to a set-theoretic formula whose elements are the stimuli. Thus, we can directly link legos to stimuli where all the interactions in the lego are activated and stimuli where none of the interactions in the lego are activated. We have applied our method to a diverse collection of data sets in *S. cerevisiae* and *H. sapiens*. We have discovered numerous network legos that are statistically significant, functionally enriched, and multi-modal.

Conserved protein interaction modules. Protein interaction networks (PINs) containing tens of thousands of interactions are now available for a number of organisms (Bork et al, 2004). We study the detection of conserved protein interactions modules (CPIMs) in the PINs of two organisms. CPIMs are small groups of interacting proteins in each organism that share a high degree of evolutionary conservation. Proteins in a CPIM are likely to have co-evolved and often perform the same function within a cell. We have developed a novel algorithm called GraphHopper (Rivera & Murali, unpublished results) that analyzes two PINs to find CPIMs. GraphHopper connects the two PINs using orthologous pairs of proteins. Starting from multiple "basis" CPIMs, GraphHopper hops from one PIN to another using orthology relationships. By keeping the two PINs separate

and by using new measures for computing the quality of a CPIM, GraphHopper is successful in finding CPIMs with a wide variety of topologies, including paths and complexes. We have applied GraphHopper to human, fly, and yeast PINs. We found a number of CPIMs conserved in all three species, enriched in fundamental processes of the cell, such as DNA replication, protein folding, and kinase activity. In addition, we found a number of CPIMs related to development and the nervous system that emerged only in the comparison between the human and fly networks. Finally, we have suggested how CPIMs can serve as the basis for judging the quality of electronically-generated functional annotations. Results obtained from our analyses are available at bioinformatics.cs.vt.edu/~cgrivera/gh.

Combinatorial control of transcription. The transcriptional regulation of a gene is an intricate, dynamic phenomenon that is often controlled by multiple transcription factors that recognize and bind to specific DNA sequences in the promoter of a gene. In the XcisClique project (Pati et al, 2006), we have assembled a collection of 276 known transcription factor-binding motifs in *Arabidopsis thaliana* and identified the occurrence of each motif in the *A. thaliana* genome. Given a list of query genes, XcisClique uses the well-known Apriori data mining algorithm (Agarwal & Srikant, 1994) to compute motifs sets, which are subsets of query genes and subsets of transcription factor-binding motifs such that each transcription factor-binding motif occurs in the promoter of each gene. The genes in a motif are potentially regulated in combination by the TFs that bind to the motifs. To check this possibility, XcisClique computes the correlation of the gene expression profiles of these genes across nine abiotic stresses in the Nottingham Arabidopsis Stock Centre (NASC) database and assesses the statistical significance of the observed correlation. XcisClique identifies new motif and gene combinations that might indicate as yet unidentified involvement of sets of genes in biological functions and processes. The system is available at bioinformatics.cs.vt.edu/xcisclique/.

Biologically-interpretable disease classification. Classification of tissues and diseases based on gene expression data is a powerful application of DNA microarrays. Many popular classifiers like Support Vector Machines, nearest-neighbour methods, and boosting (Ben-Dor et al, 2000) have been applied successfully to this problem. However, it is difficult to determine from these classifiers which genes are responsible for the distinctions between the diseases. We have developed a novel framework for classification of diseases and tissues based on the notion of condition-specific clusters of co-expressed genes called xMotifs (Grothaus & Murali, Biologically-Interpretable Disease Classification, unpublished results; Murali & Kasif, 2003). xMotifs are biologically interpretable since (i) the genes in an xMotif and the constraints on their expression values can be directly read out from a description of the xMotif and (ii) the xMotifs we compute are enriched in biological functions relevant to the diseases we study. Our xMotif-based classifier is a nearest neighbor classifier with a twist: each point in the classifier is an xMotif and its dimensionality may be different from the other points in the classifier. Our classifier achieves high accuracy comparable to that of Support Vector Machines on leave-one-out cross-validation on both two-class and multi-class data. Details of the xMotifs we compute, their functional enrichments, and the xMotif software are available at bioinformatics.cs.vt.edu/~murali/papers/xmotif-classifier.

Automatic prediction of gene functions

More than 250 complete genome sequences are now available, including those of 35 eukaryotes (Bernal et al, 2001). However, a fundamental roadblock to progress in systems biology is the poor state of knowledge about the biological functions of the genes in sequenced genomes (Enright et al, 2003). Using sequence similarity to predict gene function provides annotations only for about 40% of eukaryotic genes. A promising basis for predicting gene function identifies associations between pairs of genes that may perform the same or similar function in the cell. For instance, two genes may have the same function if their protein products interact (Bork et al, 2004) or if they have very similar patterns of gene expression (Bergmann et al, 2003; Stuart et al, 2003). A functional linkage network (FLN) (Karaoz et al, 2004; Lee et al, 2004; Marcotte et al, 1999) is a powerful

framework for representing and analysing such relationships. An FLN is a graph in which each node corresponds to a gene; the node is labeled by the set of functions that annotate the gene. An edge in an FLN connects two genes if some experimental or computational procedure suggests that these genes might share the same function.

We have developed a computational technique called Gene Annotation using Integrated Networks (GAIN) (Karaoz et al, 2004) that provides a comprehensive and robust mechanism for systematically transferring functional annotations from annotated genes to genes with unknown function across the entire FLN and for measuring the reliability of the resulting functional predictions. In its current form, GAIN constructs an FLN for a single organism by integrating functional genomic information such as gene expression data, protein-protein interactions, and protein-DNA binding data. GAIN includes a local search algorithm for systematically propagating annotations through the entire FLN. FLNs have the attractive property that they directly model functional associations between genes. Subgraphs of an FLN often correspond to biological pathways and networks. Thus, biologists can easily interpret functional predictions that are based on FLNs. In previous research, we have applied this technique to obtain several high-quality annotations in *S. cerevisiae* (Karaoz et al, 2004).

The VIRGO web server. We have developed a powerful interface called the VIRtual Gene Ontology (VIRGO) (Massjouni et al, 2006) that enables a biologist to (i) integrate gene expression data collected in the laboratory with molecular interaction networks, (ii) construct a functional linkage network (FLN) from these data sets, (iii) label the genes in the FLN with functional annotations from the Gene Ontology (GO), and (iv) systematically propagate these labels across the FLN in order to predict the functions of unlabelled genes. The biologist can query VIRGO for predictions of interest and prioritize them using confidence values assigned by VIRGO. VIRGO also provides informative propagation diagrams that trace the flow of information in the FLN. These diagrams may assist the biologist in ascertaining the rationale behind a prediction. As far as we are aware, VIRGO is the first web-server that makes an FLN-based prediction engine widely available. VIRGO is available at whipple.cs.vt.edu:8080/virgo.

Evaluation of predictions made by GAIN. In Fall 2005, graduate students in two class projects performed extensive analysis of the predictions made by GAIN. One group studied predictions for *S. cerevisiae* while the second group considered predictions for *H. sapiens*. For each organism, we collected more than 20 gene expression datasets representing a wide range of cellular states and stimuli. We integrated each gene expression dataset with a collection of protein-protein interactions and used the resulting FLN to predict gene functions. We based the predictions on GO annotations as of March 2005. We performed an exhaustive comparison of these predictions with the annotations as of March 2006, considering predictions only for those genes that had new annotations in 2006. We considered a prediction (a gene-function pair) to be correct if the gene was annotated with a function more specific than the predicted function in 2006. We found that GAIN made correct predictions for over half of the 299 human genes that had new annotations. Moreover, the fraction of correct predictions had a strong positive correlation with the confidence value estimated by GAIN with each prediction. We observed similar results for *S. cerevisiae*. All these predictions are now available in VIRGO. To our knowledge, these are the first large-scale evaluations of a function prediction algorithm performed by comparing predictions to the latest annotation databases. Papers describing these results are in preparation.

Hierarchically-consistent prediction of gene functions. The Gene Ontology (GO) has emerged as a standard for describing the functions of gene products. The power of GO comes from the fact that it allows us to specify the function of a gene at a number of different levels of detail. The true path rule of GO specifies that if a gene is annotated with a function, then the gene must be annotated with all ancestors of that function. However, the functional annotations predicted by almost all known algorithms do not obey the true path rule, primarily because these methods introduce inconsistencies when they artificially generate negative examples (gene-function pairs where the gene does not have the function).

Generating negative examples must be done carefully since the parent-child relationships in GO form a Directed Acyclic Graph (DAG) and not a tree. We have developed an FLN-based algorithm that provably makes predictions that follow the true path rule of GO (Murali, 2006). Our approach works by modifying the approach commonly used to generate negative examples. We prove that if we process the genes using the same permutation for all the functions, our Hopfield-network-based algorithm (Karaoz et al, 2004; Hopfield & Tank, 1986) will provably make hierarchically-consistent predictions of gene functions. We exploit this property to speed up the analysis of all functions in GO by traversing the GO DAG in topological order from root to leaf and propagating prediction results from parent to child. We also assign confidence estimates to our predictions by repeating our analysis with different gene permutations and counting the fraction of permutations for which we predict that a gene should be annotated with a particular function. Our modified method for generating negative examples can be used by other prediction engines.

Prediction of gene functions without negative examples. As described above, many machine-learning algorithms used for predicting gene function artificially add negative examples. In collaboration with Dr. Allan Dickerman and Dr. Brett Tyler of the Virginia Bioinformatics Institute (VBI) and in collaboration with Dr. Madhav Marathe, Dr. Henning Mortveit, and Dr. Anil Vullikanti of the Network Dynamics and Simulation Science Laboratory, we are developing methods that predict gene function based solely on positive examples. A basic challenge we must overcome is preventing positive examples from overrunning the entire FLN and predicting every gene as being annotated with every function. With Dickerman and Tyler, we are focusing on biological applications of the new methods such as simultaneously annotating the genomes of a number of sequenced microbes of interest and delineating oxidative stress response pathways in model eukaryotes such as *A. thaliana*, *S. cerevisiae*, and *Phytophthora sojae*. With Marathe, Mortveit, and Vullikanti, we are focusing on the mathematical and algorithmic aspects of the problem.

Acknowledgments

The research on ActiveNetworks is in collaboration with Richard Helm, Malcolm Potts, and Naren Ramakrishnan. We developed XcisClique in collaboration with Ruth Grene and Lenwood Heath. We thank Miguel Colon-Velez, Konstantinos Krampis, Harsha Rajasimha, and Lachelle Waller for their analysis of the *S. cerevisiae* gene expression datasets and Pallavi Sharma and Andrew Warren for their analysis of the human gene expression datasets.

References

Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In *Proceedings of 20th International Conference on Very Large Data Bases*, Bocca JB, Jarke M, Zaniolo C, Morgan Kaufmann (eds) pp. 487–499.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. *Nature Genet.* **25**(1): 25-29.

Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z (2000) Tissue classification with gene expression profiles. *J. Comput. Biol.* **7**(3-4): 559-583.

Bergmann S, Ihmels J, Barkai N (2003) Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol.* **2**(1): E9.

Bernal A, Ear U, Kyrpides N (2001) Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res.* **29**(1): 126-127.

Bork P, Jensen L, von Mering C, Ramani A, Lee I, Marcotte E (2004) Protein interaction networks from yeast to human. *Curr. Opin. Struct. Biol.* **14**(3): 292-299.

Enright A, Kunin V, Ouzounis C (2003) Protein families and TRIBES in genome sequence space. *Nucleic Acids Res.* **31**(15): 4632-4638.

Hartwell L, Hopfield J, Leibler S, Murray A (1999) From molecular to modular cell biology. *Nature* **402**(6761 Supplement): C47-C52.

Higo K, Ugawa Y, Iwamoto M, Korenaga T (1999) Plant *cis*-acting regulatory DNA elements (PLACE) database. *Nucleic Acids Res.* **27**: 297-300.

Hopfield J, Tank D (1986) Computing with neural circuits: a model. *Science* **233**(4764): 625-633.

Karaoz U, Murali TM, Letovsky S, Zheng Y, Ding C, Cantor CR, Kasif S (2004) Whole genome annotation using evidence integration in functional linkage networks. *Proc. Natl. Acad. Sci.* **101**: 2888-2893.

Lee I, Date S, Adai A, Marcotte E (2004) A probabilistic functional network of yeast genes. *Science* **306**(5701): 1555-1558.

Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D (1999) A combined algorithm for genome-wide prediction of protein function. *Nature* **402**(6757): 83-86.

Massjouni N, Rivera CG, Murali TM (2006) VIRGO: Computational Prediction of Gene Functions. *Nucleic Acids Research*, Web server issue, **34**: W340-W344.

Murali TM, Kasif S (2003) Extracting conserved gene expression motifs from gene expression data. In *Proceedings of the Pacific Symposium on Biocomputing*. vol. 8, 77-88.

Pati A, Vasquez-Robinet C, Heath LS, Grene R, Murali TM (2006) XcisClique: Analysis of regulatory bicliques in *Arabidopsis thaliana*. *BMC Bioinformatics* **7**: 218.

Stuart JM, Segal E, Koller D, Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**(5643): 249-255.

Publications

Grothaus G, Mufti A, Murali TM (2006) Automatic layout and visualisation of biclusters. *Proceedings of the 6th SIGKDD Workshop on Data Mining in Biology* (in press).

Li P, Sioson A, Mane S, Ulanov A, Grothaus G, Heath L, Murali TM, Bohnert H, Grene R (2006) Response diversity of *Arabidopsis thaliana* ecotypes in elevated CO_2 in the field. *Plant Mol. Biol.* (in press).

Massjouni N, Rivera CG, Murali TM (2006) VIRGO: Computational prediction of gene functions. *Nucleic Acids Res.* Web server issue, **34**: W340-W344.

Murali TM (2006) Hierarchically-consistent prediction of gene functions. *Proceedings of the Second Automated Function Prediction Meeting* (in press).

Pati A, Vasquez-Robinet C, Heath LS, Grene R, Murali TM (2006) XcisClique: Analysis of regulatory bicliques in *Arabidopsis thaliana*. *BMC Bioinformatics* **7**: 218.

Simulation and analysis of molecular regulatory systems in cell biology

John J. Tyson, tyson@vt.edu
University Distinguished Professor, Biological Sciences, Virginia Tech

Complex networks of interacting proteins control the physiological properties of a cell (metabolism, reproduction, motility, signaling, etc.). Intuitive reasoning about these networks is often sufficient to guide the next experiment, and a cartoon drawing of a network can be useful in codifying the results of hundreds of observations. But what tools are available for understanding the rich dynamical repertoire of such control systems? Using basic principles of biochemical kinetics, we convert network diagrams into dynamical models (ordinary differential equations, stochastic processes, or Boolean switching networks) and then explore the models by analytical and computational methods. Of particular interest to us are the mechanisms that control cell division in prokaryotes and eukaryotes, ranging from beneficial and parasitic bacteria to cancerous tumors.

Keywords: network dynamics; cell division cycle; bifurcation analysis; stochastic modeling; parameter estimation.

Network dynamics and cell physiology

The fundamental goal of molecular cell biology is to understand how the information encoded in the genome is used to direct the complex repertoire of physiological responses of the cell to its environment, in order to keep the cell alive and to propagate its genome to a new generation. At one end of this continuum, nucleotide sequences direct the synthesis of polypeptide chains, which then fold into three-dimensional structures that function as enzymes, motors, and channels. At the other end, complex assemblages of interacting proteins carry out the fundamental chores of life: energy metabolism, biosynthesis, signaling, movement, differentiation, and reproduction. The triumph of molecular biology in the last half of the twentieth century was to identify and characterize the molecular components of this machine, epitomized by the complete sequencing of the human genome. The grand challenge of post-genomic cell biology is to assemble these pieces into a working model of a living cell — a model that gives a reliable account of how the physiological properties of a cell derive from its underlying molecular machinery.

The Tyson research group is deeply involved in meeting this challenge in several areas: cell cycle control in prokaryotes and eukaryotes, signal transduction in mammalian cells, and circadian rhythms in cyanobacteria, fungi and animals. In addition to modeling, the group works closely with experimentalists (Dr. Jill Sible at Virginia Tech and Dr. Fred Cross at Rockefeller University) who are testing model predictions, and with computer scientists at Virginia Tech who are building next-generation software tools for simulation and analysis. This report summarizes research in the Tyson group that is closely allied to work at the Virginia Bioinformatics Institute (VBI).

Cell division cycle in *Caulobacter crescentus*

Caulobacter crescentus is a dimorphic bacterium inhabiting freshwater, seawater and soils, where it plays an important role in global carbon cycling by mineralizing dissolved organic

Collaborators:
Paul Brazhnik, Shenghua Li, Tongli Zhang, Rajat Singhania, Mohsen Sabouri

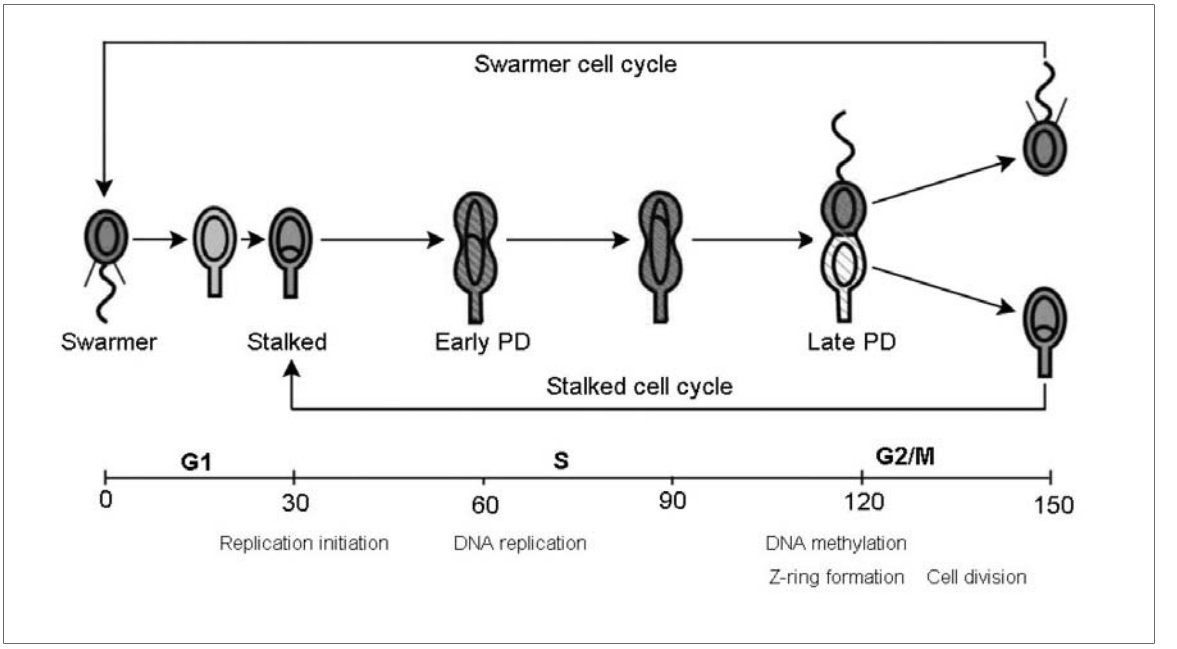


Fig.1. Physiology of the cell division cycle in *Caulobacter*. Three cell division cycle phases can be distinguished in a swarmer cell of *Caulobacter crescentus*: the growth and differentiation phase (G1) lasts about 30 min, DNA synthesis (S) phase takes about 90 min, and the ~30-min G2/M phase culminates in the separation of mother and daughter cells. The stalked cell cycle lacks the G1 phase. Proteins such as GcrA, CtrA, DnaA are expressed at different stages of the cell division cycle. The θ-like structure denotes replicating DNA during the cell cycle.

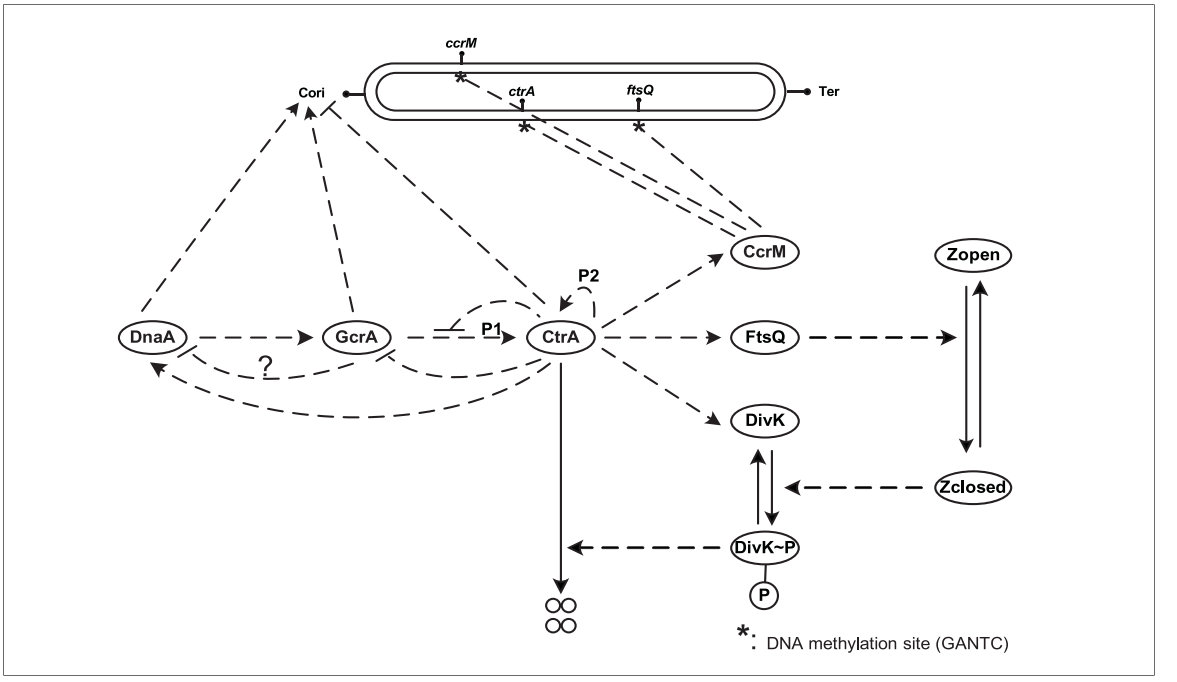


Fig. 2. Wiring diagram of the cell-cycle control model for *Caulobacter*. Solid lines represent chemical reactions. Dashed lines indicate effects on protein production or DNA methylation. All proteins are assumed to be produced and degraded with specific rates. Degradation that is not subject to regulation is not shown. The double-stranded closed curve at the top represents DNA. Cori is the origin site for DNA replication and Ter is the termination site.

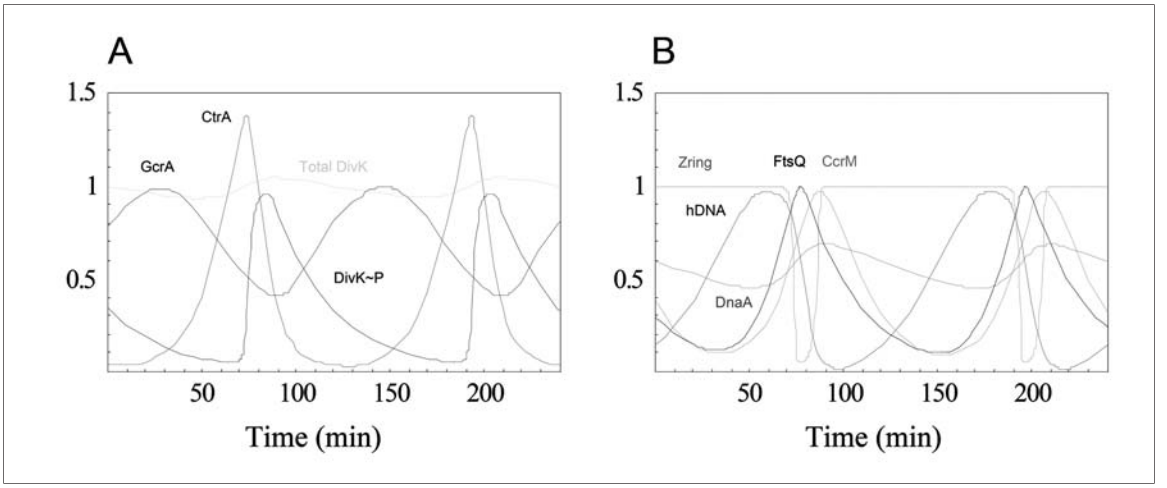


Fig.3. Modeling the cell division cycle in *Caulobacter*. (A, B) Change of protein concentrations during the *Caulobacter* cell division cycle as simulated by the model (Brazhnik & Tyson, 2006).

material. *Caulobacter* normally undergoes an asymmetric cell division cycle, producing two different progeny (Fig.1): a motile swarmer cell with a flagellum and a sessile stalked cell. The two cell types undergo different developmental programs. The nascent stalked cell immediately enters into a new round of cell division and produces, about 90–120 min later, another swarmer cell. The nascent swarmer cell swims around for 30–45 min before it differentiates into a stalked cell and initiates division.

A molecular mechanism controlling the cell division cycle in *Caulobacter* has been proposed using recent experimental evidence (Fig.2). Following standard rules of chemical kinetics, the mechanism is converted into a set of rate equations describing the dynamics of the model at the protein level. The justification of our approach has been described in detail (Brazhnik & Tyson, 2006). The model accounts for important details of the physiology, biochemistry and genetics of cell cycle control in *Caulobacter*, reproducing available protein time courses in wild-type cells (Fig.3) and mimicking correctly the phenotypes of a number of mutant strains. Since many of the proteins involved in regulating the cell cycle of *Caulobacter* are conserved among the family of α-proteobacteria, the mechanism we propose may be generic for the whole family.

In particular, we plan to extend the model to proliferation and differentiation of *Sinorhizobium meliloti*, an α-proteobacterium that lives endosymbiotically in the root nodules of legumes, where it fixes nitrogen for the plants. This project is being carried out in collaboration with Dr. Bruno Sobral's research group at VBI, which provides experimental expertise with *Sinorhizobium*. The class of α-proteobacteria also includes serious pathogens, such as *Brucella* and *Rickettsia*.

Growth, division and death in mammalian cells

In multicellular animals, such as humans, cell proliferation and differentiation are strictly regulated by the body to serve the needs of specific tissues and organs. To create a supply of differentiating cells to meet the regenerative demands of the body is the job of stem cells, which respond to positive and negative feedback signals from local tissues and from the whole organism. Cell death is also actively regulated by the body to achieve suitable sizes of tissues and organs. Much is known about the general molecular mechanisms by which animals control cell growth and division, gene expression and differentiation, and programmed cell death. These processes play predominant roles in wound healing, tissue and organ regeneration, tumorigenesis, and tissue engineering. Better understanding and control (intervention) of

cell proliferation will have huge impact on the health industry (e.g., surgery, pharmaceuticals, cancer), the defense establishment (e.g., fighter performance and repair), and the biotechnology industry (e.g., artificial tissues and organs).

The Tyson group is contributing to this enterprise by building realistic, accurate, and reliable mathematical models of the molecular mechanisms controlling cell replication, cell signaling, and cell responses (in terms of altered gene expression and, when appropriate, cell death). Such mathematical models will be necessary for a thorough "systems approach" to cell dynamics in the human body, and few research groups in the world have the experience and talents to build, analyze, simulate and interpret mathematical models that are truly useful to the biomedical research community.

Present efforts focus on the cyclin-dependent kinase network controlling DNA synthesis and mitosis, the major tumor-suppressor genes (the retinoblastoma protein, Rb, and the DNA-damage response protein, p53), the major signal transduction pathways (Mitogen Activated Protein (MAP) kinase, NF-κB, integrins, cadherins, and Smads), and the caspase network that carries out the cell suicide program. As for the *Caulobacter* cell cycle, the networks are modeled first by nonlinear ordinary differential equations. The models are developed in modular fashion and then assembled into a comprehensive and coherent, computable account of the integrated system. The goal is to understand (i.e. to compute accurately) the decision-making characteristics of individual stem cells. Ordinary Differential Equation (ODE) models of these networks are being supplemented by two other approaches: probabilistic Boolean networks, and hybrid stochastic-deterministic models suitable for modeling the responses of individual cells.

Probabilistic Boolean networks

In a Boolean network, the state variables, $X_i(t)$, take on only two values, 0 or 1, and the index t is an integer sequence (0, 1, 2, …) representing progression in time. Updating the state variables is done by rules, $X_i(t+1) = F_i(X_1(t), X_2(t), \ldots, X_N(t))$, where each F_i is a Boolean function of its arguments. Boolean networks have a long history in theoretical biology, dating back to the seminal work of Kauffman (1969). A few years ago, Tang and colleagues (Li et al, 2004) published a Boolean-network version of the Chen–Novak–Tyson model (Chen et al, 2000) of the eukaryotic cell cycle. Although appealingly simple and suggestive, Tang's model is ultimately unproductive, because it cannot be compared in any meaningful way with experimental data. The Tyson group has improved on Tang's model in two ways. First, the Boolean representation of the protein interaction network has been supplemented by realistic rules for cell growth and division. Second, time-stepping has been replaced by a continuous random process in the time dimension. That is to say, updating is done by deterministic-probabilistic rules, $X_i(t+\tau) = F_i(X_1(t), X_2(t), \ldots, X_N(t))$, where the F_i's are the same deterministic Boolean functions as before, but now t is a real number (clock time) and τ is a random number drawn from an exponential probability density function, $e^{-p\tau}$. The transition probability, p, depends on the initial state $\{X_1(t), X_2(t), \ldots, X_N(t)\}$. In this formalism, we can model the sequence of events governed by the control system (by our choice of Boolean functions) and the timing of the state transitions (by our choice of transition probabilities). This approach is closely allied to the experimental measurements of Jacobberger (Darzynkiewicz et al, 2004), who uses fluorescent labeling of proteins X_i to follow proliferating mammalian cells through their state-space transitions. If this approach proves to be successful, then Tyson and Jacobberger will team up with Rober Jackson of Cyclacel Pharmaceuticals, Inc. (Dundee, Scotland) to apply the model to data on antiproliferative drug candidates.

Stochastic models of the cell cycle

Deterministic ODE models of the cell cycle capture the average behavior of populations of cells, but cannot describe the idiosyncrasies of cell cycle progression in single cells. The probabilistic Boolean model may do a better job capturing single-cell dynamics, but its representation of state transitions in terms of Boolean functions will eventually prove to be too restrictive. The gold-standard of single-cell modeling is a fully stochastic account of the biochemical reactions taking place within the cell. The Tyson group is well underway in

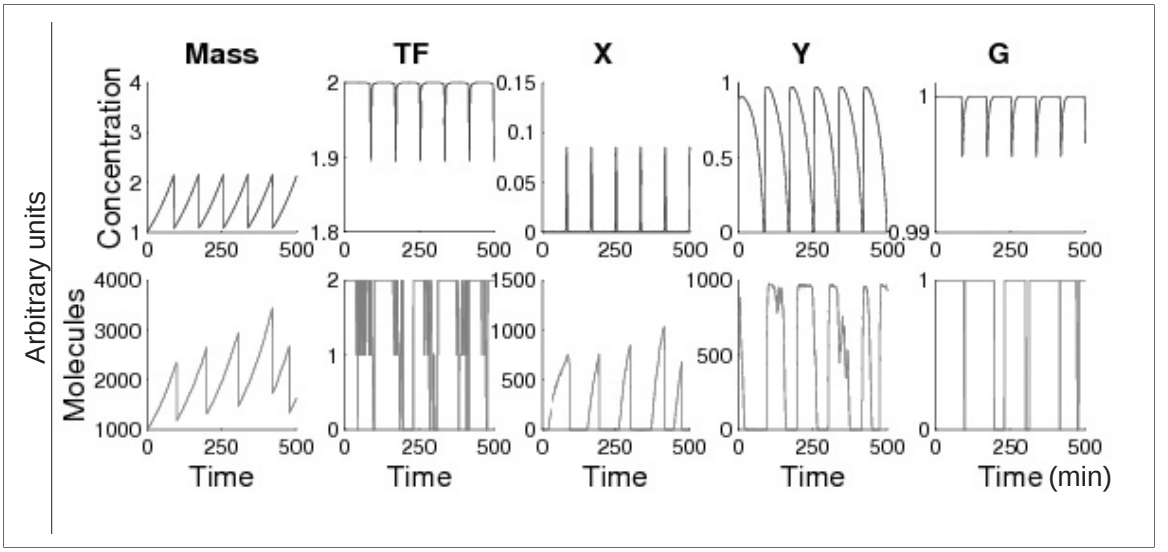


Fig. 4. Comparison of deterministic (top row) and stochastic (bottom row) simulations of a model of cell cycle progression in budding yeast. TF, X, Y and G refer to the active forms of components (proteins and a gene) in the model.

translating the phenomenological ODE models into elementary reaction networks suitable for exact stochastic simulation. An example of such simulations is given in Fig.4. This research enterprise, in collaboration with the Virginia Tech College of Engineering, has recently been funded by the National Institute of General Medical Sciences.

Conclusions

The physiological characteristics of a cell are determined by networks of interacting proteins that process energy, material and information. Confined to a few femtoliters of cytoplasm, these processing and control systems are not only as complex as a Boeing 777, but also able to make exact replicas of themselves from CO_2, NO_3^-, PO_4^{3-}, and a drop of mineral water. We would like to know how these marvelous machines work, but they do not come with instruction manuals or schematic wiring diagrams. It is the grand challenge for post-genomic life scientists to deduce the diagrams and write the manuals. This effort will take a variety of resources and approaches: genetics and biochemistry, hardware and software, high-throughput and low-throughput technologies, hypothesis-driven and discovery-driven experiments, silicon-based and myelin-based reasoning. Tyson's research group is contributing to this enterprise by careful simulation and analysis of the kinetic properties of protein interaction networks.

References

Brazhnik B, Tyson JJ (2006) Cell cycle control in bacteria and yeast: a case of convergent evolution? *Cell Cycle* **5**: 522-529.

Chen KC, Csikasz-Nagy A, Gyroffy B, Val J, Novak B, Tyson JJ (2000) Kinetic analysis of a molecular model of the budding yeast cell cycle. *Mol. Biol. Cell* **11**: 369-391.

Darzynkiewicz Z, Crissman H, Jacobberger JW (2004) Cytometry of the cell cycle: cycling through history. *Cytometry* **528A**: 21-32.

Kauffman SA (1969) Homeostasis and differentiation in random genetic control networks. *Nature* **224**: 177-178.

Li F, Long T, Lu Y, Ouyang Q, Tang C (2004) The yeast cell-cycle network is robustly designed. *Proc. Natl. Acad. Sci. USA* **101**: 4781-4786.