Virginia Bioinformatics Institute
Washington Street, MC 0477
Bioinformatics Facility
Blacksburg, VA 24061

Scientific Annual Report • 2005

2005

Scientific Annual Report

Virginia Bioinformatics Institute

Virginia Bioinformatics Institute at Virginia Tech

Scientific Annual Report 2005

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

Research Reports — VBI Fellows

Introduction

Dear friends,

I am pleased to share with you the Virginia Bioinformatics Institute's (VBI) first scientific annual report. As we celebrate our fifth anniversary as a research institution, the faculty and staff at VBI continue to focus on innovative research efforts and seek answers to diverse questions that impact all aspects of life. This report highlights the latest accomplishments of our research faculty at VBI.

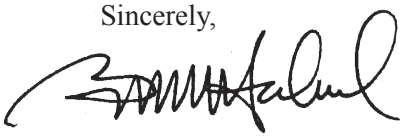
Our researchers collaborate to bring together diverse disciplines, such as mathematics, computer science, biology, plant pathology, biochemistry, statistics, and economics. These collaborations help link seemingly unrelated fields in an effort to develop and implement new innovations and discoveries in the areas of bioinformatics and systems biology.

Working in an interactive research environment, faculty at VBI utilize a variety of methods and tools, including simulation and modeling, statistical genetics, computational systems biology approaches, large-scale comparative genomics, microfluidic bioanalysis platforms, and wet chemistry experimentation. Using these applications, the research faculty are involved in a wide spectrum of work in an effort to advance infectious disease research and better understand large biological, information, social, and technological systems. Researchers focus on more specific processes, such as interactions in cells, advanced proteomics investigations, microbial genomics, metabolomics, functional interactions among pathogen and host genes, and the molecular basis for the onset and sustenance of infection.

Also included in this publication are reports from three of VBI's Virginia Tech College of Engineering (COE) fellows. These VT faculty members work closely with VBI researchers on projects involving multi-scale modeling, software/hardware engineering, microfluidics, and image and signal process modeling.

As the research highlighted in this report shows, VBI has continued its dedication to connecting the strong foundations of past scientific research with modern tools and technologies. It is our hope that this research will lead to breakthroughs in science and technology that will help create a better tomorrow.

Sincerely,



Dr. Bruno Sobral
Executive and Scientific Director
Virginia Bioinformatics Institute

Research Reports

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

2005
Research Reports
From the Faculty
at the
Virginia
Bioinformatics
Institute

Network Dynamics and Simulation Science Laboratory

Christopher L. Barrett
Research Professor, VBI
Professor of Computer Science, Virginia Tech
cbarrett@vbi.vt.edu

Karla Atkins, Richard Beckman, Keith Bisset, Stephen Eubank, U.S. Kumar, Achla Marathe, Madhav V. Marathe, Henning Mortveit, Paula Stretz

Laboratory Overview

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 simulation applications are underwritten by a theoretical program in discrete mathematics and theoretical computer science, and are sustained by more than a decade of experience with the interplay of research and application.

Biological, information, social, and technological systems consist of large numbers of interacting components that together produce a "global system" with properties that are the result of interactions among the representations of the local system elements. Examples of such global systems include urban and regional transportation systems, the United States and world-wide electrical power markets and grids, the Internet, peer-to-peer networks, ad hoc communication and computing systems, bio-signaling systems, ecologies, gene regulatory networks, the public health system, and contagious disease economics. The complicated interactions and interdependencies among the constituent biological, information, social, and technological systems are inherent because the individual components are networked, only interacting with a specified set of components within local time intervals. The interactions can be physical or a matter of convention, such as those imposed by law or social norms, and typically consist of one or more social, biological, or information networks interacting with underlying technological or physical networks. For many reasons, ranging from practical difficulty to the possibility of great harm, simulations are a uniquely capable medium in which representation and analysis can be performed.

A key feature of the lab's work is the scale and scope of the systems represented. Constructing large simulations of social and technological systems is challenging and novel, since, unlike physical systems, socio-technical systems are affected not only by physical laws, but also by human behavior, regulatory agencies, courts, government agencies, and private enterprises. Our interdependent systems simulation suite provides a controlled environment to represent interactions among socio-technical networks, such as extremely large, interdependent urban infrastructure systems consisting of millions of interacting agents. For example, our population and transportation simulation system can represent every individual and their network in an extended urban region, including areas spanning hundreds of square miles and municipalities, at a spatial resolution of meters and a temporal resolution of one second or less. There can be tens of millions of individuals each taking

roughly five trips every day. Metropolitan Chicago spans approximately 250 square miles with more than 400 municipalities and over ten million inhabitants, and runs on a cluster of approximately 100 nodes to thousands of nodes. The size, scope, and multiple time scales of these representations naturally motivate a high performance computing implementation and require new engineering design principles.

The mathematical primitive in the interaction-based setting is an iterated composition of local functions, whereas the traditional setting is grounded in recursion and grammatical rules of symbol substitution and rewrite. Moreover, the interaction-based setting emphasizes what is being computed by interacting systems, rather than how "hard" it is to compute a given procedure or class. Interaction-based computational systems are often more like operating systems than specific algorithms, as they maintain specified relationships between individual computational elements. They are infinite-state systems with programs and behavior that evolve in time as a result of interactions with other sub-systems, and provide the main source of increasing complexity of very large, ubiquitous systems.

Current Projects

We are currently pursuing projects in the following programmatic areas:

- Epidemiology and the spread of infectious diseases
- Social networks and associated social and population dynamics
- Integrated next generation telecommunication systems
- Internet economics and commodity markets
- Molecular, system, and ecological biology

Public Health/Contagious Diseases

The epidemiological simulation models the spread of disease in urban areas, allowing for the assessment of prevention, intervention, and response strategies by simulating the daily movements of individuals within an urban region. The individuals are synthetic—they do not represent specific people—but as a group are statistically indistinguishable from the actual census. The locations visited by individuals are real street addresses that reflect actual land-use patterns in a region. In conjunction with the population simulation model, mobility models represent behavioral reactions to an outbreak and official interventions.

The simulation assigns different effects to various people based on demographic characteristics, and associates a state of health with each individual. It provides detailed information about every simulated person and the significant events that happen to each person during the simulation—including infection, incapacitation, and treatment—along with a time stamp and current location. It can also produce a representation of the social network, i.e., the person-to-person contact patterns within the entire population, and a description of the outbreak path over the social network. The system additionally allows the user to introduce contamination at any location as an exogenous event in the simulation, and to specify in detail the effects of a pathogen on a specific person. By varying a few parameters, users can model many different diseases, allowing for efficient and comprehensive measuring of the structural properties of large social networks.

Transportation and Population Mobility

Large-scale, activity-based, microscopic simulations of transportation systems can provide an individual's characteristics, location, and activity, in addition to simulating traffic movements in very small time intervals within the system. Such simulations conceptually decompose the transportation planning task into many spatial and time scales, and capture the underlying social and technological interactions.

In the initial stage of regional simulation, large time scales are associated with land use and demographic distributions as characterizations of travelers. A synthetic population is endowed with demographics that match the joint distributions from sources such as census data.

Demographic information is then used to create activity information for travelers that consists of requirements for each traveler to be at certain locations at specified times, and includes information on available travel modes. Patterns derived from daily activity patterns of several thousand surveyed households are used as templates and associated with synthetic households with similar demographics. The activity locations are estimated by taking into account observed land use patterns, travel times, and dollar costs of transportation.

An intermediate time-scale method is used to satisfy the activity requirements and generates planned routes and trip chains. The module finds minimum cost paths through the transportation infrastructure that is consistent with the constraints on mode choice. An example constraint would be: "walk to a transit stop, take transit to work using no more than two transfers and no more than one bus." Finally, a very short time-scale is associated with the execution of trip plans throughout the network via a simulation that represents travelers moving through a very detailed map of the urban transportation network as cellular automata. The simulation resolves traffic down to one car length and seconds. It provides updated estimates of travel times, including the effects of traffic congestion, route finding, and location choosing algorithms, which in turn produces new plans. This feedback process iteratively continues until there is convergence into a steady state where the best path for each individual is determined in the context of all other individual's paths. Traffic patterns that result from this process imitate actual traffic.

Telecommunications

The telecommunication modeling environment is an end-to-end simulation medium for representing and analyzing complex interdependent telecommunication networks comprised of cellular networks, public switched telephone networks, Internet networks, and ad hoc mesh networks. The system is designed to be useful in multiple settings, including the design and analysis of large-scale hybrid and sensor networks, vulnerability and criticality analysis of integrated telecommunication systems, and the interaction of the telecommunication system with social, information, and other infrastructure systems. For example, the modeling system can provide information about the physical and protocol-level vulnerabilities of a telecommunication network, as well as simulate the consequences of a physical, worm, or virus attack on a network; any resulting changes in network usage; and the feasibility and effectiveness of response options.

The modeling system decomposes the telecommunication system into four basic time scales. Devices and individuals are placed throughout an urban region by the first modeling module. It generates the positions of transceivers at various times of the coarse simulation clock and allows transceivers to become idle for a period of time and to rejoin the network at a later time. The module also provides for new transceivers to join the network and existing transceivers to permanently leave the network. Wireline devices are placed at permanent locations. Each device, such as a phone or computer, is assigned data sessions that are consistent with its type, location, and users, and are statistically identical to actual sessions generated in an urban region of interest. A time-varying telecommunication network is then constructed that is dynamic with variable topology and corresponds to an intermediate time scale. Finally, at the finest time scale, packet data is moved over the dynamic network using simulation methods based on flow techniques and discrete dynamical systems. The data is stored succinctly using signal theoretic methods. Markov chain methods are used to regenerate statistically equivalent packet streams and an auxiliary module maintains construction, analysis, and regeneration of integrated telecommunication networks. The module synthesizes publicly available data with simulated population mobility information to construct a complete set of dynamic networks—wireline, wireless, ad hoc, and packet switched IP networks—produced by an interdependent telecommunication system.

Commodity Markets

Markets are sensitive indicators of infrastructure disruptions and are often used to gauge public mood and awareness in crisis situations. We have recently designed and constructed a detailed simulation-based analysis tool for simulating commodity markets, such as electricity markets, which are aimed at understanding and analyzing large commodity markets.

The system simulates market activities such as bidding, contracts, and pricing of individual market players, and is driven by dynamic demand profiles that reflect the changing needs of an urban population. It can be coupled with physical flow models for commodities that require physical clearing. It uses population dynamics and activity location data from population mobility simulations and ties the market simulations to the urban infrastructure. It is individual based and uses a bottom-up method for generating power consumption patterns that drive the market and the physical grid. Three main components form a coupled system:

1. The electrical power grid with associated elements, including generators, substations, transmission grids, and their related electrical characteristics;
2. A market consisting of market entities, including buyers, sellers, the power exchange (where electricity trades are carried out at various time and size scales), the independent system operator, and the market clearing rules and strategies; and
3. An activity-based individual power demand creator that yields spatio-temporal distribution of the power consumed.

Due to scaling requirements, such simulations have a parametric representation for buyers as well as sellers and allow for a number of realistic behavioral features that are typically omitted in classical economic literature due to mathematical intractability.

Molecular, System, and Ecological Biology

Representations of complex and evolving social and technological systems reflect the complicated interdependency between constituent individual elements and the system dynamics. Our approach captures the causes underlying socio-technical interactions and relates them to measurements of the dynamic properties of functioning socio-technical systems. These methods can easily be transferred to problems in systems biology.

Our approach of combining network and data analysis with simulation-based dynamic analysis is similar to simulation-based analysis for computational systems biology, including gene annotation. Our work on local search algorithms scheduling computations on parallel machines is closely related to the computational methods developed for understanding the evolution of single stranded RNA molecule based on neutral network theory. We are currently exploring these connections further.

Research

There are four interrelated areas of our research:

1. Theoretical foundations
2. Computational science
3. Scalable software design and development
4. Applications research and development

Mathematical and computational theory view simulations as discrete dynamical systems providing formal methods for design, specifications, and analysis of such simulations. The theoretical formation of our interaction-based approach to modeling large biological, information, social, and technological systems includes the following areas:

- Mathematics of complex interdependent dynamic networks,
- Mathematical and computational theory in a class of finite discrete dynamical systems called Sequential Dynamical Systems, and
- Design and analysis of efficient and exact sequential and distributed algorithms, and large scale efficient combinatorial optimization.

Sequential Dynamical Systems are comprised of local maps, the composition of which reflects causal relationships between individual agents abstracted as functions; the locality of the functions reflects limited interaction and knowledge of the entire system available to each agent.

Engineering principles allow us to specify, design, and analyze large system simulations and implement them on massively parallel architectures. The basis of our computational science and scalable software systems research consists of the following components:

- Efficient data manipulation, including synthesis, integration, storage, and regeneration;
- New methods for efficient modeling and simulation, including concepts of disaggregated normative agents, lightweight agent representation, and multi-scale modeling; and
- High performance computing oriented system design, development, and implementation.

Publications

Barrett C, Smith JP, and Eubank S (2005) If Smallpox Strikes Portland…, *Sci Am* **292**: 42–49

Balakrishnan H, Barrett C, Kumar A, Marathe M and Thite S (2004) The Distance 2-Matching Problem and Its Relationship to the MAC Layer Capacity of Ad Hoc Wireless Networks. In special issue of *IEEE J. Selected Areas in Communications* **22**: 1069–1079.

Atkins K, Barrett C, Homan C, Marathe A, Marathe M, and Thite S (2004) Agent Based Economic Analysis of Deregulated Electricity Markets. *Proc. 6th IAEE European Conference*, Zurich, Switzerland

Dickerman Group Activities
In Silico and *In Vitro* 2005

Allan W. Dickerman
Research Assistant Professor
dickerman@vt.edu

Johanna C. Craig, Elena Shulaeva, Yuying Tian, Eric K. Nordberg

Introduction

Genomic biology in the "post-genomic era" is flourishing through a continual reinterpretation of sequence data with the aid of new information streams and analysis methods. One approach that provides deeper exploitation of available genome sequences is the comparative method. We have developed a powerful phylogenomics analysis platform, which we call "GeneTrees," for performing large-scale comparative genomics. We expect our phylogenomics system to be a powerful tool for genome annotation as well as a wide range of evolutionary investigations.

Extrapolating gene functions requires an input of experimental data to avoid circularity. Our group has been involved in laboratory-based elucidation of gene functions in the model plant *Arabidopsis* using genome-scale mRNA profiling to discover genes uniquely involved in two developmental arenas: the differentiation of the vascular xylem and phloem tissues and the development of the embryo within the seed.

Additional smaller projects are underway that relate to both comparative and functional genomics. A collaboration with Drs. Bruno Sobral and João Setubal will apply principles of the GeneTrees system to the α-proteobacteria. Ph.D. student Eric Nordberg is working with Drs. Brett Tyler and T.M. Murali to systematically extrapolate gene function from known to unknown genes by homology and other clues. He presented a poster on this work at the TIGR Bacterial Genomes conference in April 2005. A project with Entomology Department Ph.D. candidate Marc Fisher has surveyed the taxonomic diversity of prokaryote symbiont within the guts of termites using 16S rDNA sequencing in our lab. His long-term goal is to perform microbial ecology studies on the roles of this community on termite biology, including behavior. Finally, we are collaborating with Dr. Vladimir Shulaev to provide informatics support for his large insertional mutagenesis screening for functional genomics of the wild strawberry, *Fragaria vesca*.

The GeneTrees Phylogenomics System

Summary

The GeneTrees project has the ambitious goal of providing detailed evolutionary models for all genomic components of fully sequenced genomes describing the pattern of common ancestry (homology) observed among sequences, with particular focus on distinctions between orthologous and paralogous relationships. Orthology is expected to track conserved function better than homology alone, while paralogy is expected to frequently be accompanied by functional divergence of genes. This connection between evolution and gene function is one of the primary uses of comparative genomics with the goal of inferring the functions of genes known by sequence alone by their relationship to genes of known function.

The GeneTrees database schema comprises a list of taxa (species or within-species variants such as strains), a list of sequences (protein or nucleotide), ideally representing the full genome of each included taxon, and a set of homology models. Each homology model consists of a multiple sequence alignment among subregions

of sequences, plus one or more phylogenetic trees inferred from the alignment using one or more algorithms. Our goal is to integrate visualization of phylogenetic models and their underlying sequence data in a system designed to aid comparative genomics, genome annotation, and other scientific applications. To this end, GeneTrees has been integrated into the growing ToolBus/PathPort suite of tools provided by VBI (Eckart and Sobral, 2003).

Database Contents and Implementation

The complex process of homology database construction has evolved in our group over several years. First, low-level pairwise homology statements are assembled using all-vs-all pairwise sequence comparison, typically using BLASTP. A specially written program mines pairwise alignments to find dense clusters with high regional overlap as candidate sets of mutually alignable subsequences. Our alignment procedure iterates traditional alignment with ClustalW (Thompson et al., 200X) or Muscle (Edgar, 2004) with redefining the subsequence endpoints using an HMM model (HMMer package, http://hmmer.wustl.edu). Phylogenetic trees are then constructed from the alignments by PAUP* (Swofford, 2002) using the maximum parsimony method. Each tree topology is then evaluated for a maximum likelihood score using PAML (Yang, 1997). Finally, all tree topologies, branch-lengths, and statistics are stored to the database.

GeneTrees contains multiple different databases, each targeting a specific set of taxa. The best-developed GeneTrees database is named 'Prokaryote2' and is devoted to the fully-sequenced and annotated prokaryote genomes. This contains 538276 proteins from 184 full genomes extracted from NCBI. Homology discovery found 19,095 multiple sequence alignments with at least 5 sequences (mean = 33.5), with 42 alignments having over 500 sequences. The mean length of these alignments was 268, with 26 spanning over 2,000 positions. Each alignment with 5 or more sequences has associated parsimony and likelihood trees.

A second GeneTrees database available through TB/PP is called Viral2 which represents 1321 fully sequenced and annotated viruses available at NCBI. From 34,866 viral proteins, we obtained a total of 1,840 alignments of 4 or more sequences. Other taxon-focused databases exist for *Arabidopsis*, the Rosaceae, a diverse set of plant EST assemblies, and one is being developed for eukaryote model genomes including human, mouse, fly, worm, yeast, etc.

We have begun a comparison of the Prokaryote2 homology models to the Pfam set of conserved domains (Bateman et al., 1999). We find that our alignments are longer than Pfam models (268 vs 233 alignment positions) and more numerous (19,095 vs 7,677 models). Searching a sample of our sequences with hmmpfam found 44% of Pfam models matched 73% of our models at an expectation threshold of 10^{-6} (Figure 1). Given the emphasis of Pfam on vertebrates and other eukaryotes, the large number of Pfam models that do not overlap our sets is not surprising. We are interested in further characterizing the set of prokaryote homology models that have no overlap with Pfam models as a unique contribution gene family documentation and resource for genome annotation.

XML for data exchange

An XML format called SequencePhylogenyML has been developed and described in the DTD standard to define the phylogenetic trees and their meta data. Phylogenetic trees are represented in the standard "New Hampshire eXtended" format (Felsenstein, 1993; Zmasek, 2000). The analysis information includes taxonomic information for protein sequences, sequence alignments, phylogenetic estimation, and statistical scores for the data on each tree. One emphasis of our XML format is to enable associating arbitrary numerical or string data with each tip, such as gene expression data or geographic provenance.

ToolBus Interface to GeneTrees

A ToolBus query interface allows searches of homology groups based on taxonomy and sequence description. Our tree visualization

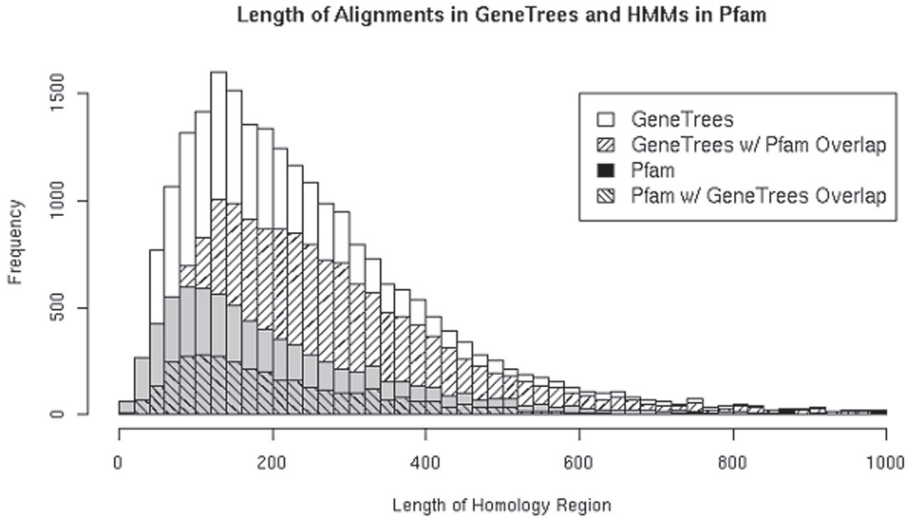


Figure 1. Comparison of Prokaryote2 versus Pfam homology models. Prokaryote2 models are longer and more numerous. The overlap regions indicates the proportion of Prokaryote2 homology groups and Pfam HMMs that matched regions in the other set at an expectation threshold of $E = 10^{-6}$.

software is a Java plug-in to TB/PP that allows simultaneous visualization of trees and the aligned sequences on which they are based (Figure 2). Tree tips are lined up horizontally with the sequence they represent. Tree tips can be labeled with a choice of species name, lineage information, gene description, or Genbank identifier. The tree can be interactively modified by re-rooting, node collapsing, node swapping, restoring to original shape, or zooming. These manipulations also affect sequences in the alignment. Re-ordering tree tips changes the order of sequences to match. If a subtree is collapsed to a single node, the affected sequences are summarized as a consensus sequence. Sequences letters are colored using a standard amino acid grouping (Devlin, 1992) or by user customization. One analysis option allows color-coding the branches of the tree according to the amino-acid states found at any specified column in the alignment. The system can also highlight all columns of the alignment that exhibit a substitution mapping to any selected branch on the tree.

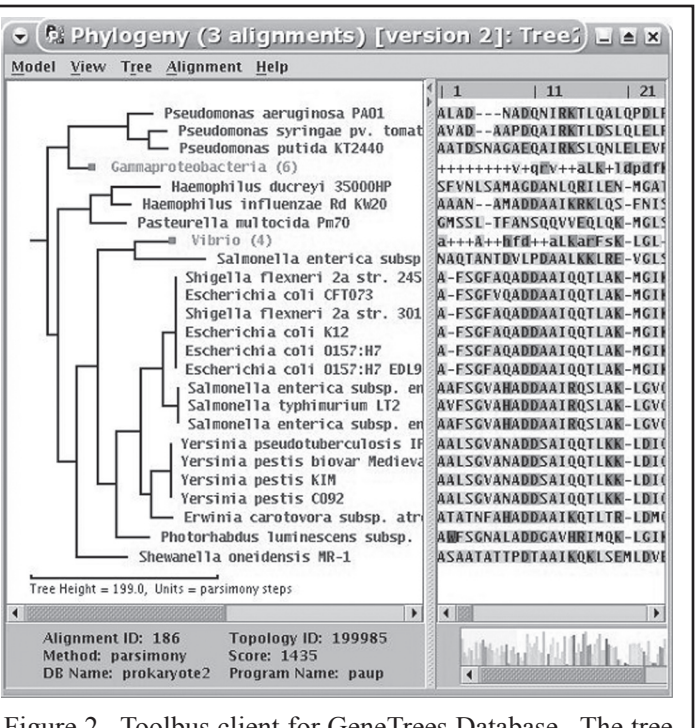


Figure 2. Toolbus client for GeneTrees Database. The tree tips, here labeled with species names, are displayed opposite correpsonding sequences in the alignment. The two nodes labeled Gammaproteobacteria and Vibrio are collapsed from 6 and 4 tips down to one and the corresponding sequences reduced to a consensus.

Alternative graphical interfaces to GeneTrees have been developed for internal research use, one as a Java application communicating directly to the database, another is a web application accessed through a standard browser. These prototypes are used to explore novel uses of GeneTrees, such as gene-tree to species-tree comparisons and mapping expression data to trees (see below).

Arabidopsis Functional Genomics

SeedGenes

The NSF-funded SeedGenes project (Tzafrir et al., 2003) maintains and evolves a website and database (www.seedgenes.org) devoted to public dissemination of all known embryonic lethal genes, those known to be essential for the life of the plant. The Principal Investigator is Dr. David Meinke of Oklahoma State University, a long-time expert on *Arabidopsis* gene nomenclature and embryo mutants. The website has been developed and is run by the Dickerman group, with data updates coming from the Meinke lab. A comparative study of essential genes in *Arabidopsis*, yeast, and worm was published recently (Tzafrir et al., 2004).

Seed and Embryo Transcriptomic

The *Arabidopsis thaliana* genome is represented on Affymetrix oligonucleotide microarrays which survey transcripts for over 22,500 genes. We are using these gene chips to ask questions about regulatory networks and molecular mechanisms of embryogenesis. This project was designed to identify embryo-specific genes that are inactive or weakly expressed in the adult tissues as candidate controlling genes in embryogenesis. We profiled gene expression in *Arabidopsis* seeds at young (globular and heart) and mature embryonic stages. We obtained comparative gene expression data from the Nottingham Arabidopsis Stock Centre web resource (www.arabidopsis.info). Multivariate analysis of this large combined data set of over 1,800 chips has yielded a number of genes with mRNA levels high in the embryos and early seed development relative to other tissues (Figure 3). Many of the embryo-specific genes encode proteins of unidentified function and may therefore be of particular interest as targets for further study using reverse genetics, promoter-reporters, or over-expression. A particularly interesting group encodes known DNA binding proteins that may regulate important developmental functions in *Arabidopsis*, by

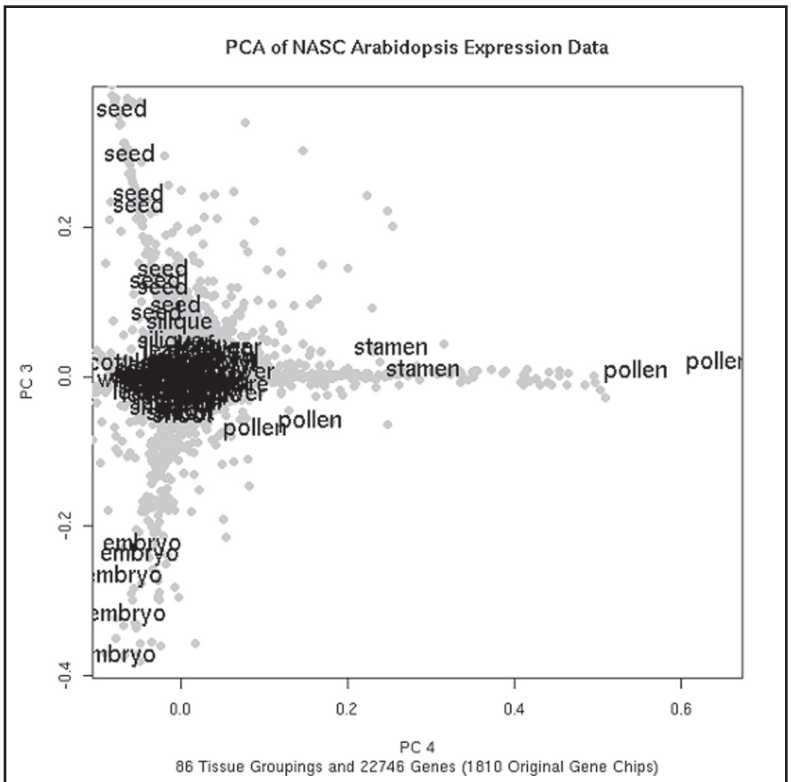


Figure 3. Principal Component analysis of a large set of Affymetrix gene chip data from the Nottingham Arabidopsis Stock Center. A total of 1810 chips were grouped into 86 tissue-by-stage categories. This matrix of 86 tissues by 22,746 genes was then explored for interesting structure. The reproductive stages of pollen, seed and embryo stand out as the most obvious groupings. The cluster in the center contains tissues such as leaves, shoots, and roots.

binding to recognized promoter sequences. The promoters of these enhancer element-containing genes may be useful for studying the timing of expression events during embryogenesis. Sequence pattern searches performed on the promoter regions revealed one or more enhancer elements, in close proximity, in most of embryo-specific genes. Future experiments to understand nearby cis-activation sequences in selected promoters and their corresponding trans-activating proteins include recombinant techniques and yeast two-hybrid screens. In this manner we can develop a second tier of hypothesis-driven tests to comprehend gene regulation networks and coordinated expression in the developing *Arabidopsis* embryo.

Vascular Tissues

Xylem and phloem comprise the water- and nutrient-conducting system in plants and have important mechanical, storage, and secondary metabolic roles. Xylem is the principal wood forming tissue in trees and is of great economic interest. Little is known about the molecular mechanisms that control growth of these tissues. The vascular cambium is a lateral meristem that produces secondary xylem and phloem. New cells produced on the inner surface of the cambium (toward the center of the stem) are destined to become secondary xylem while new cells toward the outside end up as secondary phloem, together leading to an increase in diameter of the root and stem. Understanding the way these tissues arise from the vascular cambium may suggest ways to improve their economic value through either marker assisted breeding or genetic manipulations. With Virginia Tech ASPIRES funding to Eric Beers and Allan Dickerman, we performed expression profiling of xylem and phloem isolated from the root-hypocotyl of *Arabidopsis* to identify genes with the potential to regulate vascular cell differentiation. From these transcript profiles we assembled three gene sets with expression significantly biased towards xylem, phloem-cambium, or non-vascular outer (bark) tissue. We developed a novel "triangle plot" to visualize this three-sample large dataset comparison (Figure 4). Subsequent promoter-reporter experiments performed in the Beers laboratory with five newly identified xylem- or phloem-biased genes validated their roles in vascular development. Details of this research are expected to be published in the journal *Plant Physiology* in 2005.

MicroRNA Isolation

We are also developing methods to profile the microRNA (miRNA) population in *Arabidopsis*

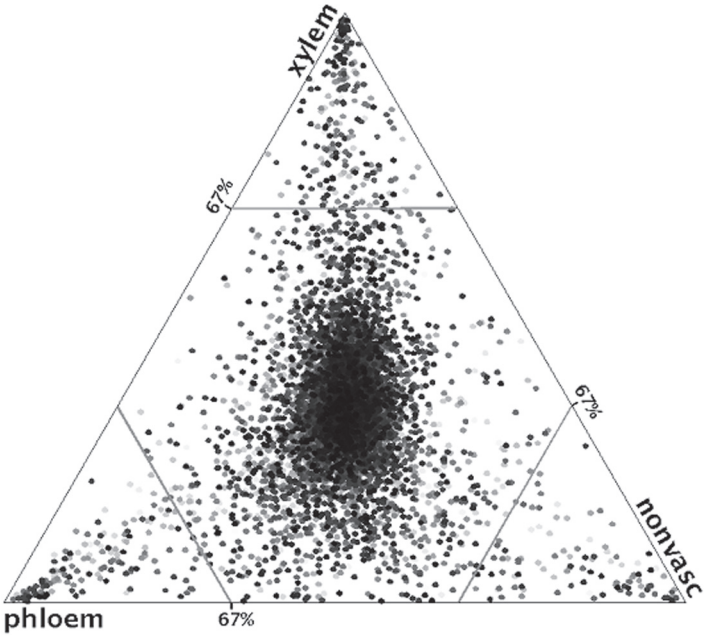


Figure 4. Triangle plot comparing Affymetrix gene expression values in three samples. Points are plotted to be close to the labeled corner in proportion to how biased they are toward one tissue or treatment. Points in the center are genes with balanced expression levels. The lines labeled '67%' delineate genes with > 2 fold expression differential versus both other conditions (i.e., 67% of the total was contributed by that tissue).

tissues. MicroRNAs (miRNAs) are short (19-21 base) non-coding RNA molecules which act in post-transcriptional regulation of gene expression. MiRNAs interact with target mRNAs by sequence complementarity, binding directly to the target mRNAs to induce mRNA degradation or repress mRNA translation. The low abundance of some miRNAs and their time- and tissue-specific expression patterns make experimental miRNA analysis difficult. We are applying special linker chemistry and HPLC techniques to sample these elusive molecules from specific tissues and plan to study miRNA population dynamics in *Arabidopsis* development.

Functional Phylogenomics

Expression Trees

The distinction between paralogs and orthologs becomes highly pertinent when reasons for functional divergence between gene family members is suspected. One reason for divergence in complex eukaryotes would be tissue specific regulation of expression. We have combined the complex datatypes of gene phylogenies and cross-tissue transcription profiles to visualize these patterns for gene families in *Arabidopsis*. We use color gradations, say from green to red, to code the expression levels of each gene represented on the tree (Figure 5). Then looking across tissues, we find branches of the tree that show very different expression levels. So far, we have found that most such changes affect single leaves of the tree rather than branches of two or more nearest neighbors. This suggests that evolution of tissue-specific regulation is rapid compared to the tempo of gene duplication. This work was presented at the TIGR Computational Genomics Conference, 2004.

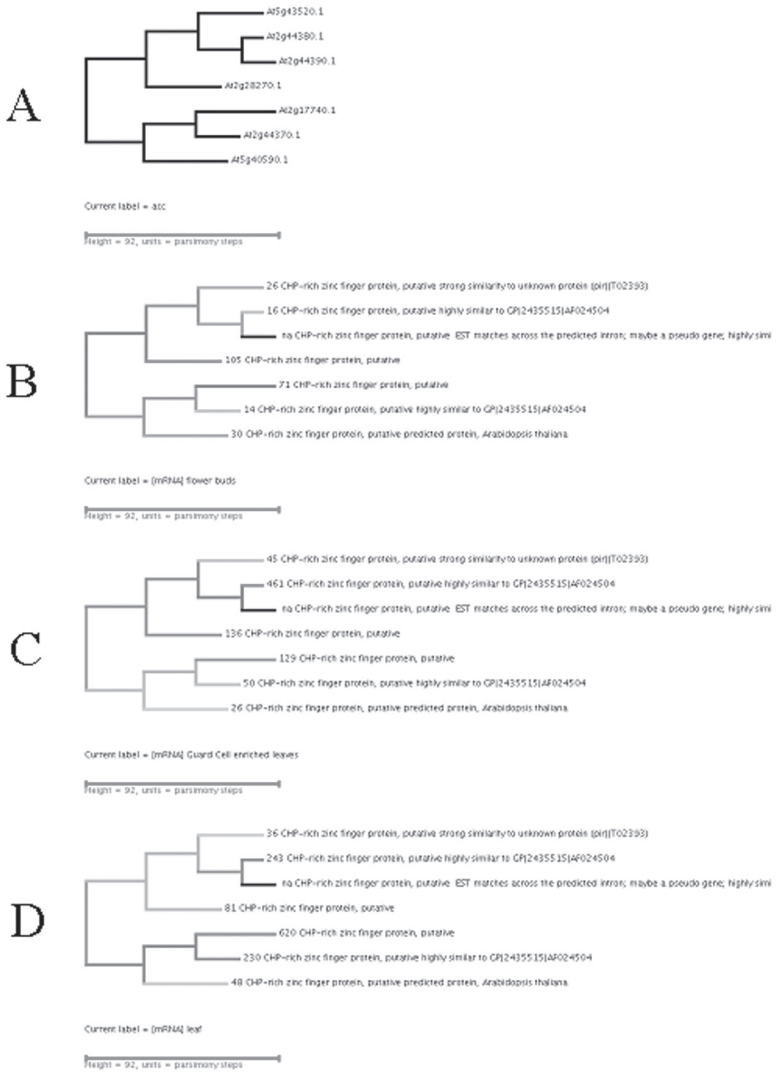


Figure 5. Tissue-specific gene expression visualized on the phylogeny of a Zinc-finger gene family in *Arabidopsis*. (A). Gene expression levels are indicated as the initial number on the label of the tree tips in parts B-D (and with color in the on-screen visualization). The most strongly expressed paralog differes in flower buds (B; At2g2820), guard-cell-enriched leaf (C; At2g44380), and normal leaf (D; At2g17740).

Acknowledgements

Dr. Eric Beers of the Horticulture Department led the xylem-phloem transcription work, with invaluable technical leadership from his graduate student Chengsong Zhao. David Meinke of Oklahoma State University is the Principal Investigator on the NSF SeedGenes grant. Bruno Sobral is the PI of the PathPort project which funds Yuying Tian as the primary developer of the Toolbus components of GeneTrees. Johanna Craig leads the expression analysis in *Arabidopsis*. Elena Shulaeva runs the laboratory.

References

Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer ELL (1999) Pfam 3.1: 1313 multiple alignments match the majority of proteins. *Nucl. Acids Res* **27**: 260–262

Devlin TM (1992) The Textbook of Biochemistry – 3rd Edition, Wiley-Liss Inc, NY

Eckart JD and Sobral BW (2003) A life scientist's gateway to distributed data management and computing: thePathPort/ ToolBus framework. *OMICS* **7**: 79–88

Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**: 113

Felsenstein J (1993) PHYLIP (Phylogenetic Inferrence Package) and manual, version 3.5c. Department of Genetics, University of Washington, Seattle

He Y, Vines RR, Wattam RA, Abramochkin,VG, Dickerman WA, Eckart JD, and Sobral WB (2005) PIML: the Pathogen Information Markup Language. *Bioinformatics* **21**: 116–121

Swofford DL (2002) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts

Tzafrir I, Dickerman A, Brazhnik O, Nguyen Q, McElver J, Frye C, Patton D, and Meinke D (2003) The Arabidopsis SeedGenes Project. *Nucleic Acids Res* **31**: 90–3

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**: 555–556

Zmasek MC and Eddy RS (2000) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics* **17**: 383–384

Publications

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

Tzafrir I, Pena-Muralla R, Dickerman A, Berg M, Rogers R, Hutchens S, Sweeney TC, McElver J, Aux G, Patton D, and Meinke D (2004) Identification of genes required for embryo development in Arabidopsis. *Plant Physiol* **135**: 1206–1220

Duca Group Computational Projects

Karen Duca
Research Assistant Professor, VBI
Adjunct Assistant Professor of Biology, Virginia Tech
kduca@vbi.vt.edu

B. Edward Fulton, Kichol Lee, Nicholas Polys, Dustin Potter, Purvi Saraiya

Project 1: Computer Models of Epstein-Barr Virus Infection of the Human Host

Collaborating Colleagues: Reinhard Laubenbacher (VBI), David Thorley-Lawson (Tufts University)

Mathematical modeling and computer simulation are playing an increasingly large role in the study of biological systems. While these approaches are not intended to replace traditional experimentation, they can provide a conceptual framework for organizing existing data, focus experiments through hypothesis generation, identify critical areas where data are missing, and permit virtual experimentation when real experiments are impractical or unethical. In the absence of applicable animal models, the most effective way to extend our understanding of EBV infection is to implement sophisticated computer simulations. This approach will allow us to define the range of parameters and initial conditions (virologic and immunologic) that produce particular long and short-term consequences for EBV infection. These predictions can then be matched to observations in areas where measurements are feasible. Our collaborator in this endeavor, Dr. David Thorley-Lawson at Tufts University Medical School, has developed sensitive quantitative techniques for measuring parameters of EBV infection capable of generating precise information that can be applied to developing accurate mathematical models.

There were several reasons for the selection of EBV. It is associated with a number of important neoplastic diseases, it provides an approachable system to study persistent virus infection in the human, and it is a paradigm for developing approaches to study important human pathogens for which no suitable animal model is available (Kieff and Rickinson, 2001; Rickinson and Kieff, 2001; Thorley-Lawson, 2001). EBV is an ubiquitous and sometimes pathogenic human γ-herpesvirus that establishes a life long persistent infection in B lymphocytes, despite

Table 1. The ODE Model for EBV Infection. Based on the standard model for HIV, this simple set of three differential equations describes the production and clearance of infected B cells and EBV virions.

Equations	Parameters	Initial Conditions
(dB/dt) = s_B - dB – kBV, naïve B cells	s_B = rate of production of naive B cells	B(0) = Bo
(dB*/dt) = kBV - fB* - pB*, infected B cells	d = rate of naïve B cell loss	V(0) = Vo
(dV/dt) = pB* - cV, free virions	c = rate of free virus clearance p = production rate of free virus	B*(0) = 0
	k = exposure (infectivity) rate constant	
	f = infected B cell death rate	

an aggressive immune response. EBV is an ideal agent for studying persistent infection in the human because: 1. The sites of persistent infection, the Waldeyer's tonsillar ring and the blood, are accessible; 2. Levels of latently infected cells, viral shedding, anti-viral antibody and CD4 and CD8 T cell immune responses can be measured in parallel; 3. Infection can be studied from an extreme state of perturbation, acute infectious mononucleosis (AIM) into resolution as persistent infection; and 4. Much is known about the biology of its host cell the B lymphocyte. Recently, considerable advances have been made in understanding the mechanism of how EBV establishes and maintains persistent infection. As there is no animal model for EBV infection, we model *in silico* how EBV interacts with the immune response to establish and maintain persistent infection in its human host. We employ an agent-based modeling approach, as well an ordinary differential equation based approach based on the HIV standard model (Perelson et al., 1993; Perelson et al., 1996).

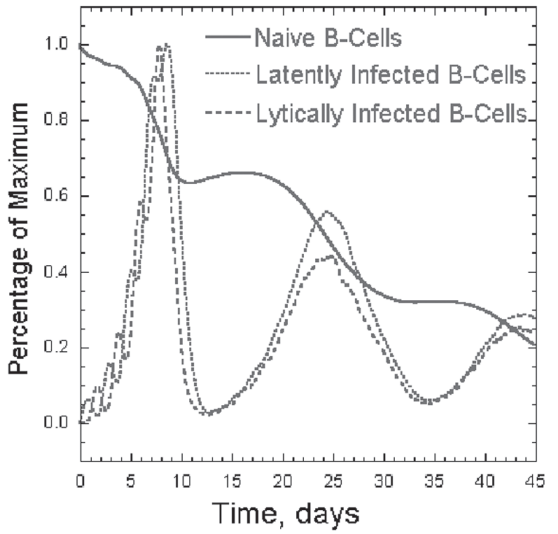


Figure 1a

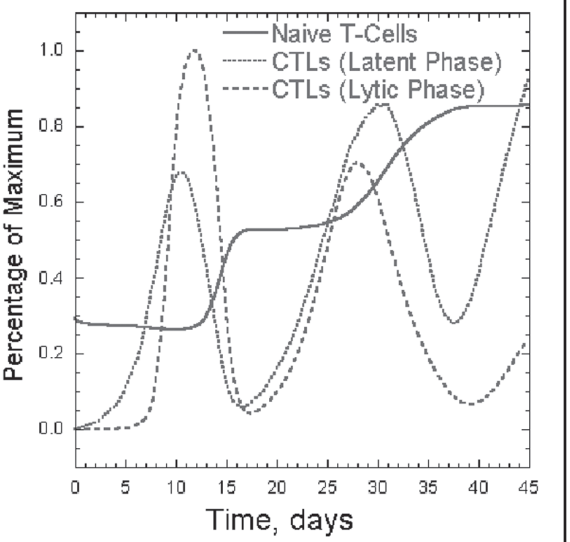


Figure 1b

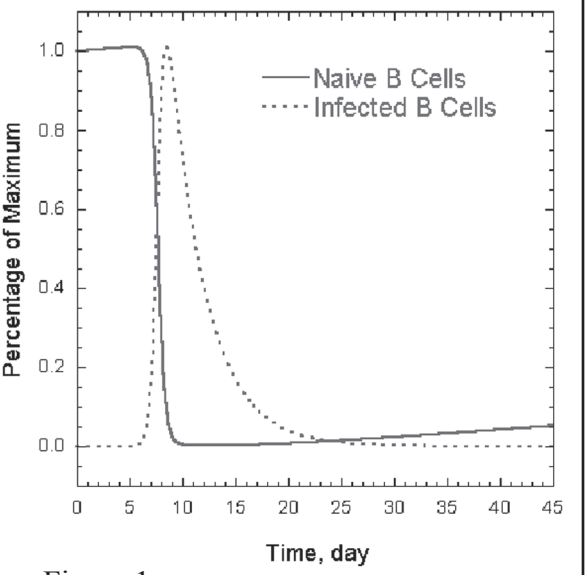


Figure 1c

Figure 1. Preliminary output comparison from the two models. Panels A and B above illustrate the total populations of agents during the acute phase of infection (AIM) relative to their maximum values. Free virus is not shown, but it roughly follows the decay of lytically infected B cells. Panel C depicts results from the standard model.

PathSim (PathogenSimulation) is a systems biology modeling tool designed for the exploration of human host responses to viral pathogens. It is based on two primary components, an agent-based simulation engine that operates at the cellular level and a multi-scale anatomical viewer that realistically represents the lymphoepithelium of the Waldeyer's ring. The current simulation engine incorporates EBV and lymphocytes and their interactions within tissues. Within the agent-based simulation, tissues are composed of meshpoints and edges that connect these meshpoints in 3-D. Each mesh point represents a small region within a tissue and is assigned to a class that further refines its properties. Edges control the local motion of agents. At each time step agent motion and interactions are controlled by a set of stochastic state transition rules. The B cell agents exist in three states: naïve, latently infected, or lytically infected. The T cell agents comprise CTLs directed against both latent and lytic phases, as well as naïve T lymphocytes. The rules represent behavioral approximations based on our current understanding of host responses to EBV, as well as the pathogen's behavior. Preliminary output from both simulations demonstrates only qualitative agreement with data reported in the literature for resolution of the acute phase of EBV infection, acute infectious mononucleosis. Currently, we are probing parameter space and refining our models for improved quantitative agreement with the data.

Project 2: The PathSim Information-Rich Virtual Environment

Collaborating Colleagues: Reinhard Laubenbacher (VBI), Chris North (VT), Doug Bowman (VT)

Computational modeling and data basing of biological phenomena provide significant benefits for the research and education communities. Our objective is to produce integrated information-rich biological databases that capture biological complexity and facilitate new discovery. Databases may be characterized as integrating diverse data types, including: spatial representations of physical phenomena, spanning levels of organization from anatomical to cellular to molecular, temporal data for time series tracking of movement or concentration, and abstract data such as functional genomics/proteomics data, annotations for sequences, or biochemical reactions. In order to derive maximum value from information-rich databases, effective user interfaces and visualizations that facilitate insight generation are absolutely critical. Maximum value will be achieved from these data resources when biomedical scientists are able to explore and navigate them in multiple ways, relating effects between spatial, temporal, abstract, and other data types. Most current virtual environments and information

Table 2. Example user activities for the PathSim IRVE.

Example Biological Questions	Example Tasks in PathSim IRVE	Advantages of the IRVE
How fast does the virus spread from the initial infection site?	View a "movie playback" of viral spread.	Ability to view simultaneous events and assess networked responses.
What is the relative proportion of immune cells to virions in particular locations?	View a heat map of changes in viral concentration as they happen in an organ.	Ability to observe events in their normal context at any level of detail.
Is the distribution of infected cells homogenous in the tissue?	Make a radial plot that evolves over time showing infection levels.	Clear presentation of spatial and temporal data.
Is the virus completely cleared? If so, how long does it take?	Show bar graphs over parts of the tissue illustrating population numbers.	Multiple ways of viewing the data in one environment.

visualizations lack the usability and support for such complex information-rich databases.

PathSim is an example of a modeling strategy that makes extensive use of complex databases for both input and output. The PathSim interface allows an end-user to explore anatomy, physiology, and, eventually, cellular biochemistry. We are constructing and evaluating an information-rich virtual environment (IRVEs) as the second key feature of the PathSim Project. Our hypothesis is that IRVEs, due to their ability to integrate heterogeneous data types and present them to the user simultaneously in appropriate spatial contexts, will enable more rapid discovery and insight generation in the life sciences. The IRVE's interface components are expressed in international standards.

An IRVE combines the capabilities of virtual environments and information visualization to support integrated exploration. Biologists can view the simulated physical structures of the Waldeyer's ring in a 3D virtual environment, interact with associated abstract data, navigate across levels of biological hierarchy, choose data for display, and manage simulation runs all within a single environment. For example, a user might decide to examine the effect of titer on the course of EBV infection. Within the IRVE, the user deposits virions in the locations to be infected. After the simulation commences, the user revisits the IRVE to view signaling events initiated by virus deposition at the molecular level. Later, the user examines how fast the virus is spreading, killing cells, or recruiting immune cells to the vicinity. All activities are viewable in the virtual environment, with interactive links and data export to a suite of analytic tools.

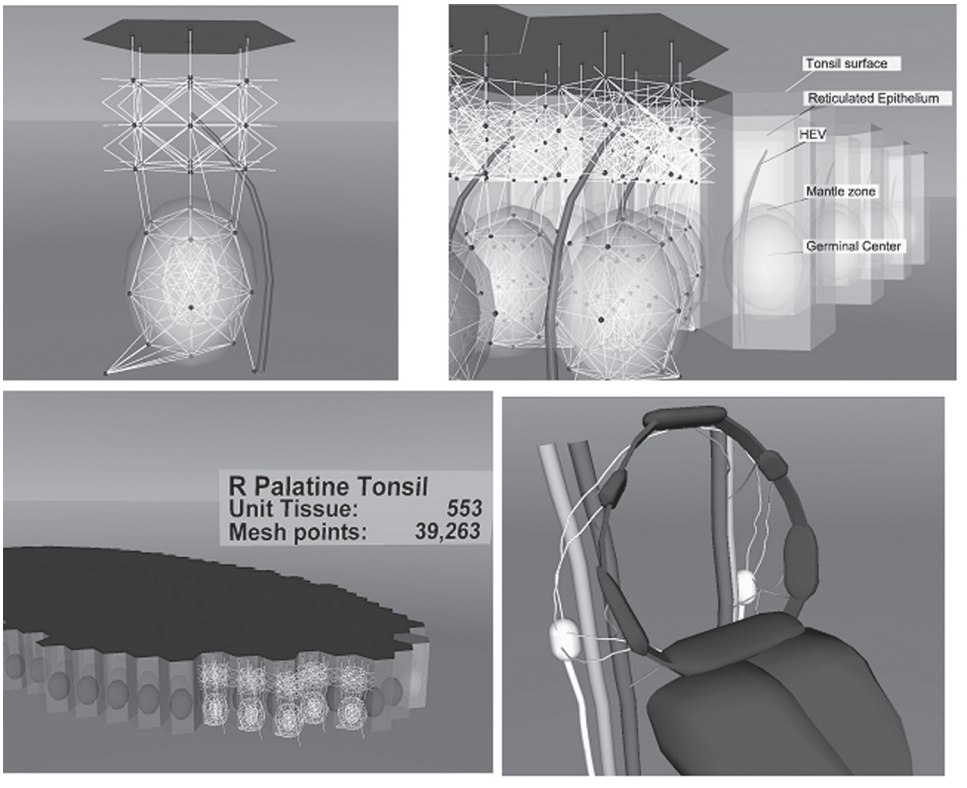


Figure 2. PathSim simulations run on anatomical meshes. Each mesh point represents a volume and type of tissue where agent interactions take place. Abstract parenchymal units and issues are generated to a hierarchical archive according to clinical knowledge. The top panels illustrate individual germinal centers that form a single tonsil, shown at the lower left. The panel at the lower right shows the Waldeyer's ring and abstracted circulation and lymphatic systems.

The IRVE interface operates on a wide range of hardware, from standard desktop displays to high-performance immersive CAVEs (CAVE Automatic Virtual Environment).

Project 3: Evaluation of Visualizations in Bioinformatics Software

Collaborating Colleague: Chris North (VT)

Microarray Viz

High-throughput experiments such as gene expression microarrays result in very large datasets that are impossible to understand and mine without sophisticated tools. Statistical analytic tools are clearly the most important for obtaining a correct biological understanding, but visualizations assist end users in direct data exploration and overviewing results. Although a wide variety of visualizations have been created to aid the data mining process, few formal evaluations have been conducted on the effectiveness of such tools. Life scientists face a dilemma in choosing the best tool for an application. The tool that works best in one context may not work well in another due to differences in the type of information sought from the data or the nature of the data itself.

A primary purpose of any visualization tool is to rapidly provide biologically-relevant insights. To assess and rank bioinformatics visualizations approaches, we developed a heuristic evaluation method that focuses on semi-quantitatively assessing insight. We accomplished this goal by defining insight, offering several measurable characteristics of insight, and methods to recognize insight. These measures are based on our observations of over thirty scientists actually doing data analysis on three very different types of microarray data. This measurement process also enables recognition of qualitative aspects of user behavior and satisfaction with the software. Clearly, true insight has a much broader meaning than our definition permits. However, although our definition is not comprehensive, it does provide an approximation of users' derived insights. This, in turn, has enabled us as evaluators to learn about the effectiveness of these visualization tools.

Based on our insight model, we empirically evaluated five microarray visualization tools (two commercial and three freeware) in a study conducted with life scientists at all levels, from students to experienced professionals. Some of our subjects were also software developers. Each subject was assigned one tool and one data set. We used three actual data sets with distinct properties, but filtered the number of genes to make exploration easier. Users were allowed to explore the data for as long as desired. The study is described in detail elsewhere and has been reviewed in *Bio1nf0rm* (Saraiya et al., 2004; Saraiya et al., 2005a; Toner, 2005).

One shortcoming of all the tools is that they did not adequately link the data to biological meaning. We noted that domain experts performed approximately on par with domain novices, generating relatively few and mainly low-quality insights. This observation suggests that the tools did not leverage the domain expertise very well. Before we conducted the study, we believed that someone more expert in the biology would gain more from visualizations than a beginner. We were also curious about whether software development experience would lead to better usage of the tools. However, these background differences did not reflect themselves strongly in the insights generated.

Pathway Viz

Due to their size and complexity, biological pathways are challenging to visually represent and analyze. Several visualization systems have been developed to assist life scientists in analyzing pathways. During structured and unstructured interviews, we found that many life scientists are reluctant to use these systems due to steep learning curves associated with using the products, as well as the amount of effort required to manually construct biologically relevant pathways. They find minimal value in a system that provides only simple visual or dynamic pictures, without providing adequate means to manipulate information to meet analytic requirements. In order to identify

critical requirements for pathway visualization systems, we conducted ethnographic field studies with four research professors and post-doctoral fellows and heuristic evaluations with six life scientists using six popular, widely-used systems. We also evaluated these and other systems to arrive at general approaches currently employed to address these requirements.

Based on our studies, a research agenda was derived identifying five critical requirements for pathway visualization systems that, if addressed effectively, prove most helpful in supporting exploratory pathway analysis. These include: 1. automated construction and updating of pathways by searching literature databases, 2. overlaying information on pathways in a biologically relevant format, 3. linking pathways to multidimensional data from high throughput experiments, 4. overviewing multiple pathways simultaneously with analysis of any interconnections, and 5. scaling pathways to higher levels of abstraction to analyze effects of complex molecular interactions at higher levels of biological organization (Saraiya et al., 2005b).

At present, we are conducting an evaluation on a range of alternate visualizations that allow life scientists to study high-throughput data within the context of the corresponding pathways. The result from such evaluations can provide guidelines to computer scientists for creating visualizations that efficiently overlay data on pathway diagrams in ways that preserve the dynamics and enable insight generation. We believe that improved evaluation leads to better visualizations that in turn can lead to richer insights for the biomedical community.

Project 4: microBLAST, A New Tool for Microarray Comparisons

Collaborating Colleague: Reinhard Laubenbacher (VBI)

A quick search of the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo) reveals that the number of profiling experiments in the public domain is increasing dramatically. At the end of the year 2000, 250 microarrays had been uploaded. By 2004, over 30,000 microarray samples had been submitted. A parallel search of the Stanford Microarray Database (http://genome-www5.stanford.edu/) revealed similar growth. As the number of such profiling experiments continues to grow, the need for tools to rapidly compare thousands of experiments to extract similarities and differences clearly increases in proportion. Moreover, with increasing amounts of data, different kinds of information may be extracted. We have progressed beyond the stage of simply identifying which single genes are up- or down-regulated in single time point snapshots of particular disease states or under certain experimental conditions. At this stage, with the plethora of data available, we may begin: 1. to rigorously explore the complex dynamics of gene expression; 2. to exploit analogies from known systems to unknown systems in order to unearth biological mechanisms in theoretical ways; and 3. to employ the collected information to infer mechanistic models of gene expression. The microBLAST tool addresses the need for large-scale comparisons, as well as these new, emerging needs.

We have used Formal Concept Analysis (FCA), a mathematical method in applied lattice theory, to construct geometric objects (represented mathematically as graphs) associ-ated with gene expression data from functional genomics experiments (or any other type of global profiling technology). Geometric measures on these objects can then be used as a signature that allows computationally tractable comparisons between other similarly represented experiments. Indeed, we call our tool microBLAST, due to its operational resemblance to the popular BLAST tool for comparing a sequence of interest to a large database of sequences. In the same manner, microBLAST implements a geometric method to take a reference microarray experiment and search for both global and local similarities in other experiments in a large database.

Twenty-one biological attributes, consisting of protein motif families, and three expression

attributes, consisting of percentile categories within an experiment, were used to construct the geometric representations of the data. A gene had a given biological attribute if the protein it coded for displayed a particular protein motif associated with one of the motif families. As a first measure of global similarity between experiments, edit distance (ED) between graphs was selected. Edit distance is very easy to implement and interpret. Basically, it represents the number of changes to vertices and edges required to transform one graph into the other. Small ED, therefore, indicates high similarity, while high ED indicates high dissimilarity.

Seventy-seven host-pathogen microarray experiments de-posited at GEO using the single channel microarray format (Affymetrix U95 chip set) were selected. Very different types of infections were chosen in order to ensure maximum biological dissimilarity. The *in vitro* experiments were obtained from three different labs and can be categorized as follows: twelve time-point expression profiles of foreskin fibroblasts infected by human cytomegalovirus (HCMV) (http://www.ncbi.nlm.nih.gov/geo/gds/gds_browse.cgi?gds=476); seven time-point expression profiles of HeLa cells infected with coxsackievirus B3 (CVB3) or mock–infected with PBS (http://www.ncbi.nlm.nih.gov/geo/gds/gds_browse.cgi?gds=477 and http://www.ncbi.nlm.nih.gov/geo/gds/gds_browse.cgi?gds=478); gene expression in macrophages and dendritic cells following exposure to 12 different pathogens that produce variable chronic infections (http://www.ncbi.nlm.nih.gov/geo/gds/gds_browse.cgi?gds=260). Only genes with ex-pression levels greater than 100 in all samples were accepted for graph building. The 500 genes with the highest variance across all experiments were chosen as the object set. Data reduction is not required for microBLAST and was performed only to facilitate data analysis during the development phase.

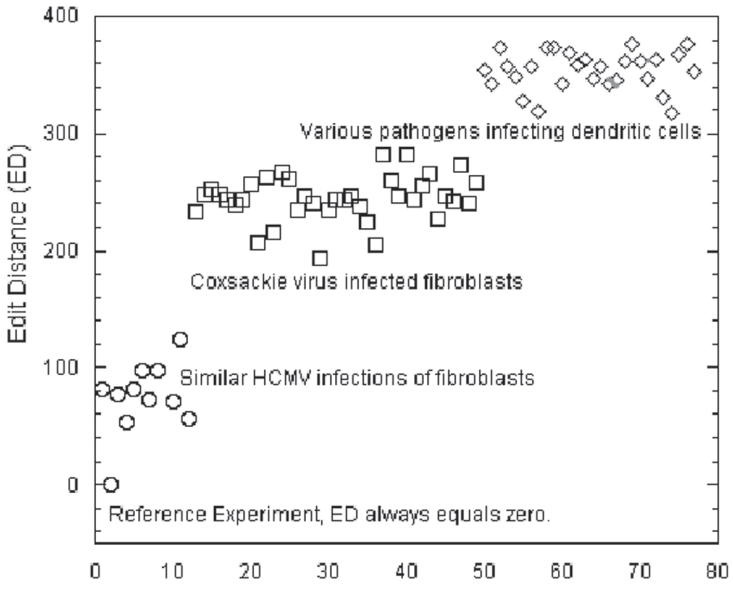


Figure 3. microBLAST successfully finds similar experiments. Three distinct host-virus interactions analyzed by single channel microarray display expected similarity with respect to a reference experiment. The reference experiment is HCMV infecting human foreskin fibroblasts. Similar HCMV experiments exhibit EDs lower than 100. Coxsackie virus infected foreskin fibroblasts exhibit greater global dissimilarity, while the dendritic cells infected with parasitic and bacterial pathogens are even more dissimilar.

References

Cruijff M, Thijs C, Govaert T, Aretz K, Dinant GJ and Knottnerus A (1999) The effect of smoking on influenza, influenza vaccination efficacy and on the antibody response to influenza vaccination. *Vaccine* **17**: 426–432

Finklea JF, Sandifer SH and Smith DD (1969) Cigarette smoking and epidemic influenza. *Am J Epidemiol* **90**: 390–399

Kark JD, Lebiush M, and Rannon L (1982) Cigarette smoking as a risk factor for epidemic (H1N1) influenza in young men. *N Engl J Med* **307**: 1042-1046

Kieff E and Rickinson AB (2001) Epstein-Barr Virus and its Replication. In Knipe, D.M. and Howley, P.M. (eds.), *Virology* pp. 2511-2574. Lippincott, Williams, and Wilkins, New York, NY

MacKenzie JS, MacKenzie IH, and Holt PG (1976) The effect of cigarette smoking on susceptibility to epidemic influenza and on serological responses to live attenuated and killed subunit influenza vaccines. *J Hyg (Lond)* **77**: 409–417

Mori I, Komatsu T, Takeuchi K, Nakakuki K, Sudo M, and Kimura Y (1995) In vivo induction of apoptosis by influenza virus. *J Gen Virol* **76** (Pt 11): 2869–2873

Perelson AS, Kirschner DE, and DeBoer R (1993) Dynamics of HIV-infection of CD4+ T-cells. *Math Biosci* **4**: 81–125

Perelson AS, Neumann AU, Markovitz M, Leonard JM, and Ho DD (1996) HIV-1 dynamics in vivo: Virion clearance rate, infected cell life-span, and viral generation time. *Science* **271**: 1582–1586

Rickinson AB and Kieff E (2001) Epstein-Barr Virus. In Knipe, D.M. and Howley, P.M. (eds.), *Virology* pp. 2575–2628. Lippincott, Williams, and Wilkins, New York, NY

Saraiya P, North C, and Duca K (2004) Evaluation of microarray visualization tools for biological insight. *IEEE Symposium on Information Visualization*, Austin, TX

Saraiya P, North C, and Duca K (2005a) An Insight-based Methodology for Evaluating Bioinformatics Visualizations. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, in press

Saraiya P, North C, and Duca K (2005b) Visualization for Biological Pathways: Requirements Analysis, Systems Evaluation and Research Agenda. *Information Visualization*, in press

Schultz-Cherry S, Krug RM, and Hinshaw VS (1998) Induction of Apoptosis by Influenza Virus. *Semin Virol* **8**: 491–495

Takizawa T, Matsukawa S, Higuchi Y, Nakamura S, Nakanishi Y, and Fukuda R (1993) Induction of programmed cell death (apoptosis) by influenza virus infection in tissue culture cells. *J Gen Virol* **74** (Pt 11): 2347–2355

Thorley-Lawson DA (2001) Epstein-Barr virus: exploiting the immune system. *Nat Immunol* **1**: 75–82

Tomita K, Caramori G, Lim S, Ito K, Hanazawa T, Oates T, Chiselita I, Jazrawi E, Chung KF, Barnes PJ, and Adcock IM (2002) Increased p21(CIP1/WAF1) and B cell lymphoma leukemia-x(L) expression and reduced apoptosis in alveolar macrophages from smokers. *Am J Respir Crit Care Med* **166**: 724–731

Toner B (2005) Will Better Usability Studies Help Swell Market for Bioinformatics Software? *Bio1nf0rm* **9**: 1

Wickenden JA, Clarke MC, Rossi AG, Rahman I, Faux SP, Donaldson K, and MacNee W (2003) Cigarette smoke prevents apoptosis through inhibition of caspase activation and induces necrosis. *Am J Respir Cell Mol Biol* **29**: 562–570

Publications

Saraiya P, North C, and Duca K (2004) Evaluation of microarray visualization tools for biological insight. *IEEE Symposium on Information Visualization*, Austin, TX

Saraiya P, North C, and Duca K (2005a) An Insight-based Methodology for Evaluating Bioinformatics Visualizations. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, in press

Saraiya P, North C, and Duca K (2005b) Visualization for Biological Pathways: Requirements Analysis, Systems Evaluation and Research Agenda. *Information Visualization*, in press

Rout S, Lam V, Bell AE, and Duca KA (2004) Epifluorescent Image Modeling for Viral Infection Analysis, Asilomar Conference on Signal Processing and Computation 2004, Asilomar, CA

Lam V, Duca KA, and Yin J (in press) Arrested spread of vesicular stomatitis virus infections *in vitro* depends on interferon-mediated antiviral activity. Journal: *Biotechnology and Bioengineering*

Polys NF, Bowman DA, North C, Laubenbacher R, and Duca K (2004) PathSim Visualizer: An Information Rich Virtual Environment Framework for Systems Biology, Proc. SIGGRAPH Web3D Session, Los Angeles, CA

Jarrah A, Vastani H, Duca K, and Laubenbacher R (2004) An Optimal Control Problem for in vitro Virus Competition, 43rd Annual IEEE Conference on Decision and Control, San Juan, Puerto Rico

Design of Four-Helix Bundle Peroxidase Mimics

Joel R. Gillespie
Research Assistant Professor, VBI
jgill@vbi.vt.edu

Kinjal Shah, Deepa Balasubramaniam, Michaela Babiceanu

Life as we know it requires the specific interactions of many different biomolecules in a complex and extremely crowded environment, namely the cellular cytoplasm. The ability of biomolecules to differentiate between one another, called molecular recognition, relies on the cooperative interaction of a number of relatively weak intermolecular forces that can be strongly influenced by other extrinsic environmental factors. Understanding the influence of these weak forces in governing the structure, dynamics, and assembly of biological macromolecules, and proteins in particular, is the central overarching theme of the research effort in the Gillespie laboratory. Here we describe progress on one of the three major research efforts underway in our laboratory—the design of simplified enzyme mimics using model-driven design and combinatorial screening. The remaining two thrusts are aimed at deciphering the physical mechanisms underlying the inappropriate aggregation of proteins in human disease and at understanding the molecular recognition and protein folding events that occur during *in vivo* chaperone-assisted protein folding.

Introduction

Protein design has emerged as an important and powerful tool for the study of the interrelationships between protein structure and protein function, and serves as a unique method for uncovering the principles that govern the roles that relatively weak intermolecular forces play in biological catalysis and self-assembly reactions (DeGrado et al., 1999; Kennedy and Gibney, 2001). The central aim of protein design is mapping of the sequence/structure phase space, i.e., to find sequences of amino acids that are compatible with a specific stable three-dimensional protein structure (or function). Understanding the underlying principles that determine structure and function, and which govern macromolecular dynamics, will ultimately allow for the reliable prediction of protein structure form sequence. Perhaps more importantly, it will also allow for the intelligent manipulation of protein structures to produce new functions. Such manipulation of protein structures has great promise for leading to novel bio-inspired supramolecular devices that will revolutionize materials science and medicine. Examples of such applications include the development of protein-based non-linear optical devices and switches for optical computing, and the design of tailored or targeted therapeutic agents (Stayton et al., 1994; Klug 2005; Takeda et al., 2004).

Practical protein design typically involves a fusion of both theoretical and experimental approaches to achieving a desired design aim. On the theoretical side, potential energy functions similar to those used in molecular dynamics (MD) simulations are used to screen a large number of amino acid sequences compatible with a fixed backbone template structure (Dahiyat and Mayo, 1996; Looger and Hellinga, 2001). The resulting limited number of high scoring sequences can then be constructed and verified in the laboratory (Kuhlman and Baker, 2004). Indeed, this methodology has been utilized to build *bona fide* designed proteins with unique folds (Kuhlman et al., 2003). Unfortunately, less success has been forthcoming in terms of the design of specific functions in proteins, though ligand binding and protein-protein self-assembly design experiments are beginning

to yield fruit (Kortemme et al., 2004). Using less ambitious techniques, the rational design of proteins using structural analysis and site-directed mutagenesis has been successfully utilized for redesigning enzymes for enhanced stability and/or altered function. Mass screening and combinatorial methods have also yielded promising results, though the tractability of screening large libraries still remains an impediment limiting the scope of sequence space that can be explored (Arnold and Volkov, 1999).

The Gillespie laboratory has initiated design projects on a variety of different enzymes including lipases, esterases, peroxidases, and carboxylases. Each of these classes of enzymes catalyzes interesting and industrially useful chemistry. The use of recombinant and designed enzymes is becoming increasingly interesting to industry as the extreme stereo- and regioselectivity of enzymes can be utilized to greatly enhance yields of desired products in chemical reactions while simultaneously using less energy and simplifying the purification of products (Schoemaker et al., 2003). Indeed, designed proteins are ubiquitous in consumer products and are found in items ranging from beer and soft drinks to laundry detergent. Here we describe our initial attempts to use model-driven rational design and directed evolution strategies to produce small protein mimics of larger, more complex peroxidase enzymes. We are interested in designing such mimics for two major reasons. First, peroxidases belong to a special class of enzymes called oxidoreductases and catalyze redox chemistry across an extremely wide spectrum of redox potentials. The unique properties of redox active enzymes, especially their selectivity and ability to be detected both optically and electrochemically, makes them especially attractive targets for protein design and biotechnological applications (Gilardi and Fantuzzi, 2001). We are interested in gaining a more thorough understanding of how protein structure influences the redox potential of prosthetic groups, especially highly conjugated groups such as heme (Schifman et al., 2000; Mazumadar et al., 2003). Second, such small protein mimics may serve as prac-tical catalysts in the industrial synthesis of bulk and fine chemicals and may find uses in bioremediation applications. Therefore we are interested in exploring the realistic use of such enzymes in catalyzing industrially important oxidation reactions.

Over the past year, our primary goal has been to rationally design a simple protein scaffold for the construction of a peroxidase active site. We are particularly interested in two types of peroxidases - the fungal heme peroxidases and vanadium peroxidases produced by *Rhodophyta*, marine red algae (Littlefield, 1999). Both of these classes of enzymes catalyze unique chemical reactions that have potential applications in bioremediation and industrial chemistry. Our first target was to design a mimic of the fungal enzyme lignin peroxidase (LiP) produced by the basidiomycete fungus *Phanerochaete chrysosporium* (Tien and Kirk, 1984*)*. LiP is a secreted heme peroxidase that degrades lignin, an irregular and complex phenylpropane polymer found in woody plants, and plays an important role in the global carbon cycle by making this otherwise indigestible substance available as food for other fungi and bacteria. In addition to degrading lignin, *P. chrysosporium* and LiP has been demonstrated to degrade a wide spectrum of recalcitrant environmental pollutants and xenobiotics including benz[a]pyrenes, PCBs, TNT, DTT, and pesticides such as methoxychlor and lindane (Abraham et al., 2002; Ohtsubo et al., 2004; Bumpus et al., 1985).

This wide spectrum of potential substrates is one of the chief reasons why LiP is of interest to us. Though LiP degrades lignin in nature, the enzyme has little substrate specificity as the substrate molecules are not physically bound to the protein. Instead, substrates are oxidized via a series of electron transfer reactions starting at the heme metal center and employing an intermediary solvent exposed tryptophan residue (Edwards et al., 1993). This tryptophan, which forms a tryptophanyl cation radical, subsequently transfers the free electron to electropositive molecules in the environment.

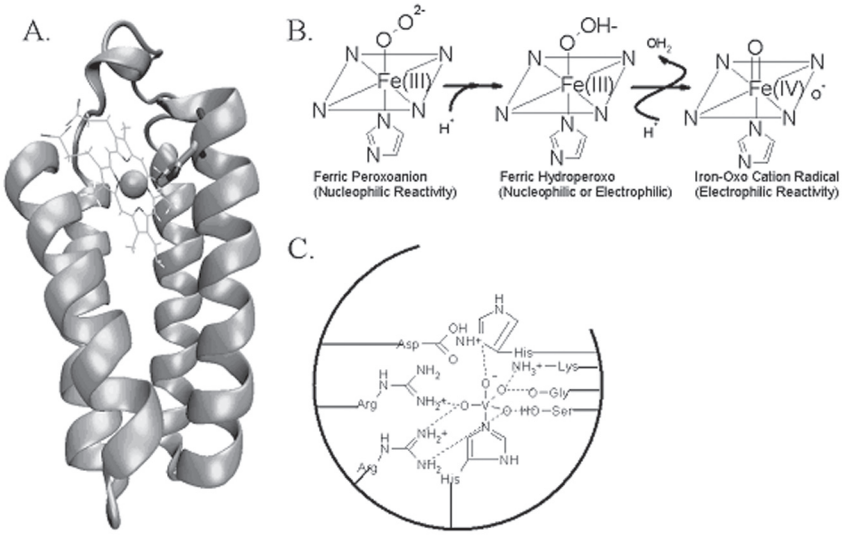


Figure 1. **A.** A four-helix bundle electron transfer protein, cytochrome b_{562}, is used as a scaffold for the inclusion of a peroxidase active site. Our initial targets include a mimic of lignin peroxidase (LiP), a secreted fungal enzyme that degrades lignin in woody plants, and a vanadium bromoperoxidase (VbP) produced by marine brown algae. The later enzyme is involved in the biosynthesis of many halogenated marine natural products including halogenated terpenes. In first generation mimics, the four-helix bundle has been modified to provide access to the heme center by hydrogen peroxide. Later generations have incorporated residues specifically designed to aid in the disproportionation of hydrogen peroxide and to stabilize heme binding which is destabilized by creation of the peroxide channel. **B**. In the LiP mimics, initial binding of peroxide produces a ferric peroxoanion. Abstraction of a proton from a neighboring histidine residue (with its pKa modified by an adjacent arginine) produces a ferric hydroperoxo center. An additional proton from the solvent (at acidic pH) allows the good leaving group OH- to dissociate (forming water) and results in a highly reactive ferryl-oxo cation radical. The radical is stabilized by the highly conjugated 18e- pi system of the heme and will transfer the electron to substrates that make physical contact with the heme group (the heme edge pathway) or may transfer the electron through space to a solvent exposed tryptophan. The resulting tryptophanyl cation radical can then further transfer the free electron to small molecule substrates in the environment. **C.** The active site of vanadium bromoperoxidase is more complex than that of LiP. Though only in the early stages of design, the four-helix bundle is still a good choice for a scaffold for design of VbP mimics as many of the amino acids stabilizing the orthovanadate are present in cytochrome b_{562} in more or less the same geometry. These include the paired arginine residues that stabilize the equatorial ligands of the vanadate by hydrogen bonding, the axially coordinating histidine, and the hydrogen bonding serine and glycine residues.

In nature, this may include aromatic secondary metabolites produced by the fungus which then go on to degrade lignin by a free radical addition mechanism. In addition to the lack of substrate specificity, the extremely high redox potential of LiP (1.4 V verses NHE) and the formation of high valent ferryl-oxy radicals make LiP an extremely interesting target for design studies.

Our second target for design has been the vanadium bromoperoxidase produced by marine red and brown algae. These enzymes catalyze the halogenation of organic substrates such as terpenes, making them attractive catalysts for the production of complex natural products, including antibiotics. This later design project, which is still in its infancy and which is considerably more complex in scope than our LiP mimics, will not be discussed further due to space limitations.

Results and Discussion

Over the past year, our initial objectives were to develop a simple protein scaffold with enough flexibility to allow the insertion and binding of a heme or other porphyrin (including phthalocyanines). The decision was made early to focus on using a four-helix bundle as such a scaffold due to the simple architecture of such proteins and their natural abundance as electron transfer heme proteins (Gibney et al., 1998). Gibney et al. (2001) have used synthetic four- helix bundles with multiple heme binding sites as models for understanding biological electron transfer. With this in mind, a synthetic strategy was developed for the Fmoc chemical synthesis of four peptides (one for each helix) with simplified amino acid sequences designed to contain two heptad repeats per helix. The scaffold would then be assembled using disulfide bonds to form dimers. The full complex would then be assembled by interaction between the helices to form paired coiled-coils, resulting in a dimer of dimers. Though the chemical synthesis approach was successful for the production of the individual helices, the yields of the peptides were between 25 and 50 percent (which is very good considering the size of the peptides), but purification of the desired sequence by RP-HPLC proved to be extremely difficult and assembly of the peptides into a four-helix bundle was abandoned in favor of an ultimately more productive genetic approach.

To facilitate our goal of designing simple

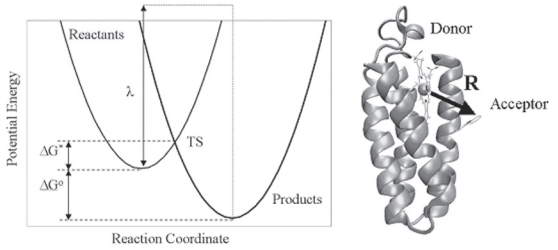


Figure 2. In the non-adiabatic limit, electron transfer in our peroxidase mimics is governed by Marcus-Hush theory. The rate of electron transfer is given by:

$$k_{et} = \left(\frac{2\pi}{\mathrm{h}}\right)\exp^{(-\beta R)}\, FC \qquad [1]$$

where R is the distance between the electron donor and electron acceptor. FC, the nuclear Franck-Condon factor, relates the driving force for the electron transfer (ΔG^o) to and the free energy of reorganization (λ) by:

$$FC = \sqrt{\left(4\pi\lambda k_B T\right)}\exp^{-\left[\frac{\left(\Delta G^o+\lambda\right)^2}{4\lambda k_B T}\right]} \qquad [2]$$

In our LiP mimics, the electron donor is the heme metal center and the acceptor is a surface exposed tryptophan residue. Ideally, this tryptophan will abstract an electron from the heme after the disproportionation of hydrogen peroxide to yield a tryptophanyl radical cation. This radical cation can then transfer the free electron to other species in the environment, effectively oxidizing them. In species where the free radical cannot be stabilized, the end result is cleavage of chemical bonds and degradation of the substrate. Biologically relevant electron transfers do not occur over distances greater than 15Å.

scaffolds for the production of peroxidase mimics, we switched to a genetic approach in which we cloned the cytochrome b_{562} from *Yersinia enterocolitica* to fill the role as a scaffold. Cytochrome b_{562} is a periplasmic electron transfer protein that does not have a direct catalytic role, serving as an electron carrier only. However, the simple architecture of the protein (a four-helix bundle with no disulfides), its ability to non-covalently bind a single heme-b moiety, and other features of its structure made it an ideal choice for our peroxidase scaffold. The *Yersinia* gene was cloned into a T7-based vector with replacement of the native *Yersinia* periplasmic leader sequence with that from *E. coli*. A protocol for the purification of the wild-type *Yersinia* protein was then worked out using periplasmic extraction to simplify the purification process. Briefly, purification of the wild-type cytochrome b_{562} utilized a combination of ion exchange and size-exclusion chromatography to produce a homogeneous product of >99% purity as measured by gel electrophoresis and the Reinheitzhal number (a ratio measurement of the heme and protein absorbance spectra at 418 and 280nm).

In our model-driven design approach, homology models of the *Yersinia* cytochrome b_{562} were generated using the *E. coli* protein's X-ray crystal structure as a template (Hamada et al, 1995). These homology models were used to direct placement of amino acid substitution mutations within the scaffold. Two initial series of mutations were made in the wild-type protein. In the first set, the heme metal center, which is normally hexacoordinate with ligation by an imidazole nitrogen from histidine 103 and pi backbonding from the sulfur of methionine 7, was made five coordinate by mutation of methionine 7 to an alanine. This mutation both opens a binding site and provides an access channel for peroxide into the heme center. A second set of mutations was also constructed in which ligation of the heme was inverted so that methionine 7 was mutated to a histidine and histidine 103 was mutated to an alanine. This arrangement was intended to preserve the peroxide access channel, but places the open coordination position of the heme in close proximity to an arginine guanido group and the imidazole group of a second histidine, both of which should act to aid the disproportionation of hydrogen peroxide bound to the heme center.

In each of the active site series, four individual single tryptophan substitution mutations were constructed in various positions in the fourth helix of the protein. These tryptophan residues were constructed to provide an acceptor for transfer of the heme centered free radical electron to a highly solvent exposed site, where an additional transfer could readily occur to substrates in the environment (thus mimicking the role of the beta-hydroxylated tryptophan-171 in native LiP). These tryptophan residues were all in solvent exposed locations (substitution of lysine residues) in the amino acid chain but differ in distance from the heme of between 8 Å and 20 Å, covering more than 5 orders of magnitude in the rate of electron transfer expected (Figure 2).

The active site alterations made in the M7A and M7H/H103A series showed marked changes in the heme binding properties of the proteins, with members of the M7H/H103A series having far lower heme binding affinity than the wild-type protein. When loaded with an excess of heme, however, all members of the M7A and M7H/H103 series showed considerable peroxidase activity comparable to that of native LiP. The differences in binding of heme were not surprising in light of the large contribution that heme makes to the overall thermodynamic stability of the protein (nearly 60% of the total, or ~ 3kcal/mol).

Each of the mimics was characterized for peroxidase activity using a POD assay in which the formation of the cation radical of ABTS (2,2'-azino-di-(3-ethyl-benzthiazoline-6-sulfonic acid)) was measured spectrophotometrically in the presence of excess hydrogen peroxide. The kinetics of ABTS oxidation for the mimics was compared to that of recombinant LiP (isozyme H8) and mixed LiP isozymes purified from fungal cultures. The kinetic progress curves showed

that the M7A and M7H/H103 mutant series had significant peroxidase activity, especially in the pH range of 4.5-6.5 (LiP has maximal activity between 2.5 and 3.5), while wild-type cytochrome b_{562} had no detectable activity at any pH. In addition, the activity of these mutants was highly dependent on the concentration of hydrogen peroxide and showed no loss of activity at the highest concentrations of peroxide used in the assays (5mM), suggesting that auto-oxidation of the heme and/or inactivation of the enzyme by formation of compound III (an $Fe(III)O_2^-$ adduct) are rare or do not occur in the mimics. Under the same conditions the half-life of LiP is less than 10 minutes do the oxidative cleavage of the heme. As shown in Figure 3, initial spectroscopic investigations of the peroxidase mimics showed the formation of *bona fide* compound I, the ferryl-oxo cation radical (depicted in Figure 1B). Formation of Compound I is characterized by the loss of the Soret band in the visible spectrum of the proteins. This band arises from π-π* transitions in the aromatic heme and are greatly reduced in intensity when the Fe(IV) distorts the planarity of the ring.

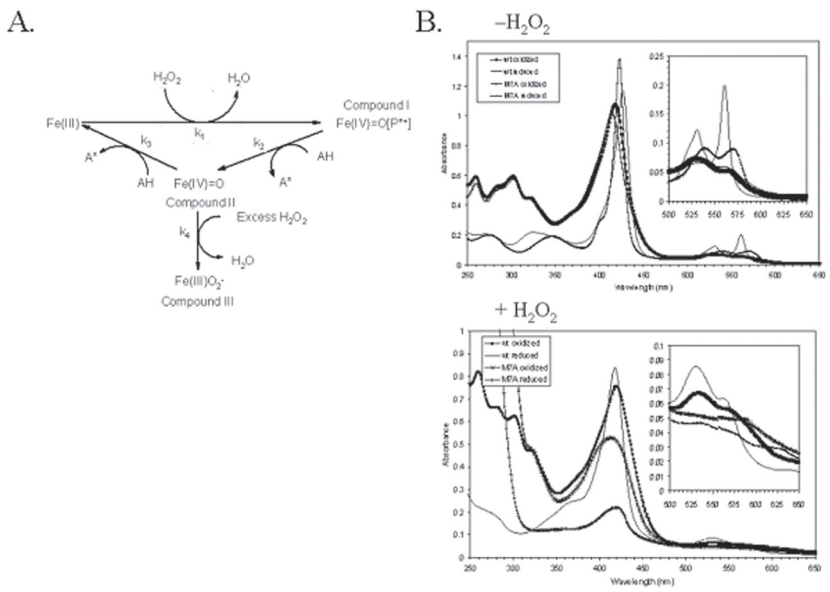


Figure 3. **A**. The classical peroxidase mechanism follows ping-pong kinetics in which the ordered binding of reactants and release of products occurs. In the peroxidases, the result is the formation of distinct intermediates with characteristic spectral and physical properties. The reduction of peroxide occurs through a two electron process, oxidizing the metal center to Fe(IV) and producing a free radical electron. This species, Compound I, is an Iron(IV)-oxo of ferryl species containing a protein centered free radical electron (either centered on the highly conjugated heme or on a tryptophan residue as in LiP). Compound I transfers a single electron to substrates in the environment to form Compound II, an Iron(IV)-oxo intermediate without the radical electron. Compound II retains significant redox potential and oxidizes an additional substrate molecule to the resting enzyme's oxidation state of Fe(III). In the presence of excess peroxide, the Fe(IV)-oxo species may form an inactive Compound III, a ferric peroxoanion. This inactivation is reversible. **B**. Equilibrium spectroscopic assessment of the heme metal center during reduction of peroxide shows the formation of a bona fide Compound I as demonstrated by the loss of the Soret band at 420nm due to distortion of the aromaticity and geometry of the heme moiety in the Fe(IV) state. In addition, the loss of pi backbonding, the result of substitution of M7 with an alanine, is readily apparent in the Q-bands.

Finally, we have begun to conduct directed evolution experiments with our peroxidase mimics to enhance the heme binding and global thermodynamic stability of the proteins. Our strategy utilizes a combination of error prone PCR and in vitro recombination (using cytochrome b_{562} genes from related enteric bacteria) to sample sequence space. When finished the mutagenized gene library will be

cloned into a T-odd bacteriophage vector for repeated rounds of screening and amplification. To facilitate this, we have developed a heme binding screen utilizing Fe(II)-porphyrin covalently attached to 200 μm oxirane-acrylic beads. The stringency of binding can be increased by washing the column with increasing concentrations of a chaotropic agent (such as urea), causing the weaker binding proteins to dissociate and wash out.

Conclusions

Using the cytochrome b_{562} protein from *Yersinia enterocolitica* as a scaffold, we have successfully modified the protein to produce a peroxidase mimic with significant peroxidase activity (similar to that of native peroxidases such as LiP) but with enhanced stability to self-oxidation of the heme group. These mimics have been demonstrated to oxidize compounds that are potential environmental pollutants such as the direct diazo dye Congo red (Figure 4). Successive generations of these mimics are planned that will have even greater catalytic activity and thermodynamic stability.

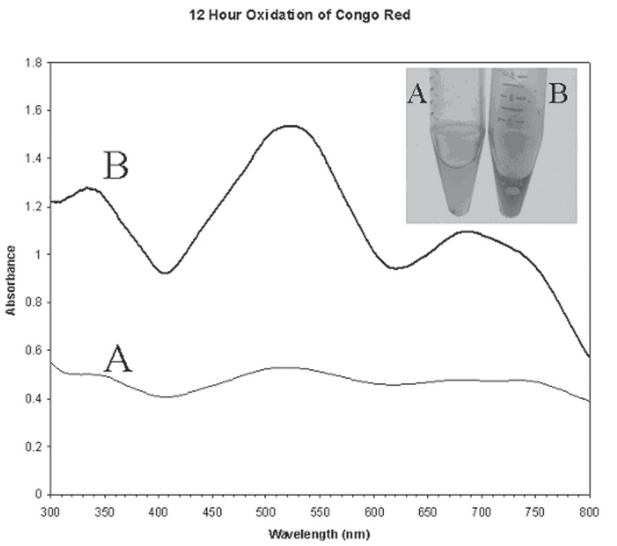


Figure 4. Peroxidase mimics have great potential as catalysts in industrial and bioremediation applications. A 1μM concentration of the M7A/K94W mimic degrades 50% of a 10mM solution of Congo red in 12 hours at pH 4.0 and at 37°C in the presence of 10mM hydrogen peroxide (A) while an untreated control (B) remains highly colored. Congo red is a common direct diazo dye used to stain cotton cloth. Annually, hundreds of thousands of tons of dyes, lakes, and mordants are released into industrial wastewater streams in the textile and clothing industries. Many of these compounds are potential mutagens and have long environmental half-lives.

Future Directions

Our investigations of heme peroxidase mimics are still in an early stage and many open questions remain. Of particular interest is proving that the tryptophan substitution mutations constructed to model the LiP reaction mechanism are actually populated as tryptophanyl cation radicals. Initial evidence using changes in the fluorescence emission spectra of the proteins in the presence of peroxide suggests that the formation of tryptophanyl cation radicals is indeed occurring, but interpretation of the spectra is made uncertain due to the fluorescence emission of the heme groups. In the near future, we will employ electron paramagnetic resonance (EPR) spectroscopy to address this issue and to directly characterize the spin state and coordination number of the heme center as well during peroxide turnover (supplemented by resonance Raman spectroscopy). In addition, we are also beginning experiments to determine if the tryptophans are hydroxylated as is the case with W171 in native LiP. In LiP, this hydroxylation occurs during the first turnover of the enzyme and may serve to alter the stability and photophysical properties of the tryptophan and thus modulate its electron transfer properties to substrates in the environment.

In addition to the above-mentioned mechanistic investigations, we also intend to continue to improve the activity of the mimics using successive generations of designs that incorporate additional features to enhance the peroxidase activity. These modifications include the addition of aromatic amino acids in the protein interior to aid in stabilization of

the heme via pi-stacking interactions. We will also initiate a more through investigation of the thermodynamic stability of the various mimics to gain an understanding of how mutations in the active site alter heme binding and overall protein stability.

References

Abraham W-R, Nogales B, Golyshin PN, Pieper DH, Timmis KN (2002) Polychlorinated biphenyl-degrading microbial communities in soils and sediments. *Curr Opin Microbiol* **5**: 246–253

Arnold FH and Volkov AA (1999) Directed evolution of biocatalysts. *Curr Opin Chem Biol* **3**: 54–59

Bumpus JA, Tien M, Wright D, and Aust SD (1985) Oxidation of persistant environmental pollutants by a white rot fungus. *Science* **228**: 1434–1436

Dahiyat BI and Mayo SL (1996) Protein design automation. *Protein Sci* **5**: 895–903

DeGrado WF, Summa CM, Pavone V, Nastri F, and Lombardi A (1999) De novo design and structural characterization of proteins and metalloproteins. *Ann Rev Biochem* **68**: 779–819

Edwards SL, Raad R, Wariishi H, Gold MH, and Poulos TL (1993) Crystal structure of lignin peroxidase. *PNAS USA* **90**: 750–754

Gibney BR, Huang SS, Skalicky JJ, Fuentes EJ, Wand AJ, and Dutton PL (2001) Hydrophobic modulation of heme properties in heme protein maquettes. *Biochem* **40**: 10550–10561

Gibney BR, Rabanal F, Reddy KS, and Dutton PL (1998) Effect of four helix bundle topology on heme binding and redox properties. *Biochem* **37**: 4635–4643

Gilardi G and Fantuzzi A (2001) Manipulating redox systems: application to nanotechnology. *Trends in Biotech* **19**: 468–476

Hamada K, Bethge PH, and Mathews FS (1995) Refined structure of cytochrome b562 from Escherichia coli at 1.4A resolution. *J Mol Biol* **247**: 947–962

Kennedy ML and Gibney BR (2001) Metalloprotein and redox protein design. *Curr Opin Struct Biol* **11**: 485–490

Klug A (2005) Towards therapeutic applications of engineered zinc finger proteins. *FEBS Lett* **579**: 892–894

Kortemme T, Joachimiak LA, Bullock AN, Schuler AD, Stoddard BL, and Baker D (2004) Computation redesign of protein-protein interaction specificity. *Nat Struct Mol Biol* **11**: 371–379

Kuhlman B and Baker D (2004) Exploring folding free energy landscapes using computational protein design. *Curr Opin Struct Biol* **14**: 89–95

Kuhlman B, Dantas G, Ireton GC, Vanni G, Stoddard BL, and Baker D (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**: 1364–1368

Littlefield, J (1999) Haloperoxidases and their role in biotransformation reactions. *Curr Opin Chem Biol* **3**: 28–34

Looger LL and Hellinga HW (2001) Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics. *J Mol Biol* **307**: 429–445

Mazumadar S, Springs SL, and McLendon GL (2003) Effect of redox potential of the heme on the peroxidase activity of cytochrome b562. *Biophys Chem* **105**: 263–268

Ohtsubo Y, Kudo T, Tsuda M, and Nagata Y (2004) Strategies for bioremediation of polychlorinated biphenyls. *Appl Microbiol Biotech* **65**: 250–258

Schoemaker HE, Mink D, and Wubbolts MG (2003) Dispelling the myths- biocatalysis in industrial synthesis. *Science* **299**: 1694–1697

Shifman JM, Gibney BR, Sharp RE, and Dutton PL (2000) Heme redox potential control in de novo designed four-a-helix bindle proteins. *Biochem* **39**: 14813–14821

Stayton PS, Olinger JM, Wollman ST, Bohn PW, and Sligar SG (1994) Engineering proteins for electrooptical biomaterials. In Birge RR (ed.), *Molecular and Biomolecular Electronics* pp 475-490. New York: American Chemical Society Press

Takeda S, Kamiya N, and Nagamune T (2004) Rational design of a protein-based molecular device consisting of blue fluorescent protein and zinc protoporphyrin IX incorporated into a cytochrome b562 scaffold. *Biotech Lett* **26**: 121–125

Tien M and Kirk TK (1984) Lignin-degrading enzyme from *Phanerochaete chrysosporium*: purification, characterization, and catalytic properties of a unique H_2O_2-requiring oxygenase. *PNAS USA* **81**: 2280–2284

Genetic Architecture of Quantitative Traits

Ina Hoeschele

Research Professor, VBI
Professor of Statistics, Virginia Tech
inah@vt.edu

Guimin Gao, Alberto de la Fuente, Hua Li, Yongcai Mao, Nan Bing, Bing Liu, Chiranjeet Chetia

Introduction

The field of statistical genetics has resulted from the merger of genetics and statistics into a quantitative theory for the interpretation of genetic data. Applications of this theory have led to substantial achievements in the fields of animal and plant breeding and human genetics. Statistical genetics is a very large field, as evidenced by the wide array of contributions to the recent "Handbook of Statistical Genetics" (Balding et al., 2003). In agreement with the "Handbook", I consider statistics to be a major contributor to the design of experiments, the management of data, the analysis and interpretation of data, and the presentation of data and results. I consider statistical genetics to include 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, and the analysis of transcriptome, metabolome, and proteome profiling data. In that sense, statistical contributions to genomics, transcriptomics, proteomics, and metabolomics are considered as new avenues within statistical genetics, rather than as new fields. This view of statistical genetics differs from the frequent consideration of genetics as the study of one or few genes at a time, which is in contrast with the joint investigation of all genes in or all proteins encoded by a genome. Our view of statistical genetics reflects the seamless transitions between these areas, exemplified by the transition from QTL analysis of traditional, organismal phenotypes with few markers to genome-wide analysis of QTL main and interaction effects on phenotypes, to QTL analysis of transcription profiles (coined 'genetical genomics' [Jansen, 2003; Jansen and Nap, 2001]), and, in the future, to an integrated whole genome analysis of gene expression combining phenomics, genetic marker, transcriptomics, proteomics, and metabolomics data.

Our group aims to contribute, within the field of statistical genetics, (i) to the basic analysis of characteristically noisy 'omics' data and of 'omics' experiments with multi-factorial treatment and covariance structures, and (ii) to the study of the genetic architecture of quantitative traits via QTL mapping of organismal phenotypes and 'omics' profiles.

Basic Analysis of Gene Expression Data

Methodology

While many computational methods have been proposed for the detection of differentially expressed genes in microarray experiments, most of the initial methods focused on comparisons between only two treatments (*e.g.*, normal versus cancerous tissues). However, wider access to the technology and a reduction in costs have made it possible to conduct much larger experiments, which have multi-factorial treatment and covariance structures. One Affymetrix experiment, for example, includes mock and pathogen inoculated groups of plants from each of six soybean cultivars sampled at five time points post-infection and at two different sites of sampling of lesions, with three overall experimental replicates (total of 144 GeneChips). While many methods can determine

whether a gene is differentially expressed between two conditions, Linear Model Analysis (LMMA) automatically uses the information in the data in an optimal way to answer more complicated questions of differential expression, such as whether differences among cultivars in differential expression between mock and pathogen infection depend on sampling site, time, or experiment. We also analyzed interwoven loop and paired cDNA experiments comparing bovine and porcine embryos generated by in-vitro fertilization versus nuclear transfer (Bing et al., 2005; Pfister-Genskow et al., 2005). In estimating differential expression between the two embryo types, LMMA automatically accounted for the increased similarity among expression values of a gene derived from the same rather than different arrays or embryos. We have implemented LMMA in three steps, using the SAS or R system: (1) Background correction via B transformation (Irizarry et al., 2003). (2) Normalization via the quantile method (Bolstad et al., 2003) or linear normalization via LMMA. (3) Gene-specific LMMA. We justify this approach (Hoeschele and Li, 2005).

Accomplishments and Results

We compared different implementations of the LMMA on the publically available Affymetrix spike-in Latin square design (Cope et al., 2004), and on several real data sets from various Affymetrix experiments. For the first data set, it was known which genes were truly differentially expressed, as gene fragments were added at known concentrations in the experiment. The model for probe-level analysis included fixed effect of array type, random array effect, and fixed probe effect. Analysis at the gene-level employed a fixed model with array type effects only. For this data set, methods were compared in terms of numbers of false positive and false negative findings. For the other data sets, we were only able to evaluate the extent to which the different implementations of LMMA produced different results. Results for the spike-in data in Table 1 show less power for gene-level compared to probe-level analyses, false positives for the probe level analyses, and only slightly better performance of the probe-level analyses in terms of total misclassification. Quantile normalization performed only slightly better than linear normalization for probe-level analysis.

Table 1: Comparison of different implementations of Linear Mixed Model Analysis of the Affymetrix spike-in Latin square design with 16 genes known to be differentially expressed (18 arrays of types D to I were used).

Method[1)]	Power	False Positives	False Negatives	Total Misclassification
Gene – MAS 5.0	6 / 16	0	10	10
Gene – MBEI	10 / 16	4	6	10
Gene – RMA	10 / 16	0	6	6
Probe – linear	16 / 16	7	0	7
Probe – inv. set	16 / 16	11	0	11
Probe – quantile	16 / 16	6	0	6

[1)] Gene – MAS 5.0: Gene-level analysis with MAS 5.0 signal as summary measure; Gene – MBEI: Gene-level analysis with Li&Wong Model Based Expression Index as summary measure; Gene – RMA: Gene-level analysis with RMA summary measure proposed by (Irizarry et al., 2003); Probe – linear: probe-level analysis using PM data only with linear global normalization; Probe – inv. set: probe-level analysis using PM data only with invariant set normalization (Li and Wong, 2001); Probe – quantile: probe-level analysis using PM data only with quantile normalization (Bolstad et al., 2003).

Our analyses of several real Affymetrix experiments using mouse and human chips showed very small differences between probe-level LMMA analyses performed with or without background correction (using the B(.) transformation of (Irizarry et al., 2003) and with linear versus quantile normalization. When compared with gene-level analyses (as described below Table 1), results were consistent for genes with large differential expression and quite variable for genes with moderate expression differences.

Conclusions and Future Direction

Performing background correction, normalization, and gene-specific LMMA in three separate steps is not optimal (although it may be the only feasible strategy for very large microarray experiments). Recently, two groups ((Wu et al., 2004), BGX group http://www.bgx.org.uk, (Hein et al., 2005)) have proposed single-step modeling of Affymetrix probe-level data, which also utilizes MM signal, up to the calculation of gene expression summary measures and evaluation of differential expression between pairs of conditions. We have initiated work trying to combine this approach with the LMMA of complex, multi-factorial experiments, and eventually, with the analysis of Genetical Genomics microarray experiments, if warranted, based on a comparison between three-step LMMA and a new single-step LMMA.

QTL Mapping in Pedigrees

Methodology

Genome-wide localization of disease genes and QTL in human and outbred animal pedigrees, using preferred multi-point mapping methods, necessitates efficient reconstruction of haplotypes from observed genotype data on multiple linked marker loci. Several methods for haplotype reconstruction have been suggested, which are either rule or likelihood based. Likelihood based methods are to be preferred (Sobel et al., 1995), but these have to be implemented with Monte Carlo methods and are computationally demanding, while rule-based methods can be much faster on large pedigrees. Here, we have developed a deterministic, likelihood based method, which is computationally efficient for large pedigrees and large numbers of linked loci (Gao and Hoeschele, 2005b; Gao et al., 2004). Our main application of this algorithm is the calculation of Identity-by-descent (IBD) matrices, which is an important step in Quantitative Trait Locus (QTL) analysis using variance component models (Gao and Hoeschele, 2005a).

Accomplishments and Results

The new method was compared with a Markov Chain Monte Carlo (MCMC) method (Loki; (Heath, 1997)) in terms of QTL mapping performance on simulated pedigrees. The methods yielded almost identical results for the estimation of QTL positions and variance parameters and for the likelihood profiles. The new method is, however, much more computationally efficient than the MCMC approach for large pedigrees and large numbers of loci. For a pedigree of size 500 and ten linked marker loci, evaluation of the IBD matrix at 27 putative QTL positions took 1 hour and 46 minutes using Loki, versus 5 minutes and 16 second with the new, deterministic method. Moreover, the new method can be used for fine-mapping via joint linkage disequilibrium (LD) and linkage analysis, which improves the power and accuracy of QTL mapping, while Loki does not incorporate non-zero IBD probabilities among founder haplotypes due to LD.

Conclusions and Future Direction

The current haplotyping method assumes that all individuals in a pedigree are genotyped at all markers. Work is underway to extend the method to allow for missing marker data.

We have partially developed an alternative method for fine-mapping in human and outbred populations, a fully-parametric distribution method implemented in a Bayesian inference framework. Current methods for such populations are either based on variance components analysis (Meuwissen and Goddard, 2000; Meuwissen and Goddard, 2001), or on the assumption of a bi-allelic QTL (Perez-Enciso,

2003). While linkage analysis does not have power to distinguish between QTL with two or more alleles, we believe that the incorporation of linkage disequilibrium provides information on the number of mutations in the population history.

Genetical Genomics Analysis for Gene Network Inference

Methodology

Genetic analysis of gene expression in a segregating population, which is expression profiled and genotyped at DNA markers throughout the genome, can reveal regulatory networks of polymorphic genes. We propose an analysis strategy with several steps: (1) Genome-wide QTL analysis of all expression profiles to identify eQTL confidence regions, followed by fine-mapping of identified eQTL if needed; (2) identification of regulatory candidate genes in each eQTL region; (3) correlation analysis of the expression profiles of the candidates in any eQTL region with the profile of the gene affected by the eQTL, to reduce the number of candidates to one or few; (4) construction of an 'encompassing' network by drawing directed edges from each eQTL and from its associated, retained regulatory candidate genes to the genes affected by the eQTL; and (5) identification of an optimal, sparser network of only direct influences, assumed to be embedded in the 'encompassing' network, via Structural Equation Modeling (SEM), a technique similar to Bayesian networks, but able to deal with cyclic structures and continuous data.

Accomplishments and Results

We applied an initial implementation of this Genetical Genomics analysis to a yeast population of 40 segregants (Brem et al., 2002). We were able to show that a gene network can be partially reconstructed using genetic data on a small population (Bing and Hoeschele, 2005). We identified eQTL by nonparametric single marker analysis and by sliding three-marker regression (to avoid unnecessarily large QTL regions due to multiple linked QTL). We estimated the false discovery rate (FDR) and the proportion of false positives from the p-values of all tests across markers and expression profiles and obtained essentially identical results. Realizing that these approaches may not provide reliable estimates of the FDR, we are now conducting a simulation study to compare these and alternative FDR controlling methods. Confidence intervals were computed for each significant eQTL via a bootstrap resampling method. For each eQTL (confidence) region, a list of candidate regulatory genes was compiled. If the list contained more than one gene, correlation analysis was performed by first determining and retaining the candidate regulatory gene having the strongest (largest in absolute value), significant (Spearman or Pearson) expression correlation with the regulated gene (*i.e.*, the gene affected by the eQTL), subsequently retaining the gene with the strongest, significant 1st order partial correlation coefficient (conditional on the gene retained in the previous step), next searching for the gene with the strongest, significant 2nd order partial correlation, and so on, until no significant (partial) correlations were found (de la Fuente et al., 2004).

The median size of eQTL regions from single marker analysis was 93.5 Kbp; the median number of candidate regulatory genes was 49; in 65 percent of all eQTL regions a single candidate gene was retained after correlation analysis; in 7 percent of all regions no candidate gene was retained; the largest number of retained genes was six; overall, 768 regulatory candidate genes were retained; 848 eQTL intervals were obtained from sliding three-marker regression; the median distance of QTL intervals from sliding three-marker regression analysis was 6 Kbp; 331 of the 768 candidate genes were located at or in one of these intervals and considered as the strongest candidate regulatory genes; of all eQTL regions with a single, retained, regulatory candidate gene, 45 percent were *cis*-regulations (a gene is located in its own eQTL region) and 55 percent *trans*-regulations (gene B is located in an eQTL region for gene A, and gene A and B's expression profiles are significantly correlated). One or several biological processes were statistically significantly over-represented in independent network structures or in highly

interconnected sub-networks. Most of the transcription factors in the inferred network had a putative regulatory link to only one other gene or exhibited *cis*-regulation. Most of the biological processes represented in the inferred network are metabolism pathways, whose gene constituents should carry different genetic variants resulting in phenotypic differences between the two yeast strains involved in the experimental cross.

In order to evaluate Structural Equation Modeling in the last step of our Genetical Genomics analysis for the identification of an optimal, more sparse network embedded within the "encompassing" network resulting from eQTL analysis, we are now performing a simulation study (de la Fuente et al., 2005). The main questions are whether Linear SEMs can adequately describe gene regulatory networks characterized by nonlinear interactions and feedback processes, and whether useful information can be obtained given the unresolved issues of identification and equivalence in cyclic SEMs.

Genotypes of 400 Recombinant Inbred Lines (RILs) were simulated using QTLcartographer (Basten et al., 1996). Gene expression levels were simulated with Gepasi (Mendes, 1997) using non-linear ordinary differential equations. Genetic polymorphisms were incorporated by setting the basal transcription rate of one allele equal to 50 percent of the other. For three different network topologies each containing 10 genes and several cycles, and three different experimental noise levels (10 percent, 25 percent, and 50 percent of the realized genetic variance of each gene expression profile), 100 data sets were simulated. QTL analysis was carried out using Interval Mapping in QTL cartographer, and an "encompassing" network was constructed.

For the small, ten-gene networks investigated initially, we searched the structure space embedded in the "encompassing" network using a simple search strategy in which each edge is tested, starting from edges with lowest likelihood ratio in the QTL analysis, using Mx (Neale et al., 2003). If the removal of the edge from the model yielded a smaller value of the Bayesian Information Criterion (BIC), the removal was performed. Networks constructed in this way were compared to the true network known from simulation. In all simulations performed to date, power (correctly discovered direct influences as a percentage of number of true direct influences) remained above 90 percent, and false discovery rate (number wrongly discovered direct influences over all discovered influences) below 15 percent. Although these initial results are encouraging, the utility of SEM for inference of cyclic gene networks depends critically on the issues of identifiability and equivalence. At the least, the inclusion of QTL nodes in the networks improves identifiability and reduces the equivalence problem, as there are no edge direction reversals in the structure search.

Conclusions and Future Direction

We and other groups have very strong evidence that Genetical Genomics is an extremely powerful approach to gene and protein network reconstruction. We will therefore continue to implement a Genetical Genomics analysis package applicable to very large, real data sets, and to evaluate this strategy, in particular the Structural Equation Modeling component, by simulation of artificial gene networks. We will also evaluate Structural Equation modeling for other specific perturbation experiments.

References

Balding DJ, Bishop M, and Cannings, C. (eds.). (2003) *Handbook of Statistical Genetics*. John Wiley, New York

Basten CJ, Weir BS, and Zeng ZB (1996) QTL Cartographer: A Reference Manual and Tutorial for QTL Mapping. Department of Statistics, North Carolina State University, Raleigh, NC, p. http://statgen/ncsu.edu/qtlcart/

Bing N and Hoeschele I (2005) Genetical genomics analysis of a yeast segregant population for transcription network inference. *Genetics*, in press

Bing N, Hoeschele I, Ye K, and Eilertsen KJ (2005) Finite mixture model analysis of microarray expression data on samples of uncertain biological type with application to reproductive efficiency. *Vet Immunol Immunopathol*, in press

Bolstad BM, Irizarry RA, Astrand M, and Speed TP (2003) A comparison of normalization methods for high density oligonucleotide data based on variance and bias. *Bioinformatics* **19**: 185–193

Brem R, Yvert G, Clinton R, and Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**: 752–755

Cope LM, Irizarry RA, Jaffee HA, Wu Z, and Speed TP (2004) A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics* **20**: 323–331

de la Fuente A, Bing N, Hoeschele I, and Mendes P (2004) Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* **20**: 3565–3574

de la Fuente A, Liu B, and Hoeschele I (2005) A genetical genomics approach to inferring gene networks. *1st FEBS Advanced Lecture Course "Systems Biology: From Molecules and Modeling to Cells"*, Gosau, Austria

Gao G and Hoeschele I (2005a) Approximating identity-by-descent matrices using multiple haplotype configurations on pedigrees. *Genetics*, in press

Gao G and Hoeschele I (2005b) A note on a conditional enumeration haplotyping method in pedigrees. In *Lecture Notes in Bioinformatics*. Springer-Verlag, New York, Vol. in press

Gao G, Hoeschele I, Sorensen P, and Du FX (2004) Conditional probability methods for haplotyping in pedigrees. *Genetics* **167**: 2055–2065

Heath SC (1997) Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *Am J Hum Genet* **61**: 748–760

Hein AMK, Richardson S, Causton HC, Ambler GA, and Green PJ (2005) BGX: A fully Bayesian gene expression index for Affymetrix GeneChip data. *Biostatistics*, in press

Hoeschele I and Li H (2005) A note on joint versus gene-specific mixed model analysis of microarray gene expression data. *Biostatistics* **6**: 183–186

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, and Speed TP (2003) Exploration, normalization, and summaries of high-density oligonucleotide array probe level data. *Biostatistics* **4**: 249–264

Jansen R (2003) Studying complex biological systems using multifactorial perturbation. *Nat Rev Genet* **4**: 145–151

Jansen R and Nap J (2001) Genetical genomics: the added value from segregation. *Trends Genet* **17**: 388–391

Li C and Wong WH (2001) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error applications. *Genome Biol* **2**: 1–11

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

Meuwissen THE and Goddard ME (2000) Fine mapping of quantitative trait loci using linkage disequilibrium with closely linked marker loci. *Genetics* **155**: 421–430

Meuwissen THE and Goddard ME (2001) Prediction of identity by descent probabilities from marker haplotypes. *Genet Sel Evol* **33**: 605–634

Neale MC, Boker SM, Xie G, and Maes HH (2003) Mx: Statistical Modeling. Department of Psychiatry., Richmond, VA, p. http://www.vcu.edu/mx/

Perez-Enciso M (2003) Fine mapping of complex trait genes combining pedigree and linkage disequilibrium combined information: A Bayesian unified framework. *Genetics* **163**: 1497–1510

Pfister-Genskow M, Myers C, Childs L, Lacson J, Betthauser J, Gouleke J, Forsberg EP, Zheng Y, Leno G, Schult, R, Liu B, Chetia C, Yang X, Hoeschele I, and Eilertsen KJ (2005) Identification of differentially expressed genes in individual bovine preimplantation embryos produced by nuclear transfer: Improper reprogramming of genes required for trophoblast development. *Biol Reprod*, **72**: in press

Sobel E, Lange K, O'Connell JR, and Weeks DE (1995) Haplotyping algorithms. In Speed, T.P. and Waterman, M.S. (eds.), *Genetic Mapping and DNA Sequencing, IMA Volumes in Mathematics and Its Applications(edited by Friedman A, Gulliver R)*. Springer, New Yor.

Wu Z, Irizarry RA, Gentleman R, Murillo FM, and Spencer F (2004) A model based background adjustment for oligonucleotide expression arrays. *J Amer Stat Assoc* **99**: 909–917

Publications

Bing N and Hoeschele I (2005) Genetical genomics analysis of a yeast segregant population for transcription network inference. *Genetics*: in press

Bing N, Hoeschele I, Ye K, and Eilertsen KJ (2005) Finite mixture model analysis of microarray expression data on samples of uncertain biological type with application to reproductive efficiency. *Vet Immunol Immunopathol* **105**: 187–196

de la Fuente A, Bing N, Hoeschele I, and Mendes P (2004) Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* **20**: 3565–3574

Gao G and Hoeschele I (2005) A note on a conditional enumeration haplotyping method in pedigrees. In *Lecture Notes in Bioinformatics*. Springer-Verlag, New York, Vol. in press

Gao G, Hoeschele I, Sorensen P, and Du FX (2004) Conditional probability methods for haplotyping in pedigrees. *Genetics* **167**: 2055–2065

Hoeschele I and Li H (2005) A note on joint versus gene-specific mixed model analysis of microarray gene expression data. *Biostatistics* **6**: 183–186

Pfister-Genskow M, Myers C, Childs L, Lacson J, Betthauser J, Gouleke J, Forsberg EP, Zheng Y, Leno G, Schultz R, Liu B, Chetia C, Yang X, Hoeschele I, and Eilertsen KJ (2005) Identification of differentially expressed genes in individual bovine preimplantation embryos produced by nuclear transfer: Improper reprogramming of genes required for trophoblast development. *Biol Reprod* **72**: 546–555

Modeling and Simulation of Biological Systems: The Applied Discrete Mathematics Group

Reinhard Laubenbacher
Research Professor, VBI
Professor of Mathematics, Virginia Tech
reinhard@vbi.vt.edu

Stephen Buck, Omar Colón-Reyes, Miguel Colon-Velez, Edgar Delgado-Eckert, Elena Dimitrova, Abdul Jarrah, John McGee, Brandilyn Stigler, Paola Vera-Licona

Introduction

The central objective of computational systems biology is the construction of explanatory mathematical models to understand the design and function of biological systems. New technologies are beginning to provide the kind of quantity and quality of data that make this task increasingly feasible. While a wide variety of computational tools are available from mathematics, computer science, statistics, and engineering, it is clear that the complexity of biological systems and their unique features make it necessary to develop a new set of computational tools for this purpose. The central goal of the Applied Discrete Mathematics Group is the development of such tools within the area of discrete mathematics. Through a network of collaborators, the group applies these mathematical methods to the modeling of several types of biological networks, including biochemical networks, gene regulatory networks, the immune system response to certain viral pathogens, and neural response networks. Furthermore, the group is developing graph theoretic tools for the analysis of large-scale social contact networks.

Modeling of Networks from Data

An important problem in computational systems biology is the construction of models from high-throughput data, such as those obtained from large-scale molecular profiling. A new modeling approach is needed to best carry out the transformation of such data into an understanding of reaction network structure. Such a "top-down" approach will start with little knowledge about the system, capturing at first only a coarse-grained image of the system with only a few variables. Then, through iterations of simulation and experiment, the number of variables in the model is increased. At each iteration, novel experiments will be suggested by simulations of the model, which, when carried out, will provide data to improve the model further, leading to a higher resolution in terms of mechanisms.

The choice of modeling method for regulatory networks depends in part on whether one views the network as a chemical reaction network or an information-processing network. The former view suggests the use of systems of differential equations, whereas the latter suggests the use of models that resemble logical switching networks, which is the appropriate one for our modeling approach. The modeling framework is that of time-discrete dynamical systems over finite sets. Let $x_1, \ldots, x_n$ be variables, which might represent concentrations of molecular species such as mRNA, cell types, or other biological quantities. Each of these variables is allowed to take values in a finite set X, e.g., $X=\{-1, 0, 1\}$, corresponding to *downregulated, unchanged, upregulated,* in the case of mRNA measuring gene activity. That is, X contains a set of qualitative states of the variables. The variables influence each other's state over time. The resulting network of dependencies can be

expressed as a dynamical system

$$F = (f_1, f_2, \ldots, f_n): X^n \to X^n$$

where f_i represents the function that expresses the change over time of variable x_i, as a function of the other variables.

The goal of top-down modeling is to start with a collection of time series of measurements of the variables x_i and then find the "best" model F that explains the data. During the previous reporting period, we developed a prototype of an algorithmic method to achieve this goal. During this period, we have refined the method, now published in (Laubenbacher and Stigler, 2004), and have implemented it (the paper was published in October 2004 and was Number 11 on the list of 25 most downloaded papers by February 2005). The computational engine for the software is the symbolic computation program Macaulay2 (Grayson and Stillman, 1999), and we are collaborating with one of the developers, M. Stillman, from Cornell University, to customize it for our purposes.

A crucial step in applying this method to experimental data is the initial discretization of the data. The number of allowed states and their assignment needs to be chosen so that key dependencies among the variables and network dynamics are preserved. This year we have developed such a method, and have implemented and tested it. The paper (Dimitrova et al., 2005) is almost ready for submission. Besides discretization of time series, the method can serve as a clustering method, with the novel feature that the number of clusters need not be specified as input.

The method in (Laubenbacher and Stigler, 2004) relies on an exact data fitting algorithm. This feature makes the method very sensitive to noisy data and leads to overfitting. To counteract this tendency, we have developed a prototype of a genetic algorithm that optimizes the model with respect to robustness to noise. This prototype is working well for binary data, but needs improvements in its mathematical foundation to be useful for larger state sets.

We have further improved the prototype of our analysis software package DVD (http://dvd.vbi.vt.edu) for discrete dynamical systems to be able to handle large linear systems, thanks to a new mathematical result (Hernandez, 2004), which we have transformed into an algorithm. A paper is in preparation. We have also made further progress in elucidating the relationship between the structure of models and their dynamics. In (Colón-Reyes et al., 2004) we determined the structure of special families of Boolean systems that do not have any oscillatory limiting behavior. We have now extended this result to systems over arbitrary finite fields. A paper is in preparation.

Collaborators on these projects include B. Pareigis, Dept. of Mathematics, University of Munich, Germany; M. Stillman, Dept. of Mathematics, Cornell University; and B. Sturmfels, Dept. of Mathematics, University of California, Berkeley.

Yeast Systems Biology

The method development described in the previous section is partially funded by NIH Grant R01 GM068947-01, *A new mathematical modeling approach to biochemical networks, with an application to oxidative stress in yeast.* The goal of the project is to develop novel methods from both continuous and discrete mathematics for the reverse-engineering of biochemical networks. These methods will be applied to genomics, proteomics, and metabolomics data from experiments specifically designed for this modeling approach, focusing on the regulatory network in *S. cerevisiae* responsible for oxidative stress response. The project is a collaboration between our group and the Mendes and Shulaev groups at VBI.

One main advance in this project during the reporting period is the generation and analysis of the transcription data from the experiments. A detailed description of the data can be found in the report from the Mendes and Shulaev groups. On the modeling side we have begun to work together with the Mendes group on simulated data from a network generated by

Mendes. It has many of the characteristics of the experimental system we are trying to model and includes over 60 different molecular species consisting of mRNAs, proteins, and metabolites. More details about the simulated network can be found in the Mendes report. We have made the first steps in trying to fuse the advantages of the continuous and discrete modeling approaches. Our first joint publication on this general topic is in press (Laubenbacher and Mendes, 2005).

Modeling of Neural Response Networks

Our most recent collaboration in the life sciences focuses on the application of the modeling methods described above to neural response networks to external stimuli. We are working with B. Komisaruk from the Rutgers University Radiology Department since November 2004. The goal of the project is to use our discrete top-down modeling techniques with time series of fMRI data to construct dynamic models of the interaction of different brain regions to external stimuli. Our collaborator's goal is to study methods for pain management in patients with completely severed spinal cords (Komisaruk et al., 2004). As a first step we have designed an experiment that generates data suitable for our modeling method from the well-known neural response network of the fingers reacting to temperature change. The network of brain regions involved in the response is well-understood and can serve as a test case for our method. Existing modeling approaches have mostly been statistical, using Bayesian networks and other statistical methods. Our approach, if applicable, can provide dynamic models of such response networks.

PathSim: Modeling the Immune Response to Viral Pathogens

The goal of this ongoing project is to make a three-dimensional spatial model of the immune response to Epstein-Barr virus in the Waldeyer's ring. The simulation is stochastic and agent-based. The project has recently been funded by a grant from NIAID, and is a collaboration with the Duca group, D. Thorley-Lawson from the Tufts University Medical School, and F. Castiglione from the National Research Council in Rome, Italy. Since the project is described in detail in the report from the Duca group, we will focus here on the more mathematical aspects. One long-term goal is to develop systematic methods for determining interventions that modify immune response in order to achieve desired outcomes. Our approach is to imitate the methodology of optimal control theory. The first step is to develop a state-space model from the stochastic agent-based model, using our reverse-engineering techniques. In engineering terminology, this corresponds to system identification. The next step is to use control theoretic techniques developed for polynomial systems over finite fields (Marchand and LeBorgne, 1998a and 1998b).

This year, as a proof-of-concept case study, we have carried out this program for a similar but much smaller system, namely the control of *in vitro* competition of two virus strains. During the last reporting period we succeeded in solving the problem posed by Duca initially, namely to explain the spatial differentiation of the two strains. The observed phenomenon of strong segmentation is unexpected. We determined that the underlying mechanism is a "winner-take-all" principle that is at work in both stochastic and deterministic agent-based models of the process.

We then applied our system identification techniques to the simulation to obtain an equation-based model into which we introduced control variables based on feasible external interventions to help one strain out-compete the other. This work was presented at the 2004 IEEE Conference on Decision and Control (Jarrah et al., 2004).

We are now working on methods to determine appropriate sampling techniques for the model building process applied to the much larger PathSim simulation.

MicroBlast: A Combinatorial Tool for Microarray Analysis

Molecular mechanisms are the principal components upon which cellular functions rely, and are the primary objects of investigation when studying cellular activities. Methods capable of tractably analyzing and integrating large data sets of heterogeneous information are needed to unravel the complex interactions between these molecular mechanisms in order to understand different cellular functions. Potential types of information include sequence data (DNA, RNA, protein), expression data (microarray data, proteomic data), structural classification (genetic networks, metabolic pathways), and functionality (protein motifs, homology). This project, a collaboration with the Duca group, is focused on the development of a method, which we call "MicroBlast," that integrates gene expression values with biological functional information in order to make global comparisons of samples. The integrated data is represented as a graph and, using appropriate graph measures, a reference experiment can be compared to samples from a database of similar experiments, and a ranking of similarity is returned. The validity of our method is supported by its implementation on data sets of both simulated and reported microarray experiments. The project is described in more detail in the report from the Duca group. We are now ready to resubmit a paper on the method, satisfying a journal request to provide a software implementation available to referees. We have also submitted an R21 grant application to the NIH in support of further development.

The main progress achieved during the reporting period on the mathematical aspects of the project consists of a better understanding of the mathematical foundation that underlies the use of Formal Concept Analysis to the analysis of biological datasets. This understanding has led to additional mathematical tools that can be used to analyze the combinatorial objects associated to an experiment. We are exploring those in collaboration with H. Barcelo from the Arizona State University Mathematics Department. We have also begun a collaboration with the Tyler and Mendes groups to use MicroBlast to study time scale differences in oxidative stress response of several organisms, including yeast.

The Topology of Large-Scale Graphs

The analysis of large graphs is an important problem in several application areas, such as communications, biochemical network analysis, and the analysis of social networks. An exploratory project we are pursuing is the development of tools from combinatorial topology for the analysis of graph connectivity. Since the tools are to be applied to graphs with millions of nodes, algorithm implementation is of crucial importance. The origin of our approach is the mathematical theory of Q-analysis (Atkin, 1974), which measures different levels of connectivity in labeled (directed or undirected) graphs. One can also associate higher dimensional geometric structures to such graphs and compute measures of their topology, using an extension of Q-analysis to higher dimensions (Barcelo et al., 2001).

Our project is focused on the analysis of large-scale social contact networks, such as those that are generated by the EpiSims project of the Barrett group. These graphs on average have in excess of a million nodes and several million edges. Our goal is to compute measures of the connectivity and persistence of the contact network for all nodes in the graph (or an appropriate sample), with the aim of identifying "unusual" network profiles. One possible application is to the identification of groups of people in urban areas that merit particular attention in the case of an infectious disease outbreak.

During the reporting period, we have completed an extremely fast implementation of all relevant algorithms in C++. We have designed a grid architecture to utilize a distributed computation environment and are running benchmark tests.

A Computational Algebra Approach to Magnetosphere Physics

The most recent project to report on is a collaboration with H. Karimabadi, a space physicist at the University of California, San Diego. The ultimate goal in space physics is to understand how solar wind transfers its mass, momentum, and energy to the magnetosphere. This deceptively simple question has kept scientists at bay for over 50 years. The interaction of solar wind with the Earth's magnetosphere has turned out to be much more complex than originally thought and involves interconnected processes spanning many orders of magnitude in spatial and temporal scales. Aside from its intellectual merits, the understanding of the details of this interaction is of great practical relevance. One of the main obstacles to a major breakthrough in this problem remains the issue of how to incorporate the microphysics of the reconnection process into the global magnetospheric codes.

We are pursuing a unique approach to this problem that relies on two innovations: (1) *Empirical approach*-perform full particle simulations under a variety of geometries and generate a large dataset describing the time evolution of the reconnection electric field as a function of local variables and (2) *Reverse Engineering*-Adapt and apply recently developed computational algebra techniques to the simulation dataset generated above in order to generate dynamical system equations (analytical) describing the reconnection electric field based on the state of the system at any given time. Our preliminary application of this technique has been quite promising. A joint paper, to be submitted to *J. Geophys. Res.,* is in preparation. We have also submitted a grant proposal to the NSF that would support the further development of our mathematical tools.

References

Atkin R (1974) An algebra for patterns on a complex. *Intl J Man-Machine Studies* **6**: 285–307

Barcelo H, Kramer X, Laubenbacher R, and Weaver C (2001) Foundations of a connectivity theory for simplicial complexes. *Annals of Appl Math* **26**: 97–128

Colón-Reyes O, Laubenbacher R, and Pareigis B (2004) Boolean monomial dynamical systems *Annals of Combinatorics* **8**: 425-439

Dimitrova E, McGee J, and Laubenbacher R (2005) A graph-theoretic method for the discretization of gene expression measurements 18pp., *Bioinformatics*: to be submitted

Grayson D and Stillman M (1999) *Macaulay2, http://www.math.uiuc.edu/Macaulay2/.*

Hernandez Toledo RA (2004) Linear finite dynamical systems, preprint

Jarrah A, Vastani H, Duca K, and Laubenbacher R (2004) An optimal control problem for *in vitro* virus competition. Proc. 43rd IEEE Conference on Decision and Control, The Bahamas

Komisaruk BR, Whipple B, Crawford A, Grimesa S, Liuc W, Kalnin A, and Mosier K (2004) Brain activation during vaginocervical self-stimulation and orgasm in women with complete spinal cord injury: fMRI evidence of mediation by the Vagus nerves. *Brain Research* **1024**: 77–88

Laubenbacher R and Mendes P (2005) A discrete approach to top-down dynamical modeling of biochemical networks, in *Comp Sys Biol,* A. Kriete and R. Eils (Eds.), Elsevier, in press

Laubenbacher R and Stigler B (2004) A Computational Algebra Approach to the Reverse Engineering of Gene Regulatory Networks. *J Theor Biol* **229**: 523–537

Marchand H and LeBorgne M (1998) On the Optimal Control of Polynomial Dynamical Systems over Z/p. *in Fourth workshop on Discrete Event Systems*, Cagliari, Italy, IEEE

Marchand H and Le Borgne M (1998) Partial order control of discrete event systems modeled as polynomial dynamical systems. *in IEEE International Conference on Control Applications*, Trieste, Italy

Publications

Polys NF, Bowman DA, North C, Laubenbacher R, and Duca K (2004) PathSim Visualizer: an information rich virtual environment framework for systems biology. Proc. SIGGRAPH Web3D Session, Los Angeles

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

Jarrah A, Vastani H, Duca K, and Laubenbacher R (2004) An optimal control problem for *in vitro* virus competition, Proc. Of the 43rd IEEE Conference on Decision and Control, Bahamas

Colón-Reyes O, Laubenbacher R, Pareigis B (2004) Boolean Monomial Dynamical Systems. *Annals of Combinatorics* **8**: 425–439

Laubenbacher R and Mendes P (2005) A discrete approach to top-down modeling of biochemical networks. Book chapter in *Comp Sys Biol,* R. Eils and L. Kriete (Eds.), Elsevier, in press

Barcelo H and Laubenbacher R (2004) Perspectives in A-theory. *Disc Math,* in press

Functional Genomics of Fungal-Host Interactions

Christopher B. Lawrence
Research Associate Professor
Associate Professor of Biology, Virginia Tech
lawrence@vbi.vt.edu

Yangrae Cho, Joshua Davis, Carlos Mauricio La Rota

Introduction

Among the most destructive plant diseases are the so-called "rots", caused by necrotrophic fungi that 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 percent of fungal diversity, they cause ~80 percent of foliar losses due to fungal diseases in some parts of the world (R. Oliver, Director, The Center 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 the production of low molecular weight, host-specific/selective, phytotoxic secondary metabolites. Moreover, there are specific examples of species 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 and do not produce host-specific toxins, (although some produce other, non host- specific phytotoxins). 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, with the recent completion of the *Magnaporthe grisea* and *Fusarium graminearum* whole genome sequencing and associated high-throughput functional genomics projects of international proportion, more insight has been gained into virulence mechanisms employed by these fungi for rice and wheat infection, respectively.

As mentioned above, 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 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*.

The *Alternaria brassicicola* – Brassicaceae Pathosystem

Brassicaceae, the crucifer plant family, consists of approximately 3,500 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 *B. oleracea* (vegetables), *B. rapa* (vegetables, oilseeds, and forages), *B. juncea* (vegetables and seed mustard), and *B. 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 and Earle, 1998; 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 and Earle, 1998; 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 and 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* have 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 and Hammerschmidt, 2002, reviewed by Thomma, 2003; Lawrence et al., unpublished). 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 et al. (1996) 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*, an antifungal defensin-like peptide. 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 et al. (1998) demonstrated that the methyl jasmonate insensitive mutant, *coi1-1*, is more susceptible to *A. brassicicola* than wild type Columbia (Col-0). Mutation of *PAD3*, a gene encoding a cytochrome P450 monooxygenase essential for synthesis of the Arabidopsis phytoalexin, camalexin, are also more susceptible (Thomma et al., 1999; Zhou et al., 1999). These studies clearly suggest that jasmonic acid (JA) 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 et al., 2003). Thomma et al. (1998) also found that *NahG* plants (which have markedly reduced salicylic acid (SA) levels) remain resistant to *A. brassicicola*, indicating that SA is not required for resistance. 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 et al. (2000) 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. (2003) have also examined gene expression in distal uninoculated tissues during *A. brassicicola* infection, finding 35 genes with altered expression. Van Wees et al. (2003) 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 early stages of defense signaling. Interestingly, 265 of the 645 *A. brassicicola* induced genes identified by Van Wees et al. (2003) required *COI1* for full expression, suggesting a major role of JA signaling in responses to *Alternaria* infection. 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.

Alternaria-Brassicaceae Functional Genomics

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 two major areas: 1) large-scale generation and analysis of expressed sequence tags (ESTs) derived from various *A. brassicicola*-host interactions and 2) functional analysis of fungal pathogenicity. In this regard we have generated over 17,000 ESTs from various cDNA libraries (Table 1).

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 (http://www.broad.mit.edu/). We have also performed Interpro analysis (http://www.ebi.ac.uk/interpro/). InterPro is a database of protein families and includes domains and functional sites in which identifiable features found in known proteins can be applied to unknown protein sequences. Results of these analyses and reviewing 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

Table 1. EST Libraries.

Tissue Source	Library Type	Approximate # of ESTs	Approximate # of unisequences (singletons + contigs)
A. brassicicola spores on Arabidopsis leaf	Subtracted cDNA	96	41
Early stage of canola infection (*B. napus*)	cDNA	3,500	Not determined
Late stage rapeseed infection (*B. rapa*)	cDNA	6, 900	2, 450
Mixture of various stages of cabbage infection (*B. oleracea*)	Subtracted cDNA	4, 225	3, 112
Nitrogen starved *A. brassicicola* mycelia	cDNA	3, 260	1, 660

host genes is only just beginning. However, we are particularly interested in host genes involved in programmed cell death (PCD), since this process is thought to be required for infection by necrotrophic pathogens.

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, 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 (KO). Gene knockout plays a critical role in identification of fungal pathogenicity/virulence factors. Unlike 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.

To improve the efficiency, as well as to expedite gene knockout construct production, we used a relatively short (<3 kilobasepairs), PCR amplified linear construct with minimal components, an antibiotic resistance marker gene (hygromycin resistance), and a 250-600 nucleotide long partial target gene truncated at 5' and 3' ends. Using standard PEG-mediated transformation of protoplasts combined with a heatshock step, the minimal construct consistently produced stable transformants for diverse categories of genes. At least 80 percent of transformants were targeted gene disruption mutants, compared to inconsistent transformation and less than 20 percent targeted gene disruption with circular plasmid constructs. Targeted gene disruption with the linear construct occurred by a single crossover event, following a circularization of the linear construct. Each mutant has a unique molecular signature thought to originate via endogenous exonuclease activity. This method is advantageous for high throughput gene knockout, overexpression, and reporter gene introduction within target genes, especially for asexual filamentous fungi like *Alternaria* where genetic approaches are unfavorable. The diagram (Figure 1) shows a pictorial of this process.

Using Southern blot analysis and PCR we have confirmed that over 80 percent of transformants using this method are typically disrupted at the locus of interest. Figure 2 shows an example of these results for one of the genes targeted for functional analysis. In this case, 100 percent KO efficiency was obtained. Figure 3 shows an example of a mutant with a complete loss of pathogenicity when inoculated onto Brassica

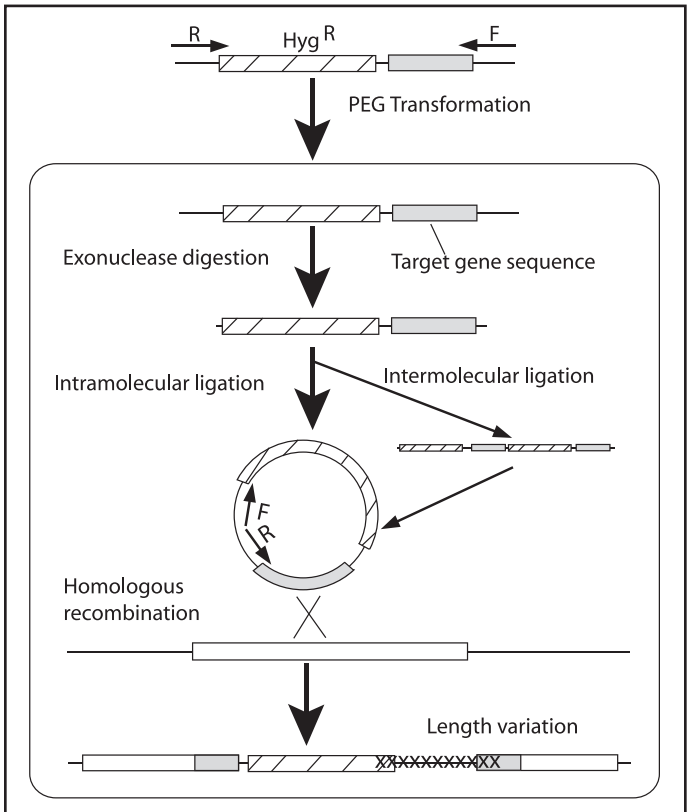


Figure 1. Schematic of mechanism of targeted gene knockout (KO) in *Alternaria brassicicola*.

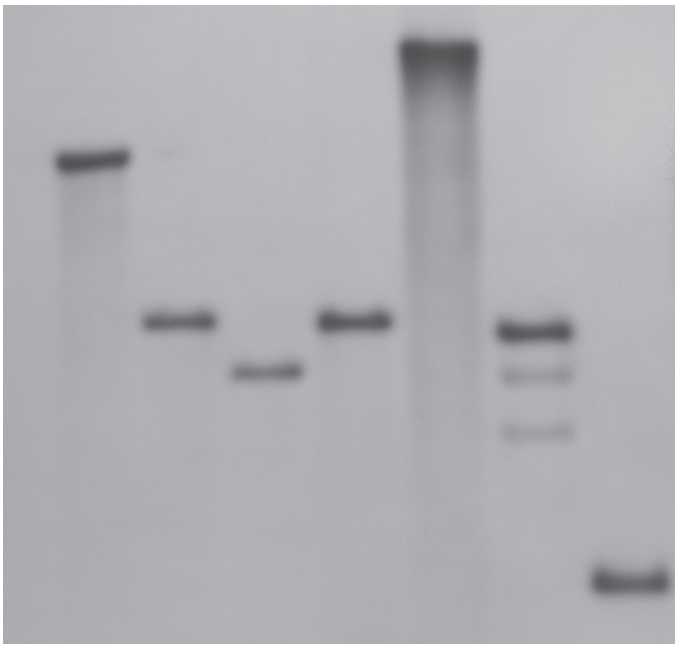


Figure 2. Southern blot analysis demonstrating gene disruption. W=wildtype genomic DNA digested with PstI and probed with labeled cDNA corresponding to gene of interest. 1-6 = digested genomic DNA from individual mutants. Notice change in MW of hybridizing bands in mutants compared to wildtype indicating insertion at target locus.

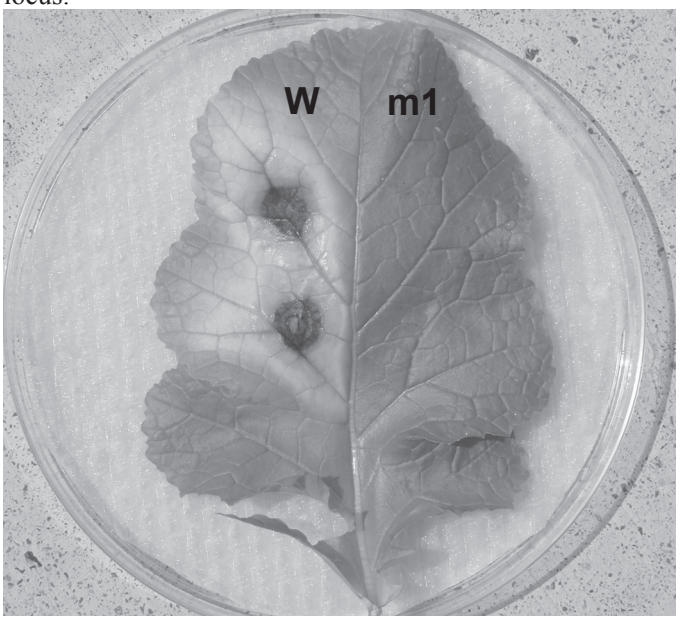


Figure 3. Pathogenicity assay of gene disruption mutant. Left side of leaf was inoculated with wildtype (w) fungus, right side with a mutant (m1) disrupted in a gene of interest (20 ul of 10^5 spores/ml for w and mutant). Notice the complete lack of lesion formation by mutant.

leaves demonstrating that we can obtain and identify non-pathogenic mutants.

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. With the recent funding of the *Alternaria brassicicola* genome sequencing project by the 2005 NSF-USDA Interagency Microbial Genome Sequencing Program (PI Lawrence), we anticipate that our high throughput functional analysis pipeline will be extremely useful to researchers worldwide interested in analyzing genes involved in necrotrophic fungal pathogenicity.

Other Projects

Alternaria, Allergy, and Asthma

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 et al., 1992; Halonen et. al., 1997). In addition, up to 70 percent 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 et al., 1992; Halonen et. al., 1997; O'Halleren 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 US 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 the surveys highlighting the ubiquitous nature of this genus. Fungal exposure differs from pollen exposure in quantity (airborne spore counts are often 1,000-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, *Alternaria alternata* (Sanchez and Busch, 2001). Our laboratory was the first to identify the major allergen homolog Alta1 in *A. brassicicola*, a species other than *A. alternata* (Cramer and Lawrence, 2003). Moreover, we have recently determined 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 a current project in our laboratory, the secretome 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 profiled such as the production of antimicrobial proteins, chemokines, cytokines, and the expression patterns of Toll-receptor genes. In addition, we will investigate gross ultrastructural changes in treated cells. We believe the interaction of secreted *Alternaria* allergens 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 *Alternaria alternata* spores been shown to contain the major allergen, Alta1, secretion of this protein has been reported to dramatically increase during germination (Mitakakis, 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. There is clearly a need to elucidate the role of *Alternaria* immunoreactive proteins in the development of allergy/asthma from both diagnostic and immunotherapeutic perspectives.

References

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

Cramer R and 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

Gergen PJ and 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, and 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: AL-Doory, Domison (Eds.), Mould allergy Philadelphia, Lea & Febinger 104-116

Hong SG, Cramer RC, Lawrence CB, and 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 and 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, and Ciuffetti LM (2002) Chromosome- based molecular characterization of pathogenic and non-pathogenic wheat isolates of *Pyrenophora tritici repentis*. *Fungal Gen and Biol* **37:** 180–189

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

O'Hallaren MT, Yunginger JW, and Offord KP (1991) Exposure to an aeroallergen as a possible precipitating factor in respiratory arrest in young patients with asthma. *N Engl J Med* **324:** 359–363

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

Peat JK, Tovey CM, Mellis CM, Leedre SR, and 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.

Penninckx IA, Eggermont K, Terras FR, Thomma BP, De Samblanx GW, Buchala A, Metraux JP, Manners JM, and 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 and 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, and 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, and Maclean DJ (2003) Systemic gene expression in Arabidopsis during an incompatible with *Alternaria brassicicola. Plant Phys* **132:** 999–1010

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

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

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, Eggermont K, Penninckx IAMA, Mauch-Mani B, Vogelsang R, Cammue BPA, and 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 BPHJ (2003) Alternaria spp. from general saprophyte to specific parasite. *Mol Plant Pathol* **4:** 225–236

Tierens KFMJ, Thommam BPHJ, Barim RP, Garmier M, Eggermont K, Brouwer M, Penninckx IAMA, Broekaert WF, and 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 and 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, and 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, and Glazebrook J (1999) Arabidopsis *PAD3*, a gene required for camalexin biosynthesis, encodes a putative cytochrome P450 monooxygenase. *Plant Cell* **11:** 2419–2428

Publications

Pruss GJ, Lawrence CB, Bass WT, Li QS, Bowman LH, and Vance V (2004) The potyviral suppressor of RNA silencing confers enhanced resistance to multiple pathogens. *Virology* **320**: 107–120

Cramer R and Lawrence CB (2004) Identification of *Alternaria brassicicola* genes expressed *in planta* during pathogenesis of *Arabidopsis thaliana*. *Fung Genet Biol* **41**: 115–128

Gent DH, Schwartz HF, Ishimaru CA, Louws FJ, Cramer RA, and Lawrence CB (2004) Polyphasic characterization of onion strains of Xanthomonas campestris. *Phytopathol* **94**: 184-195

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

Funnell DL, Lawrence CB, Pedersen JF, and 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* in press

Cramer RA, Thon M, Cho Y, Craven KD, Knudson DL, Mitchell TK, and Lawrence CB (2005) Analysis of Expressed Sequence Tags Derived from a Compatible Alternaria brassicicola—Brassica oleracea Interaction. *Mol Plant Pathol* accepted for publication

Novel Microfluidic Architectures for Proteomics and Mass Spectrometric Detection

Iuliana M. Lazar

Research Assistant Professor, VBI
Assistant Professor of Biology, Virginia Tech
lazar@vbi.vt.edu

Hetal Sarvaiya, Phichet Trisiripisal, Jung Hae Yoon

Introduction

One of the greatest challenges in biological research today is providing a reliable and comprehensive characterization of all protein constituents in a cell. The recent past has witnessed intensive efforts to develop novel technologies that can handle the complexity (thousands of proteins/sample), a wide range of concentrations (dynamic range of $1:10^6$), low level expression (less than 1000 copies/cell), and the dynamic composition (different sets of proteins are expressed in various stages of cell development) of cellular protein extracts. Mass spectrometry (MS) has evolved into an irreplaceable tool for the characterization of protein samples as it combines the benefits of specificity, sensitivity, and resolving power. The ability to characterize the expressed proteins within a cell involves lengthy protocols and a substantial amount of work. The preparation of these samples for MS analysis concludes, generally, with the generation of tens or hundreds of sample sub-fractions in the 5-10 μL and low pM-μM concentration range. Novel technologies with parallel processing capabilities that enable fast and sensitive investigations have not been developed so far.

Microfabricated devices have evolved into ideal analysis platforms for minute amounts of samples, and present promising applications for proteomic investigations. Miniaturized devices enable process integration, multiplexing, automation, high-speed analysis, and ultimately high-throughput sample processing. We are developing microfluidic bioanalysis platforms for advanced proteomic investigations that are compatible with electrospray ionization (ESI) and matrix assisted laser desorption ionization (MALDI)-MS detection. These microfabricated architectures integrate an array of functional elements, have a standalone configuration and enable con-tamination free analysis. It can be foreseen that, as a result of the superior analytical capabilities of microfabricated devices, this technology may become the optimum platform for fast, sensitive, and high-throughput handling of minute amounts of proteomic samples. The applicability of microfluidic devices can be envisioned in practically any field of biological, biomedical, biochemical, and pharmaceutical sciences.

Research in the Lazar lab is focused on two major topics:

1. The development of mass spectrometric methods and protocols for qualitative and quantitative mapping of protein components in cellular extracts, differential protein expression analysis, and characterization of posttranslational modifications (phosphorylation and glycosylation). The MCF7 cancer cell line is used as a model system.
2. The development of microfluidic platforms with mass spectrometric detection for proteomic applications, which includes the design, development, and integration

of functional elements (sample propulsion elements, microreactors, mixers, 2D-separation systems, MS interfaces, multiplexed architectures, etc.);
and bioanalytical process implementation on the chip, which includes sample cleanup, prefractionation, preconcentration, labeling, digestion, and separation. Emphasis is placed on developing chemistries based on affinity interactions for capturing, purifying, labeling, and immobilizing peptide or protein components.

Methods

Microfluidic chips are fabricated in-house from glass substrates pre-coated with chrome and photoresist (Nanofilm). Microchannels are etched to a depth of 1.5-50 micrometers in both substrate and cover plate. Fluidic propulsion is accomplished using electrically and pressure driven mechanisms. Fluidic manipulations are optimized with a Nikon epi-fluorescent microscope. Capillary separation columns are prepared from reversed phase C18 packing using 3 µm and 5 µm particles. Mass spectrometric detection is accomplished using electrospray ionization with an ion trap LTQ system (Thermo Electron). MCF7 cancerous cells are typically cultured to 70 percent confluence, harvested, lysed, and the soluble fraction is digested with trypsin and fractionated using strong cation exchange (SCX) separation columns. The SCX fractions are analyzed using an Agilent micro liquid chromatography (LC) system and microfluidic chips.

Results and Discussions

In this presentation, we are reporting on a microfluidic analysis platform that integrates a 2D separation system and that is being developed for the characterization of the MCF7 cancerous cell line. The soluble fraction of the extract was processed according to a shotgun protocol and, after digestion with trypsin and SCX pre-fractionation, the sample sub-fractions were analyzed with a benchtop LC system to evaluate the optimum conditions that generate a maximum number of proteins identified with high confidence. Optimization parameters were sample related (6), data acquisition related (8), and database search related (4). Sequest filtering parameters were as follows: singly charged peptides must have a cross correlation score Xcorr > 1.9, doubly charged tryptic peptides must have Xcorr > 2.2, and triply charged tryptic peptides will be accepted if their Xcorr > 3.75. High quality MS/MS spectra that will meet the above-described criteria will be stored in an in-house generated database, and will be used for future reference in the analysis of other samples. A typical complex peptide mixture separation that is generated using this analysis system is shown in Figure 1.

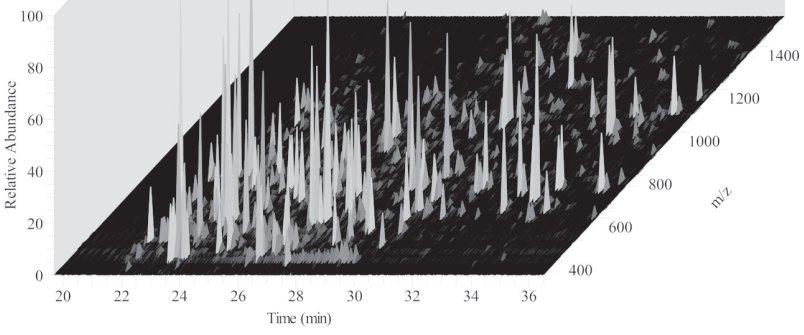


Figure 1. LC-MS/MS analysis of a complex peptide mixture.

Low femtomole detection levels from pico/nanomolar level solutions is

accomplished routinely using the above described parameters. To date, we have identified over 200 proteins in the MCF7 extract, of which about 20 were described in the literature as potential biomarkers that were differentially expressed between normal and cancerous cells. A representative MS/MS spectrum of the 1024.48 $^{2+}$ parent ion that was present in relatively low abundance, is shown in Figure 2. Once the overall analysis parameters are optimized, the protocols are transferred to a microfluidic chip. The ultimate goal is to design completely stand-alone devices that can handle a large number of analytical sample processing steps. An integrated microfluidic system with MS detection is shown in Figure 3.

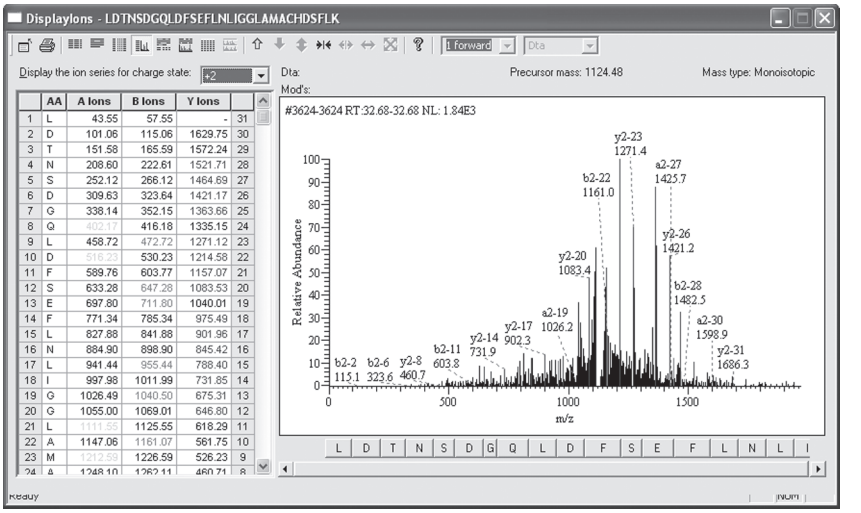


Figure 2. Tandem mass spectrum of the 1124.48^{2+} parent ion from the MCF7 cellular extract. Highlighted in the spectrum are doubly charged "a," "b," and "y" ions. Complementary information is provided by the singly and triply charged ions that are shown in alternate spectral views.

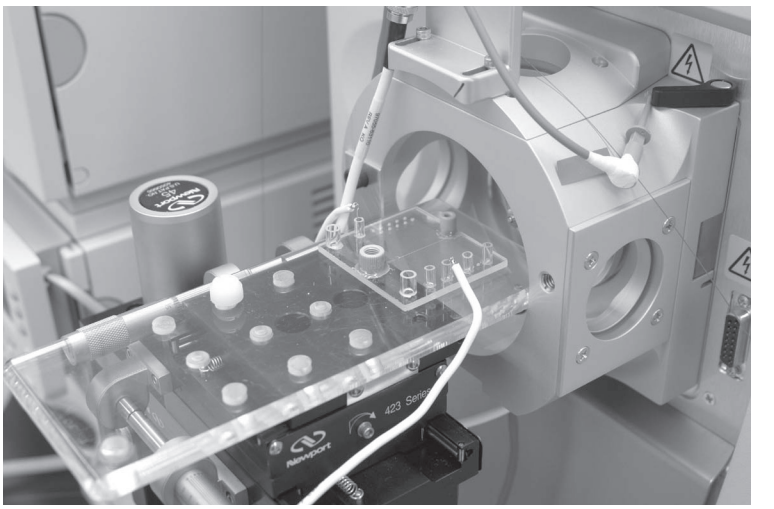


Figure 3. Photograph of a microfluidic bioanalysis system.

The novel aspects that differentiate our microfluidic platform from previous designs are the following: (1) the proposed microfluidic devices will lead to a new paradigm for investigating complex samples and designing novel interfaces and sample introduction approaches to MS; they will comprise not just a few but a series of functional elements that perform sample preparation prior to MS detection (sample clean-up, preconcentration, affinity selection, digestion, and separation); part of this analysis scheme will be a fully integrated LC separation system; electrically and pressure driven fluidic manipulations will be possible; (2) the microfluidic chips will be compatible with both ESI and MALDI-MS detection; this is an essential feature that will enable the interfacing of these chips to a variety of mass spectrometers capable of performing MS/MS investigations and delivering peptide/protein sequence information; (3) the microfluidic chips

will have a stand-alone configuration to enable the full exploitation of benefits associated with miniaturization; all components necessary to perform electrically and pressure driven fluidic manipulations will be integrated on the chip; no external assistance will be necessary to operate these devices; (4) the microchip platforms will have a multiplexed layout to enable the preparation of high-throughput, contamination free, and fully-disposable devices; (5) simplicity will result in reduced cost and labor.

Essential to have stand-alone configuration, chips are the mechanisms used for fluidic propulsion. Fluidic manipulations on our chips is accomplished using electrically and pressure driven mechanisms. Pressurized fluid flows are created with the aid of EOF pumps. The choice for these pumps was dictated by two reasons: first, the EOF pumps are the only miniaturized pumping systems that can generate high pressures (tens/hundreds of bars), and second, their manufacturing is extremely simple, necessitating only wet/dry etching protocols. Accurate calculations for the pumping system allow an appropriate design that can sustain the flows and pressures that are necessary for moving the fluids through the entire microfluidic network. The EOF pump has a multiple open-channel configuration that consists of hundreds of parallel, small diameter microchannels. Pumps with microchannels (1-2 μm in depth, 20-50 mm in length) that deliver flow rates of 10-200 nL/min at pressures of 5-10 bar are usually constructed (Figure 4).

Presently, the micro-pump that operates the LC system is comprised of 400 pumping channels, 20 mm long and ~ 1.5 μm deep, and delivers flow rates at approximately 80 nL/min. The valving system contains 100 microchannels with similar dimensions to the pump. The separation channel is 20 mm long, ~ 50 μm deep, and is typically filled with Poros packing material. With a volatile LC eluent (NH_4HCO_3 in H_2O/CH_3OH), peptide separations as shown in Figure 5 can be accomplished.

Previous experience with MS detection for complex peptide samples has shown that both ESI and MALDI-MS should be used since complementary information is provided by the two techniques. In addition, novel developments in MALDI-MS, i.e., the introduction of tandem MS and atmospheric pressure ionization capabilities, significantly increase the power of MALDI-MS investigations. Moreover, MALDI-MS is an ideal detection tool for high-throughput microchip applications, since it enables simultaneous collection of samples from parallel structures on the chip. Consequently, microchip configurations are developed to facilitate MALDI-MS as well. This task is accomplished by using the chips developed for ESI-MS, with the distinction that the sample is electrosprayed or deposited in discrete spots creating arrayed structures on the MALDI plate.

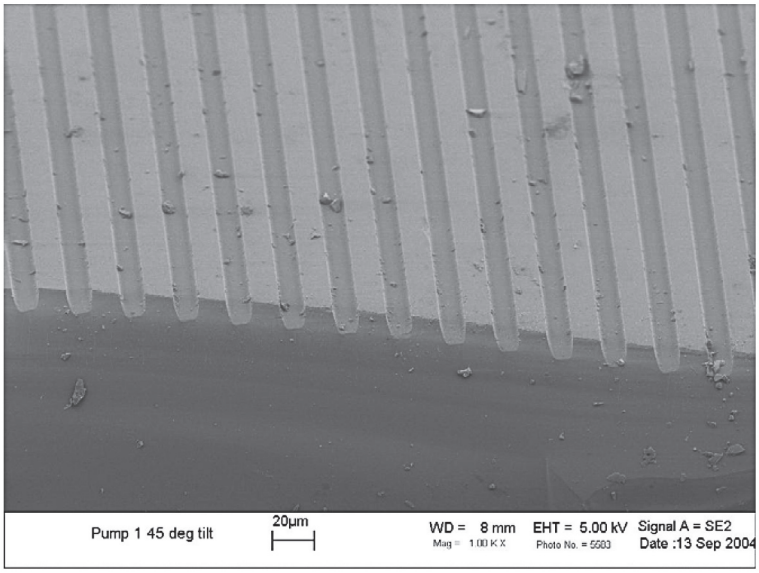


Figure 4. SEM representation of a multichannel electroosmotic pumping system.

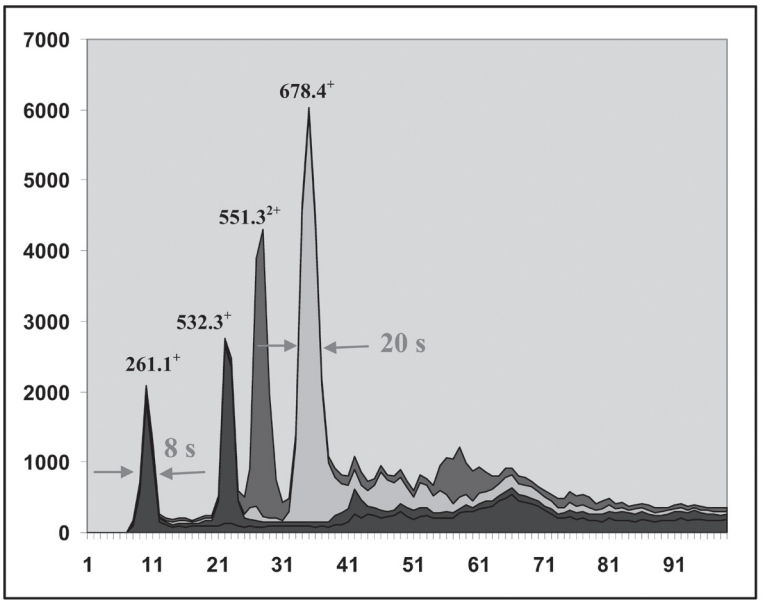


Figure 5. Microfluidic LC-MS separation of peptides. Peak assignments: 261.1^{+} (GGK), 532.3^{+} (AAWGK551.3^{2+}), 551.3^{2+} (VDEVGGEALGR), and 678.4^{+} (YIPGTK).

For typical flow rates for nano-LC of ~100 nL/min, and say, average peak widths of 30 s, one peak/spot is deposited in 30 s. One hundred spots, or the whole MALDI plate, can be generated in approximately 1 hour. Of course, according to these calculations, for a MALDI plate of 96 spots, the LC run will be limited to 1 hour, while for a MALDI plate of 384 spots, to 4 hours, respectively.

A number of analytical tools are designed to be disposable in order to prevent sample contamination, carryover, and false positive identifications. While at the present time commercially available microfluidic devices are rather expensive, and their use as disposable devices may be considered exotic, the experience provided by the microelectronic industry points otherwise. The apparatus that is necessary for the fabrication of these chips is similar to the one that is utilized for the fabrication of microelectronic equipment, and parallelization and large-scale integration of analytical processing steps will result in a similar decrease of investment in the manufacturing of these chips. Once an efficient and reliable workflow is designed, implemented, tested and demonstrated to perform adequately on the chip, an entire process can be replicated with the same ease and effort as required for the fabrication of a single, isolated, processing component. Thus, the idea of a "disposable Lab-on-a-Chip" device becomes acceptable.

Conclusions

The research outlined in this presentation will lead to the development of novel micro-analytical systems and detection strategies that will enable the generation of high quality proteomic information in practically any field of biological, biomedical, biochemical, and pharmaceutical sciences. The value of this approach arises from the superior analytical capabilities of microfabricated devices that can provide the optimum platform for fast, sensitive, and high-throughput handling of minute amounts of proteomic samples.

Developing Strategies for Systems Biology

Pedro Mendes
Research Associate Professor, VBI
Adjunct Associate Professor of Biochemistry, Virginia Tech
mendes@vbi.vt.edu

Jessica Caldwell, Diogo Camacho, Alberto de la Fuente, Stefan Hoops, Aejaaz Kamal, Christine Lee, Xing Jing Li, Ana Martins, Bharat Mehrotra, Wei Sha, Helen Shibru, Gaurav Singh, Anurag Srivastava, Sameer Tupe

Introduction

Research is moving away from the molecular biology paradigm to an approach characterized by large-scale molecular profiling of living cells. Technologies for transcript, protein, and metabolite profiling produce measurements that are large-scale characterizations of the state of biological material. What makes these different from traditional molecular biology experiments is that, by nature of measuring so many components of the cell, one does not have to decide a priori which ones are more likely to be related with phenomena of interest. Instead, we now have the luxury of making unbiased observations. Application of "omics" technologies has been largely discussed in terms of identifying the function of open reading frames—the functional genomics agenda. In this sense "function" usually means the molecular action of that protein (Oliver, 1996). An alternative view takes gene "function" to be the collection of interactions that its products have in the cell (Brazhnik et al., 2002). Accordingly, a major research objective is the identification, quantification, and modeling of those interactions—the systems biology agenda.

Research in this Biochemical Networks Modeling Group is in line with the systems biology agenda. Early on, we started researching how best to approach it from various independent angles, including software for storage, visualization, and analysis of data; reverse-engineering methods; and modeling and simulation tools. During last year we began tracing strategies for convergence of these methods, to result in a set of methods that will allow the extraction of knowledge from omic data in the form of quantitative models. The remainder of this manuscript describes our research and accomplishments during the period from April 2004 to March 2005. References published by us in this period are indicated in citations with # and listed separately.

Modeling and Simulation of Biochemical Networks

Software for modeling and simulation - COPASI

Our biochemical simulator Gepasi (www.gepasi.org) continued to be widely used by the scientific community, with 55 articles published in this period using it as a research tool. But Gepasi is the product of evolutionary development and is hard to extend to new functions. In a collaboration with the Kummer group at EML Research (Heidelberg), we are developing a new simulator, COPASI, to replicate the functionality of Gepasi and to be able to be extended with new functions. The first public test version of COPASI was released in October 2004 at the 5th International Conference on Systems Biology, and new versions have been released monthly thereafter. The software was demonstrated in March 2005 at the 3rd Symposium on Computational Cell Biology with great success. It has now been downloaded more than 1,600 times from our web site (www.copasi.org).

The February 2005 release of COPASI already includes all the functionality of Gepasi, except optimization and fitting. It also includes new features, like stochastic integration by the Gillespie algorithm (Gillespie, 1976), sliders

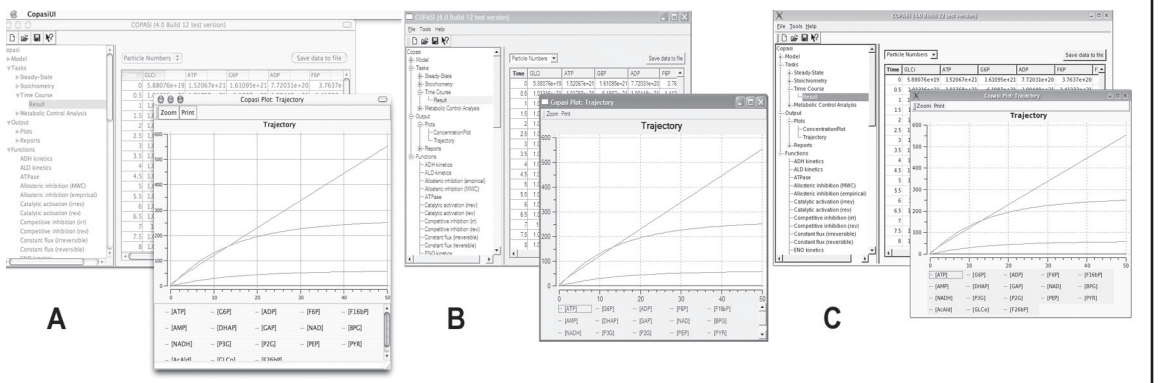


Figure 1. COPASI biochemical simulator under three different operating systems: Apple OS X (A), Microsoft Windows XP (B), and Linux (C).

for interactive parameter adjustments, and new plotting options. COPASI is a multi-architecture application and versions are supplied for Linux, Solaris, OS X, and Windows. The portability was achieved by using the C++ QT library (Trolltech AS, Oslo) with excellent results: not only does this allow for different developers to be working on different environments, but also the software behaves exactly the same on all platforms while keeping the look and feel of each one (Fig. 1).

Software interoperability is a major concern in bioinformatics, and this is true also in modeling and simulation. We have been involved in a community effort that resulted in the specification of the Systems Biology Markup Language (SBML, Hucka *et al*., 2003). This continued in 2004, when we participated in the 9th SBML Workshop and in the 2nd SBML Hackathon. COPASI imports and exports SBML level 2 (Finney and Hucka, 2003) and so it is able to exchange models with some other 75 software packages.

Synthetic data sets for method development

A large number of algorithms has been proposed for analysis for large-scale "omic" data (Quackenbush, 2001). Often, their application and expected outcomes are poorly justified, and they are not compared to other analyses objectively. To overcome these difficulties, it becomes important to establish reference data, both experimental and simulated. We have previously published some synthetic data sets: one for testing parameter estimation algorithms (3-enzyme pathway, Mendes, 2001; Moles *et al*., 2003) and a larger one for gene expression data analyses (AGN, Mendes *et al*., 2003). The 3-enzyme pathway contains all three levels of biochemical organization but is small (3 transcripts, 3 proteins, and 2 metabolites), while the AGN data sets are larger (up to 1,000 transcripts), but contain only gene transcripts. Intermediate networks containing the three levels of organization are also needed.

A new abstract biochemical network, known as the Claytor Network and depicted in Figure 2, was created to produce simulated systems biology experiments. The Claytor Network includes common biochemical structures, like redox chains, biosynthesis, competitive substrate usage for energy and reducing equivalents, and signaling. This network was inspired by the glutathione oxidative stress response system of *S. cerevisiae*, which we research in our NIGMS-funded systems biology project. An interesting fact in the development of the Claytor Network was that, although we had perfect knowledge of its details, it was very hard to predict which changes to effect to obtain particular behaviors. This is humbling since it demonstrates that current knowledge about biochemical dynamics is very incomplete and much is still to be learned—even for an abstract network like in Figure 2. Data from this network was used in activities described below and by the Laubenbacher group.

Towards a top-down modeling strategy

It has been recognized (e.g. Loomis and Sternberg, 1995) that modeling is necessary to transform omic data into knowledge. While biochemical network modeling methods have existed for many decades, they are not optimal for large-scale data sets. A new modeling approach is needed to best suit large-scale profiling experiments, and we argue that this should be conducted top-down. The idea is to first capture a coarse-grained image of the system and then, by iterations of simulation and experiment, increase the mechanistic detail of the model. The development of methods enabling this new approach forms the backbone of our research. We view this strategy as a stepwise approach; at each step, a new characteristic of the system is identified, which will then determine the structure of the problem to be solved in the following step. Table 1 delineates this top-down modeling strategy.

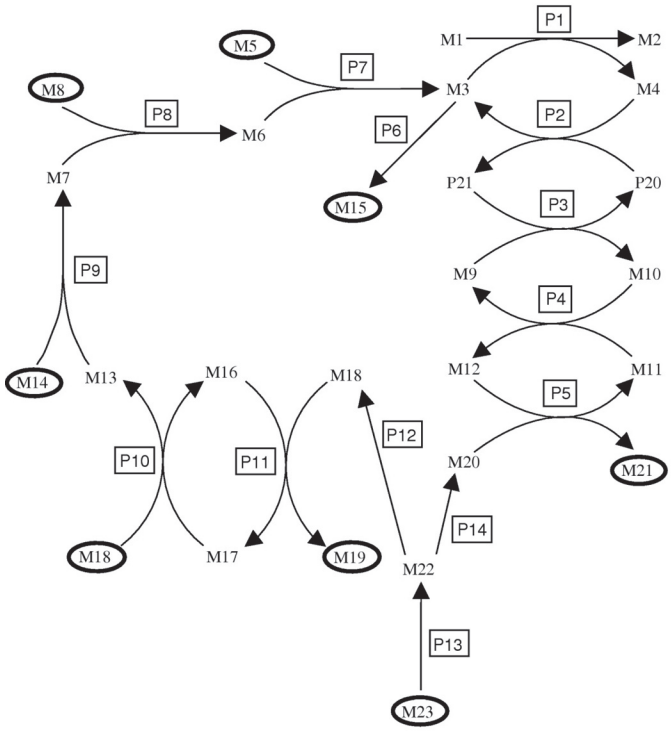


Figure 2. The Claytor Biochemical Network. M1 is a toxic compound and is detoxified through a redox chain (catalyzed by P1, P2, P3, P4 and P5 proteins). Compound M3 is needed for detoxification but also has alternative uses (reaction of P6). It is synthesized through a series of steps catalyzed by P7-P10, requiring energy. M17/M16 is a pair of charged/uncharged energy compounds, and energy is obtained from the M23 substrate. M23 is also used for recharging the final redox acceptor pair M12/M11. In addition to this metabolic network, there are also signaling and gene networks ovelayed on top (data not shown).

We started by studying the first iteration through steps 1-4. A first-order approximation to the dynamics (time series) is first used to model the system. Step 1 is simple, and consists of eliminating variables that were essentially unchanged. Assuming that each molecule can interact with any of the others eliminates step 3 from this first iteration. A system of n coupled ODEs (Eq. 1) is then formed, taking also in consideration the p perturbations as parameters of each ODE. Steps 2 and 4 are then carried out simultaneously through nonlinear least squares determination of the $n(n+p)$ coefficients of the

Table 1. A proposal for a top-down modeling strategy.

Step #	Activity
1	Identification of relevant variables for modeling
2	Identification of the interactions between the variables
3	Identification of the functional form of the interactions (i.e. kinetics)
4	Estimation of the parameter values
5	Model validation
6	Design of further experiments to increase model resolution
7	Continue iterating from step 1 until satisfied

differential equations. The least squares fit is carried out with Gepasi (since fitting is not yet available in COPASI). The best solutions have been obtained using evolutionary algorithms and the Hooke and Jeeves method.

$$dx_k/dt = \sum_l^n j_{kl} x_l + \sum_m^p w_m$$

Equation 1

Use of the synthetic data from the Claytor Network described above allowed the inferences to be compared with the original system. Two hard problems dominated the effort. One relates to the predetermined boundaries for the coefficients being estimated. Because the integral of Equation 1 depends exponentially on the j_{kl} coefficients, it is easy for the stochastic algorithms to generate candidate solutions that overflow. This is avoided by initializing the fit with tight boundaries around 1, then increasing only the ones that are touched in intermediate sub-optimal solutions, and re-estimating all parameters again. Another issue is that the number of parameters depends on the square of the number of variables, and becomes unmanageable in moderately large systems. To overcome this, the fit is preceded by calculation of correlation coefficients, and then setting j_{kl} to zero when the correlation between x_k and x_l is low. This resulted in moderate improvements, but we predict that it will be much more effective in larger systems.

Large-Scale Data Sets and Analysis

Software for management, analysis, and visualization of systems biology data

Systems biology experiments produce data from different technologies and for biological entities of a different nature. The experiments of our collaborations contain data from microarrays (two-dye and Affymetrix), 2D gels and mass spectrometry, and chromatography-mass spectrometry of metabolites. We have been developing a system to manage and analyze these data. It is primarily composed of a relational database, DOME, storing not only the actual data, but also metadata and background knowledge. The system is operated through a web-based interface from which the data can be analyzed directly on the server, or be downloaded to the user's computer in formats suitable for further analysis in other software. DOME was under intense development in the previous year and is now in a stable state. The software allows the user to quickly construct a complex query by filling simple forms. DOME produces data visualizations using colored metabolic networks, constructed with the Brome software that we built earlier. Data can also be visualized in tables that are colored for rapid identification of features, and filters can be created for several criteria (e.g. *p*-value). DOME currently provides analysis by PCA and k-means clustering. The first was implemented in the C program Ometer (by PM), while the second is carried out with the R statistical package. These form an expandable backbone for processing analyses on the server. We can expand the analyses provided in DOME with R routines or by coding algorithms in Ometer. In addition, Ometer provides correlation and partial correlation analysis (see below), and discriminant analyses.

We have also been active in community efforts relating to the creation of standards for data exchange and communication of metabolomics data (#Bino et al., 2004), having subscribed to a proposal of a suitable data model (#Jenkins et al., 2004).

Progress in data analysis methods

Several statistical methods have been proposed for detection of differential expression levels. We have used synthetic data from the AGN data set (Mendes et al., 2003) to study a few of these, namely Welch t test, ANOVA with fixed model for genes, mixed model for genes, or mixed model for the whole experiment. The latter is used as a best-case scenario since it is impossible to carry out for even moderate numbers of genes. Noise was added to the synthetic data at different levels and simulating different sources, and results of differential expression detection were compared to the noiseless data. We found out that all methods produced few false positives, but that the rate of false negatives for the *t* test grows

linearly with the coefficient of variation (being 80 percent at around CV of 120 percent). At low levels of noise, all ANOVA models perform equally well, while at higher noise levels (CV > 100 percent) the ANOVA mixed model for genes (Wolfinger et al., 2001) produced nearly as few false negatives as the mixed model for the whole experiment, while the fixed gene model was considerably worse.

The inference of gene networks from microarray data has been in our interests for a while (e.g. de la Fuente *et al*., 2002). This continued in the last year in collaboration with the Hoeschele group. A new method was proposed based on partial correlation coefficients to infer direct interactions between genes. First the pairs that have Pearson correlation above some significance threshold are enumerated, then each of these correlations is conditioned to every other gene in the network, and then to each pair of other genes. Whenever these higher-order partial correlations are above a threshold, the correlation is deemed to have originated from indirect interactions and the connection between those two genes is no longer considered. Theoretically all other higher order partial correlations (trios, etc.) should be considered but this would become NP-hard. By application to synthetic data sets we observed that conditioning only up to pairs is enough to decrease the false positive rate considerably making this a useful method. Its great advantage is that it is ready to analyze data from virtually any kind of experimental design, as long as there are enough samples. This method was published (#de la Fuente *et al*., 2004) and implemented in two software packages (ParCorA and Ometer).

Metabolomics has been under rapid development, but appropriate bioinformatics methods have been lacking (Mendes, 2002). We have proposed a set of rules to interpret an intriguing phenomenon present in metabolomics data, where very significant correlations between a few metabolite pairs occur in replicate measurements. A study of noise propagation in biochemical networks proved that these correlations are not due to the stoichiometric network (Steuer *et al*., 2003), but did not clearly establish their origin. Using simulation and metabolic control analysis, we established four conditions in which these correlations arise from the overall regulatory properties of the network (#Camacho *et al*., 2005): mass conservation, chemical equilibrium, disproportionate control of the two metabolites by a single enzyme, and disproportionate variance in a single enzyme level. The change in metabolite correlations between two states of the system can be used as a diagnostic of significant changes in overall regulation.

Biological Discovery

Yeast systems biology

Together with the Shulaev group, we have carried out a systems biology study of two yeast physiological states. Exponentially growing cultures were compared against 4-day old cultures using global transcript and metabolite profiles. The data was analyzed using the ANOVA mixed model for genes (see above) and metabolite correlations were calculated. As a way of demonstrating the power and limits of the traditional bottom-up approach, we combined two separate bottom-up models of glycolysis (Teusink *et al*., 2000) and glycerol synthesis (Cronwright *et al*., 2002) to predict the metabolite correlations. We observed good agreement between model and experiment in the case of exponentially growing cultures. However, we were unable to match the correlations observed in the 4-day old cultures, after adjusting the model to the lower levels of sucrose of this growth stage. This was not too surprising since the two original models were calibrated with enzyme levels measured in growing cultures. It clearly shows, though, that bottom-up metabolic models are limited because they are unable to explain large changes in enzyme levels, even though we collected trascriptomic data that should have been useful had the models used it. This work was published in a special issue on Yeast Systems Biology of the journal *Current Genomics* (#Martins *et al*., 2004).

Plant functional genomics projects

Work on the NSF-funded project on *Medicago*

truncatula functional genomics continued, and it is now in its third year of execution. We have now started to publish results, mostly from the Noble Foundation collaborators, showing different responses to the different elicitors at the metabolite (#Broeckling *et al*., 2005) and transcript levels (#Suzuki *et al*., 2005). Data have been uploaded to the VBI DOME system. Data from the NSF-funded *Vitis vinifera* project, led by collaborators at the University of Nevada, Reno, has also been uploaded to the VBI DOME server. In both cases, new data meant that the schema had to suffer some (predicted) adjustments.

Ascorbate biosynthesis in plants

This collaboration with the Nessler and Chevone groups (VT's Plant Pathology and Physiology Department) has produced exciting results. Our bioinformatics work predicted the existence of four *myo*-inositol oxygenase (MIOX, EC 1.13.99.1) genes in *Arabidopsis thaliana*. Cloning of these genes and purification of their products led to *in vitro* demonstrations that the enzymes are effectively MIOX. Furthermore, over-expressors of these genes display higher levels of ascorbate, while knockout mutants have lower levels. This suggests a route from *myo*-inositol to ascorbic acid (#Lorence *et al*., 2004) in addition to previously proposed pathways (Agius *et al*., 2003; Wheeler *et al*., 1998). Further bioinformatics work in our group has revealed 24 genes that could be possible glucuronate reductases (EC 1.1.1.19), the metabolic step after MIOX. These are now being cloned and tested *in vitro* by the Nessler group.

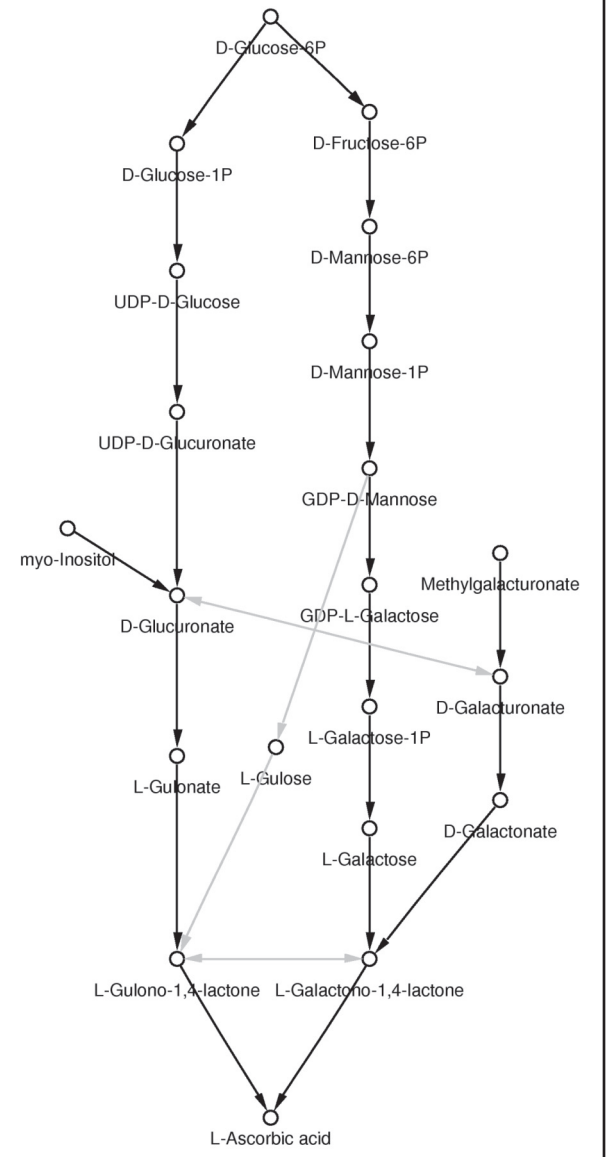


Figure 3. Plausible ascorbic acid biosynthetic network in plants. Reactions in grey are speculations. The step that converts *myo*-inositol to D-glucuronate is catalyzed by the *myo*-inositol oxygenase (MIOX) enzyme discovered by informatics means by us.

Conclusion

Our activities during last year have started converging the diverse methods used towards a common goal of top-down modeling of biochemical networks. In the near future, we will make the first release of COPASI and DOME, and hope to establish a robust method of deriving phenomenological dynamic models through linear dynamics approximations. We are also pursuing the application of machine learning and statistics to the problem of variable selection for the subsequent iterations of the top-down modeling strategy.

Acknowledgments

This work was supported by NSF grants DBI-0109732, DBI-0217653, and IBN-0118612, NIGMS grant GM068947, USDA/CREES grant 2002-3S321-11600, and by VBI. We are

grateful to the following for very productive collaborations: R. Laubenbacher, V. Shulaev, I. Hoeschele, C. Nessler, B. Chevone, A. Lorence, R. Dixon, G. May, L. Sumner, T. Smith, G. Cramer, J. Cushman, D. Schooley, J. Banga, U. Kummer, S. Sahle, R. Gauges, S. Akman, F. Torti, J. Snoep, and D. Sullivan.

References

Agius F, Gonzalez-Lamothe R, Caballero JL, Munoz-Blanco J, Botella MA, and Valpuesta V (2003) Engineering increased vitamin C levels in plants by overexpression of a D-galacturonic acid reductase. *Nat Biotechnol* **21**: 177–181

Brazhnik P, de la Fuente A, and Mendes P (2002) Gene networks: how to put the function in genomics. *Trends in Biotechnol* **20**: 467–472

Cronwright GR, Rohwer JM, and Prior BA (2002) Metabolic control analysis of glycerol synthesis in *Saccharomyces cerevisiae*. *Appl Environ Microbiol* **68**: 4448–4456

de la Fuente A, Brazhnik P, and Mendes P (2002) Linking the genes: inferring quantitative gene networks from microarray data. *Trends in Genet* **18**: 395–398

Finney A and Hucka M (2003) Systems biology markup language: Level 2 and beyond. *Biochem Soc Trans* **31**: 1472–1473

Gillespie DT (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J comput Phys* **22**: 403–434

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, and Wang J (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**: 524–531

Loomis WF and Sternberg PW (1995) Genetic networks. *Science* **269**: 649

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

Mendes P (2002) Emerging bioinformatics for the metabolome. *Brief Bioinform* **3**: 134–145

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

Moles CG, Mendes P, and Banga JR (2003) Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Res* **13**: 2467–2474

Oliver SG (1996) From DNA sequence to biological function. *Nature* **379**: 597–600

Quackenbush J (2001) Computational analysis of microarray data. *Nat Rev Genet* **2**: 418–427

Steuer R, Kurths J, Fiehn O, and Weckwerth W (2003) Observing and interpreting correlations in metabolomic networks. *Bioinformatics* **19**: 1019–1026

Teusink B, Passarge J, Reijenga CA, Esgalhado E, van der Weijden CC, Schepper M, Walsh MC, Bakker BM, van Dam K, Westerhoff HV, and Snoep JL (2000) Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. *Eur J Biochem* **267**: 5313–5329

Wheeler GL, Jones MA, and Smirnoff N (1998) The biosynthetic pathway of vitamin C in higher plants. *Nature* **393**: 365–369

Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, and Paules RS (2001) Assessing gene significance from cDNA microarray expression data via mixed models. *J Comput Biol* **8**: 625–637

Publications

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, and Sumner LW (2004) Potential of metabolomics as a functional genomics tool. *Trends in Plant Sci* **9**: 418–425

Broeckling CD, Huhman DV, Farag MA, Smith JT, May GD, Mendes P, Dixon RA, and Sumner LW (2005) Metabolic profiling of *Medicago truncatula* cell cultures reveals the effects of biotic and abiotic elicitors on metabolism. *Exp Bot* **56**: 323–336

Camacho D, de la Fuente A, and Mendes P (2005) The origin of correlations in metabolomics data. *Metabolomics* **1**: 53–63

de la Fuente A, Bing N, Hoeschele, I, and Mendes P (2004) Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* **20**: 3565–3574

de la Fuente A and Mendes P (2004) Book Review: Computational analysis of biochemical systems, by Eberhart O. Voit. X. *Bull of Math Biol* **66**: 195–197

de la Fuente A, Brazhnik P, and Mendes P (2004) Regulatory strength analysis for inferring gene networks. In Kholodenko, B.N. and Westerhoff, H.V. (eds.), *Metabolic engineering in the post-genomics era*. Horizon Bioscience, Wymondham, UK, pp. 107–137

Jenkins H, Hardy N, Beckmann M, Draper J, Smith AR, Taylor J, Fiehn O, Goodacre R, Bino RJ, Hall R, Kopka J, Lane GA, Lange BM, Liu JR, Mendes P, Nikolau BJ, Oliver SG, Paton NW, Rhee S, Roessner-Tunali U, Saito K, Smedsgaard J, Sumner LW, Wang T, Walsh S, Wurtele ES, and Kell DB (2004) A proposed framework for the description of plant metabolomics experiments and their results. *Nat Biotechnol* **22**: 1601–1606

Lorence A, Chevone BI, Mendes P, and Nessler CL (2004) myo-inositol oxygenase offers a possible entry point into plant ascorbate biosynthesis. *Plant Physiol* **134**: 1200–1205

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

Suzuki H, Reddy MS, Naoumkina M, Aziz N, May GD, Huhman DV, Sumner LW, Blount JW, Mendes P, and Dixon RA (2005) Methyl jasmonate and yeast elicitor induce differential transcriptional and metabolic re-programming in cell suspension cultures of the model legume *Medicago truncatula*. *Planta* **220**: 696–707

Methanogenic Archaea, Coalbed Methane and Mycobacteria

Biswarup Mukhopadhyay
Research Assistant Professor, VBI
Adjunct Assistant Professor of Biochemistry and Biology, Virginia Tech
biswarup@vt.edu

Christopher L. Case, Eric F. Johnson, Jessica L. Kraszewski, Endang Purwantini (collaborator)

Environmental biology and novel metabolism of methanogenic archaea

Discovery of a new type of sulfite reductase in *Methanocaldococcus jannaschii*

M. jannaschii is a deeply rooted hyper-thermophilic methanogenic archaeon (Jones et al., 1983). This strict anaerobe is an inhabitant of the submarine hydrothermal vents (1). It grows in the 48-94 ^{o}C range with an optimal growth temperature of 85°C (Jones et al., 1983). *M. jannaschii* is an obligate hydrogenotroph. It derives energy exclusively by oxidizing hydrogen. The genome of this extremophile is only 1.66-megabase pair in size (Bult et al., 1996), which is relatively small in the microbial world**.** On the other hand it synthesizes all cellular components from H_2 and CO_2 in a mineral salts medium (Jones et al., 1983). Thus, *M. jannaschii* might represent a minimum requirement for a life form to exist independently. These observations suggest that the expression of most of the genes of this archaeon would not be under regulation. This deduction is consistent with the finding that the genome of *M. jannaschii* does not carry an obvious homolog of a sensory histidine kinase or a response regulator with a PAS domain, which are hallmarks of two-component regulation in the bacteria (Kennelly, 2002; Taylor and Zhulin, 1999). In contrast, other archaeal genomes carry these elements (Kennelly, 2002; Klenk et al., 1997; Taylor and Zhulin, 1999). We expect that in *M. jannaschii* only a very small number of genes will be expressed conditionally and the organism would use these genes for surviving or taking advantage of certain environmental changes that are specific to the hydrothermal vent environments. We are interested in identifying and studying these regulated genes, for they might be associated with novel regulators and mechanisms. The DNA replication, transcription, and stress response machineries of the archaea are more similar to that of the eukaryotes than the bacteria (Bell and Jackson, 2001; De Biase et al., 2002; Giraldo, 2003; Grabowski and Kelman, 2003; Macario et al., 1999; Omer et al., 2003) and *M. jannaschii* is a deeply rooted organism (Wheelis et al., 1992). Therefore, our investigations will help to understand how regulatory mechanisms evolved on earth and how complexity arose in the eukaryotes. To take advantage of this opportunity we have developed a strategy for locating the regulated genes in *M. jannaschii* (Mukhopadhyay et al., 1999; Mukhopadhyay et al., 2000). Our approach involves developing hypotheses on conditions that the archaeon would face in its natural habitat, exposing the organism to those conditions and tracking the consequent intracellular activities via comparative proteomics analyses (Mukhopadhyay et al., 2000). In our hands this approach has been successful in unraveling novel behaviors in *M. jannaschii* (Mukhopadhyay et al., 2000). We report below another such observation with this archaeon.

The temperature of the hydrothermal fluid is about 350 ^{o}C (Corliss et al., 1979; Jannasch and Mottl, 1985). A mixing with cold seawater that permeates through the chimney wall creates

cooler zones where *M. jannaschii* can grow (Corliss et al., 1979; Jannasch and Mottl, 1985; Jones et al., 1983; McCollom and Shock, 1997). The oxygen present in the seawater is neutralized via a reaction with sulfide that is present in the vent fluid (Corliss et al., 1979; Jannasch and Mottl, 1985; McCollom and Shock, 1997). This reaction ensures that the areas with conducible temperatures are also anaerobic, a condition that is required for the growth of a methanogen (Zinder, 1993). However, this process has the potential of producing sulfite. Sulfite is an inhibitor of methanogenesis (Balderston and Payne, 1976), the energy metabolism of the methanogenic archaea (Wolfe, 1992). To understand how *M. jannaschii* handles this problem, we conducted growth experiments. Our data show that this methanogen not only tolerated sulfite, but also can use this oxyanion as a sole sulfur source. Since *M. jannaschii* has an obligate requirement for sulfide (Jones et al., 1983), we hypothesized that the organism alleviated sulfite toxicity by converting sulfite into sulfide. However, the search of the genome sequence database did not show the existence of a sulfite reductase gene in this archaeon (Bult et al., 1996). Therefore, we performed a proteomic analysis. The results showed that during growth on sulfite, *M. jannaschii* expressed a 70 kDa polypeptide that was absent in the cells raised with sulfide as the sulfur source. From a database search with the mass spectrometry data for an unseparated tryptic digest, we identified this polypeptide as MJ0870. Although this ORF has been annotated in the NCBI database as the β-subunit of a coenzyme F_{420}-reducing hydrogenase (Bult et al., 1996), a closer look using comparative sequence analysis showed that we have discovered a novel type of sulfite reductase, which uses coenzyme F_{420} as an electron carrier (Figure 1). We call this enzyme an F_{420}-dependent sulfite reductase (Fsr). F_{420} is a naturally occurring 5-deazaflavin-derivative that was originally discovered in the methanogenic archaea (Eirich et al., 1978). It is an obligatory 2-electron or hydride transferring coenzyme (DiMarco et al., 1990; Eirich et al., 1978). The standard midpoint potential of this coenzyme is –350 mV (DiMarco et al., 1990; Thauer et al., 1977). Consequently, under standard conditions, H_2F_{420} is a potent reductant for reducing sulfite (SO_3^{-2}) or bisulfite (HSO_3^-) and the overall reaction is exergonic (DiMarco et al., 1990; Thauer et al., 1977).

$$F_{420} + 2e + 2H^+ \longrightarrow H_2F_{420}, \qquad \Delta E^{o'} = -350 \text{ mV}$$
$$HSO_3^- + 6e + 6H^+ \longrightarrow HS^- + 3H_2O, \qquad \Delta E^{o'} = -116 \text{ mV}$$
$$HSO_3^- + 3H_2F_{420} \longrightarrow HS^- + 3H_2O + F_{420}, \qquad \Delta G^{o'} = -135 \text{ kJ/mol}$$

The N-terminal half of MJ0870 was a homolog of the electron input unit (H_2F_{420} dehydrogenase, FqoF or FpoF) of a membrane-based electron transport system that is found in *Archaeoglobus fulgidus*, a sulfate reducing archaea, and in methylotrophic methanogens belonging to the

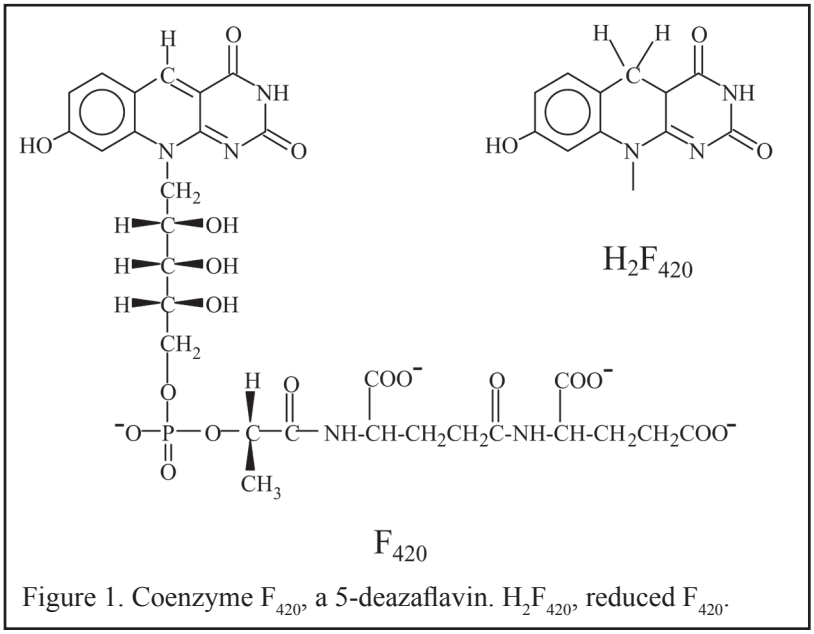


Figure 1. Coenzyme F_{420}, a 5-deazaflavin. H_2F_{420}, reduced F_{420}.

genus *Methanosracina* (Baumer et al., 2000; Bruggemann et al., 2000). The C-terminal half was similar to the B subunit of sirohaeme-containing dissimilatory sulfite reductases or Dsr (Crane and Getzoff, 1996). Therefore, Fsr was a chimera and from the known functions of FqoF/FpoF and Dsr we hypothesized the following mechanism for the new enzyme. The N-terminal domain of Fsr receives reducing equivalents from H_2F_{420} in the form of hydride which are transferred to the C–terminal sirohaeme-containing domain as electrons for the reduction of sulfite to sulfide. This scheme is similar to the mechanism used by the *E. coli* sulfite reductase, where a flavoprotein subunit (SirFP) derives electrons from NADPH, a 2e-restricted donor, and passes those via FAD, FMN, and [4Fe-4S] centers to the siroheme of a hemoprotein subunit (Sir-HP) (Crane and Getzoff, 1996). However, in primary structure, Fsr is unrelated to the *E. coli* enzyme.

Fsr homologs were found in *Methano-thermobacter thermautotrophicus* (MTH280) and *Methanopyrus kandleri* (ORF MK0799), which, similar to *M. jannaschii,* are thermo-philic and strictly hygenotrophic methanogens (Slesarev et al., 2002; Smith et al., 1997). MTH280 has been annotated as β-subunit of a coenzyme F_{420}-reducing hydrogenase or FrhB (Smith et al., 1997) and MK0799 as a protein with FrhB and nitrite reductase characteristics (Slesarev et al., 2002), respectively. Similar to MJ0870 or Fsr, each of these homologs possesses an N-terminal H_2F_{420} dehydrogenase domain and a C-terminal dissimilatory sulfite reductase domain. Interestingly, *Methanococcus maripaludis*, a close relative of *M. jannaschii* and a mesophile, lacked an Fsr homolog (Hendrickson et al., 2004). It has been previously shown that *Methanothermococcus thermolithotrophicus* and *M. thermautotrophicus,* can use sulfite as a sole sulfur source (Daniels et al., 1986). We expect *M. thermolithotrophicus* to carry an Fsr homolog; the genome sequence of this thermophile is yet to be determined.

We have followed up our discovery of Fsr with enzyme assays-based studies. Our cell extract activity data supported the hypothesis that MJ0870 is an F_{420}-dependent sulfite reductase (Figure 2). Extracts of cells grown with sulfite oxidized H_2F_{420} with sulfite as the electron acceptor. This activity was absent in cells grown with sulfide as the sulfur source. Fsr activity was associated with the membrane (data not shown). These observations indicated that Fsr was poised to protect *M. jannaschii* from the toxic effect of sulfite. It is likely that the membrane resident Fsr converted sulfite rapidly into sulfide and consequently the intracellular level of sulfite never reached an inhibitory value.

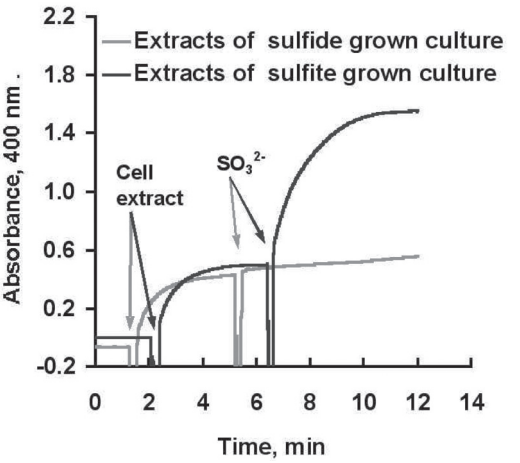


Figure 2. F_{420}-dependent sulfite reductase activity in *M. jannaschii* cell extracts. Oxidation of H_2F_{420} (reduced F_{420}) with sulfite was monitored spectrophotometrically at 400 nm (appearance of F_{420}). Starting Assay mixture: 50 mM potassium phosphate buffer, pH 7, and 40 µM H_2F_{420} (F_{420} reduced with $NaBH_4$). Further additions: cell extracts, 40 µg protein/ml; Na_2SO_3, 1.5 mM (similar results with 50 µM Na_2SO_3). Extracts: Cells grown with sulfite or sulfide were lysed osmotically. The lysate was centrifuged at 9,000 x g to obtain the extracts.

Note: At pH 7, Coenzyme F_{420} exhibits an absorbance maximum at 420 nm and substantial absorbance at 400 nm (the isosbestic point). The reduced form of this coenzyme does not absorb at either 420 nm or 400 nm.

Fsr is a possible ancestor of the FqoF and FpoF of the late evolving archaea and the A and B subunits (DsrA or DsrB) of dissimilatory sulfite reductase of the bacteria and sulfate reducing archaea (Crane and Getzoff, 1996). It is equally possible that Fsr was generated from the fusion of the *fqoF/fpoF* and *dsrA*/*dsrB* genes. *dsrA* and *dsrB* are believed to have originated from a gene duplication event (Stahl et al., 2002). Fqo and Fpo are homologous to the NADH dehydrogenase or complex I of the bacteria and mitochondria (Deppenmeier, 2004).

The highly inducible nature of the *fsr* and an undetectable level of Fsr protein in cells grown with sulfide indicates that the *fsr* gene is regulated via an all-or-nothing, or autocatalytic control. In all, Fsr brings new topics of study in the areas of enzymology, gene regulation, and evolution of metabolism. We are currently pursuing these opportunities.

A novel archaeal type phosphoenolpyruvate carboxylase

We have been investigating the CO_2-fixation pathways in the methanogenic archaea. Methanogens determine the success of the mineralization of complex polymers in anaerobic niches (Zinder, 1993). They convert H_2 + CO_2, acetate, and simple methylated compounds into methane and thereby remove the thermodynamic block in the overall degradation process (Zinder, 1993). This role is essential for the operation of carbon cycle in nature and the functioning of anaerobic sewage digestors and biogas generators (Zinder, 1993). Therefore, a detailed knowledge of methanogen physiology is of practical interest. Compared to many anaerobic and aerobic microorganisms, these archaea obtain very little energy from their catabolic activities (Zinder, 1993). Hence a methanogen must use the available energy very efficiently for cell biosynthesis. Since CO_2-fixation at the level of oxaloacetate (OAA) provides precursors for amino acid and tetrapyrrole biosynthesis (Simpson and Whitman, 1993), we have been studying this step in methanogens in detail. During this investigation we have discovered a new type of phosphoenolpyruvate carboxylase (Ppc) (Patel et al., 2004). A Ppc carboxylates phosphoenolpyruvate (PEP) and produce OAA and inorganic phosphate (Izui et al., 2004). In bacteria, this enzyme serves an anaplerotic function; in plants, it provides carbon for photosynthesis (Izui et al., 2004). For these reasons, Ppcs from these sources have been investigated in detail (Izui et al., 2004). The plant and bacterial Ppcs are highly homologous to each other both at the primary and three dimensional structure levels (Izui et al., 2004). Therefore, we were surprised to find that the phosphoenolpyruvate carboxylase of *Methanothermobacter thermautotrophicus* has very little amino acid sequence similarity to the known Ppcs (Patel et al., 2004). Also, unlike the plant and bacterial enzymes, the archaeal enzyme is not significantly regulated by metabolites (Patel et al., 2004). The subunit size of the *M. thermautotrophicus* enzyme is about the half of that of known Ppcs (Patel et al., 2004). We call the archaeal enzyme PpcA (A, for archaeal). PpcA homologs are present in most archaea (Ettema et al., 2004; Patel et al., 2004; Sako et al., 1997; Sako et al., 1996) (Table 1). Interestingly, only three bacteria, *Clostridium perfringens*, *Oenococcus oeni* and *Leuconostoc mesenteroides*, carry a PpcA homolog and they do not possess a bacterial and plant type Ppc (Patel et al., 2004). We are investigating the structural and kinetic properties and physiological roles of PpcA. Our preliminary data suggests that this enzyme is essential in the methanogens belonging to the genus *Methanosarcina* (J. L. Kraszewski, J.K. Zhang, W.W. Metcalf, and B. Mukhopadhyay, work in progress). *Methanosarcina* species carry out acetoclastic methanogenesis, the major mode of biological methane production in nature (Zinder, 1993). These findings show that PpcA could be a major player in the global carbon cycle. Similarly PpcA might determine the pathogenicity of *C. perfringens* and efficacy of *O. oeni* and *L. mesenteroides* in wine production and industrial and food fermentations, respectively.

Table 1. Distribution of archaeal-type phosphoenolpyruvate carboxylase (PpcA) in three domains of life

Organism	PpcA ORF	Accession Number
Archaea		
Methanogenic Archaea		
Methanopyrus kandleri	MK0190	NP_613477.1
Methanothermobacter thermoautotrophicus	MTH943	NP_276080.1
Methanothermus sociabilis	Present	-[a]
Methanosarcina acetivorans	MA2690	NP_617588.1
Methanosarcina barkeri	METH1932	ZP_00296000.1
Methanosarcina mazei	MM3212	NP_635236
Non-methanogenic Archaea		
Halobacterium sp NRC-1	VNG2259c	NP_280898.1
Archaeoglobus fulgidus	AF1486	NP_070315.1
Pyrococcus furiosus	PF1975	NP_579704.1
Pyrococcus horikoshii	PH0016	NP_142039.1
Pyrococcus abyssi	PAB2342	NP_125707
Ferroplasma acidarmanus	FACI0253	ZP_00305878.1
Pyrobaculum aerophilum	PAE3416	NP_560717.1
Sulfolobus acidocaldarius	Present	-[b]
Sulfolobus solfacaricus	SSO2256	NP_343633.1
Solfulobus tokadaii	ST2101	NP_378096.1
Bacteria		
Clostridium perfringens	CPE1094	NP_562010
Oenococcus oeni	Ooen1256	ZP_00070236
Leuconostoc mesenteroides	Lmes0541	ZP_00063059
Eukarya	None	

[a] Purified enzyme data (Sako et al., 1996).

[b] From an analysis of unfinished genome sequence and enzyme data (Sako et al., 1997).

In situ conversion of coal to methane

In collaboration with Altuda Energy Corporation (San Antonio, TX), we are investigating the possibility of converting coal to methane in situ by use of microorganisms that are found in the coalbed. Our work is based on the available geological data. Often coal beds contain large amounts of methane-rich gas called coalbed methane (CBM). Isotope analysis indicates that in part CBM is biogenic (Rice and Claypool, 1981). The biogenic part has been generated due to microbial activities at two different ages. Coal has originated from the burial and further processing of plant materials. For a period immediately following the burial, the plant materials were held at a temperature that was conducive for microbial activities (Rice and Claypool, 1981). Oxygen-dependent microbial processes quickly depleted oxygen in the buried materials and allowed ananerobic microorganisms to degrade the complex polymers and to generate methane (Rice and Claypool, 1981). As the burial process continued, the temperature in the decaying materials increased, and the conversion process shifted from biogenic to thermogenic, yielding thermogenic methane (Rightmire, 1984). The second biogenic phase was initiated due to

recharge of the coalbed with freshwater that brought mineral nutrients and microorganisms for further conversion of the organic content of the coal (Scott, 1999). This latter aspect is the basis of our current research.

Our work is highly relevant to the energy economy and security of the U.S. According to the US Energy Information Administration (Caruso, 2003), "by 2025 natural gas consumption is expected to increase to almost 35 trillion cubic feet (Tcf), or 26% of US delivered energy consumption. Such a demand represents an increase of about 52% from the expected 2003 level." In contrast, "domestic gas production is expected to increase more slowly, rising from 19.5 Tcf in 2001 to 26.4 Tcf in 2025." Imports, particularly in the form of liquefied natural gas (LNG), are expected to cover the gap between consumption and production throughout the forecast. The projected increase in the U.S. natural gas production through 2025 will be due to unconventional sources (such as CBM) and deliveries from Alaska. The National Petroleum Council ("NPC") 2003 report on natural gas, "Balancing Natural Gas Policy- Fueling the Demands of a Growing Economy" (http://www.npc.org/) shows that CBM will continue to grow as an important factor in the domestic natural gas supply. Between 1989 and 2002, the CBM proved reserves have risen from 3.676 Tcf to 18.491 Tcf (Eia, 2002) and in the same period the production of CBM has increased from a mere 0.091 Tcf to 1.614 Tcf (Eia, 2002). Thus, CBM accounts for about 9.89 percent of the total natural gas reserve and 8.34 percent of the total natural gas production in the United States (Eia, 2002). Therefore, a development that will increase CBM production will have a major beneficial impact on the U.S. energy security.

Development of population-specific vaccines for tuberculosis in Indonesia

About 10 percent of the worldwide cases of *Mycobacterium tuberculosis* infections are found in Indonesia (Corbett et al., 2003). The eventual goal of this project is to deliver effective vaccines for tuberculosis to the citizens of Indonesia, where the disease is endemic, and to develop a strategy that can be applied elsewhere. Our team includes researchers from the Institut Teknologi Bandung or ITB (Bandung, Indonesia), Rotinsulu Pulmonary Hospital (Bandung, Indonesia), and the Center for Tuberculosis Research, Johns Hopkins University School of Medicine (Baltimore, MD). Dr. Endang Purwantini of the Virginia Bioinformatics Institute is the lead for this project. Dr. Biswarup Mukhopadhyay serves as the proteomics specialist and international coordinator. The study is focused on the Kiara Condong area of Bandung, Indonesia, where the population density is very high, and about 70 percent of the population is infected with *M. tuberculosis* (F. S. Tanoerahardjo, unpublished data). A part of our current effort is dedicated to the development of research infrastructure at the Institut Teknologi Bandung and Rotinsulu Pulmonary Hospital and to the training of Indonesian students and professionals in microbiology, molecular biology, immunology, proteomics, and bioinformatics techniques. The research activities include the following: 1. Isolation of *M. tuberculosis* strains from TB patients; 2. Determining the genotypes of the isolates; and 3. Identifying the antigens in the culture filtrates and cell surfaces of the isolates that can stimulate IFN-γ production in Peripheral blood mononuclear cells (PBMC) isolated from Kiara Condong subjects. Thus far, 30 clinical isolates have been obtained and crude antigen preparations for these isolates have been generated. Genotyping experiments and IFN-γ production assay are in progress. One of the unique scientific aspects of this project is that we are using an archae based expression system for *M. tuberculosis* antigens. Some of these antigens can not be expressed in soluble forms in *E. coli*. Interestingly, certain methanogenic archaea carry the homologs of these proteins.

Oxidative damage defense of the mycobacteria

This project is a collaboration with Dr. Endang Purwantini of the Virginia Bioinformatics Institute. *Mycobacterium tuberculosis* and

Mycobacterium leprae are intracellular pathogens, which survive in macrophage. This survival is partly due to their abilities to defend themselves against damage by free radicals such as reactive oxygen intermediates (ROI) or reactive nitrogen intermediates (RNI) produced by the macrophage. The interactions of a mycobacterial cell with the RNI have been studied in detail, but the aspects of sensitivity to ROI remain poorly understood (Flynn and Chan, 2001). In the literature often the role ROI in controlling *M. tuberculosis* infection has been questioned (Flynn and Chan, 2001). It is possible that the observed apparent insensitivity suggests an ability of the mycobacteria to deal with ROI efficiently. With this hypothesis we have been investigating the mechanisms that these organisms use in coping an exposure to reactive oxygen species. Our strategy is to work with *Mycobacterium smegmatis,* for identifying the genes that a mycobacterial cell could use for defense against ROI and to carry out gene knockout and further physiological studies with *M. tuberculosis. M. smegmatis* is a fast growing nonpathogen closely related to *M. tuberculosis.* Therefore, it allows in vitro studies with great speed. We have investigated the effect of paraquat, menadione, and plumbagin, which are intracellular superoxide generators, on the growth of *M. smegmatis*. The behavior of *M. smegmatis* was different from that of *Escherichia coli.* We found that *M. smegmatis* was highly sensitive to plumbagin and moderately sensitive to menadione, but very insensitive to paraquat. In contrast, *E. coli* is highly sensitive to paraquat, but less sensitive to menadione and plumbagin (Cardullo and Gilroy, 1975; Hassan and Fridovich, 1978; Imlay and Fridovich, 1992). This difference can be attributed to an array of specialized cell wall components in the former (Chan et al., 1991; Chan et al., 1989; Daffe and Etienne, 1999). To understand the mode of action of plumbagin in killing *Mycobacterium smegmatis*, we conducted 2-D gel experiments with extracts of cells grown in the presence and absence of plumbagin. The differentially expressed proteins were identified via mass spectrometry and de-novo amino acid sequence analyses. The results show that some of these proteins are involved in lipid biosynthesis, translation, protein folding and stress response. Our random mutagenesis and screening of the mutants for plumbagin hypersensitivity have also identified more such targets. Our data provides hypotheses on the processes that a mycobacterial cell uses to combat oxidative stress. A detailed study on the roles of the identified proteins in combating superoxide stress is currently in progress.

Acknowledgements

The work in our laboratory is supported by a start up fund from the Virginia Bioinformatics Institute, a Johns Hopkins University-Virginia Bioinformatics Institute Collaboration grant, U.S. Department of Energy grants DE-FG02-03ER83605 (Phase I), DE-FG02-03ER83606 (Phase I), DE-FG02-03ER83605 (Phase II) and DE-FG02-03ER83606 (Phase II) (in partnership with Andrew Scott, Altuda Energy Corporation, San Antonio, TX) to BM, and a RUTI grant 20A/SPK/RUTI/KRT/IV/2004 from the Indonesian International Joint Research Grant Program, Government of Indonesia, to EP.

References

Balderston WL and 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, and 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

Bell SD and Jackson SP (2001) Mechanism and regulation of transcription in archaea. *Curr Opin Microbiol* **4**: 208–213

Bruggemann H, Falinski F, and 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, and Venter JC (1996) Complete genome sequence of the methanogenic archaeon. *Methanococcus jannaschii Science* **273**: 1058–1073

Cardullo MA and Gilroy JJ (1975) Inhibition of oxidative metabolism in *Escherichia coli* by d-camphor and restoration of oxidase activity by quinones. *Can J Microbiol* **21**: 1357–1361

Caruso GF (2003) Natural Gas Supply and Demand Issues. *http:// energycommercehouse gov/108/Hearings/ 06102003hearing944/Caruso1517htm*

Chan ED, Chan J, and Schluger NW (2001) What is the role of nitric oxide in murine and human host defense against tuberculosis? Current knowledge. *Am J Respir Cell Mol Biol* **25**: 606–612

Chan J, Fan XD, Hunter SW, Brennan PJ, and Bloom BR (1991) Lipoarabinomannan, a possible virulence factor involved in persistence of Mycobacterium tuberculosis within macrophages. *Infect Immun* **59**: 1755–1761

Chan J, Fujiwara T, Brennan P, McNeil M, Turco SJ, Sibille JC, Snapper M, Aisen P, and Bloom BR (1989) Microbial glycolipids: possible virulence factors that scavenge oxygen radicals. *Proc Natl Acad Sci USA* **86**: 2453-2457

Corbett EL, Watt CJ, Walker N, Maher D, Williams BG, Raviglione MC, and 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, and van Andel TH (1979) Submarine thermal springs on Galápagos Rift. *Science* **203**: 1073–1083

Crane BR and Getzoff ED (1996) The relationship between structure and function for the sulfite reductases. *Curr Opin Struct Biol* **6**: 744–756

Daffe M and Etienne G (1999) The capsule of Mycobacterium tuberculosis and its implications for pathogenicity. *Tuber Lung Dis* **79**: 153–169

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

De Biase A, Macario AJ, and Conway de Macario E (2002) Effect of heat stress on promoter binding by transcription factors in the cytosol of the archaeon. *Methanosarcina mazeii Gene* **282**: 189–197

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

DiMarco AA, Bobik TA, and Wolfe RS (1990) Unusual coenzymes of methanogenesis. *Annu Rev Biochem* **59**: 355–394

Eia (2002) US Crude Oil, Natural Gas, and Natural Gas Liquids Reserves Annual Report US Crude Oil, Natural Gas, and Natural Gas Liquids Reserves Annual Report. http://www.eia.doe.gov/oil_gas/ natural_gas/data_publications/crude_oil_

natural_gas_reserves/reserves_historical.html

Eirich LD, Vogels GD, and Wolfe RS (1978) Proposed structure for coenzyme F_{420} from Methanobacterium. *Biochemistry* **17**: 4583–4593

Ettema TJ, Makarova KS, Jellema GL, Gierman HJ, Koonin EV, Huynen MA, de Vos WM, and van der Oost J (2004) Identification and functional verification of archaeal-type phosphoenolpyruvate carboxylase, a missing link in archaeal central carbohydrate metabolism. *J Bacteriol* **186**: 7754–7762

Flynn JL and Chan J (2001) Immunology of tuberculosis. *Annu Rev Immunol* **19**:93–129

Giraldo R (2003) Common domains in the initiators of DNA replication in Bacteria, Archaea and Eukarya: combined structural, functional and phylogenetic perspectives. *FEMS Microbiol Rev* **26**: 533–554

Grabowski B and Kelman Z (2003) Archeal DNA replication: eukaryal proteins in a bacterial context. *Annu Rev Microbiol* **57**: 487–516

Hassan HM and Fridovich I (1978) Superoxide radical and the oxygen enhancement of the toxicity of paraquat in *Escherichia coli. J Biol Chem* **253**: 8143–8148

Hendrickson EL, Kaul R, Zhou Y, Bovee D, Chapman P, Chung J, Conway de Macario E, Dodsworth JA, Gillett W, Graham DE, Hackett M, Haydock AK, Kang A, Land ML, Levy R, Lie TJ, Major TA, Moore BC, Porat I, Palmeiri A, Rouse G, Saenphimmachak C, Soll D, Van Dien S, Wang T, Whitman WB, Xia Q, Zhang Y, Larimer FW, Olson MV, and Leigh JA (2004) Complete genome sequence of the genetically tractable hydrogenotrophic methanogen *Methanococcus maripaludis. J Bacteriol* **186**: 6956–6969

Imlay J and Fridovich I (1992) Exogenous quinones directly inhibit the respiratory NADH dehydrogenase in *Escherichia coli. Arch Biochem Biophys* **296**: 337–346

Izui K, Matsumura H, Furumoto T, and Kai Y (2004) PHOSPHOENOLPYRUVATE CARBOXYLASE: A New Era of Structural Biology. *Annu Rev Plant Biol* **55**: 69–84

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

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

Kennelly PJ (2002) Protein kinases and protein phosphatases in prokaryotes: a genomic perspective. *FEMS Microbiol Lett*, **206**: 1–8

Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, Ketchum KA, Dodson RJ, Gwinn M, Hickey EK, Peterson JD, Richardson DL, Kerlavage AR, Graham DE, Kyrpides NC, Fleischmann RD, Quackenbush J, Lee NH, Sutton GG, Gill S, Kirkness EF, Dougherty BA, McKenney K, Adams MD, Loftus B, Venter JC, and et al (1997) The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon. *Archaeoglobus fulgidus Nature* **390**: 364–370

Macario AJ, Lange M, Ahring BK, and De Macario EC (1999) Stress genes and proteins in the archaea. *Microbiol Mol Biol Rev* **63**: 923–967, table of contents

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

Mukhopadhyay B, Johnson EF, and Wolfe RS (1999) Reactor-scale cultivation of the hyperthermophilic methanarchaeon *Methanococcus jannaschii* to high cell densities. *Appl Environ Microbiol* **65**: 5059–5065

Mukhopadhyay B, Johnson EF, and Wolfe RS (2000) A novel pH2 control on the expression of flagella in the hyperthermophilic strictly hydrogenotrophic methanarchaeaon *Methanococcus jannaschii Proc Natl Acad Sci USA* **97**: 11522–11527

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

Omer AD, Ziesche S, Decatur WA, Fournier MJ, and Dennis PP (2003) RNA-modifying machines in archaea. *Mol Microbiol* **48**: 617–629

Patel HM, Kraszewski JL, and Mukhopadhyay B (2004) The Phosphoenolpyruvate Carboxylase from *Methanothermobacter thermautotrophicus* Has a Novel Structure. *J Bacteriol* **186**: 5129–5137

Rice DD and Claypool GE (1981) Generation, accumulation, and resource potential of biogenic gas. *Am Assoc Petrol Geol Bull* **65**: 5–25

Rightmire CT (1984) Coalbed methane resources In Rightmire CT, Eddy GE, and Kirr JN (eds), *Coalbed methane resources of the United States: American Association of Petroleum Geologists Explorer* **21**: 16, 18–20, 22–23

Sako Y, Takai K, Nishizaka T, and Ishida Y (1997) Biochemical relationship of phosphoenolpyruvate carboxylases (PEPCs) from thermophilic archaea. *FEMS Microbiol Lett* **153**: 159–165

Sako Y, Takai K, Uchida A, and Ishida Y (1996) Purification and characterization of phosphoenolpyruvate carboxylase from the hyperthermophilic archaeon Methanothermus sociabilis. *FEBS Lett* **392**: 148–152

Scott AR (1999) Improving Coal Gas Recovery with Microbially Enhanced Coalbed Methane. In *Coalbed Methane: Scientific, Environmental, and Economic Evaluations*, pp 89–111

Simpson PG and Whitman WB (1993) Anabolic pathways in methanogens. In Ferry, JG (ed), *Methanogenesis: ecology, physiology, biochemistry, and genetics*, Chapman & Hall, New York, NY. pp 445–472

Slesarev AI, Mezhevaya KV, Makarova KS, Polushin NN, Shcherbinina OV, Shakhova VV, Belova GI, Aravind L, Natale DA, Rogozin IB, Tatusov RL, Wolf YI, Stetter KO, Malykh AG, Koonin EV, and Kozyavkin SA (2002) The complete genome of hyperthermophile *Methanopyrus kandleri* AV19 and monophyly of archaeal methanogens. *Proc Natl Acad Sci USA* **99**: 4644–4649

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, Reeve JN, and et al (1997) Complete genome sequence of *Methanobacterium thermoautotrophicum* DH: functional analysis and comparative genomics. *J Bacteriol* **179**: 7135–7155

Stahl DA, Fishbain S, Klein M, Baker BJ, and Wagner M (2002) Origins and diversification of sulfate-respiring microorganisms. *Antonie Van Leeuwenhoek* **81**: 189–195

Taylor BL and Zhulin IB (1999) PAS domains: internal sensors of oxygen, redox potential, and light. *Microbiol Mol Biol Rev* **63**: 479–506

Thauer RK, Jungermann K and Decker K (1977) Energy conservation in chemotrophic anaerobic bacteria. *Bacteriol Rev* **41**: 100–180

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

Wolfe RS (1992) Biochemistry of methanogenesis. *Biochem Soc Symp* **58**: 41–49

Zinder SH (1993) Physiological ecology of methanogens. In Ferry, JG (ed), *Methanogenesis: ecology, physiology, biochemistry, and genetics* Chapman and Hall, NY, pp 128–206

Publications

Guss AM, Mukhopadhyay B, Zhang JK, and Metcalf WW (2005) Genetic analysis of mch mutants in two Methanosarcina species demonstrates multiple roles for the methanopterin-dependent C-1 oxidation/reduction pathway and differences in H(2) metabolism between closely related species. *Mol Microbiol* **55**: 1671–1680

Patel HM, Kraszewski JL, and Mukhopadhyay B (2004) The Phosphoenolpyruvate Carboxylase from *Methanothermobacter thermautotrophicus* Has a Novel Structure. *J Bacteriol* **186**: 5129–5137

Patrie SM, Charlebois JP, Whipple D, Kelleher NL, Hendrickson CL, Quinn JP, Marshall AG, and Mukhopadhyay B (2004) Construction of a hybrid quadrupole/Fourier transform ion cyclotron resonance mass spectrometer for versatile MS/MS above 10 kDa. *J Am Soc Mass Spectrom* **15**: 1099 –1108

Molecular Mechanisms of Pathogenesis in Malaria and Cryptosporidiosis: Exploiting Genomic Information Towards the Identification of New Targets for Intervention

Dharmendar Rathore
Research Assistant Professor, VBI
rathore@vbi.vt.edu

Rana Nagarkatti, Dewal Jani

Summary

Malaria and Cryptosporidiosis are two important human diseases affecting the world population. Malaria, a blood-borne infection caused by *Plasmodium* parasites, is a major health issue in the tropics, with 300-500 million clinical episodes of this disease occurring each year. A licensed vaccine against malaria is not available and the parasite is developing resistance against most of the currently available antimalarials. There is an urgent need to develop a vaccine against malaria, which will reduce the morbidity and mortality associated with this disease. Similarly, Cryptosporidiosis, a water-borne infection caused by *Cryptosporidium* parasites, is life-threatening in AIDS patients and other immunocompromised individuals. There is no known cure for this disease. The genomes of these two parasites have been sequenced. Our efforts are focused on exploiting available genome sequence from *Plasmodium falciparum* and *Cryptosporidium* parasites towards understanding the molecular basis of the onset and sustenance of infection by these pathogens. Deciphering these mechanisms will unravel the complex interplay between the troika of host, pathogen, and its environment, which will be vital for identifying new targets for intervention. Here we describe the identification and characterization of a *Plasmodium* parasite protein, which is intimately involved in the onset of malaria infection, hence an ideal target for intervention. We also describe the characterization of a *Cryptosporidium parvum* protein that interacts with the host cells and could be associated with disease pathogenesis.

Investigations in Malaria Pathogenesis

Background & Significance

Malaria infection starts with the introduction of *Plasmodium* sporozoites into the blood stream of its human host, when it is bitten by an infected mosquito. Of the four *Plasmodium* species that infect humans, *P. falciparum* is the most virulent—resulting in severe anemia and cerebral malaria, which could be fatal. Fewer than 200 sporozoites are introduced and even fewer succeed in invading liver cells, the target organ for the onset of malaria infection in host. A successful adhesion and liver cell invasion by the sporozoite is critical for this onset and is therefore, the Achilles heel of the parasite. Once inside the liver cell, the parasite rapidly multiplies, and within a few days releases thousands of parasites, which lead to the clinical pathology of this disease. Therefore, an ideal approach to control malaria is to develop a vaccine or therapeutic, which either prevents the sporozoite from infecting liver cells or destroys the parasite during liver stages of its life cycle. Such a vaccine is feasible as animals and human volunteers immunized with *Plasmodium* sporozoites that have been attenuated by exposure to X-Ray or gamma radiation, are protected when subsequently challenged with infectious sporozoites (Hoffman et al., 2002; Nussenzweig et al., 1967). While this groundbreaking discovery clearly indicates

that it is feasible to make a vaccine against malaria, the biggest stumbling block for malaria researchers worldwide has been to decipher the parasite antigens recognized by the host and to understand the immune mechanisms underlying this protection. Extensive immunological studies with known sporozoite antigens have concluded that this protection is not conferred due to a dominant immune response against a single antigen, but is mediated by the summation of many modest humoral and cell-mediated immune responses against a large variety of antigens, many of which are currently not known (Hoffman, 1996). Identification of these antigens is not only a major challenge, it is vital for the development of a successful vaccine against malaria.

Historically, antigen(s) selected as a vaccine candidate in a given pathosystem are (i) present on the surface of the pathogen, and (ii) generally involved in host-pathogen interactions and are therefore, one of the first molecules that are recognized by the host immune system (Moxon and Rappuoli, 2002). These criteria are also valid for malaria parasite as the two major vaccine candidates viz., Circumsporozoite (Cerami et al., 1992) and Thrombospondin-related anonymous protein (TRAP) (Robson et al., 1995) are involved in the invasion of liver cells by the parasite. With the availability of the genome sequence of *P. falciparum*, the complete genetic blueprint of the parasite is now known and, using existing recombinant DNA technologies, it is now possible to identify new antigens with potential for vaccine development.

Accomplishments

We have taken a systematic approach to identify parasitic proteins that are expressed on the sporozoite surface, can be involved in pathogenesis, and are possible targets for intervention. Using a combination of in-silico algorithms, available microarray and proteomic analysis of *P. falciparum* sporozoites, we initially selected 17 genes representing sporozoite stage *P. falciparum* proteins. The selection criteria included predicted antigenicity (Jameson and Wolf, 1988), presence of adhesive motifs (Marchler-Bauer and Bryant, 2004), presence of a secretory signal sequence using SignalP 3.0 (Bendtsen et al., 2004), transmembrane anchor sequence (using TMHMM 2.0), and homology to known surface antigens in other

Table 1. List of selected *P. falciparum* antigens expressed at the sporozoite stage of the parasite's lifecycle. Names in bold represent hypothetical proteins that have been cloned and expressed, and the purified protein has been obtained for further analysis.

	Name	Chromosome	Annotation	
			Size (AA)	Exons
1.	**PF11_0069**	**11**	**276**	**8**
2.	**PF14_0446**	**14**	**205**	**3**
3.	**PFL0870w**	**12**	**316**	**1**
4.	**PF11_0349**	**11**	**99**	**1**
5.	**PFA0200w**	**1**	**163**	**3**
6.	**PFL0800c**	**12**	**182**	**1**
7.	MAL8P1.45	8	891	3
8.	PFC0215c	3	474	4
9.	PF14_0467	14	248	3
10.	PFI0580c	9	413	2
11.	PF11_0394	11	186	2
12.	PFA0490w	1	234	1
13.	PFE0565w	5	133	1
14.	PF07_0089	7	467	1
15.	PFI0595c	9	453	1
16.	PF14_0016	14	107	1
17.	PF14_0344	14	993	1

Figure 1. Expression of PfFRAP during sporozoite stage of the parasite's life cycle. *P. falciparum* sporozoites were incubated with anti-FRAP antibodies (1:2560 dilution) (Panel A), normal mouse serum (Panel B), and monoclonal antibody directed against the repeat region of CS protein (1:5180 dilution) (Panel C), which was followed by the addition of anti-Mouse FITC conjugate.

pathosystems. All the selected genes have been annotated as hypothetical proteins in the published sequence (Gardner et al., 2002). Most of these genes are coded on multiple exons and therefore, to study the role of these hypothetical proteins, stage-specific RNA was isolated. To start with, DNA encoding only the coding sequence for six of these genes was amplified by RT-PCR, using RNA isolated *P. falciparum* parasites. These genes have been shown in bold in Table 1. The amplified products were cloned in pET101/TOPO, a T7 promoter-based *E. coli* expression vector. The recombinant expression of these constructs was performed in *E. coli*. *P. falciparum* has the highest known (81%) AT content in its genome. Expression of such a biased coding sequence in *E. coli* was achieved by standardizing culture conditions and utilizing *E. coli* strains that were more suited towards the expression of some of the codon. The DNA construct used for expression provides an inframe polyhistidine tag at the carboxyl terminus of the protein. Therefore, the proteins were purified to homogeneity by a combination of affinity and gel permeation chromatography.

While RT-PCR indeed verifies that a given gene is being transcribed by the parasite, it does not always directly correlate to its expression. Expression of a given protein in the parasite can only be detected using antibodies that specifically recognize the protein. Therefore, an effort to raise specific antibodies against each of the protein was undertaken following a standard protocol of subcutaneous immunization and boosting, in CD1 outbred mice. Antibodies were successfully raised against PF14_0446, PF11_0069, and PFL0870w and were subsequently used for detection of expression of the native proteins in the parasite. While expression of all these proteins could be easily detected in the parasites, protein PF14_0446 showed an interesting profile on the sporozoites (Figure 1). This protein showed very high expression on the sporozoite surface (Figure 1a) and its expression was comparable to the expression of Circumsporozoite protein (Figure 1c), a parasite surface antigen known for its role in the onset of malaria infection. The recognition was specific as serum from an unimmunized mouse did not detect the expression of the protein (Figure 1b).

To investigate the biological significance of protein PF14_0446 expression, we investigated the possibility that the protein could be involved in the adhesion and invasion of sporozoites to liver cells. The binding activity of the protein was evaluated on HepG2 cells, a human hepatocyte line that supports parasite adhesion and invasion in vitro (Rathore et al., 2002). The protein showed robust and dose dependent binding on liver cells, suggesting that it is involved in the adhesion of parasites to the liver cells. This adhesion activity was comparable to the activity of CS protein, the predominant surface antigen involved in malaria pathogenesis and currently being investigated as a vaccine candidate in Phase II trials. Sequence analysis of this protein had revealed that the carboxyl terminus of the protein has a weak homology to fasciclin 1,

(Figure 2a) a domain present in a large number of proteins with adhesive properties (Huber and Sumper, 1994). Therefore, we named this protein as FRAP (Fasciclin domain Related Adhesive Protein). Surprisingly, though the protein encodes this domain, the binding activity of the protein was not dependent on this domain as when we recombinantly prepared a truncated

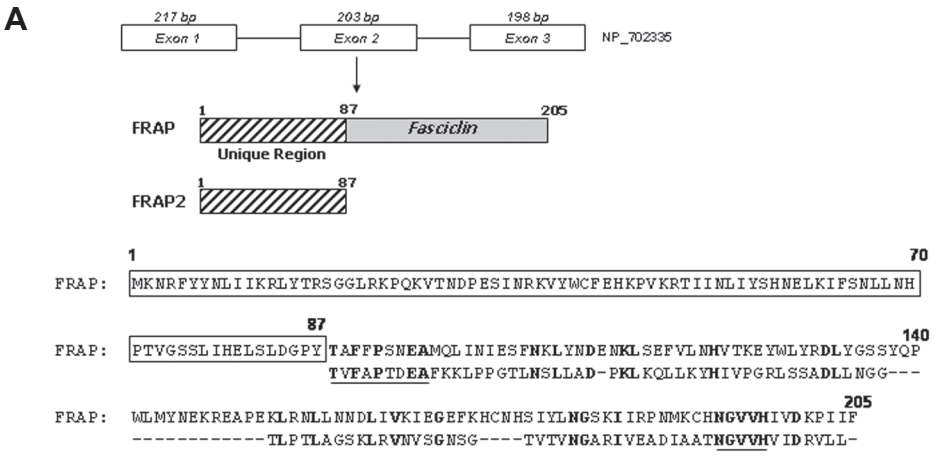


Figure 2. (a) Schematic representation of *P. falciparum* FRAP proteins. FRAP represents the full length protein encoded by 205 amino acids. FRAP2 represents a truncated version of the full length protein containing only amino acids 1-87, which have been boxed in the sequence. FAS1 domain of FRAP has been aligned with the consensus sequence of FAS1 domain and has an e-value of 2e-10. Residues shown in bold are the ones that are identical in both the FAS1 domain of FRAP and the consensus sequence. The two conserved motifs have been underlined. (b) Binding analysis of PfFRAP proteins on HepG2 cells.

version of this protein—FRAP2 (Figure 2b)—that lacked the fasciclin domain; there was no loss of binding to liver cells (Figure 2b).

We subsequently investigated the role of this protein in the process of host cell invasion. Both proteins and antibodies raised against the protein could inhibit the invasion of liver cells by *P. falciparum* sporozoites by as much as 94 percent (Table 2). The invasion activity of the protein resides in the amino terminus, as FRAP2, the truncated version of FRAP inhibited invasion with equal intensity (Table 2). This potent and highly effective blockade of liver cell invasion demonstrated that not only was FRAP critical for the onset of malaria infection, but antibodies against the protein could successfully prevent the parasite from invading liver cells, thus making this protein an ideal candidate for vaccine development.

Future Directions

The two malaria vaccine trials that have successfully moved to the advanced stages are based on the two known sporozoite proteins involved in liver cell invasion (Alonso et al., 2004; Moorthy et al., 2004). As FRAP shows biological properties that are at par with these vaccine candidates, investigations are currently underway to do a preliminary analysis of FRAP as a vaccine candidate for malaria in an animal model. The FRAP ortholog from *P. yoelii,* the rodent malaria parasite, has been cloned and the recombinant protein has recently been purified. The purified protein will be used for immunization and the animals will be challenged with *P. yoelii* sporozoites to investigate the potential of this protein to provide protection against malaria. The immunological analysis is being performed in collaboration with Prof. Fidel Zavala of Johns Hopkins University. Recognition of FRAP by malaria-infected subjects in the malaria-endemic regions of Mali will be investigated and a collaboration has been worked out with Dr. Kirsten Lyke at the Center for Vaccine Development, University of Maryland School of Medicine in Baltimore, MD.

We will continue to screen selected malarial proteins for their role in pathogenesis, as well as their recognition by the host immune system using serum from volunteers that participated in an irradiated sporozoite immunization and challenge trial. This work will be undertaken in collaboration with Dr. Rana Chattopadhyay, at the Malaria Program of Naval Medical Research Center in Silver Spring, MD. This will not only provide an antigenic map of the *Plasmodium* sporozoites, it will be of tremendous help in the development and optimization of a multi-antigen vaccine, designed to mimic the complexity of responses elicited against malaria infection.

Table 2. FRAP is involved in invasion of liver cells by *P. falciparum* sporozoites. Invasion of HepG2 cells by *P. falciparum* sporozoites was evaluated in the presence of different concentrations of free proteins or anti-FRAP2 serum and compared with the invasion in the presence of culture medium. Percentantage of inhibition represents the decrease in the number of sporozoites that invaded liver cells in comparison to the invasion level in cells incubated with medium.

Treatment	Concentration	% Inhibition
Medium		-
FRAP	20 µg/ml	89.5 ± 1.0
	10 µg/ml	80.9 ± 1.0
FRAP2	20 µg/ml	92.4 ± 3.5
	10 µg/ml	88.1 ± 4.6
CS Protein	20 µg/ml	92.6 ± 2.0
Anti-FRAP2 serum	1:100	94.6 ± 1.2
Anti-CS monoclonal	100 µg/ml	97.4 ± 0.7

Cryptosporidiosis

Background and significance

Transmission of a pathogen by contaminated drinking water supply is a major health issue for communities and governments alike. Combine it with a pathogen that is resistant to commonly used methods of water treatment and it becomes a significant public threat. These very properties of *Cryptosporidium* have made this parasite a Category-B priority pathogen. *Cryptosporidium* parasites are the third major cause of profuse diarrhea, mucosal inflammation, and gastroenteritis in humans. There is no effective treatment against cryptosporidiosis, which highlights the need to understand the underlying mechanism behind the onset of disease that will lead to effective methods for its control. The infection starts with the ingestion of parasitic oocysts in contaminated drinking water, which excyst in the intestinal milieu of the host and releases infective sporozoites in the gastrointestinal environment. The sporozoites immediately attach to the gastrointestinal mucosa, where they undergo development leading to clinical pathology of this disease. Attachment of *Cryptosporidium* parasite to the gastrointestinal mucosa is a prerequisite for the pathophysiological events in infection. At the molecular level, cryptosporidiosis is poorly understood due to limited knowledge of the interactions that take place between the host and the parasite. Recently, it has been shown that healthy individuals develop antibodies against surface antigens in *Cryptosporidium* (Okhuysen et al., 2004), suggesting that antibody responses against these proteins could be important in disease control. As genome sequence of *Cryptosporidium* parasites have now become available, several novel classes of cell surface and secreted proteins, which could possibly play a role in host-parasite interaction and pathogenesis, have been detected (Abrahamsen et al., 2004; Xu et al., 2004). We are investigating the contributions of some of the newly identified *C. parvum* genes in the pathogenesis.

Accomplishments

During the course of evolution, parasites have acquired a set of domains that are also expressed by their hosts. While such domains in hosts (humans) perform vital biological functions, in the case of parasites, proteins encoding these domain play an important role in host-parasite interactions (Cerami et al., 1992; Sultan et al., 1997). The domains by themselves do not contribute to this activity (Rathore et al., 2002) and the "domain free" regions of the protein are critical for the biological activities of the protein. This fact, which has recently come to light by our investigations in *Plasmodium* parasite, as described above for FRAP and shown elsewhere for CS protein (Rathore et al., 2002), led to the selection of 3 *C. parvum* proteins for investigating their role in pathogenesis. These proteins viz., CpTSP4, CpTSP7, and CpTSP10, encode TSP1 domains in combination with Apple, EGF, and Kringle domains respectively (Figure 3a).

To identify the role of the three *C. parvum* proteins in pathogenesis, DNA encoding of the mature protein sequence was amplified by PCR (Figure 3b) using gene specific primers and cloned in a pET101/D-TOPO, a T7 promoter-based *E. coli* expression vector (Figure 3a). On expression, all the three proteins were localized in inclusion bodies. Initially, purification conditions were standardized and CpTSP7 was purified (Figure 3c) by an in-vitro denaturation-renaturation protocol under redox conditions (Rathore et al., 2003), followed by a two step column chromatography. While parasite adhesion and development can be achieved in vitro in human cell lines of epithelial origin (Joe et al., 1998; Upton et al., 1994), an assay to evaluate the role individual *Cryptosporidium* antigens in pathogenesis was not available. Therefore, using Caco2, a human gastrointestinal cell line, we developed an in vitro cell-based assay system, to evaluate the host cell adhesion properties of the parasitic proteins in vitro. Protein CpTSP7 showed a dose dependent binding to Caco-2 cells which suggested that CpTSP7 is involved in *Cryptosporidium* pathogenesis (Figure 3d). Work is currently underway to purify the remaining proteins and to evaluate their role in disease pathogenesis.

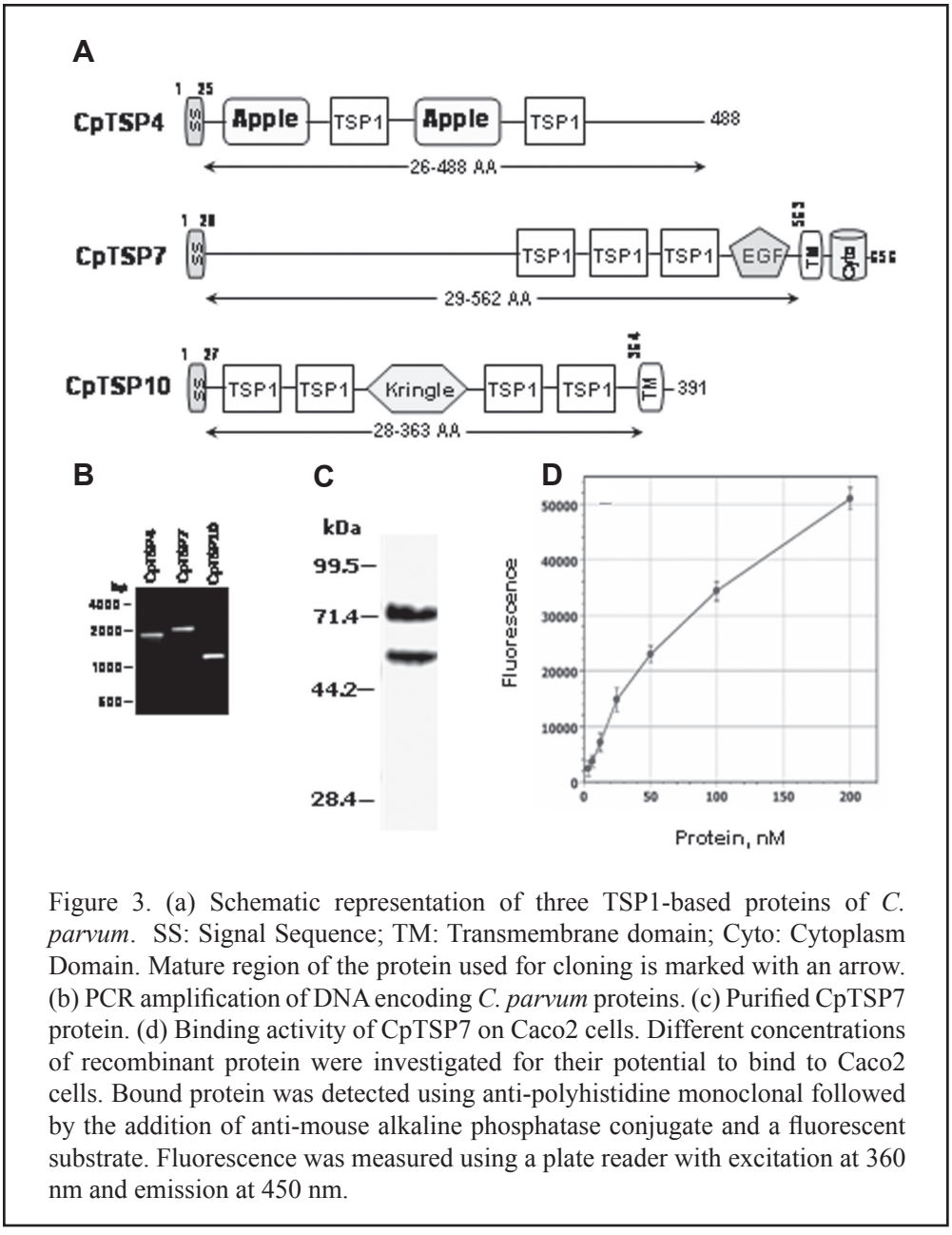


Figure 3. (a) Schematic representation of three TSP1-based proteins of *C. parvum*. SS: Signal Sequence; TM: Transmembrane domain; Cyto: Cytoplasm Domain. Mature region of the protein used for cloning is marked with an arrow. (b) PCR amplification of DNA encoding *C. parvum* proteins. (c) Purified CpTSP7 protein. (d) Binding activity of CpTSP7 on Caco2 cells. Different concentrations of recombinant protein were investigated for their potential to bind to Caco2 cells. Bound protein was detected using anti-polyhistidine monoclonal followed by the addition of anti-mouse alkaline phosphatase conjugate and a fluorescent substrate. Fluorescence was measured using a plate reader with excitation at 360 nm and emission at 450 nm.

Future Directions

The recombinantly expressed and purified proteins will be characterized by several mechanisms to assess their role in pathogenesis. We have recently raised specific antibodies against these proteins in mice. In collaboration with Prof. David Lindsay at the College of Veterinary Medicine, Virginia Tech, we will investigate if these antibodies can block the parasitic infection in a rodent model.

One of the possible mechanisms of infection control will be to block the interaction of *Cryptosporidium* sporozoites with host cells. While evidence suggests that *Cryptosporidium* sporozoites bind to epithelial cells of the intestine by specific receptor-ligand interactions (Joe et al., 1994; Joe et al., 1998; Thea et al., 1992), we have limited knowledge about the molecular events that precede the visible pathological symptoms associated with Cryptosporidiosis. Recombinantly purified *Cryptosporidium* proteins will be utilized for the identification and characterization of host receptors exploited

by the parasites. The identification of the host receptor(s) will be an important step in designing chemical analogs that can mimic these receptors and prevent the parasite from attaching to the target host cells, thus acting as a prophylactic agent.

This investigation will advance research on many fronts including pathogenesis, discovery of new potential drug targets, development of new genetic tools, and perhaps development of new disease control strategies.

References

Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto CA, Deng M, Liu C, Widmer G, Tzipori S, Buck GA, Xu P, Bankier AT, Dear PH, Konfortov BA, Spriggs HF, Iyer L, Anantharaman V, Aravind L, and Kapur V (2004) Complete genome sequence of the apicomplexan Cryptosporidium parvum. *Science* **304**: 441–445

Alonso PL, Sacarlal J, Aponte JJ, Leach A, Macete E, Milman J, Mandomando I, Spiessens B, Guinovart C, Espasa M, Bassat Q, Aide P, Ofori-Anyinam O, Navia MM, Corachan S, Ceuppens M, Dubois MC, Demoitie MA, Dubovsky F, Menendez C, Tornieporth N, Ballou WR, Thompson R, and Cohen J (2004) Efficacy of the RTSS/AS02A vaccine against Plasmodium falciparum infection and disease in young African children: randomised controlled trial. *Lancet* **364**: 1411–1420

Bendtsen JD, Nielsen H, von Heijne G, and Brunak S (2004) Improved prediction of signal peptides: SignalP 30. *J Mol Biol* **340**: 783–795

Cerami C, Frevert U, Sinnis P, Takacs B, Clavijo P Santos MJ, and Nussenzweig V (1992) The basolateral domain of the hepatocyte plasma membrane bears receptors for the circumsporozoite protein of Plasmodium falciparum sporozoites. *Cell* **70**: 1021–1033

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, and Barrell B (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* **419**: 498–511

Hoffman S (1996) Malaria Vaccine Development: A multi immune response approach. ASM press Washington DC

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, and Richie TL (2002) Protection of humans against malaria by immunization with radiation- attenuated Plasmodium falciparum sporozoites. *J Infect Dis* **185**: 1155–1164

Huber O and Sumper M (1994) Algal-CAMs: isoforms of a cell adhesion molecule in embryos of the alga Volvox with homology to Drosophila fasciclin I. *Embo J* **13**: 4212–4222

Jameson BA and Wolf H (1988) The antigenic index: a novel algorithm for predicting antigenic determinants. *CABIOS* **4**: 181–186

Joe A, Hamer DH, Kelley MA, Pereira ME, Keusch GT, Tzipori S, and Ward HD (1994) Role of a Gal/GalNAc-specific sporozoite surface lectin in Cryptosporidium parvum-host cell interaction. *J Eukaryot Microbiol* **41**: 44S

Joe A, Verdon R, Tzipori S, Keusch GT, and Ward HD (1998) Attachment of Cryptosporidium parvum sporozoites to human intestinal epithelial cells. *Infect Immun* **66**: 3429–3432

Marchler-Bauer A and Bryant SH (2004) CD-Search: protein domain annotations on the fly. *Nucleic Acids Res* **32**: W327–331

Moorthy VS, Imoukhuede EB, Keating S, Pinder M, Webster D, Skinner MA, Gilbert SC, Walraven G, and Hill AV (2004) Phase 1 evaluation of 3 highly immunogenic prime-boost regimens including a 12-month reboosting vaccination for malaria vaccination in Gambian men. *J Infect Dis* **189**: 2213–2219

Moxon R and Rappuoli R (2002) Bacterial pathogen genomics and vaccines. *Br Med Bull* **62**: 45–58

Nussenzweig RS, Vanderberg J, Most H, and Orton C (1967) Protective immunity produced by the injection of x-irradiated sporozoites of plasmodium berghei. *Nature* **216**: 160–162

Okhuysen PC, Rogers GA, Crisanti A, Spano F, Huang DB, Chappell CL, and Tzipori S (2004) Antibody response of healthy adults to recombinant thrombospondin-related adhesive protein of cryptosporidium 1 after experimental exposure to cryptosporidium oocysts. *Clin Diagn Lab Immunol* **11**: 235–238

Rathore D, Hrstka SC, Sacci JB, Jr De la Vega P, Linhardt RJ, Kumar S, and McCutchan TF (2003) Molecular mechanism of host specificity in Plasmodium falciparum infection: role of circumsporozoite protein. *J Biol Chem* **278**: 40905–40910

Rathore D, Sacci JB, de la Vega P, and McCutchan TF (2002) Binding and invasion of liver cells by Plasmodium falciparum sporozoites Essential involvement of the amino terminus of circumsporozoite protein. *J Biol Chem* **277**: 7092–7098

Robson KJ, Frevert U, Reckmann I, Cowan G, Beier J, Scragg IG, Takehara K, Bishop DH, Pradel G, Sinden R, and et al (1995) Thrombospondin-related adhesive protein (TRAP) of Plasmodium falciparum: expression during sporozoite ontogeny and binding to human hepatocytes. *Embo J* **14**: 3883–3894

Sultan AA, Thathy V, Frevert U, Robson KJ, Crisanti A, Nussenzweig V, Nussenzweig RS, and Menard R (1997) TRAP is necessary for gliding motility and infectivity of plasmodium sporozoites. *Cell* **90**: 511–522

Thea DM, Pereira ME, Kotler D, Sterling CR, and Keusch GT (1992) Identification and partial purification of a lectin on the surface of the sporozoite of Cryptosporidium parvum. *J Parasitol* **78**: 886–893

Upton SJ, Tilley M, and Brillhart DB (1994) Comparative development of Cryptosporidium parvum (Apicomplexa) in 11 continuous host cell lines. *FEMS Microbiol Lett* **118**: 233–236

Xu P, Widmer G, Wang Y, Ozaki LS, Alves JM, Serrano MG, Puiu D, Manque P, Akiyoshi D, Mackey AJ, Pearson WR, Dear PH, Bankier AT, Peterson DL, Abrahamsen MS, Kapur V, Tzipori S, and Buck GA (2004) The genome of Cryptosporidium hominis. *Nature* **431**: 1107–1112

Publications

Mahajan B, Jani D, Chattopadhyay C, Nagarkatti R, Zheng H, Weiss W, Kumar S, and Rathore D (2005) Cloning, Expression of Characterization of a Plasmodium knowlesi Protein Containing an Altered Thrombospondin Repeat Domain. *Infect Immun* In press

McCutchan TM, Grim KC, Li J, Weiss W, Rathore D, Sullivan M, Graczyk TK, Kumar S, and Cranfield MR (2004) Measuring the Effects of an Ever-Changing Environment on Malaria Control. *Infect Immun* **72**: 2248–2253

Rathore D, Nagarkatti R, Jani D, Chattopadhyay R, de la vega P, Kumar S, and McCutchan TM (2005) An immunologically cryptic epitope within the Circumsporozoite protein of *Plasmodium falciparum* sporozoites facilitates liver cell recognition and induces protective antibodies that block liver cell invasion. *J Biol Chem* In press-available online

Tewari R, Rathore D, and Crisanti A (2005) Motility and infectivity of Plasmodium berghei sporozoites expressing avian *Plasmodium gallinaceum* circumsporozoite protein. *Cell Microbiol* **7**: 699–707

Rathore D, McCutchan TF, and Kumar S (2005) Antimalarial Drugs: Current Development and the Future. *Expert Opin Investig Drugs* In press

Bioinformatics Applied to Mitochondrial Medicine

David C. Samuels
Research Assistant Professor, VBI
dsamuels@vbi.vt.edu

Summary

Our group focuses on applications of bioinformatics and computational biology to research medical conditions involving mitochondria. This is a wide-ranging medical field, with applications to diabetes, numerous myopathies, and neurological diseases. Mitochondria contain their own genome, called mtDNA, separate from the nuclear DNA. In vertebrates, the mtDNA only encodes 13 proteins, 22 tRNAs and two rRNAs, but these genes are all essential for the survival of the organism. In this report, I give an overview of our research, primarily through descriptions of the projects that have been published by our group over the last year. All of these projects relate to some aspect of mitochondrial DNA.

Genome Analysis

In this field, we are fortunate to have a large number of completely sequenced mitochondrial genomes over a wide range of species. NCBI currently contains the complete mtDNA sequences of 670 species, a wealth of data for comparative genome analysis. Many researchers use this data for evolutionary studies. Our focus is on using this data for medical research. We have published two papers in this area over the last year, both in *Trends in Genetics* and both related to aging.

"Mitochondrial DNA repeats constrain the lifespan of mammals" DC Samuels, *Trends in Genetics*, 20 (5) 226-229, May 2004.

In this paper, I analyzed the mtDNA sequences of 61 mammalian species for which we have both complete mtDNA genomes and longevity data. The mitochondrial theory of aging (a version of the free radical theory of aging) relates the rate of aging to the rate of damage to mitochondria, and in particular the rate of damage to mtDNA (Balaban et al., 2005; Chinnery et al., 2002; Elson et al., 2001; Trifunovic et al., 2004). There are many research groups working on the relationship between aging and the rate of production of reactive oxygen species (ROS) or the effectiveness of intracellular anti-oxidants (Barja, 2004). Our research is taking a different, and I believe unique, approach to this question. We are considering how the *susceptibility* of the mtDNA to ROS damage may vary with species, and how this may relate to longevity in that species. In this approach, we are evaluating how certain DNA sequence properties that are known to allow mutation (by known mechanisms) vary across species. In this study, I began the research program with a study of the frequency in the mtDNA of direct repeats, a risk factor for deletion. I found that there is a relationship of the longer direct repeats (> 10 base pairs) with life span in mammals. Life span constrained the number of longer direct repeats, with the short-lived species allowed to have a wide range of repeat numbers, low to high, but the long-lived mammals only having a low number of repeats. By comparing the values in the actual sequences to the same analysis of randomly shuffled sequences, I showed that, in general, a species' mitochondrial genome contains more direct repeats than would be expected in a random sequence of the same nucleotide content, indicating some unknown function for these direct repeats. Interestingly, humans contain the same number of direct repeats as would be expected in a random sequence.

"Two direct repeats cause most human mtDNA deletions" DC Samuels, EA Schon, PF Chinnery, *Trends in Genetics*, 20 (9) 393-398, September 2004.

In this paper, we focused on human aging. Starting in their 40's, every human begins to accumulate a particular mitochondrial mutation; the common deletion and the amount of this mutation in your cells increases as you age. This is a severe mutation, removing almost 1/3 of the mitochondrial genome. It always occurs in the same sequence location, between a pair of 13 bp direct repeats. In addition to the ubiquitous common deletion, there have been hundreds of other different deletions reported in human mtDNA. Usually, but not always, these other deletions are also flanked by direct repeats, but these are much smaller repeats, typically about 5 bp in length. The standard explanation has been that these rare deletions are formed directly due to these small direct repeats. In this paper we reject that explanation. We show that the distributions of the 5' and 3' ends of the rare deletions are not uniform, in contrast to the distribution of the small direct repeats, which is uniform. We consider a series of alternative formation mechanisms for these deletions that have been proposed in the literature, and show that none of them can explain the non-uniform distribution of the deletion ends. Instead, we show that this distribution can be explained by the common deletion itself (the distributions of the ends of the rare deletions are centered on the common deletion ends). We propose that the initial trigger for all of these deletions is the 13 bp direct repeat that is also responsible for the common deletion. In most cases the deletion process removes all the sequence between the direct repeats, forming the common deletion. In rare cases the deletion process that begins with the 13 bp repeats stops in a metastable state, often at a shorter direct repeat.

Future Directions

We are continuing the mtDNA mutation susceptibility work with genome analysis of other sequence properties associated with mutability. In work in progress, we are comparing the average binding energy of the mtDNA strands to longevity and are seeing a strong direct relationship. We are interpreting this as the effect of DNA breathing, where thermal fluctuations separate the DNA strands, temporarily exposing the mtDNA to a greater risk of oxidative damage. We are also analyzing the data on observed human mtDNA mutations in tumors to determine hotspots for mtDNA mutation and to develop hypothetical mutation mechanisms to explain these hotspots.

Metabolism Modeling

MtDNA molecules are copied about once every 10 days, even in post-mitotic cells where nuclear DNA replication has ceased. So mitochondria in all cells must contain a deoxyribonucleotide metabolism to support this replication process. This is a new area of research for us, as our previous work has focused on the mtDNA itself, not the metabolism that supports it. Our first paper in this area was accepted in January 2005 and even though the paper has not yet appeared in print, the online preprint has already sparked three new collaborations with groups at the medical schools of Emory and Indiana University and with another modeling group at Case Western. We expect this to develop into a major area of research for us.

"A computational model of mitochondrial deoxynucleotide metabolism and DNA replication" PC Bradshaw and DC Samuels, to appear in *The American Journal of Physiology: Cell Physiology*.

This paper presents the definition of the model, along with its first application to basic questions about nucleotide metabolism in mitochondria. The model includes transport of nucleosides and nucleotides between the organelle and the cytoplasm, a chain of phosphylation events, and polymerization of the triphosphate nucleotides into mtDNA during replication. It would have been more traditional, not to mention easier, to have published this initial paper in the mathematical biology literature since it does focus on the definition of the model. However, in order to get the right audience for the work, we aimed this paper at a mainstream biology journal and the rapid, positive response has

justified the effort.

As part of the description of the simulation model, we define and name a new concept in this metabolism regarding the flow of deoxynucleotide material between the mitochondrion and the cytoplasm. We define three possible states for the metabolism. In the "phosphorylating" state, the net flow of material into the mitochondrion comes through the nucleoside transporter, phosphorylated nucleotides are exported from the mitochondrion into the cytoplasm, and some of this flow is diverted to mtDNA replication. In the "dephosphorylating" state, the material flow is the reverse, again with some of the flow diverted to mtDNA replication. Only in the "efficient" state does the net flow of material enter the mitochondrion through both the nucleoside and nucleotide transporters with all the flux going into the replicating mtDNA. In the simulations we show that the efficient state occurs only within a small parameter range (the cytoplasmic concentrations of the four deoxynucleotides). Our philosophy is that the definition of qualitative concepts such as this is a valuable end product of any simulation and is often more robust and more informative to the biomedical community than are the straightforward quantitative results of the simulations (though these are important too!).

We ended this paper with a statement of five testable hypotheses that we derived from the results of the simulation. This kept the focus on the model as a "hypothesis generator" for experiment, not just a calculator of certain numbers. This emphasis was an important point in getting this paper published in the biological literature and in attracting the new collaborators.

For details of the modeling procedures and the quantitative results, I refer you to the paper. I have deliberatively kept this discussion more on the philosophical side, as an example of how we in the mathematical and computer science fields should structure our modeling work to communicate well with our biomedical collaborators. I believe that this project is an example of where we have the philosophy right. Our progress over the next few years will test this.

Future Directions

This area is a basic metabolism in all cells, and therefore there is a wealth of potential projects. We are currently applying this model to the antiviral medication AZT, used in AIDS treatment. Nucleoside analogs, including AZT, are a major group of antiviral drugs. However, this class of drugs also suffers from mitochondrial toxicity (Lewis and Dalakas, 1995), which is fatal in some cases. The nucleoside analogs follow the same metabolic paths as their natural nucleoside counterparts, and therefore they are a natural extension of the model discussed above. We have a paper in preparation on AZT metabolism in mitochondria and an R21 application submitted to NIH. In another direction, there are many genetic diseases due to mutations of the enzymes involved in this metabolism. This will be a major area of application of this model, and we have an R01 application submitted to NIH for the further development of the model with application to these genetic diseases. Finally, part of this model involves the action of the mtDNA polymerase. We intend to develop this model further to include polymerase fidelity data from experiment, in order to model the polymerase errors leading to mtDNA mutations. This will tie the nucleotide metabolism model into our research efforts on the mechanisms of mutation in mtDNA.

Hematopoietic Stem Cell Modelling

People who carry pathogenic mtDNA mutations generally have a mixture of mutant and wild-type mtDNA in all their cells, a condition called heteroplasmy. The proportion of mutant mtDNA in a cell changes over time, due to the dynamic and random nature of mtDNA replication and destruction. Since the wild-type mtDNA is keeping the cell alive, if the amount of wild-type mtDNA in the cell ever drops too low it is assumed that the cell dies. The loss of post-mitotic cells is the major pathogenic mechanism in most mitochondrial diseases. However, in

dividing cell populations this same process of cell loss may clear the mtDNA mutation from the system.

"A simulation methodology in modeling cell divisions with stochastic effects" HK Rajasimha, RE Nance, and DC Samuels, *Proc. Of the 2004 Winter Simulation Conference*, pg 2032-2038.

The progression and severity of a patient's mitochondrial disease is usually determined through the mtDNA mutation level in a muscle biopsy, a painful and potentially damaging procedure. It has long been noted that the measurement of mtDNA mutation levels in a patient's blood has little relevance for determining the progression of the disease. MtDNA mutation levels in blood are usually lower than that in muscle in the same patient, and elderly patients may show no mtDNA mutation at all in a blood sample. In this paper, we begin to apply our modeling to hematopoietic (blood) stem cells, building on our earlier work on modeling colon crypt stem cells (Taylor et al., 2003). We model a population of approximately 200,000 hematopoietic stem cells (about 1/10 the total number of these cells in a human) over 100 years. The model includes mtDNA replication, destruction and stem cell division, which occurs only about once per year in these cells. The stem cell division may be asymmetric (resulting in one stem cell and one progenitor cell that continues to divide rapidly and differentiate into blood cells), or symmetric (either two stem cells or two progenitor cells). In a separate model, also discussed in this paper, we model the series of rapid divisions (about 20) of the progenitor cells to form the blood cells.

The final mechanism in the model is that any stem cells with an mtDNA mutation level greater than a threshold (typically 90 percent mutant) is removed from the stem cell population. Over long time scales, the random drift in the stem cell mutation level and the loss of stem cells with high mutation level naturally leads to an exponential loss of the mutation from the total stem cell population, and therefore from the blood.

Future directions

While the idea that mtDNA mutation levels decrease in blood over time is a well accepted one, all the clinical measurements have attempted to measure a linear rate of decrease, not an exponential one. It is no surprise then (using our understanding from these simulations) that the measured linear rates of decrease have always been inconsistent between patients. We have analyzed the measurements from a published clinical study (Rahman et al., 2001) and a new clinical study by our collaborators and have shown that an exponential decay, in the range predicted by the simulation, does fit all the clinical data from different studies. The new clinical data was taken in response to our simulation results, an example where we have tested and validated in the lab a hypothesis generated from our simulations. The paper on this is in preparation now.

In this project, we have developed a very general tool for modeling mtDNA in large populations of dividing cells, and we will be applying this tool to other systems. The simplest application is to cell cultures, where much experimental data is available. A more medically relevant application of this work is to early embryogenesis, following the development of the embryos from the single fertilized egg cell to a mass of a few tens of millions of cells (the current practical limit of the simulation). This simulation would deal with questions about the non-uniform distribution of mtDNA mutations throughout the body (which may cause the highly variable phenotypes of many mtDNA mutations) and will tie in with our longstanding and continuing work on mtDNA inheritance (Brown et al., 2001; Chinnery et al., 2000b).

Other Work and Future Projects

This report is primarily organized by the published work from our group over the past year. In addition to this, much of our research efforts have been devoted to developing new projects and new collaborations that have not yet progressed to the publication stage, and this work deserves at least a brief discussion.

Muscle modeling

Muscle fibers are post-mitotic, giant multi-nucleated cells that can reach centimeters in length. In an earlier experimental study (Elson et al., 2002) we showed that in two patients with myopathy due to mtDNA mutations, the mitochondrial dysfunction occurred in numerous isolated segments on each muscle fiber, interspersed with normally functioning segments of the fiber. These dysfunctional sections were of variable length and the total amount of dysfunctional fiber was largest in the more severely affected patient. From this data, we proposed the hypothesis that the progression of the myopathy was primarily driven by the growth in length of these dysfunctional segments and was only secondarily due to increased formation of new dysfunctional segments. In related work (Taivassalo et al., 1998; Taivassalo et al., 2001), clinical studies of exercise therapy for patients with mitochondrial myopathies have had the unpleasant surprise that these therapies actually were harmful to most patients, resulting in an increased mtDNA mutation level in muscle biopsies. These results have made exercise therapy a very controversial area in this field, and an increased understanding of this process is urgently needed to determine the proper patient care.

This year we designed a simulation of mtDNA in skeletal muscle fibers to further develop our hypothesis about the progression of mitochondrial myopathies. We will model a single muscle fiber as a linear series of a few hundred compartments. Our standard mtDNA simulation will run in each compartment with the addition of movement of mtDNA between neighboring compartments (modeled as diffusion). We will apply this simulation to the normal progression of myopathy and also to models of exercise therapy (both aerobic and strength training). Since the muscle fiber segments with dysfunctional mitochondria are also a feature of normal aging, we propose to apply this model to mtDNA mutations acquired in the aging process, and to the question of whether exercise therapy in the elderly may suffer from the same drawbacks as it does in mitochondrial myopathy patients. We have made an R01 application to NIH on this project, and the primary researchers on the clinical studies of exercise therapy for the myopathy patients are co-investigators on that application.

Inheritance and epidemiology of mtDNA mutations

The inheritance of mtDNA is strictly maternal. However, this is complicated in the case of people carrying mixtures of wild-type and mutant mtDNA. In these cases, each person is represented by a single value of percent mutant mtDNA, called the *heteroplasmy level*, usually measured through a muscle biopsy (though this value can both change over time and vary across tissues in the body). The heteroplasmy levels of a mother and offspring can differ by a large and apparently random amount. In an earlier study (Chinnery et al., 2000b) of mothers and offspring, we quantified this heteroplasmy shift. In our current work, we are using that data as the basis for a model of mtDNA heteroplasmy shifts over multiple generations in a maternal pedigree, to investigate how mtDNA mutations progress from the *de novo* mutation in an asymptomatic individual, to the appearance of the disease in later generations, and finally to the loss of the pathogenic mutation from the pedigree (a similar principal to the loss of the mutation from stem cells, discussed above). We have applied to NIH for an R01 grant to investigate this, with an expert in mtDNA epidemiology (Chinnery et al., 2000a; Chinnery and Turnbull, 2001) (and a long time collaborator) as co-PI. We are preparing the preliminary studies for publication.

Diabetes and beta cell function

The secretion of insulin from pancreatic beta cells in response to a glucose challenge is controlled by changes in the membrane potential of the mitochondria in those cells. The disruption of this insulin secretion is an early sign of type II diabetes. Modeling of this process has been carried out previously; however, these models invariably use a single mitochondrial compartment. In reality, a single beta cell contains a few hundred separate mitochondria. These mitochondria have variable properties

that will affect the rate that their membrane potential will change in response to the glucose challenge. Therefore, insulin secretion from a beta cell is not driven by a single mitochondrial oscillator, but is instead driven by hundreds of inhomogeneous oscillators. There has been a great deal of study of entrainment in coupled inhomogeneous oscillators in the physics literature over the past 20 years. We have proposed to apply this theory from physics to this biomedical problem in an application to the ASPIRES program. This work would be carried out with a new collaborator from Tufts medical school who is measuring the inhomogenieties in beta cell mitochondria.

An Internet resource for pathogenic mt-tRNA mutations

In this project, we have applied for funding from IBM and AFM (the Association Francaise contre les Myopathies) to develop a website to report a "pathogenicity score" for all known mt-tRNA mutations (McFarland et al., 2004). This project will also have a bioinformatics component to assess the impact of every possible point mutation in each mt-tRNA on the set of predicted secondary structures, including suboptimal structures.

References

Balaban RS, Nemoto S, and Finkel T (2005) Mitochondria, oxidants, and aging. *Cell* **120**: 483–495

Barja G (2004) Free radicals and aging. *Trends Neurosci* **27**: 595–600

Brown DT, Samuels DC, Michael EM, Turnbull DM, and Chinnery, PF (2001) Random genetic drift determines the level of mutant mtDNA in human primary oocytes. *Am J Hum Genet* **68**: 533–536

Chinnery PF, Johnson MA, Wardell TM, Singh-Kler R, Hayes C, Brown DT, Taylor RW, Bindoff LA, and Turnbull DM (2000a) The epidemiology of pathogenic mitochondrial DNA mutations. *Ann Neurol* **48**: 188–193

Chinnery PF, Samuels DC, Elson J, and Turnbull DM (2002) Accumulation of mitochondrial DNA mutations in ageing, cancer, and mitochondrial disease: is there a common mechanism? *Lancet* **360**: 1323–1325

Chinnery PF, Thorburn DR, Samuels DC, White SL, Dahl HHM, Turnbull DM, Lightowlers RN, and Howell N (2000b) The inheritance of mitochondrial DNA heteroplasmy: random drift, selection or both? *Trends Genet* **16**: 500–505

Chinnery PF and Turnbull DM (2001) Epidemiology and treatment of mitochondrial disorders. *Am J Med Genet* **106**: 94–101

Elson JL, Samuels DC, Johnson MA, Turnbull DM, and Chinnery PF (2002) The length of cytochrome c oxidase-negative segments in muscle fibres in patients with mtDNA myopathy. *Neuromuscul Disord* **12**: 858–864

Elson JL, Samuels DC, Turnbull DM, and Chinnery PF (2001) Random intracellular drift explains the clonal expansion of mitochondrial DNA mutations with age *Am J Hum Genet* **68**: 802–806

Lewis W and Dalakas MC (1995) Mitochondrial Toxicity of Antiviral Drugs. *Nat Med* **1**: 417–422

McFarland R, Elson JL, Taylor RW, Howell N, and Turnbull DM (2004) Assigning pathogenicity to mitochondrial tRNA mutations: when 'definitely maybe' is not good enough. *Trends Genet* **20**: 591–596

Rahman S, Poulton J, Marchington D, and Suomalainen A (2001) Decrease of 3243 A -> G mtDNA mutation from blood in MELAS syndrome: A longitudinal study. *Am J Hum Genet* **68**: 238–240

Taivassalo T, De Stefano N, Argov Z, Matthews PM, Chen J, Genge A, Karpati G, and Arnold DL (1998) Effects of aerobic training in patients with mitochondrial myopathies. *Neurology* **50**: 1055–1060

Taivassalo T, Shoubridge EA, Chen J, Kennaway NG, DiMauro S, Arnold DL, and Haller RG (2001) Aerobic conditioning in patients with mitochondrial myopathies: Physiological, biochemical, and genetic effects. *Ann Neurol* **50**: 133–141

Taylor RW, Barron MJ, Borthwick GM, Gospel A, Chinnery PF, Samuels DC, Taylor GA, Plusa SM, Needham SJ, Greaves LC, Kirkwood TBL, and Turnbull DM (2003) Mitochondrial DNA mutations in human colonic crypt stem cells. *J Clin Invest* **112**: 1351–1360

Trifunovic A, Wredenberg A, Falkenberg M, Spelbrink JN, Rovio AT, Bruder CE, Bohlooly-Y M, Gidlof S, Oldfors A, Wibom R, Tornell J, Jacobs HT, and Larsson NG (2004) Premature ageing in mice expressing defective mitochondrial DNA polymerase. *Nature* **429**: 417–423

Bioinformatics For and From Microbial Genomics

João Carlos Setubal
Research Associate Professor, VBI
Associate Professor of Computer Science, Virginia Tech
setubal@vbi.vt.edu

Introduction

This paper outlines the results of the research work that I have done since 2004. Most of the work has to do with bioinformatics support for microbial genome sequencing projects, a theme with which I have been heavily involved since 1997.

The words "for" and "from" in the title suggest the fascinating give-and-take game that biology and informatics have played for the past 10 years or so. Initially, it was the task of informatics to provide the tools with which genome projects could be done; from there was born the term *bioinformatics*. But the wealth of data generated in these projects created new and interesting problems, which stimulated the advancement of bioinformatics far beyond a mere tool-providing service. It is in this interplay that I locate most of my research.

Microbial genome projects

Microbial genome sequencing projects have become almost routine in the past few years. The NCBI website currently (March 2005) lists 225 complete bacterial and archaeal genomes. It is expected that in the next few years, there will be well over a thousand complete microbial genomes.

In the following, I describe genome projects that have recently been completed using an annotation infrastructure developed under my guidance at the Laboratory of Bioinformatics at the University of Campinas, Brazil. In this section, I give the major biological insights gained by these projects. The highlights of the bioinformatics work are given in the next section.

***Leptospira interrogans* serovar Copenhageni (LIC).** This genome was described by Nascimento et al. (Nascimento et al., 2004; Nascimento et al., 2004). The genus *Leptospira* comprises a heterogeneous group of pathogenic and saprophytic species belonging to the phylum Spirochaetales. *Leptospira interrogans* is the causative agent of the important human disease leptospirosis. Because of the large spectrum of animal species that serve as reservoirs, leptospirosis is considered to be the most widespread zoonotic disease in the world. In this project, the strain Fiocruz L1-130 was sequenced, annotated, and compared to the genome of *Leptospira interrogans* serovar Lai (Ren, et al., 2000), and to the genomes of other completely sequenced spirochetes (*Borrelia burgdorferi* and *Treponema pallidum*). The genome has two chromosomes 4.3 Mbp and 350 Kbp long. Even though the Copenhageni and Lai genomes are very similar, a large chromosomal inversion was observed between the Copenhageni and Lai large chromosomes. A broad array of transcriptional regulation proteins and two new families of afimbrial adhesins (which contribute to host tissue colonization in the early steps of infection) were identified in the LIC genome. Differences between the Copenhageni and Lai serovars in genes involved in the biosynthesis of lipopolysaccharide O-side chains were identified offering an important starting point for the elucidation of the organism's complex polysaccharide surface antigens. Differences in adhesins and in lipopolysaccharides might be associated with the adaptation of serovars Copenhageni and Lai to different animal hosts. Hundreds of genes encoding surface-exposed lipoproteins and outer membrane proteins were identified as candidates for development as

vaccines for prevention of leptospirosis.

***Leifsonia xyli* subsp. *xyli* (Lxx).** This genome was described by Monteiro-Vitorello et al (Monterio-Vitorello, et al., 2004). Lxx is a member of the Gram-positive GC-rich Order Actinomycetales. Lxx is the causative agent of ratoon stunting disease in sugarcane. In this project, strain CTCB07 was sequenced and annotated. The genome consists of one 2.6 Mbp circular chromosome. Its analysis revealed more pseudogenes than any bacterial plant pathogen sequenced to date. Many of these pseudogenes, if functional, would likely be involved in the degradation of plant heteropolysaccharides, uptake of free sugars, and synthesis of amino-acids. The numbers of predicted regulatory genes and sugar transporters found in CTCB07 are more typical of those seen in free-living organisms even though it is a xylem-limited bacterium that has not been identified outside of sugarcane. Many of the predicted pathogenicity genes identified appear to have been acquired by lateral transfer. Among these are a cellulase, a pectinase, a wilt-inducing protein, a lysozyme, and a desaturase. The presence of the latter may contribute to stunting since it is likely involved in the synthesis of abscisic acid, a hormone that arrests growth. The findings are consistent with the nutritionally fastidious behavior exhibited by Lxx and suggest an ongoing adaptation to the restricted ecological niche it inhabits.

In 2001, the genome of *Agrobacterium tumefaciens* C58 was published (Wood, Setubal, et al., 2001). C58 is known as biovar 1 for the genus *agrobacterium.* As a follow-up to that project a consortium of institutions is sequencing biovars 2 (*A. radiobacter* K84) and 3 (*A. vitis* S4). My group at VBI is providing the bioinformatics support for this project, with some details given in the next section.

Genome annotations in projects, such as the ones described above, are a first pass in extracting and understanding the wealth of information that such projects collect. Following the publication of the sequence and its annotation, more in-depth studies usually follow. Recently I have been part of two such studies, focusing on the organisms *Xylella fastidiosa* 9a5c (a citrus pathogen), *Xylella fastidiosa* temecula 1 (a grapevine pathogen), *Xanthomonas axonopodis* pv *citri* (another citrus pathogen), and *Xanthomonas campestris* pv. *campestris* (a brassica pathogen). They are all gamma-proteobacteria of the *Xanthomonadaceae* family. The studies were published in references (Moreira et al., 2005) and (Moreira et al., 2004). In (Moreira et al., 2004) the focus was on the citrus pathogens. Several genes and operons that may be relevant for adaptation to citrus were identified. In (Moreira et al. 2005) this analysis was expanded to include the other *Xylella* and the other *Xanthomonas*, thereby yielding a more complete comparison of all fully sequenced *Xanthomonadaceae.*

In addition to the two studies mentioned above, I co-authored two reviews of genomics of phytopathogens (Wood et al., 2004; Setubal and da Silva, 2004).

Bioinformatics analyses of microbial genomes

In a genome project, there are three major tasks: sequencing, assembly, and annotation. Assembly and annotation are tasks that require sophisticated bioinformatics support. An annotation infrastructure has the following basic components: 1) processing pipeline for automatic annotation, 2) database to house the sequence and all annotation data, 3) genome analysis tools, and 4) interface to the database for viewing and editing annotation data, and for running the various analysis tools. The availability of genomes of closely related species has led to the need for an annotation infrastructure that is able to deal with several genomes at the same time; this has an impact primarily on the database and on the analysis tools, which should include programs that provide some kind of comparison among genomes of interest.

The annotation infrastructure used for the LIC and Lxx projects used my SQL as the database, an interface based on perl CGI, and various

third-party genome analysis tools. In the LIC project, there was a need to develop special software for detecting lipoproteins. A group composed of myself, my student, Marcelo Reis, and David Haake from UCLA developed the tool SPLIP (spirochaete lipoprotein detector). The program is based on a set of 28 experimentally confirmed lipoproteins in spirochaetes and on a set of rules for lipobox recognition in these organisms, as described in Haake (Haake, 2000). SPLIP uses position specific scoring matrices to evaluate potential lipoboxes in the first 70 amino acids of every predicted coding sequence. The ones that have a putative lipobox (positive score) are then evaluated in terms of their putative hydrophobic region amino acid composition and cleavage site position, resulting in a re-scoring. Putative lipoproteins are classified as "probable" and "possible," reflecting what is known about the occurrence of certain amino acids in the detected lipobox. A manuscript describing SPLIP is in preparation.

In the Lxx project, the bioinformatics highlights were pseudogene and genomic island detection. Pseudogenes were detected using a combination of manual curation and automatic parsing of BLASTX results (Carriao et al., 2004). We are currently studying this problem in depth and hope to come up with better methods for automatic detection of pseudogenes.

In the *Agrobacterium* biovar project, the main bioinformatics development has been the development of a new curation infrastructure called GAT: Genome Annotation Tool. This tool is a significant improvement over the earlier one in the following respects: 1) It allows storage of multiple genomes instead of only one. This is accomplished by the usage of the Genomics Unified Schema (http://www.gusdb.org), developed by the CYBIL group at the University of Pennsylvania. 2) For genome browsing, it uses Gbrowse, a module part of GMOD. 3) In addition to "standard" features for gene editing and navigation, it includes the capability for linking orthologous and paralogous genes and anomalous regions. Development of GAT is ongoing and two abstracts from this meeting describe it in more detail. One (Jhaveri and Setubal) is about GAT proper; the other (Zheng and Setubal) describes methodology and results for detection of genomic islands in *A. vitis* S4.

PATRIC

Bruno Sobral and I are coPIs in the PathoSystems Resource Integration Center (PATRIC) project (http://patric.vbi.vt.edu). This is a 5-year contract (2004-2009) awarded by NIH-NIAID to establish a multi-organism database of curated genomics information and additional computational resources. The organisms whose genomes will be curated are *Rickettsia* spp., *Brucella* spp., *Coxiella burnetii* (Bacteria), and Coronaviruses, Caliciviruses, Rabies viruses, Hepatitis A virus, and Hepatitis E virus. The goal is to provide a resource that will help researchers in understanding the genomic basis of disease and in the development of improved vaccines, diagnostics, and therapeutics.

This project is in its early stages as of this writing. But it has allowed us to reflect upon the nature of genome curation. What is curation and how does it differ from annotation? My definition is that curation is a much broader process than what is usually associated with genome annotation. Curation implies comprehensiveness (no features are missed), accuracy (the annotation is correct), and enrichment using other sources of information. This last part has been referred to by the PATRIC team as the use of "post-sequence data" (PSD) in the annotation process. Examples of PSD are gene expression data, proteomics data, and, of particular importance, literature data. As of this writing, the PATRIC team is developing the curation infrastructure that will allow curation goals to be met. Sobral's paper in this volume as well as the abstract by Czar et al. give additional details about PATRIC.

EST projects

In the previous sections the basic genomic data is the *complete sequence.* For bacterial genomes, the cost of obtaining such sequences is relatively low. For larger genomes it is still a major undertaking, especially for academics

outside the U.S., Europe, and Japan. For this reason, EST projects are an attractive alternative for lean budgets. In these projects a certain number of expressed sequence tags, or end sequences of cDNAs obtained from mRNAs, are collected and sequenced. This collection almost always represents a sample of the *transcriptome* of the target organism, but it nevertheless provides many of the same valuable insights about the organism's genes as a complete genome project. In fact, the two kinds of project are complementary, given the still difficult problem of predicting genes in eukaryotes. ESTs can be used to check and complement gene predictions.

The EST approach was used in the *Schistosoma mansoni* project, described in (Verjovski-Almedia, et al. 2003). *S. mansoni* is the major causative agent of schistosomiasis, which affects 200 million individuals in 74 countries. In the project, 163,000 ESTs were generated from normalized cDNA libraries from six selected developmental stages of the parasite, resulting in 31,000 assembled sequences and 92 percent sampling of an estimated 14,000 gene complement. A secondary result of this project was the identification of four novel *S. mansoni* retrotransposons (DeMarco et al., 2004).

The bioinformatics support for this project was in some ways similar to a regular genome project, but with one major difference: almost all annotation was automatic, given that the number of sequences to be annotated was about 10 times the usual number of genes in a bacterial project. For this reason, substantial effort went into the automatic classification of genes based on the Gene Ontology scheme (Ashburner et al., 2000). This was accomplished by using as BLAST databases a careful selection of GO-annotated public sequence sets.

An interesting bioinformatics research topic derived from this project came from the problem of detecting contaminants. This is an instance of the more general problem of computational determination of the organismal origin of DNA sequences. In the case of EST projects, such a problem is usually dealt with using similarity-based approaches. Such methods have a number of disadvantages. They will not classify sequences without matches in the databases; simple use of BLAST will not give false positive/negative error estimates; and a particular sequence classification is not attached to a confidence value in the prediction. My student, J. Piazza, and I developed a novel method based on the extraction of intrinsic information from the DNA sequences themselves. Briefly, the method relies on several feature extractors (some of them third party) and on feature combination and sequence classification. Different standard classifiers were employed, ranging from majority voting to support vector machines. The resulting tool was used on the *S. mansoni* assembled sequences (previously filtered of contaminants using a similarly-based method), suggesting that perhaps an additional 1,700 of these sequences are contaminants. A preliminary description of this work appeared in (Piazza and Setubal, 2003).

Other work

In this section, I outline additional work in which I participated. In (Digiampietri et al., 2003) we presented a simple data model for comparative genomics that was used to create a multi-genome database of phythopathogens.

Computing homologous families based on similarity information is at the heart of many bioinformatics analyses. Several methods have been proposed, but to my knowledge there is no single method that performs generally well in practice. In (Almeida et al., 2004) a new methodology for computing orthologous sequences was proposed. This methodology combines a graph-clique clustering strategy with profile HMM searches. This method has not been compared to existing methodologies. In the automated annotation pipeline that feeds GAT, we have used a combination of Tribe-MCL (Enright et al., 2003) results with direct pairwise BLAST alignments and synteny data.

An interesting exploratory project that I have just started in collaboration with Boris Vinatzer is

the study of variable genomic regions of several *Pseudomonas syringae* strains. The initial focus will be in the regions that contain effector genes, since these are particularly important for *P. syringae* pathogenicity.

A crucial part of any genome annotation is classification of gene products. For this, one needs a well-designed, controlled vocabulary and relationship graph among the terms; this is termed an ontology. The Gene Ontology Consortium (Ashburner, M. et al., 2000) has been developing on-tologies for annotation for several years now. However, this ontology is still not broad enough; in particular, it lacks a good system of terms to annotate genes involved in host-pathogen interactions. This lack has led a group of researchers led by Brett Tyler to create the Plant-associated microbes Gene Ontology (http://pamgo.vbi.vt.edu). The group has already developed high level terms, and plans to use this preliminary ontology to annotate genes in the genomes of a few fully-sequenced plant-associated microbes.

Future directions

In the upcoming years, I expect to stay involved with microbial sequencing projects, and at the same time continue to focus on new genome analysis tools. One topic that I plan to focus on is bacterial evolution. There seems to be abundant data about lateral transfers, genome rearrangements, pseudogenes, and other signs of genome evolution to allow novel and detailed reconstructions and models of evolutionary history for several classes of bacteria.

References

Almeida Jr N, Setubal JC, de Carvalho M, and Viana C (2004) A method to determine homologous protein families based on graph cliques and profiles. Proceedings of the III Brazilian Workshop on Bioinformatics (extended abstract), Brasília

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

Carriao P, Perez M, and Setubal JC (2004) Parsing BLAST results to find pseudogenes in bacterial genomes. Proceedings of the III Brazilian Workshop on Bioinformatics (extended abstract), Brasília

DeMarco R, Kowaltowski A, Machado A, Bento Soares M, Gargioni C, Kawano T, Rodrigues V, Madeira A, Wilson RA, Menck C, Setubal JC, Dias-Netto E, Leite L, Verjovski-AlmeidaS (2004) Saci-1, -2, and -3 and Perere, Four novel retrotransposons with high transcriptional activities from the human parasite *Schistosoma mansoni*. *J Virol* **78**: 2967–2978

Digiampietri LA, Medeiros CB, Setubal JC (2003) A data model for comparative genomics. *Revista Tecnologia da Informação* **3**: 35–40

Enright AJ, Kunin V, and Ouzounis CA (2003) Protein families and TRIBES in genome sequence space. *Nucleic Acids Res* **31**: 4632–8

Haake DA (2000) Spirochaetal lipoproteins and pathogenesis. *Microbiology* **146**: 1491–1504

Monteiro-Vitorello CB, Camargo LEA, ..., Setubal JC (44 authors) (2004) The Genome Sequence of the Gram-Positive Sugarcane Pathogen. *Leifsonia xyli* subsp. *xyli*. *Mole Plant Microbe Interact* **17**: 827–836

Moreira LM, de Souza RF, Almeida Jr NF, Setubal JC, Oliveira JCF, Furlan LR, Ferro JA, da Silva ACR (2004) Comparative genomics analyses of citrus-associated bacteria. *Annu Rev Phytopathol* **42**: 163–84

Moreira LM, de Souza RF, Digiampietri L, Rasera da Silva AC, and Setubal JC (2005) Comparative analyses of Xanthomonas and Xylella complete genomes. *OMICS* **9**: (in press)

Nascimento ALTO,... Setubal JC, van Sluys MA (47 authors) (2004) Comparative genomics of two *Leptospira interrogans* serovars reveals novel insights into physiology and pathogenesis. *J Bacteriol* **186**: 2164–2172

Nascimento ALTO, Verjovski-Almeida S, Van Sluys MA, Monteiro-Vitorello CB, Camargo LEA, Digiampietri LA, Harstkeerl RA, Ho PL, Marques MV, Oliveira MC, Setubal JC, Haake DA, and Martins EAL (2004) Genome features of Leptospira interrogans serovar Copenhageni. *Braz J Med and Biol Res* **37**: [serial on the Internet]

Piazza JP and Setubal JC (2003) New ways for automatic detection of contaminants in EST projects. *Revista Tecnologia da Informação* 3(2)

Ren SX, Fu G, Jiang XG, Zeng R, Miao YG, Xu H, Zhang YX, Xiong H, Lu G, Lu LF, Jiang HQ, Jia J, Tu YF, Jiang JX, Gu WY, Zhang YQ, Cai Z, Sheng HH, Yin HF, Zhang Y, Zhu GF, Wan M, Huang HL, Qian Z, Wang SY, Ma W, Yao ZJ, Shen Y, Qiang BQ, Xia QC, Guo XK, Danchin A, Saint Girons I, Somerville RL, Wen YM, Shi MH, Chen Z, Xu JG, and Zhao GP (2003) Unique physiological and pathogenic features of Leptospira interrogans revealed by whole-genome sequencing. *Nature* **422**:888–93.

Setubal JC (2004) Bioinformatics. In L. Mir (ed), *Genomics*, Editora Atheneu, (this is an introductory book chapter, in Portuguese), pp. 105–118

Setubal JC, and da Silva ACR (2004) Genomic approaches to the study of plant pathogenic bacteria. In Robert M. Goodman (ed), *Encyclopedia of Plant and Crop Science*, Marcel Dekker, pp. 524–526

Verjovski-Almeida S,... Setubal JC, Leite LCC, Dias-Neto E (37 authors) (2003) Transcriptome analysis of the acoelomate human parasite Schistosoma mansoni. *Nat Genet* **35**: 2

Wood DW, Setubal JC, and Nester EW (2004) Genome sequence analysis of prokaryotic plant pathogens. In M. Gillings and A. Holmes (eds.), *Plant microbiology*, pp. 223-241, BIOS Scientific Publishers (Oxford, UK) in cooperation with Marcel Dekker

Wood DW, Setubal JC, *et al.* (51 authors) (2001) The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. *Science* **294**: 2317–2323

Publications

Moreira LM, de Souza RF, Digiampietri L, Rasera da Silva AC, and Setubal JC (2005) Comparative analyses of Xanthomonas and Xylella complete genomes. *OMICS* **9**: (in press)

Nascimento ALTO,..., Setubal JC, van Sluys MA (47 authors) (2004) Comparative genomics of two *Leptospira interrogans* serovars reveals novel insights into physiology and pathogenesis. *J Bacteriol* **186**: 2164–2172

Nascimento, ALTO, Verjovski-Almeida S, Van Sluys MA, Monteiro-Vitorello CB, Camargo LEA, Digiampietri LA, Harstkeerl RA, Ho PL, Marques MV, Oliveira MC, Setubal JC, Haake DA, and Martins EAL (2004) Genome features of Leptospira interrogans serovar Copenhageni. *Braz Med Biol Res* **37**: [serial on the Internet]

Monteiro-Vitorello CB, Camargo LEA, ..., Setubal JC (44 authors) (2004) The Genome Sequence of the Gram-Positive Sugarcane Pathogen *Leifsonia xyli* subsp. *xyli*. *Mol Plant-Microbe Interact* **17**: 827–836

Moreira LM, de Souza RF, Almeida Jr NF, Setubal JC, Oliveira JCF, Furlan LR, Ferro JA, da Silva ACR (2004) Comparative genomics analyses of citrus-associated bacteria. *Ann Rev Phytopathol* **42**: 163–84

Wood DW, Setubal JC, and Nester EW (2004) Genome sequence analysis of prokaryotic plant pathogens. In M. Gillings and A. Holmes (eds.), *Plant microbiology*, pp. 223-241, BIOS Scientific Publishers (Oxford, UK) in cooperation with Marcel Dekker

Setubal JC (2004) Bioinformatics. In L. Mir (ed), *Genomics*, Editora Atheneu, (this is an introductory book chapter, in Portuguese), pp. 105–118

Setubal JC, and da Silva ACR (2004) Genomic approaches to the study of plant pathogenic bacteria. In Robert M. Goodman (ed), *Encyclopedia of Plant and Crop Science*, Marcel Dekker, pp. 524–526

DeMarco R, Kowaltowski A, Machado A, Bento Soares M, Gargioni C, Kawano T, Rodrigues V, Madeira A, Wilson RA, Menck C, Setubal JC, Dias-Netto E, Leite L, Verjovski-Almeida S (2004) Saci-1, -2, and -3 and Perere, Four novel retrotransposons with high transcriptional activities from the human parasite *Schistosoma mansoni*. *Virol* **78**: 2967–2978

Carriao P, Perez M, and Setubal JC (2004) Parsing BLAST results to find pseudogenes in bacterial genomes. Proceedings of the III Brazilian Workshop on Bioinformatics (extended abstract), Brasília

Almeida Jr N, Setubal JC, de Carvalho M, and Viana C (2004) A method to determine homologous protein families based on graph cliques and profiles. Proceedings of the III Brazilian Workshop on Bioinformatics (extended abstract), Brasília

Digiampietri L, Medeiros CMB, and Setubal JC (2004) A framewok based on Web services orchestration for bioinformatics workflow management. Proceedings of the III Brazilian Workshop on Bioinformatics (full paper), Brasília

Metabolomics for Systems Biology and Gene Function Elucidation

Vladimir Shulaev
Research Associate Professor, VBI
Associate Professor of Horticulture, Virginia Tech
vshulaev@vbi.vt.edu

Vishal Arora, Aaron Baxter, Leslie Blischak, Autumn Clapp, Ryan Cohill, Diego Cortes, Nigel Deighton, Hope Gruszewski, Beth Henry, John Chia Jiming, Hy Martin, Erica Mason, Ana Martins, Jianghong Qian, Teruko Oosumi, Joel Shuman, Vladimir Tolstikov, Leeipka Tuli, Andrew Warren, Selene Willis

Introduction

Metabolomics has emerged recently as another powerful tool for functional genomics (Trethewey et al., 1999; Fiehn et al., 2000; Hall et al., 2002; Sumner et al., 2003). Metabolic footprinting has been used to classify unknown phenotypes of yeast genetic mutants by characterization of the extracellular metabolites (Allen et al., 2003). Use of a sulfur auxotroph in *M. tuberculosis* identified novel sulfated metabolites and their pathways (Mougous et al., 2002). A combination of metabolome and transcriptome analysis of the hypoxia induced vascular remodeling process identified a therapeutic low molecular weight ligand, taurine as well as a transcriptional hypoxia-inducible factor 1 activation of a promoter for a gene, S100, important for vascular remodeling (Amano et al., 2003). Metabolomics consists of a comprehensive analysis in which all metabolites in a biological sample are to be identified and quantified (Fiehn, 2002). Identification is a one-time process for each metabolite, consisting of capturing its physico-chemical properties in sufficient detail. Quantification is the determination of the metabolite's concentration (absolute or relative) in a specific cell type under specific environmental conditions. Identification of all metabolites of a species (e.g., yeast) is akin to whole-genome sequencing, quantification of metabolites is similar to microarray or 2D gel protein analysis. Similar to metabolomics is metabolite profiling, where a subset of all metabolites is identified and/or quantified. Often metabolite profiles are specific for a chemical class of compounds. In some applications, one does not attempt to identify or quantify metabolites of a sample, but rather just to obtain a fingerprint of that sample, *i.e.* a unique pattern that reflects the chemical composition of the tissue. Fingerprints are often done with infrared spectroscopy (Oliver et al., 1998) or direct-infusion mass spectrometry (Goodacre et al., 2002).

Our research is focused on developing methods for high throughput metabolite profiling and application of metabolomics to study stress response in microorganisms, plants and animals.

Analytical Techniques for Metabolite Profiling

There are multiple techniques that can be used for metabolite profiling. Each technique has associated advantages and drawbacks. Thus, a combination of different analytical technologies has to be used to gain a broad perspective of metabolome of a tissue. Analytical techniques that are most often used for metabolite profiling include NMR, Gas Chromatography—Mass Spectrometry (GC-MS), Liquid Chromatography—Mass Spectrometry (LC-MS), and Capillary Electrophoresis—Mass Spectrometry (CE-MS). Clearly, no single extraction methodology is ideal for all of the many thousands of metabolites within a cell; instead, our aim is to be non-selective in the extraction and apply the most appropriate separation

technique in front of mass spectrometer to identify as many diverse metabolites as possible from a single biological sample.

Gas Chromatography – Mass Spectrometry (GC-MS)

Currently, the most developed technology for rapid profiling of large number of metabolites is gas chromatography coupled to electron impact (EI) quadrupole mass spectrometry (GC/MS) (Fiehn et al., 2000; Roessner et al., 2000). Using this approach it is possible to simultaneously profile several hundred chemically diverse compounds including organic acids, most amino acids, sugars, sugar alcohols, aromatic amines, and fatty acids. These compounds can be separated and quantified by GC/MS either directly or following simple chemical derivatization procedures.

Liquid Chromatography – Mass Spectrometry (LC-MS)

LC-MS has one advantage over GC-MS in that there is no need for chemical derivatization of compounds (which is needed for GC-MS analysis of non-volatile compounds). This removes one step from the technique, which may decrease the possibility of technical artifacts. As LC-MS becomes more routine and available to the non-specialist, the need for assistance in spectral interpretation continues to grow. In this regard, the easiest and most accessible interpretation would, through choice, come from an on-board, searchable library. The creation of these compounds libraries for GC-MS has been facilitated through the nature of the ionization process, and this remains unchanged after five decades. We have developed a new approach for creating searchable LC-MS-MS libraries that are transferable between different instruments and are populating our custom libraries with mass spectra of many physiological metabolites.

Capillary Electrophoresis – Mass Spectrometry (CE-MS)

Despite the fact that CE-MS is a relatively new technique compared to GC-MS and LC-MS, it has been widely used for both targeted and non-targeted analysis of metabolites (Cohen et al., 1987; Soga et al., 2002). CE-MS provides several advantages over existing separation techniques. It combines high resolving power (plate number of 100,000 to 1,000,000) and small sample requirements (average injection volume of 1 to 20 nL) with short analysis times. CE has been used to analyze a wide range of compounds including inorganic ions, organic acids, amino acids, nucleotides and nucleosides, vitamins, thiols, carbohydrates, and peptides. Due to its ability to separate cations, anions and uncharged molecules in the same run, CE can also be used for simultaneous profiling of various metabolite classes which makes it a very attractive analytical platform for high-throughput non-targeted metabolite profiling (Soga et al., 2002). We have developed methods for non-targeted CE-MS profiling of metabolites isolated from yeast, Arabidopsis, strawberry and *Plasmodium falciparum*. Our protocols are based on cationic or anionic polymer-coated capillaries to reverse the electroosmotic flow. Using this technique in a single 30 minute CE-MS analysis of an Arabidopsis leaf sample, for example, we can isolate more than 135 distinct components in positive polarity mode and more than 85 distinct components in negative polarity mode.

Metabolomics for Yeast Systems Biology

Baker's yeast *Saccharomyces cerevisiae* is a well established model eucaryotic system with easily accessible resources. It is easy to culture and manipulate genetically as well as biochemically, and there is a vast knowledge of this organism at the physiological, genetic, and molecular levels. *S. cerevisiae* has been used to study many processes, including primary metabolism, cell cycle control, apoptosis, aging, and response to a variety of stresses. Complete sequencing of the *S. cerevisiae* genome (Goffeau et al., 1996) identified around 6,000 open reading frames, of which approximately 45 percent do not have well-defined function (Oliver, 1996). A nearly complete collection of gene-deletion mutants of *S. cerevisiae* has recently been made available (Giaever et al., 2002), making this model organism a powerful tool for systematic

functional analysis and systems biology research.

We use a systems biology approach to study oxidative stress response in *S. cerevisiae* working on the project funded by NIH Grant R01 GM068947-01 "*A new mathematical modeling approach to biochemical networks, with an application to oxidative stress in yeast.*" The goal of the project is to develop novel methods from both continuous and discrete mathematics for the reverse-engineering of biochemical networks. These methods will be applied to genomics, proteomics, and metabolomics data from experiments specifically designed for this modeling approach, focusing on the regulatory network in *S. cerevisiae* responsible for oxidative stress response. The project is a collaboration between our group and the Mendes and Laubenbacher groups.

Our group is generating experimental data sets for this project. We have established methods for metabolite extraction and non-targeted metabolite profiling of yeast cells using GC-MS, which is based on the methodology described by Roessner et al. (Roessner et al., 2000). We have successfully applied GC-MS technique in parallel with transcript profiling to compare gene expression and metabolite profiles of two distinct growth phases of *S. cerevisiae* cultures and identified a series of genes and metabolites that changed significantly between growth phases (Martins et al., 2004).

To study *S. cerevisiae* response to organic peroxide cumene hydroperoxide (CHP), we performed analysis of changes in gene expression in *S. cerevisiae* strain BY4743 following exposure to CHP. Treatment with CHP resulted in rapid and transient activation of genes involved in ROS detoxification, including γ–glutamylcysteine synthase (*GSH1*) and glutathione synthase (*GSH2)*, glutathione peroxidase *GPX2*, and glutaredoxins *GRX1* and *GRX*. By applying K-Means Clustering algorithm, we separated all the array features into 20 different clusters based on their response to perturbation. The clusters ranged in size from 53 to 1154 genes and presumably include genes having distinct physiological functions in protecting against oxidative damage or maintaining cellular homeostasis. These data will be used for continuous and discrete modeling by the Mendes and Laubenbacher groups.

Using Metabolomics to Study *Plasmodium falciparum* Infection of Red Blood Cells in the Absence and Presence of Antimalarial Drugs

Malaria remains a scourge of the developing world, killing over a million people each year and infecting around 500 million. The situation has worsened over recent years as resistance has developed against chloroquine and sulphadoxine-pyrimethamine, the two drugs most commonly used to treat malaria. The parasites are small 'protozoan' cells that enter their human host through a mosquito bite. Changes in metabolism may be one of the first responses to drug action followed by transcriptional and protein changes (Watkins and German, 2002). In the study of *Plasmodium* biology, analysis of metabolite changes may be even more important because of the "apparent" decrease in transcriptional responses (Hartl et al., 2002; Le Roch et al., 2003).

For the past two years, our group has been working with David Sullivan's group at John Hopkins University on developing methodology for metabolite profiling of *P. falciparum*-infected and uninfected human red blood cells. We have optimized and validated the methodology for rapid profiling of a large number of metabolites using GC-MS, LC-MS, and CE-MS. While testing several methods for metabolite extraction and analysis from infected and non-infected erythrocytes and from isolated parasites, we have found that, for non-targeted profiling, best results were achieved using the Covaris E100 High Performance Extraction and Dissolution System (Covaris, Inc.) that utilizes ultra-high frequency sound waves for metabolites extraction. Using a combination of GC-MS, LC-MS, and CE-MS, we can distinguish and integrate several hundred distinct peaks, many of which can be positively identified using a

custom library of common metabolites. We are expanding our custom library by adding a series of standard compounds described for erythrocytes and *P. falciparum*.

For both the host erythrocyte devoid of mitochondria and the *Plasmodium* intra-erythrocytic stage with a single acristate mitochondria, glycolysis and its metabolites account for one of the major intraerythrocytic metabolic pathways (Gardner et al., 2002; Becker et al., 2003). High outputs of extracellular lactate are produced from *Plasmodium* glycolysis (Elliott et al., 2001; Ginsburg, 2002). 2,3-Diphosphoglycerate, a glycolytic metabolite that affects oxygen binding to hemoglobin and osmotic fragility of erythrocytes, is decreased by *Plasmodium* erythrocyte invasion (Dubey et al., 2003). A recent report demonstrated that chloroquine induced binding of heme to glycolytic enzymes might account for the inhibition of glycolysis by this drug (Campanale et al., 2003). With an enhanced loss of glutathione, the intraerythrocytic *Plasmodium* is dependent on *de novo* synthesis of glutathione (Deharo et al., 2003). Analysis of untreated and drug treated Percoll purified trophozoite infected erythrocytes using GC-MS highlights some general trends in this metabolic analysis. Shown in Figure 1A are changes in lactate between uninfected erythrocytes, infected erythrocytes, and diverse drug treatments. Precise quantitation methods utilizing single reaction monitoring (SRM) can be further used to quantitate individual metabolites affected by drug treatment (for example, quantitation of oxidation products and thiols like glutathione as shown in Figures 1B and C). A combination of untargeted and

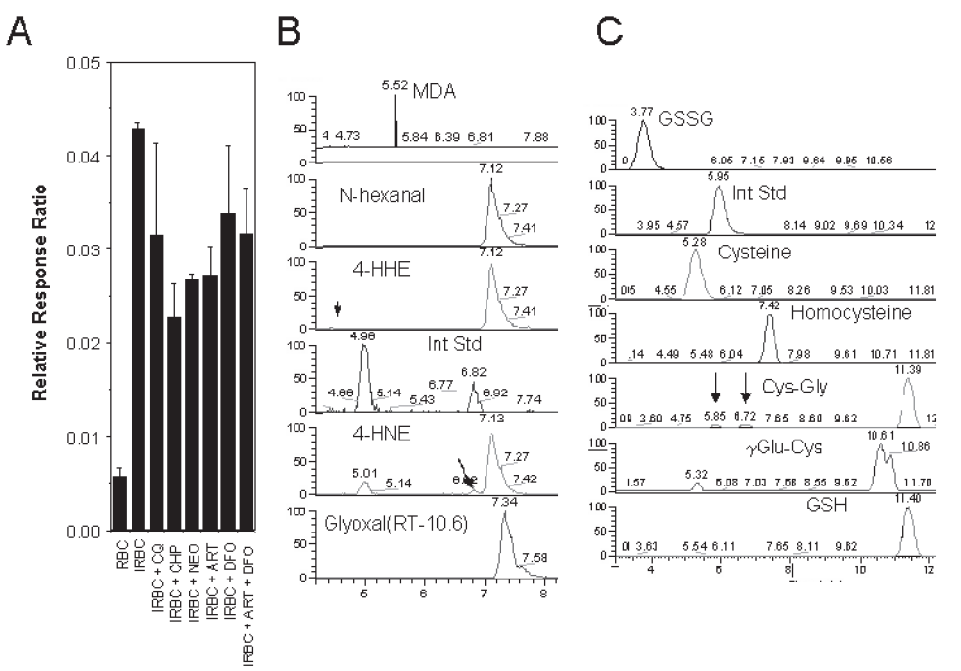


Figure 1. Metabolite profiling of the uninfected and *P. falciparum*-infected erythrocytes. (A) Lactate in intact uninfected erythrocytes (RBC), infected erythrocytes (IRBC) and drug treated chloroquine (CQ), cumene hydrogen peroxide (CHP), neocuproine (NEO), artemisinin (ART), Desferoxamine (DFO). The values for each molecule are an average of triplicate biological experiments performed on separate weeks. (B) LC MS MS with single reaction monitoring (SRM) of oxidative products from trophozoite infected erythrocytes (C) LC MS MS (SRM) of N-ethylmaleimide derivatized thiols from trophozoite infected erythrocytes.

targeted metabolite profiling will allow us to detect fluxes in specific pathways and identify most likely metabolic consequences of various antimalarial drugs, as well as discover potential metabolic targets for novel antimalarials.

Application of Metabolomics to Study Gene Functions in Arabidopsis

Current developments in genomics resulted in new tools to study biological processes (Bouchez & Hofte, 1998). Complete sequencing of *Arabidopsis* and rice has identified thousands of new genes in these organisms. However, the biological function of about half of all computationally identified genes is unknown. A major challenge of genomics is to identify the functions of these genes and use this knowledge for improving economically important agronomic and quality traits in major crops. Metabolomics can provide researchers with new tools to identify the functions of unknown genes in Arabidopsis and other plants.

In collaboration with Ron Mittler, we applied metabolite profiling to study metabolic responses in plants affected by the combination of different stresses. By analyzing plants subjected to drought, heat stress, or a combination of drought and heat stress, we have found that plants subject to a combination of drought and heat stress accumulated sucrose and other sugars such as maltose and gulose. In contrast, proline that accumulated in plants experiencing drought stress did not accumulate during a combination of drought and heat stress (Rizhsky et al., 2004). Heat stress was found to ameliorate the toxicity of proline to cells, suggesting that during a combination of drought and heat stress, sucrose replaces proline in plants as the major osmoprotectant.

In another collaborative project funded by the NSF Arabidopsis 2010 Program, together with Eran Pichersky and Joseph Noel, we 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 SABATH family. The *Arabidopsis thaliana* genome contains 24 related genes encoding enzymes that belong to the SABATH family of methyltransferases (MTs). Preliminary experiments suggest that SABATH MTs convert several important hormones and other plant constituents into their methyl esters, thereby exerting important effects on the biological activity of these molecules and consequently on important physiological processes. Three of the SABATH enzymes have been recently characterized as jasmonic acid methyltransferase (*AtJAMT1*), benzoic and salicylic acid methyltransferase (*AtBSMT1*), and indole-3-acetic acid methyltransferase (*AtIAMT1*). Preliminary metabolite profiling of the Arabidopsis mutant lines showed that CE-MS has the lowest variability among different profiling techniques (Figure 2). Recently, using gene expression profiling of plants overexpressing JMT, we have identified an unknown gene with homology to esterases that is highly induced in this line. This homology to esterases suggested that this gene may encode a methyljasmonate esterase.

Metabolomics for Fruit Functional Genomics

Fruit is important to the human diet as a major source of numerous phytochemicals, such as flavonoids and other phenolic compounds, cyanogenic glucosides, phytoestrogens, and phenols that could yield potential health and disease-fighting benefits (Macheix et al., 1991; Swanson, 1998; Selmar, 1999). Ellagic acid, L-ascorbic acid, quercetin, kaempferol, myricetin, p-coumaric acid, and gallic acid are well-known antioxidant or cancer-inhibiting compounds isolated from fruits. Ellagic acid, a potent cancer inhibiting compound, is present in strawberry, red raspberry, arctic bramble, cloudberry and other rosaceous berries. Some strawberry cultivars contain the highest levels of L-ascorbic acid among fruits (Haffner and Vestrheim, 1997). Numerous other phenolic compounds belonging to distinct chemical classes have been isolated from fruit (Macheix *et al.*, 1991), many of which have also been shown to be potential

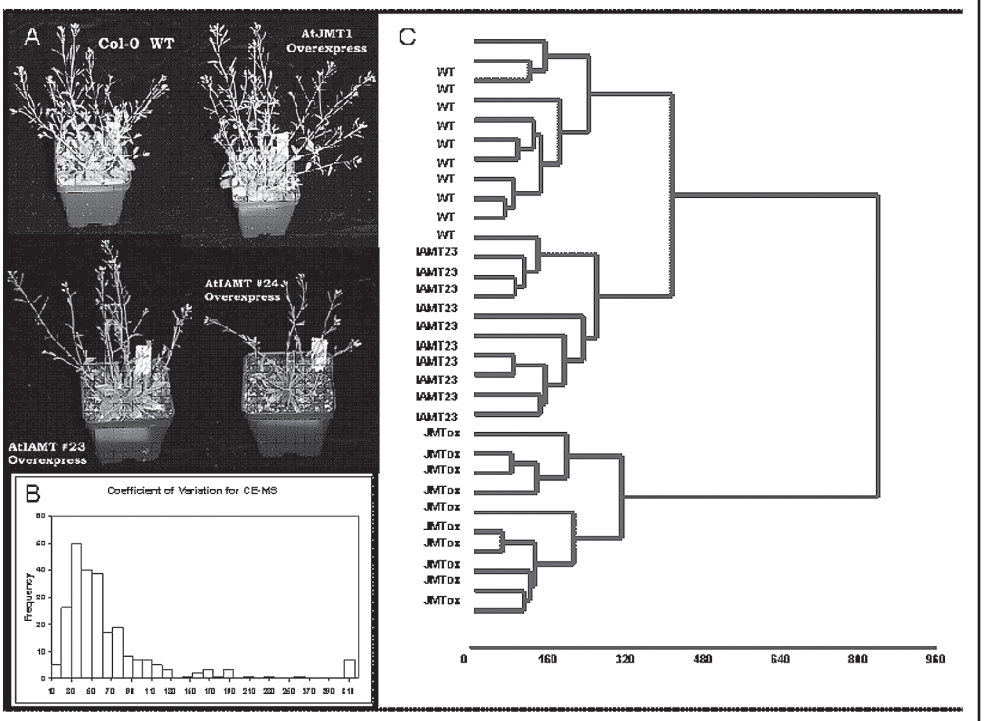


Figure 2. Metabolite profiling of the Arabidopsis IAMT and JMT mutant lines. (A) Visual phenotype of the mutant lines; (B) Distribution of the coefficient of variation among ten replications per sample analyzed by CE-MS; (C) Hierarchical Agglomerative Clustering (HCA) of metabolic phenotype profiling for arabidopsis Wilt Type Col-0, AtJMT and AtIAMT.

antioxidants and anticancer agents (Schieber et al., 2001; Eichholzer et al., 2001).

Identification of novel fruit-derived, biologically active compounds and genes involved in their biosynthesis and metabolism will provide a better understanding of phytochemicals' metabolism, yielding tools to engineer fruit with enhanced levels of biologically active compounds and fruit quality traits. We have developed a high throughput platform for reverse and forward genetics in woodland strawberry *Fragaria vesca* based on insertional mutagenesis, gene identification and phenotype profiling. To facilitate the development of insertional mutagenesis, we have developed a new and highly efficient transformation protocol that can be used for systematic production of T-DNA tagged insertional mutants. We used this protocol to produce 1,000 independent T-DNA-tagged lines in *F. vesca* (Figure 3). Segregation analysis of the transgene inheritance in the T_1 generation showed 3:1 Mendelian segregation of the transgene indicating single T-DNA insertion site in the genome for many lines. Additional molecular analysis of randomly selected transformants by Southern hybridization confirmed that most transgenic lines have single or low copy number of T-DNA vector integrated in the unrelated sited in the genome.

We used thermal asymmetric interlaced (TAIL)-PCR to amplify and sequence genomic regions flanking T-DNA insertions in transformants. In one of the mutant lines translated genomic sequence was similar to c-myc-binding protein. In our effort to facilitate gene isolation in *F. vesca*, we performed sequencing of random EST clones from a full-length enriched *F. vesca* cDNA library. To date, we have sequenced a total of 1,553 EST clones and processed all the raw sequences through the ESTAP pipeline. To store woodland strawberry sequence information, we created *Fragaria vesca* EST

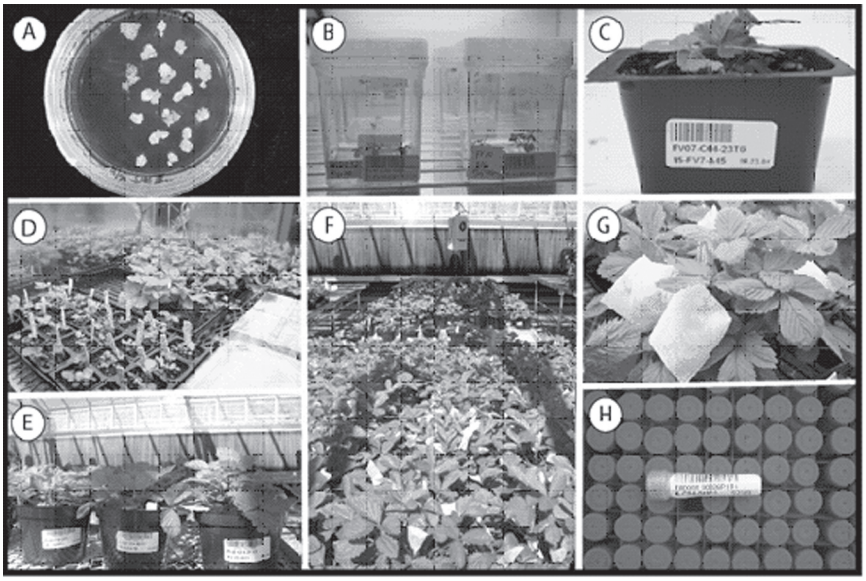


Figure 3. Pipeline for T-DNA tagged line production in *F. vesca*. (A) regenerating GFP$^+$ calli on shoot induction media after 12 weeks in culture; (B) rooted GFP$^+$ plantlets in Magenta boxes; (C) a healthy, rooted plant in a plug cell, each label contains a unique, global barcode for line tracking; (D) putative transgenic plants in trays in controlled-environment chamber, (E & F) plants in BL2-P greenhouse for seed generation; (G) hermaphroditic flowers are bagged early each day to ensure self-pollination; (H) T_1 seeds in storage tubes (mature fruit are collected, seed extracted, and desiccated prior to storage).

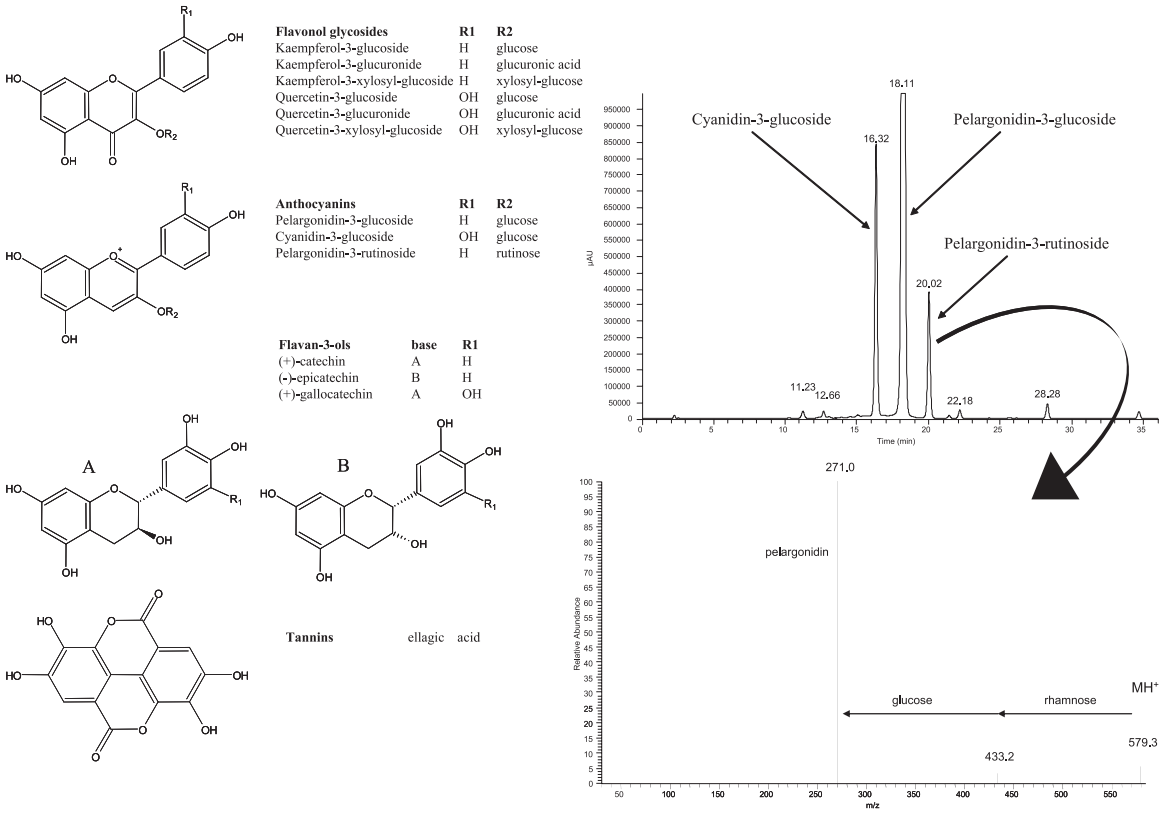


Figure 4. Abundant health related molecules of strawberry and LC-PDA-MS/MS analysis of strawberry anthocyanins (arrow indicates the MS/MS spectrum of the selected peak).

database (FVdbEST), which was made publicly available through the ESTAP website (http://staff.vbi.vt.edu/estap/).

Epidemiological evidence suggests that diets rich in fruit and vegetables can significantly reduce cancer risk (Caragay, 1992; Ziegler et al., 1996). *In vitro* and *in vivo* studies with animal models provide evidence that fruit and leaf extracts from many Rosaceae inhibit various forms of cancer or have pronounced antioxidant action (Yau et al., 2002). We have developed analytical assays for many strawberry metabolites, in-cluding flavonoids and other health related metabolites (Figure 4). High throughput screening assays utilize both diode array detection (DAD) and mass spectrometry (Figure 4). Screening collection of the T-DNA–tagged mutants using these analytical techniques will help us to identify genes involved in biosynthesis of many health-related compounds. The generation of a T-DNA mutant collection of strawberry will impact functional genomics research and gene discovery in Rosaceae and other fruit crops.

Conclusions

Metabolomics is emerging as powerful high-throughput platform complementing other genomics platforms like transcriptomics and proteomics. Combination of these high throughput data generation techniques with mathematical modeling of biochemical and signaling networks is essential for systems biology and will help us to deeper understand how biological systems work as a whole.

Acknowledgements

Our work was supported by the 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," Virginia Tech ASPIRES grant "Strawberry functional genomics," Virginia Bioinformatics Institute, and Virginia Tech Horticulture Department. We thank our collaborators 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 also thank Susan Martino-Catt, Clive Evans, and all VBI CLF staff.

References

Allen J, Davey HM, Broadhurst D, Heald JK, Rowland JJ, Oliver SG, and Kell DB (2003) High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. *Nat Biotechnol* **21**: 692–696

Amano H, Maruyama K, Naka M, and Tanaka T (2003) Target validation in hypoxia-induced vascular remodeling using transcriptome/metabolome analysis. *Pharmacogenomics J* **3**: 183–188

Becker K, Rahlfs S, Nickel C, and Schirmer RH (2003) Glutathione--functions and metabolism in the malarial parasite Plasmodium falciparum. *Biol Chem* **384**: 551–566

Bouchez D and Hofte H (1998) Functional genomics in plants. *Plant Physiol* **118**: 725-732

Campanale N, Nickel C, Daubenberger CA, Wehlan DA, Gorman JJ, Klonis N, Becker K, and Tilley L (2003) Identification and characterization of heme-interacting proteins in the malaria parasite, Plasmodium falciparum. *J Biol Chem* **278**: 27354–27361

Caragay AB (1992) Cancer-Preventive Foods and Ingredients. *Food Technology* **46**: 65–68

Cohen AS, Terabe S, Smith JA, and Karger BL (1987) High-performance capillary electrophoretic separation of bases, nucleosides, and oligonucleotides: retention manipulation via micellar solutions and metal additives. *Anal Chem* **59**: 1021–1027

Deharo E, Barkan D, Krugliak M, Golenser J, and Ginsburg H (2003) Potentiation of the antimalarial action of chloroquine in rodent malaria by drugs known to reduce cellular glutathione levels. *Biochem Pharmacol* **66**: 809–817

Dubey ML, Hegde R, Ganguly NK, and Mahajan RC (2003) Decreased level of 2,3-diphosphoglycerate and alteration of structural integrity in erythrocytes infected with Plasmodium falciparum in vitro. *Mol Cell Biochem* **246**: 137–141

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

Elliott JL, Saliba KJ, and Kirk K (2001) Transport of lactate and pyruvate in the intraerythrocytic malaria parasite, Plasmodium falciparum. *Biochem J* **355**: 733–739

Fiehn O (2002) Metabolomics--the link between genotypes and phenotypes. *Plant mol biol* **48**: 155–171

Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey RN, and Willmitzer L (2000) Metabolite profiling for plant functional genomics. *Nat. Biotechnol* **18**: 1157–1161

Gardner MJ, Hall N, Fung E, White O, et al (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* **419**: 498–511

Giaever G, Chu AM, Ni L, Connelly C, et al (2002) Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418**: 387–391

Ginsburg H (2002) Abundant proton pumping in Plasmodium falciparum, but why? *Trends Parasitol* **18**: 483–486

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, and Oliver SG (1996) Life with 6000 Genes. *Science* **274**: 546–567

Goodacre R, Vaidyanathan S, Bianchi G, and Kell DB (2002) Metabolic profiling using direct infusion electrospray ionisation mass spectrometry for the characterisation of olive oils. *Analyst* **127**: 1457–1462

Haffner K and Vestrheim S (1997) Fruit quality of strawberry cultivars. *Acta Hort* **439**: 325–332

Hall R, Beale M, Fiehn O, Hardy N, Sumner L, and Bino R (2002) Plant metabolomics: the missing link in functional genomics strategies. *Plant Cell* **14**:1437–1440

Hartl DL, Volkman SK, Nielsen KM, Barry AE, Day KP, Wirth DF, and Winzeler EA (2002) The paradoxical population genetics of Plasmodium falciparum. *Trends Parasitol* **18**: 266–272

Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK, Haynes JD, De La Vega P, Holder AA, Batalov S, Carucci DJ, and Winzeler EA (2003) Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* **301**: 1503–1508

Macheix JJ, Sapis JC, and 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, and Shulaev V (2004) A systems biology study of two distinct growth phases of Saccharomyces cerevisiae cultures *Cur Genomics* **5**: 649–663

Mougous JD, Leavell MD, Senaratne RH, Leigh CD, Williams SJ, Riley LW, Leary JA, and Bertozzi CR (2002) Discovery of sulfated metabolites in mycobacteria with a genetic and mass spectrometric approach. *Proc Natl Acad Sc. U.S.A.* **99**: 17037–17042

Oliver SG (1996) From DNA sequence to biological function. *Nature* **379**: 597–600

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

Rizhsky L, Liang H, Shuman J, Shulaev V, Davletova S, and 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, and 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, and Carle R (2001) By-products of plant food processing as a source of functional compounds - recent developments. *Trends in Food Science & Technology* 12:401–413

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

Soga T, Ueno Y, Naraoka H, Ohashi Y, Tomita M, and Nishioka T (2002) Simultaneous determination of anionic intermediates for Bacillus subtilis metabolic pathways by capillary electrophoresis electrospray ionization mass spectrometry. *Anal Chem* **74**: 2233–2239

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

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

Trethewey RN, Krotzky AJ, and Willmitzer L (1999) Metabolic profiling: a Rosetta Stone for genomics? *Curr Opin Plant Biol* **2**: 83–85

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

Yau MH, Che CT, Liang SM, Kong YC, and Fong WP (2002) An aqueous extract of Rubus chingii fruits protects primary rat hepatocytes against tert-butyl hydroperoxide induced oxidative stress. *Life Sci* **72**: 329–338

Ziegler RG, Mayne ST, and Swanson CA (1996) Nutrition and lung cancer. *Cancer Causes Control* **7**: 157–177

Publications

Davletova S, Rizhsky L, Liang H, Shengqiang Z, Oliver DJ, Coutu J, Shulaev V, Schlauch K, and Mittler R (2005) Cytosolic ascorbate peroxidase is a central component of the reactive oxygen gene network of Arabidopsis. *Plant Cell* **17**: 268–281

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

Rizhsky L, Shulaev V, and Mittler R (2004) Measuring programmed cell death in plants. *Methods Mol Biol* **282**: 179–189

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

The α-Proteobacteria and Prokaryotic Life Inside Eukaryotic Cells

Bruno W.S. Sobral
Executive and Scientific Director; Research Professor, VBI
Professor of Plant Pathology, Physiology, and Weed Science, VT
sobral@vt.edu

PathoSystems Biology Group: Trevor Charles, Raymond Equi, Yongqun He, Jocelyn Kemp, Chunhong Mao, Erica Mason, Endang Purwantini, Jing Qui, Xiaoyan Sheng, Chunxia Wang
Cyberinfrastructure Group: George Abramochkin, Cory Byrd, Stephen Cammer, Jeff Chen, Balaprasuna Chennupati, Oswald Crasta, Mike Czar, Chitti Dharmanolla, Darius Dziuda, Dana Eckart, Zhangjun Fei, Herman Formadi, Mark Hance, Yongqun He, Joseph Horton, Kiran Indukuri, Ranjan Jha, Pradhuman Jhala,Yongjian Guo, Nithiwat Kampanya, Ron Kenyon, Chaitanya Kommidi, Raju Lathigra, Bryan Lewis, Jian Li, Hui Liu, Hanning Ni, Eric Nordberg, Anjan Purkayastha, Balaji Rajagopalan, Harsha Rajasimha, Joao Setubal, Bruce Sharp, Maulik Shukla, Eric E. Snyder, Jeetendra Soneja, Wei Sun, Yuying Tian, Nishantsinh Vaghela, Nirali Vaghela, Rebecca Wattam, Eric Wu, Tian Xue, Boyu Yang, GongXin Yu, Qiang Yu, Chengdong Zhang, Fengkai Zhang, Yan Zhang, Zhihong Zhang, Jing Zhao

Introduction

Infectious diseases are the result of interactions between hosts, pathogens, and their environments. Most infectious disease research and development has been done by disciplinary groups. Areas such as microbiology, immunology, plant pathology, epidemiology, and others have contributed to this R&D.

The public health infrastructure of many nations has suffered from a significant lack of investment and innovation over the last 40 or so years (Garrett, 2000). Changing approaches to infectious disease research should parallel changes in life sciences research (Fauci, 2004). Some important changes include the implementation of multidisciplinary teams (Anonymous, 2003); the development and deployment of high performance laboratory and Cyberinfrastructure (CI) (Atkins et al, 2003); and the full integration of modeling and simulation as virtual analogs of theory and wet chemistry experimentation. For the purposes of this article, the approach of integration of modeling, simulation, theory, and wet chemistry experimentation to infectious disease research is called PathoSystems Biology (Eckart et al, 2003).

The α-proteobacteria form a monophyletic group of organisms that have, among other characteristics, successfully explored life within diverse eukaryotic cells, ranging from mammals, to plants and insects. Mitochondria are likely to have originated from an ancestral α–proteobacterium (Emelyanov, 2001). Approximately 22 α–proteobacterial genomes are completed and 19 are in progress. The majority of the sequenced α–proteobacteria represent organisms that have at least one phase of their life cycle that occurs within a eukaryotic cell. These organisms provide a unique opportunity to apply pathosystems biology and cyberinfrastructure to life sciences in an effort to increase our understanding of infectious diseases and hopefully innovate and accelerate the development of countermeasures such as

vaccines, therapeutics, and diagnostics.

This article summarizes some of the collective efforts in our CI and PathoSystems Biology Groups over the last year to use the α-proteobacteria as a central set of organisms to increase our knowledge of (bacterial) infectious disease systems and to deploy CI for PathoSystems Biology.

Pathosystems Biology of α-proteobacteria

One of the subgroups of α-proteobacteria is the Rhizobiales. From a wet chemistry experimental agenda, we focus primarily on *Sinorhizobium meliloti*, a dinitrogen-fixing symbiont of leguminous plants (*Medicago* spp.), and, secondarily, on one of its closest relatives, *Brucella* spp., the microbial agent associated with brucellosis. Although the hosts are very different, the bacterial agents have a number of intriguing similarities. For example, both bacteria are capable of entering their host cells (plant root cells and mammalian macrophages), evading their response mechanisms and establishing themselves within those cells. Both bacteria have genomes that are partitioned into multiple, large replicons, some with chromosomal style replication origins and others with plasmid-style origins. At the genome sequence similarity level, both bacteria are much more similar to each other than are their hosts.

We have the long-term goal of extending the genome-level comparisons of *Brucella* spp. and *S. meliloti* to comparisons of the dynamic responses of these bacteria as they infect their hosts and establish themselves intracellularly. The same applies to host responses, since disease phenotypes depend upon interactions between responses of each organism. These comparisons require the development of high-performance, robust platforms to assay mRNAs, proteins, and metabolites throughout developmental time-courses of infection. We have designed and are using an Affymetrix GeneChip to analyze the mRNA readout of *S. meliloti* 1021 in developmental time-courses of infection on *Medicago truncatula* and *M. sativa* (alfalfa) hosts (Mao et al, unpublished, 2005). Our design targets both the annotated ORFs and the intergenic regions from both strands of *S. meliloti* 1021 genome, which offers unique opportunities to investigate gene expression patterns, identify new genes and update the annotation of the genome, and assess genomic variations. Other transcriptional profiling resources have been applied to *S. meliloti* (Barnett et al, 2004; Becker et al, 2004) and all are compared in Table 1. We use VBI's Core Laboratory Facility for standardization of operating procedures regarding profiling experiments and are developing and deploying informatics resources to support community efforts to use these technologies and provide access to data sets for comparative analyses (Mao et al, unpublished, 2005; http://rhizobia.vbi.vt.edu). We are working with the *Brucella* community to establish similar GeneChip resources for *Brucella* spp.

In any host-pathogen-environment interaction, comprehension of the interaction requires an integrative approach that considers factors in the hosts, pathogens, and the environments in which the interaction occurs. There are publicly available GeneChip resources to assay macrophage host responses and we have employed these to investigate host responses to *B. melitensis* infection in a time-course (He et al, unpublished, 2005). These results showed that the most significant changes in macrophage gene transcription happened early following infection, which is consistent with the rapid clearance of *B. melitensis* 16M from macrophages during the first 24 hours of infection. This was followed by growth of the surviving bacteria inside replicative phagosomes at the late infection stage. *Medicago* spp. have yet to offer a complete genomic sequence for the same level of analysis of host response, although preliminary analysis using a Symbiosis GeneChip, which incorporates over 9,000 EST-based assemblies from *M. truncatula* has yielded some insights (Barnett et al, 2004).

Table 1. A comparison *S. meliloti* Gene Chips

	Rm1021 Chip **This work**	Symbiosis Chip **Barnett *et al*. (2004)**	Sm6K Chip **Becker *et al*. (2004)**
Total # probes	511,888	501,170	6223
Sm intergenic regions	Evenly tiled 5193 IG regions from both strands (>24nt). The spacing between tiled probes is ~ 43nt	Evenly tiled intergenic sequences (>150nt) on both strands. The spacing between tiled probes is 26-33nt	
Sm genes	6209 (protein genes) 64 (RNA control)	All annotated ORFs (6270?)	6207
Probe length (nt)	25	25	70
Mt genes	13 (protein genes) 6 (control genes)	9,935 TCs from root library (TIGR)	
Array format	11-μm feature size 7.9-mm chip size	18-μm feature size and12.8-mm chip size	
Array design	Affymetrix GeneChip	Affymetrix GeneChip	Spotted array
#probe pair per set	13	13 for Sm, 11 for Mt	1

mRNA, protein, and metabolite profiles represent dynamic responses of hosts and pathogens to their encounters and environmental contexts; thus, comparison of those profiles over developmental time-courses will extend the paradigm of sequence comparison to include the comparison of the profiles themselves through biochemistry, as well as through novel mathematical techniques (Laubenbacher, personal communication, 2005). The underlying hypothesis is that in spite of the differences in the hosts themselves, the architectures of the response and their dynamic signatures may be conserved at levels not revealed by sequence similarity alone. This would further the case that an integrative and comparative view of the α–proteobacteria will be productive as proposed.

Cyberinfrastructure for PathoSystems Biology

To handle heterogeneous data and distributed communities involved in α–proteobacterial research, there needs to be coordination of information resources. At the level of bioinformatics, models have been proposed (Stein, 2003), but we feel it goes beyond this to include sharing of high-performance, wet chemistry experimental platforms and coordination of restricted vocabularies and ontologies across groups, with contributions by service providers, scientists, and engineers as described in the Atkins Report vision for CI (Atkins et al, 2003). There are crucial questions of "social engineering", which will not be developed here, but should be central to the

interactions if there is to be a positive outcome for the communities engaged.

The CI Group supports infectious disease R&D through the design, development, and deployment of computational and information system tools starting with the molecular biology data types (following the "central dogma" of DNA, mRNA, proteins, and metabolites, with integrative views provided through pathways, for example). Most of the effort is currently at the DNA, mRNA, and protein levels. In the long-term, handling epidemiology and phenotypic (at the disease description level) attributes will become increasingly necessary.

The Pathogen Portal Project: A Service-Oriented Architecture to Support Interoperability

Software component reuse benefits organizations by reducing development, testing, and maintenance efforts while improving quality and reducing time to market. Inevitable component customization often robs software component reuse of these promised benefits and social aspects still must be addressed. We have used a combination of Web services (Brittain & Darwin, 2003; Haddad et al, 2004) and XML (Graham et al, 2004) to build reusable software components that provide the core functionality of our Service Oriented Architecture (SOA) (Erl, 2004). As a test of the architecture, these components have been incorporated into multiple and diverse projects described briefly here.

The Pathogen Portal (PathPort) project utilizes a client-side interconnect application, ToolBus, to contact Web services and make results available via highly interactive visualization plug-ins (Eckart & Sobral, 2003). ToolBus-friendly Web services provide additional capabilities, including information for GUI construction, detailed service description information, polling, and time to finish estimates which, when combined with ToolBus' Web service failover support, enables load balancing across multiple Web service hosts. Systems to support interoperability across data and analysis tools must be extensible to allow new data domains to be added without having to redesign the system, be accessible to a wide OS platform base to allow life scientists to use multiple OS, enable data from different information and analysis resources to be studied in concert without the need for a dedicated bioinformatician, and scale in both the number of users the system could support, and the services and visualizations provided. An SOA was chosen for the project, consisting of a client-side application with a plug-in architecture utilizing SOAP (Snell et al, 2001) over HTTP(S) (Gourley & Totty, 2002)-based Web services to provide access to data and analysis tools. This separation into three primary component types of ToolBus, plug-ins, and Web services enabled much of the early development to occur in parallel with a prototype of a minimal ToolBus/PathPort system (functioning within four months). Results from Web services are returned as XML documents. Many tools (e.g., gene predictions, sequence alignments) share common attributes, supporting a small number of XML document types that enable back-ends (i.e., web services) and front-ends (i.e., visualization plug-ins) to communicate via a common format.

In SOAs, services can be built on top of other services. For example, the MUMmer (Kurtz et al, 2004) based service provided within PathPort utilizes the genome database access service to enable easy comparison of genomes contained within that database without needing the data to pass through the client, which would pose a significant bottleneck for data flow. Using the Web service configuration files, it is also easy to change the genomic database access service that the MUMmer service utilizes, thus enabling multiple versions of the MUMmer service to be rapidly deployed. This capability has been used to provide tailored MUMmer services for multiple projects. In a similar fashion, other services can be combined to form composite services of arbitrary depth. This is an improvement over typical N-tier architectures that *a priori* limit the number of tiers. This

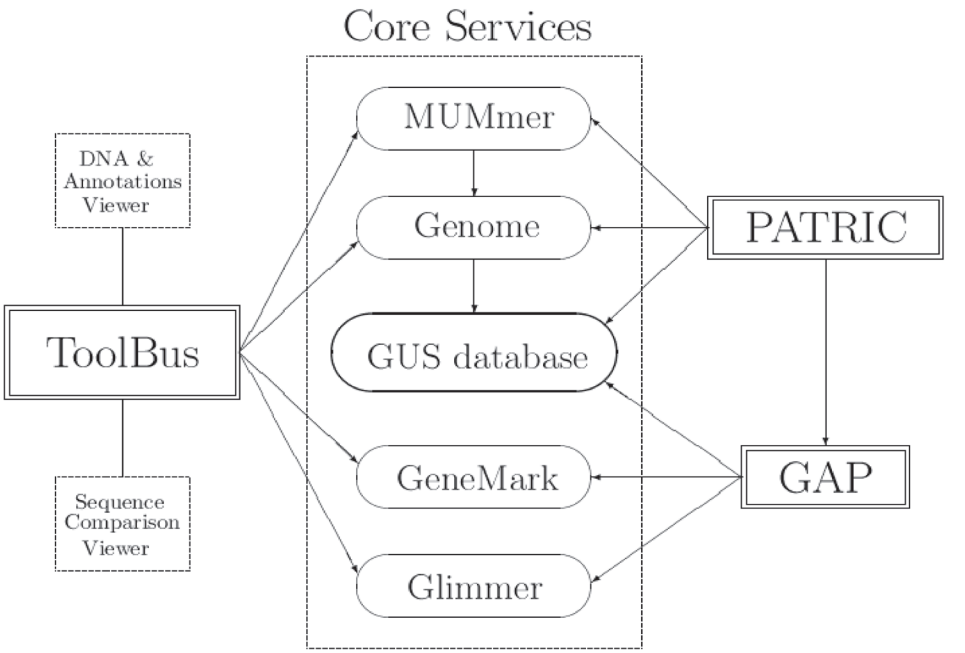


Figure 1. SOA Provides Re-Usable Components to Projects. The relationship between different high-level components and projects and core reusable components made available through the SOA in a programmatic manner. Note that projects external to VBI could just as easily use these same components programmatically.

aspect of SOA is illustrated in Figure 1, as is the relationship of the ToolBus (http://pathport.vbi.vt.edu), GAP, and PATRIC (http://patric.vbi.vt.edu) projects (described below) to the core services.

A novel aspect of ToolBus is the ability of users to create groups of associated data. Such data need not share anything with one another other than that the researcher wanted to associate them. These data can reside within different visualizations for different data types. Such groups can be compared with one another by the interactive creation of Venn Diagrams. Figure 2 illustrates this capability with a comparison of significant differential expression of human genes infected with *E. coli*, influenza virus, and *Candida* yeast (Huang, et. al, 2001). Any mix of regions within the diagrams can be highlighted (or hidden), with the corresponding data within the visualizations being highlighted (or hidden). This capability is further supported by a special class of visualization plug-in known as "group suggestors". Group suggestors can snoop ToolBus' bus and discern information that might be useful to associate with one another. Group suggestors are free to use a variety of means to determine possible associations. Gene ontology terms, corresponding identifiers (e.g., SWISS-PROT and GeneBank identifiers), pre-computed orthologs, or any other mechanism can be used to identify data (e.g., genes) for which a useful association seems likely. The ability to create and compare groups is the foundation upon which ToolBus enables integration of data from distributed, previously non-interoperable resources.

Nodulation Mutant Database, Pathogen Information Markup Language, and Molecular Interaction Network: Integrating Literature Curation and Controlled Vocabularies

One of the challenges life scientists face when trying to compare large-scale experimental data sets is the lack of standardized terminology to

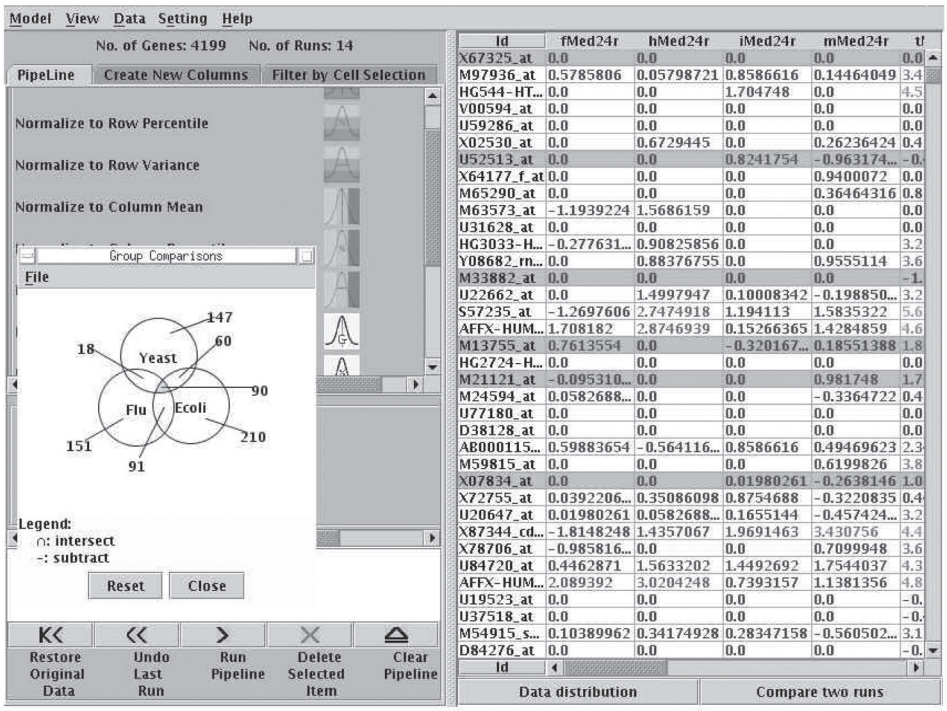


Figure 2. ToolBus' Group Suggestor Interface. An example comparison of data groups using Group Suggestor (see text for details). Note that any data type can be used to make groups, including multiple data types.

describe their data. Within the α–proteobacteria, this creates difficulty for the comparison of data sets that involve dynamic states of the organism. We have started addressing this through the development of NodMutDB (Mao et al, 2005), PIML (He et al, 2005), and MINet (He et al, unpublished, 2005) resources. We implemented a relational database, NodMutDB (Nodulation Mutant Database), for depositing, organizing and retrieving information on symbiosis-related genes, mutants, and published literature. 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. We are further developing a Rhizobiales Bioinformatics Resource Center (RhizobialesBRC) to provide the framework for an integrated and comprehensive resource that will incorporate current NodMutDB and GeneChip™ projects and future program modules. RhizobialesBRC currently allows researchers to view, query, download, analyze, and compare genomic information of a number of Rhizobiales. RhizobialesBRC will eventually become an integrated, comprehensive, and accurate resource of Rhizobiales and their host plants. Finally, RhizobialesBRC leverages the information system developed for PATRIC, furthering our experiences with software component re-use.

PIML was structured to handle information of interest to a wide variety of users and specify topics, including pathogen taxonomy and life cycle, genome sequence information, epidemiology, host-pathogen interaction, disease prevention and treatment, laboratory isolation, and diagnostic methods. PIML allows for a portable, system-independent, machine-parseable, and human-readable representation of general information for any pathogen. To date, 31 pathogens have been described using extensible PIML documents. This body of pathogen information can be queried and

displayed graphically from a Web service (http://staff.vbi.vt.edu/pathport/services/wsdls/piml.wsdl) accessible by ToolBus, which provides a custom graphical visualization module for PIML documents, and interoperability with other types of infectious disease data (e.g. genomic). A Web-based query and display system was also developed to query the complete pathogen information or a specific topic across multiple pathogens (http://www.vbi.vt.edu/pathport/pathinfo/query.html).

MINetML, an XML-based Molecular Interaction Network Markup Language, was designed to represent pathways implicated in microbial pathogenesis at the levels of detail typical of the relevant literature, including from specific biochemical reactions to general interactions of large, diffuse, or poorly understood components (He et al, unpublished, 2005). This language forms the basis for a web service and a visualization system available through ToolBus and on the web (URL). The database currently stores curated pathogenesis data from peer-reviewed publications for 22 pathogens of high priority to public health and biological defense.

Bioinformatics and Genomics Research Core for the Mid-Atlantic Regional Center of Excellence in Biodefense and Emerging Infectious Diseases (MARCE)

NIAID's Regional Centers of Excellence (RCEs) are aimed at the development of diagnostics, vaccines, and therapeutics for NIAID Category A, B, and C pathogens. Region III's Mid-Atlantic Regional Center of Excellence (MARCE) is a consortium of 14 universities, seven government partners, and 11 corporate partners, and is currently in its third year of a five-year award. MARCE is comprised of six research projects - *Anthrax*, *Emerging Viruses*, *Poxviruses*, *Tularemia*, *Enteric Pathogens* and *Public Health Response*; and three cores – *Non-Human Primate Core*, *Clinical Core* and *Bioinformatics and Genomics Research Core (BGRC)*. The CI Group, together with VBI's Cores (Laboratory–CLF, and computational–CCF), serve as the BGRC, providing high-throughput data generation, high performance computational and bioinformatics support, and collaborative research support for data analysis, curation, and bioinformatics services.

In contrast with the MARCE research projects, BGRC does not have funding for direct scientific development and discovery. BGRC funds are directed towards providing core services and training to the MARCE consortium. This role places BGRC members in direct contact with leading infectious disease researchers developing diagnostics, therapeutics, and vaccines. Thus, we have opportunities to conduct collaborative research, use a mechanism to collect and understand researchers' bioinformatics needs, and make tools available for the researchers' evaluation and use. We also have the opportunity to integrate and serve data to MARCE under standard formats and using standard tools. This last point is crucial if one thinks of comparative analyses across different pathosystems; it is also socially complex to find drivers for data sharing. Thus far, VBI's CLF has generated experimental data for different MARCE research groups by using Affymetrix Genechip arrays to identify genes and pathways that are transcriptionally regulated during early responses to hemorrhagic fever in rhesus macaque using a PF2D liquid chromatography system and LC-MS/MS to identify protein differences between *E. coli* O157 wild type strain vs. regulatory mutant strain and annotate differentially expressed peaks, and using MALDI-TOF mass spectrometry to identify recombinant fusion proteins from *Coxiella burnetii* expressed in *E. coli*. The CI Group has also submitted joint applications with other MARCE participating institutions for MARCE and NIAID-funded supplemental grant opportunities. The first supplemental funding awarded to BGRC is to support the creation of a non-human primate laboratory and pre-clinical database management system for the University of Pittsburgh School of Medicine. The development of this system, called the Non-Human Primate Information Management System (NHP-IMS), leverages the existing NIAID Division of AIDS (DAIDS) Simian

Vaccine Evaluation Unit System (SVEU). The NHP-IMS will provide the electronic infrastructure to support the animal management, protocol scheduling, and operational reporting functions of the University of Pittsburgh facility. Through this effort, we expect to learn about and internalize information regarding workflows, IT needs, and procedures used in a vaccine and therapeutic development pre-clinical environment, furthering our ability and opportunities to collaborate with and support researchers in this field, who are our target users. The system design is modular and extensible to facilitate future enhancements to support integration with clinical data and deployment in other animal cores. We will enable Health Level Seven (HL7, www.hl7.org) messaging standards for the exchange, management, and integration of clinical data to ensure interoperability with other healthcare management systems.

PathoSystems Resource Integration Center

NIAID also created eight Bioinformatics Resource Centers (BRCs) to develop the bioinformatics expertise and infrastructure to support the analysis of the genomes of a specific set of bacterial and viral pathogens with the goal of aiding development of vaccines, diagnostics and therapeutics. One such BRC, VBI's PATRIC (PathoSystems Resource Integration Center), is tasked with analyzing all available genome sequences for the α-proteobacteria, *Brucella* and *Rickettsia*; *Coxiella* and the following viral systems: rabies, hepatitis A, hepatitis E, SARS/coronaviruses and caliciviruses. The scope of the program is very broad. In the first year, we built a relational database for genomic data based on the Genomics Unified Schema (Anonymous, 2005a) (GUS) developed at the University of Pennsylvania. The database was populated with information from GenBank records for PATRIC organisms and NCBI RefSeq annotation (Pruitt et al, 2005), when available. Basic tools have been constructed for browsing this data. Our current focus is on creating a system for automated genome analysis and an interface for data visualization and curation.

The first goal of PATRIC's genome analysis effort is a thorough understanding of each organism's genomic sequence and protein repertoire. The creation of the Genome Annotation Pipeline (GAP) is the first step in building the infrastructure to support this goal. While no single pre-existing software package could support the entire range of activities and data types envisioned by the PATRIC project, many such as GUS (Anonymous, 2005a) and BioPerl (Stajich et al, 2002) contain functionality that could be used directly in GAP or curation tool development. Others such as ASAP (Anonymous, 2005b) and BioPipe (Hoon et al, 2003) provided inspiration for ideas that were ultimately implemented independently. Ultimately, the PATRIC pipeline benefited the most from our earlier work on PathPort and other projects. The former wrapped many of the sequence analysis applications as Web services, an important enabling aspect of the pipeline architecture. Other projects sponsored the development of a pipeline for annotation of ESTs and genomic DNA from *Nicotiana*, laying much of the groundwork for the PATRIC pipelines for prokaryotic and viral annotation.

In its implementation, the Genome Analysis Pipeline is actually a series of pipelines dedicated to specific types of analysis. The architecture is based on a hub and spoke model in which a central job-control program reads the pipeline configuration from an XML document and coordinates data input, analysis, and output through a series of Web services. The configuration document contains the name, location, and parameters for each program, as well as the data flow between them. As shown in Figure 3, submitting a genomic sequence to the GUS database triggers the invocation of the first GAP component, GSAP (Genomic Sequence Analysis Pipeline). GSAP programs annotate DNA-level features, some of which imply derivative sequences such as proteins or RNAs. The work of the curation team begins at this point, using the curation tool to review these results. Submission of curated protein sequences to the database triggers execution of the Protein Analysis Pipeline (PAP), which adds

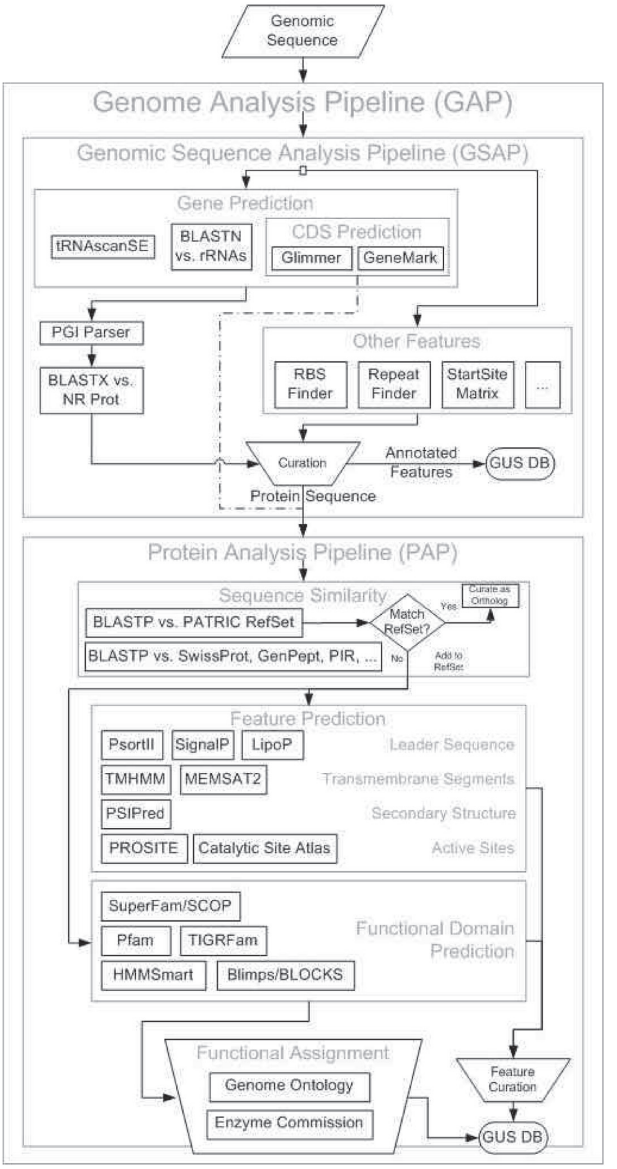


Figure 3. PATRIC's Genome Analysis Pipeline (GAP). GAP is an automated system for annotating prokaryotic (and viral) genomes. It consists of two conceptual units, the Genomic Sequence Analysis Pipeline (GSAP) and Protein Analysis Pipeline (PAP), and is configured using GAPML, an XML-based pipeline description language (Jeetendra Soneje, unpublished). Submission of a genomic sequence to the database triggers pipeline execution. Analysis begins in the GSAP with programs to identify tRNA, rRNA, and protein-coding genes. The programs tRNAscanSE (Lowe and Eddy, 1997), BLASTN (Altschul et al., 1990), Glimmer (Eckart and Sobral, 2003; He et al., 2005), and GeneMark (Lukashin and Borodovsky, 1998; McIninch et al., 1996), respectively, make the gene predictions. These are processed by the "putative gene interval" (PGI) Parser (Snyder, unpublished, 2005) to segment the genome into fragments containing a single gene. This breaks the genome into a manageable size for similarity searches and simplifies interpretation of their results. The PGIs are annotated with putative ribosome binding sites, repetitive sequences, and other features and queued for curatorial review. Curators make the final call on the predicted gene coordinates and translation, and review the other results prior to submission to the GUS database. The translations are passed to the PAP where they are first classified with respect to the Reference Protein Set, a collection of canonical proteins for each category of PATRIC organisms. If the protein is found to be similar to an existing entry in this database based on BLASTP search and very stringent cutoff, it is binned with that sequence. If no match is found, the protein is added to the Reference Set. Characterization continues with similarity searches to other databases such as SwissProt (Boeckmann et al., 2003), GenPept, and PIR (Wu et al., 2003). The sequence is then analyzed to identify physical characteristics such as signal peptides, transmembrane segments, secondary structure, and PROSITE motifs. The final step involves characterization of functional domains using Pfam (Bateman et al., 2004; Sonnhammer et al., 1998) and TIGRfam (Zhang et al., 2003) HMM libraries, SCOP (Andreeva et al., 2004), SMART (Letunic et al., 2004), and BLOCKS (Henikoff and Henikoff, 1996; Henikoff et al., 1999). These programs/databases not only predict features, but are also used by the curators to infer functions. Functions are encoded using terms from Genome Ontology (Harris et al., 2004) (GO) and Enzyme Commission (EC) numbers. Features and functional assignments are then written to the database where they are used to infer pathway membership.

structural and functional annotation. These results are also curated and form the basis of downstream studies such as pathway analysis and functional genomics.

The annotation produced by GAP is curated using the Curation Tool (CT, Figure 4). Following the addition of a new genome to the database (and subsequence pipeline execution), curators are greeted with a list of uncurated PGIs (putative gene intervals) on login. The curator selects an interval and opens a "PGI editor" that provides a graphical overview of GAP features followed by an "evidence table" with more detailed information on those features. The page contains another table of "curated features". Initially, the table is populated with GAP-suggested features that must be reviewed, edited if necessary, then approved before submission to the database. The curator is free to add additional features to either list. As curated features are added to the database, they will appear in the graphical display. Some edits will have downstream consequences. For example, editing the coordinates of a CDS feature will change the inferred translation, invalidating the PAP annotation, and triggering its re-execution.

After completing a PGI, the curator may return to the PGI list to annotate another or continue to curate the protein sequence derived from that PGI. Protein curation begins with a "protein editor" similar in many respects to the "PGI editor". However, whereas PGI curation focuses on features, protein curation involves features *and* attributes. Some tools, such as Pfam, explicitly provide both types of information. A protein may have a *feature* representing an alignment to an HMM representing a collection of domains with a common function. From that similarity, the protein is assigned an *attribute* that represents that function (such as GO terms). This connection is invaluable for rapid functional characterization. In many cases, however, a curator must combine information from a variety of sources to make difficult functional assignments. Making the correct assignment will have a critical impact on the validity of downstream analyses such a pathway assignment and conclusions based on comparative genomics.

The PATRIC pipelines and curation tool are aimed at high-throughput genome annotation, however, this sort of batch processing is not conducive to exploratory studies where the user needs to choose tools and parameters interactively. ToolBus is an ideal platform for work in this area. Using ToolBus, an optimal pipeline configuration can be determined then applied using the automated system described earlier. The hope is that by exploiting ToolBus' flexibility, it can complement many of the approaches used in the PATRIC project. In doing so, lessons learned in building the Curation Tool can be applied to building a portable, reusable genome annotation and curation system. In time, the synergy between PATRIC and PathPort projects can develop into a standardized, yet extensible system that will allow genomics laboratories to attack problems that have hitherto been out of their reach without substantial bioinformatics support. This holds the promise of increasing the capabilities of the biomedical research community.

Administrative Resource Center for Proteomics Biodefense

NIAID's Administrative Resource Center (ARC) for proteomics biodefense team provides support to seven Proteomics Research Centers (PRCs) by developing an information system that contains data and technology protocols generated by each of the seven distributed Proteomics Research Center (PRC) sites. The ARC is a collaboration between Social and Scientific Services, Inc., Georgtown University, and the CI Group. The CI Group project team has prototyped a public accessible database system (http://proteinbank.vbi.vt.edu:8080/bprc), which supports functionalities with account/document management and data downloading/uploading. The backend of this Web site is an Oracle-based relational database with multiple schemas modeled with data types of administration, 2-D gel, and Mass spectrometry (MS). There are three schemas being deployed so far for testing the database

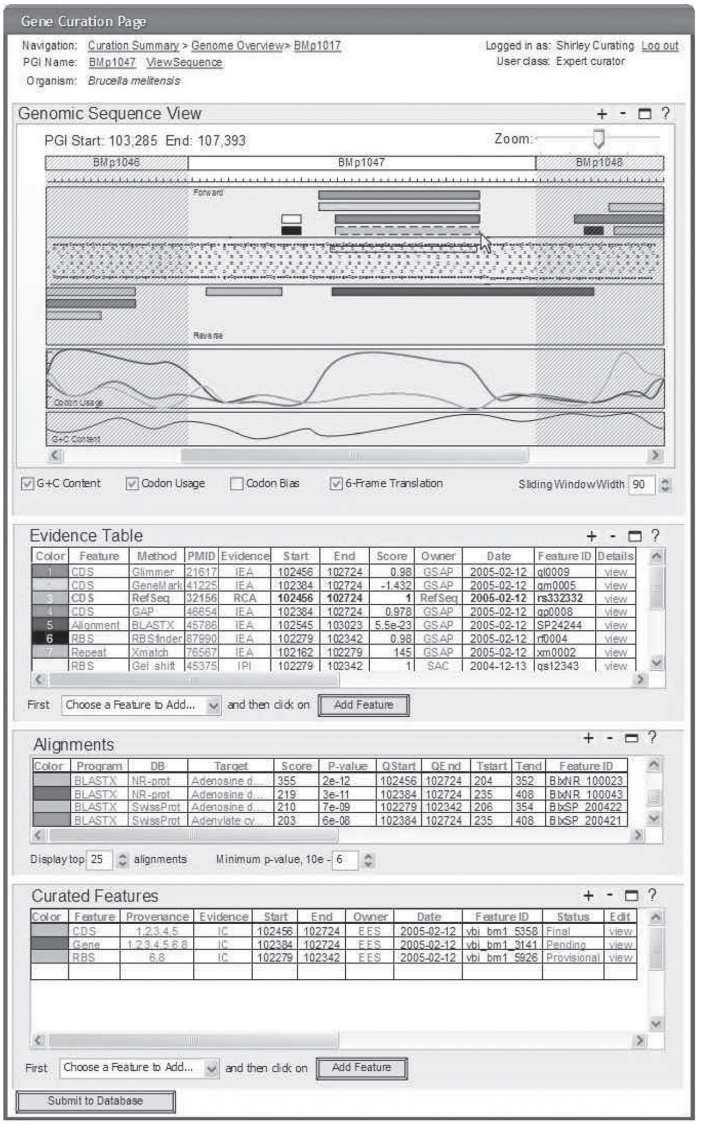


Figure 4. PATRIC Curation Tool Prototype Gene Edit Page. This prototype of the gene edit page illustrates the complexity of the information that the curators must interpret before creating their final "curated features". The "navigation bar" (top) gives basic information on the organism, putative gene interval (PGI), and curator. The Genomic Sequence View presents a graphical overview of the DNA segment corresponding to that PGI. Features are displayed as colored horizontal bars and as text in the Evidence and Curated Feature tables below. The sequence view also displays sequence properties such as G+C content, codon usage, and 6-frame translations. Features generated by the GSAP are shown in the Evidence Table along with experimental information from the literature. This data is considered "immutable" and is not subject to change by the curators. Alignments from programs such as BLASTX also fall into this category. The bottom table is reserved for curated features. Here, the curator indicates the evidence on which each curated feature is based in the "provenance" column. For example, the curated ribosome binding site (RBS) feature is based on the information from the RBSfinder (Suzek et al., 2001) program and a gel shift experiment reported in a journal article (shown in the Evidence Table). Similarly, the CDS feature is based on evidence from features 1–5. When the CDS feature is entered into the database, the corresponding translation will be sent to the Protein Annotation Pipeline (PAP).

functionalities. The first is the administrative schema that supports user profile, query history, data uploading, and file management. The second is a schema to support public 2-D gel and MS proteomics data downloading. The third is the schema for PRC and internal 2-D gel and MS data storage in highly decomposed fashion. A total of 12 MS datasets from four organisms with 13772 protein hits and a total of six sets of 2-D gel with 2936 spots list have been populated into this database schema during the first year.

Acknowledgements

We are grateful to the members of VBI's Core Facilities for their strong support of our work and team spirit. Special thanks to Tracy Wilkins, Minnis Ridenour, Linwood McCoy, and Charles Steger from Virginia Tech, and to VBI's administrative, financial, outreach and communications staff for their professionalism and service. BWSS specifically thanks Ms. Shannon Worringham for making it all possible for him. The authors are grateful to the US Department of Defense (contracts DAAD 13-02-C-0018 and W911SR-04-0045), Philip Morris Corporation, and the National Institute of Allergy and Infectious Diseases (Cooperative Agreement 1 U54 AI057168-01 and 1 R21 AI057875-01, and Contracts HHSN266200400035C and HHSN266200400061C); two IBM Shared University Research (SUR) equipment donation awards, and one SUN Center of Excellence equipment donation award for the generous funding they have given to support the ToolBus, GAP, and PATRIC projects respectively.

References

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

Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, and Murzin AG (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* **32**: D226–229

Anonymous (2005a) The GUS Platform for Functional Genomics

Anonymous (2005b) Automatic Sequence Analysis and Annotation Pipeline (ASAP)

Anonymous (2003) Who'd want to work in a team? *Nature* **424**: 1

Atkins DE, Droegemeier KK, Feldman DI, Garcia-Molina H, Klein ML, Messerschmitt DG, Meddina P, Ostriker JP, and Wright MH (2003) Revolutionizing science and engineering through cyberinfrastructure: report of the National Science Foundation blue-ribbon advisory panel on cyberinfrastructure. Arlington, Va: National Science Foundation

Barnett MJ, Toman CJ, Fisher RF, Long SR. (2004) A dual-genome Symbiosis Chip for coordinate study of signal exchange and development in a prokaryote-host interaction. *Proc Natl Acad Sci U S A* **101**: 16636–41

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, and Eddy SR (2004) The Pfam protein families database. *Nucleic Acids Res* **32**: D138–141

Becker A, Berges H, Krol E, Bruand C, Ruberg S, Capela D, Lauber E, Meilhoc E, Ampe F, de Bruijn FJ, Fourment J, Francez-Charlot A, Kahn D, Kuster H, Liebe C, Puhler A, Weidner S, Batut J (2004) Global changes in gene expression in *Sinorhizobium meliloti* 1021 under microoxic and symbiotic conditions. *Mol Plant Microbe Interact* **17**: 292–303

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, and Schneider M (2003) The

SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31**: 365–370

Brittain J and Darwin IF (2003) Tomcat: The Definitive Guide. O'Reilly, 1st edition

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

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

Erl T (2004) Service-Oriented Architecture: A Field Guide to Integrating XML and Web Services. Prentice Hall PTR

Fauci AS (2001) Infectious diseases: considerations for the 21st century. Clin Infect Dis **32**: 675–85

Gourley D and Totty B (2002) HTTP: The Definitive Guide. O'Reilly, 1st edition

Graham S, Simeonav S, Boubez T, Daniels G, Davis D, Nakamura Y, and Neyma R (2004) Building Web Services with JAVE: Making Sense of XML, SOAP, WSDL, and UDDI. Sams, 2nd edition

Harris MA, Clark J, et al (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32**: D258–261

Haddad C, Bedell K, Brown P, Srinivas D, and Almaer D (2004) Programming Apache Axis. O'Reilly and Associates, 1st edition

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

Henikoff JG and Henikoff S (1996) Blocks database and its applications. Methods Enzymol **266**: 88–105

Henikoff S, Henikoff JG, and Pietrokovski S (1999) Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics* **15**: 471–479

Hoon S, Ratnapu KK, Chia JM, Kumarasamy B, Juguang X, Clamp M, Stabenau A, Potter S, Clarke L, and Stupka E (2003) Biopipe: a flexible framework for protocol-based bioinformatics analysis. *Genome Res* **13**: 1904–1915

Huang Q, Liu D, Majewski P, Schulte LC, Korn JM, Young RA, Lander ES, Hacohen N (2001) The plasticity of dendritic cell responses to pathogens and their components. *Science* **294**:870-875

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, and Salzberg SL (2004) Versatile and open software for comparing large genomes. *Genome Biol* **5**: R12

Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP, and Bork P (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res* **32**: D142–144

Lowe TM and Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**: 955–964

Lukashin AV and Borodovsky M (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res* **26**: 1107–1115

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

Pruitt KD, Tatusova T, and Maglott DR (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **33**:D501–D504

Snell J, Tidwell D, and Kulchenko P (2001) Programming Web Services with SOAP. O'Reilly, 1st edition

Sonnhammer EL, Eddy SR, Birney E, Bateman A, and Durbin R (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res* **26**: 320–322

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, and Birney E (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* **12**: 1611–1618

Stein LD (2003) Integrating biological databases. *Nat Rev Genet* **5**: 337-345

Wu CH, Yeh LS, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu Z, Kourtesis P, Ledley RS, Suzek BE, Vinayaka CR, Zhang J, and Barker WC (2003) The Protein Information Resource. *Nucleic Acids Res* **31**: 345–347

Zhang Y, Yin Y, Chen Y, Gao G, Yu P, Luo J, and Jiang Y (2003) PCAS--a precomputed proteome annotation database resource. *BMC Genomics* **4**: 42

Publications

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

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

Sobral BW (2005) Cyberinfrastucture for PathoSystems Biology. In Proceedings of the Brazilian Symposium in Bioinformatics (Sao Leopold), Lecture Notes in Bioinformatics volume 3594, Springer Verlag

Genomic and Bioinformatic Analysis of Phytophthora-Host Interactions

Brett Tyler
Research Professor, VBI
Professor of Plant Pathology, Physiology and Weed Science,
Virginia Tech
bmtyler@vt.edu

Sucheta Tripathy, Dianjing Guo, Yinghui Dan, Lecong Zhou, Trudy Torto-Alalibo, Daolong Dou, Xuemin Zhang, Hua Li, Brian Smith, Konstantinos Krampis, Lachelle Waller, Bing Liu, Venkataraghavan Srinivasan, Regina Hanlon, Elizabeth Bush, Felipe Arredondo, Tejal Kharkhanis, Nathan Bruce, Varun Pandey, Samantha Chandrasekar, Bing Fang, Ken Tian, Angela Ko, Nikolaus Galloway, Andrew Kincaid, Rajat Singhania, Shomir Wilson
TM Murali (CS; VBI Fellow), Ina Hoeschele, Saghai Maroof (CSES), Keying Ye (Statistics), Reinhard Laubenbacher, Pedro Mendes, Vladimir Shulaev, Allan Dickerman, John McDowell (PPWS); Anne Dorrance (Ohio State), Steven St. Martin (Ohio State), Jeffrey Boore (DOE JGI), Daniel Rokhsar (DOE JGI), Nik Grunwald (USDA-ARS, Oregon), Sandra Clifton (Washington University), Jim Beynon (Warwick University, UK)

Introduction

Plant pathogens from the genus *Phytophthora* cause destructive diseases in an enormous variety of crop plant species as well as forests and native ecosystems (Erwin and Ribiero, 1996). The potato pathogen, *P. infestans*, was responsible for the Irish potato famine, and is still a destructive pathogen of concern for biosecurity. The newly emerged *Phytophthora* species, *P. 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, *P. sojae*, is also causing serious losses to the US soybean crop, of the order of $1-2 million annual damage (Erwin and Ribiero, 1996). This species has been used for many basic studies of *Phytophthora* because it is easy to genetically manipulate. Superficially, *Phytophthora* pathogens resemble fungi, but they in fact belong to a kingdom of life called Stramenopiles that are most closely related to algae such as kelp and diatoms. Hence conventional fungus control measures often fail against these pathogens.

The overall thrust of our research is to identify and characterize the genes in *Phytophthora* species that enable these pathogens to recognize their plant hosts and to 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 approaches, namely using a combination of high throughput experimental methods and bioinformatic approaches 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.

Genome Sequencing of *Phytophthora Sojae* and *Phytophthora Ramorum*

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. To aid the assembly of the sequence, a physical map of *P. sojae* was constructed from two Bacterial Artificial Chromosome (BAC) libraries. The BAC clones were digested with seven enzymes and the fragments labeled with four fluorescent dyes to obtain detailed fingerprints of each BAC (Luo et al., 2003). Contigs were assembled from the fingerprints using FPC software (Soderlund et al., 1997). We have also been developing a new algorithm for fingerprint contig assembly based on Hopfield networks (Karaoz et al., 2004). A total of 8,681 clones were assembled into 257 contigs. The largest contig spans a 2.2 Mb region. A minimum tiling path consisting of 1400 clones was assembled and BAC end sequencing has been initiated to integrate the physical map with the genome sequence.

P. sojae and *P. ramorum* are diploid, so that sequences of two haplotypes were obtained from both species. The *P. ramorum* sequence contains approximately 200,000 single nucleotide polymorphisms (SNPs). In order to determine whether these SNPs may be useful in genetically fingerprinting *P. ramorum* isolates from the California epidemic, we have designed an Affymetrix SNP GeneChip containing probes for 880 of the SNPs. We will use these chips to assay differences in the genomes of *P. ramorum* isolates caused by gene conversion and mitotic crossing over. These kinds of changes occur frequently in the genome of *P. sojae* (Chamnanpunt et al., 2001). The *P. sojae* genome contained only 7,000 SNPs, consistent with the fact that *P. sojae* is homothallic and inbreeds prolifically to produce long-lived oospores, whereas *P. ramorum* is heterothallic and outbreeding.

The JGI genome annotation pipeline was used to predict and annotate 19,027 genes in the genome of *P. sojae* and 15,743 genes in the genome of *P.ramorum.* 9,768 putative pairs of orthologs were identified between *P. sojae* and *P. ramorum* gene models using the criterion of best bi-directional Blast hits. 7,850 ESTs from *P. sojae* have been mapped onto the genomic assembly of *P. sojae* and used for validation, correction, and extension of predicted gene models. 7,088 models were supported at least partially. Additional annotation has been provided by the *Phytophthora* research community via an Annotation Jamboree held in August 2004 at the DOE JGI, and via a community gene annotation interface developed at VBI.

Whole-genome DNA sequence alignment demonstrated a high level of similarity between the two species; 75.8 percent of all *P. ramorum* and 79.7 percent of all *P. sojae* exons were covered by the alignment. Based on single-linkage clustering analysis, the majority of predicted genes from both genomes form groups of homologous proteins. Only a small number of genes, 1,755 in *P. sojae* and 624 in *P. ramorum,* did not have a homolog in the other genome when a significance threshold of 1e-8 was used. The overall higher number of predicted genes in *P. sojae* results from greater expansion of many gene families in *P. sojae*. Overall, about 80 percent of the genes in both genomes have homology to known proteins or known protein domains. 1,563 pairs of *Phytophthora* orthologs showed no homology to any other species than *Phytophthora*. In order to expand the comparative analysis of the genome sequences of oomycete pathogens, we recently began sequencing the genome of the *Arabidopsis* pathogen *Hyaloperonospora parasitica*, in collaboration with Washington University Genome Sequencing Center.

Analysis of the genes encoded in the genome sequences supports the hypothesis that heterotrophic Stramenopiles such as the oomycetes, and photosynthetic Stramenopiles such as diatoms share a secondarily photosynthetic ancestor (Fast et al., 2001), but suggest that the heterotrophic and photosynthetic lineages diverged very early in Stramenopile evolution. The genomes show rapid diversification of

proteins associated with plant infection such as hydrolases, ABC transporters, protein toxins, proteinase inhibitors, and avirulence gene products (see below for more details), but relatively little diversification of biosynthetic genes for metabolite toxins. The *P. sojae* and *P. ramorum* genomes show substantial gene colinearity (synteny) except in regions encoding putative pathogenicity genes, where there is evidence for accelerated genome evolution.

VBI Microbial Database and Genome Community Annotation Tool

Analysis and visualization of gene prediction and annotation for these two genomes are available from JGI Genome Portal (www.jgi.doe.gov/genomes) and at the VBI Microbial Database (VMD) (http://phytophthora.vbi.vt.edu). The database schema for VMD is based on the Genomics Unified Schema (GUS) developed at the Center for Bioinformatics, University of Pennsylvania (http://www.gusdb.org/). GUS is an object-relational schema that provides fully integrated support not only for genome sequence data, but also EST, microarray, and proteomics data. GUS is an open source project. In implementing GUS, we significantly improved the documentation for installation of GUS. In addition, we built a new genome browser for viewing genome sequence data stored in GUS and also a new interface (Genome Community Annotation Tool; GCAT) for community annotation of the sequences. GCAT enables users to submit additions or modifications to sequence assemblies, gene models, or functional annotations of gene products, including Gene Ontology term assignments. Submitted annotations are held in temporary tables until reviewed and approved by a sequence curator.

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. Until recently, there were few terms for annotating the contributions of genes to interactions between microbes and hosts, making it very difficult to systematically identify similarities among plant-associated microbes. In collaboration with João Setubal at VBI and researchers at Cornell University (Alan Collmer), the University of Wisconsin (Nicole Perna), Wells College (Candace Collmer), North Carolina State University (Ralph Dean and David Bird), and the Institute for Genome Research (Michelle Gwinn-Giglio, Owen White and Robin Buell), we have developed a set of 33 new terms specifically to describe biological processes carried out by host-associated microbes and reorganized 13 other terms and their sub-terms. The changes (see http://pamgo.vbi.vt.edu) were accepted by the GO consortium in January 2005.

Function Of Avirulence Genes During Phytophthora Infection

At least 14 disease resistance genes (*Rps* genes) have been identified in soybean that protect against *P. sojae* infection (Burnham et al., 2003). All of these Rps genes are dependent for their effectiveness on the presence of corresponding avirulence genes in the pathogen. For example, a soybean cultivar containing the resistance gene Rps1a is not protected against *P. sojae* strains that do not express avirulence gene Avr1a. At least 10 avirulence genes have been identified genetically in *P. sojae* (Tyler, 2002). Avirulence genes are also found in plant pathogenic bacteria, fungi, nematodes and viruses. In bacteria, there is extensive evidence that many proteins encoded by avirulence genes enter the cytoplasm of plant cells via a type III secretion pathway, and once inside the plant cell, act to increase the susceptibility of the plant tissue to infection by the pathogen (Chang et al., 2004). There is indirect evidence that proteins encoded by fungal and oomycete avirulence genes can enter into plant cells (Tyler, 2002), but nothing is known about how this might occur.

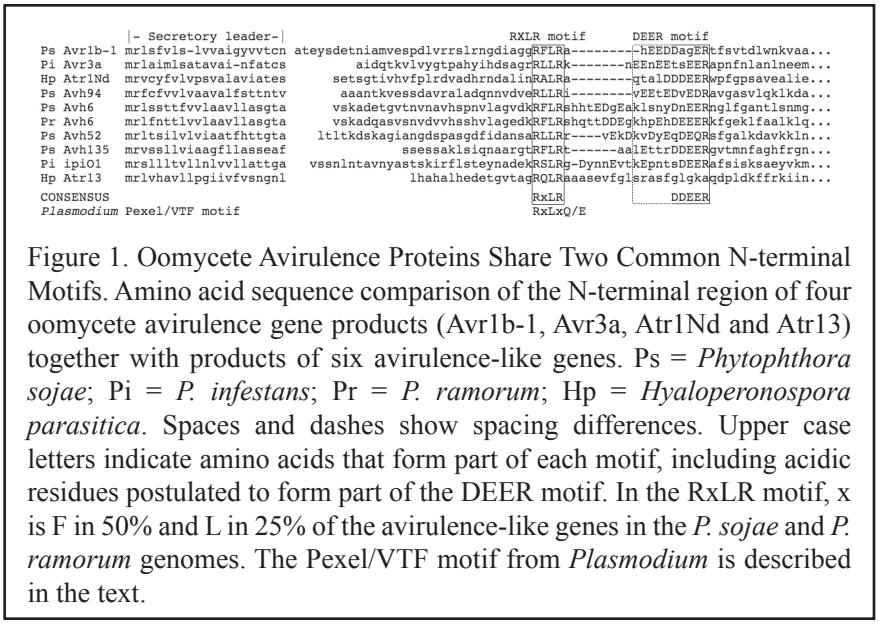


Figure 1. Oomycete Avirulence Proteins Share Two Common N-terminal Motifs. Amino acid sequence comparison of the N-terminal region of four oomycete avirulence gene products (Avr1b-1, Avr3a, Atr1Nd and Atr13) together with products of six avirulence-like genes. Ps = *Phytophthora sojae*; Pi = *P. infestans*; Pr = *P. ramorum*; Hp = *Hyaloperonospora parasitica*. Spaces and dashes show spacing differences. Upper case letters indicate amino acids that form part of each motif, including acidic residues postulated to form part of the DEER motif. In the RxLR motif, x is F in 50% and L in 25% of the avirulence-like genes in the *P. sojae* and *P. ramorum* genomes. The Pexel/VTF motif from *Plasmodium* is described in the text.

We previously cloned the *Avr1b*-1 gene of *P. sojae* and showed that it encoded small secreted protein that could trigger a defense response in soybean plants that carry the Rps1b resistance gene (Shan et al., 2004). Since the soybean resistance gene *Rps*1b encodes an intracellular protein we infer that the Avr1b protein can enter soybean cells, and can do so without the aid of the pathogen. We recently collaborated with researchers at Warwick University (UK) who have cloned the Atr1 avirulence gene from the *Arabidopsis* pathogen *Hyaloperonospora parasitica* and researchers at the Scottish Crops Research Institute who have cloned the *Avr3a* avirulence gene from *Phytophthora infestans*. *Avr3a* is of interest because it is one of the genes in *P. infestans* most similar to *Avr1b-1* (Rehmany et al., 2005; Shan et al., 2004). *Atr1* is of interest because it is located in a region of the *H. parasitica* genomes that is homeologous to the region containing *Avr3a* in *P. infestans* (Rehmany et al., 2005). Comparison of the sequences of the three avirulence genes with each other (Rehmany et al., 2005) and with over a hundred genes in the *P. sojae* and *P. ramorum* genomes encoding secreted proteins similar to Avr1b has identified two conserved motifs called RXLR and dEER a short distance from the end of the secretory leader (Figure 1). All three of the avirulence proteins, and most of the *P. sojae* and *P. ramorum* paralogs contain both motifs, but a second *H. parasitica* avirulence protein, *Atr13* contains just the RXLR motif. Since all four avirulence proteins are inferred to interact with intracellular plant disease resistance proteins, they are also inferred to have the ability to enter the plant cell, we hypothesize that the RXLR motif, with or without the dEER motif, is involved in the ability of these proteins to enter the plant cell. This hypothesis is encouraged by recent reports that proteins secreted by the malaria parasite *Plasmodium* have the ability to cross the membrane enclosing the parasite (the parasitiphorous vacuolar membrane) into the lumen of the red blood cell if they 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 (Figure 1).

We are currently characterizing the contribution of Avr1b proteins and their paralogs to the virulence of *P. sojae*, and the function of the RXLR motif, using DNA transformation of *P. sojae*. In addition to conventional protoplast methods, we have developed a moderate throughput procedure for transformation of *P. sojae* using *Agrobacterium tumefaciens*. We are also exploring methods for high efficiency gene silencing.

Counter-Play of Plant and Pathogen Genes During Phytophthora Infection of Soybean

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 (also called 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, because multigenic resistance is created by many genes making small contributions, this kind of resistance is much harder to improve by conventional breeding (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. The initial objective of the project is to assay the transcriptional profiles of both soybean and *P. sojae* at two timepoints during infection of each of 12 cultivars of soybean differing in quanti-tative resistance. The transcriptional profiles will be assayed using Affy-metrix GeneChips that contain probes for 38,000 soybean genes and 15,800 *P. sojae* genes. We have completed a series of pilot experiments in order to optimize the statistical design for the assay of the 12 cultivars and have constructed a Laboratory Information Management System (LIMS) in order to track all of the samples and data related to the experiments. This objective is close to completion. The second objective will be to carry out transcriptional profiling over a detailed time course of infection of two selected resistant and two selected susceptible cultivars. 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.

Design of a Soybean PathoChip

A key resource for this project is the Affymetrix Soybean "PathoChip" GeneChip containing 11 probes each for 35611 soybean genes and 15421 *P. sojae* genes, as well as 7431 soybean cyst nematode transcripts. Project staff were closely involved in the design of this chip, in collaboration with staff from Affymetrix and a committee representing the soybean genomics community. Project staff contributed all of the gene sequences for *P. sojae*, including 7200 EST unigene assemblies and preliminary gene predictions from the *P. sojae* genome sequence. The project also contributed 1700 soybean unigenes derived from infection of susceptible soybean tissue by *P. sojae*. Project staff also designed 24 probes specific for each of the soybean and *P. sojae* 18S rRNAs, to enable quantitation of the level of *P. sojae* RNA in infected tissue.

Cross-Kingdom Comparative Genomics Of The Oxidative Stress Response

Damaging reactive oxygen species such as superoxide, hydrogen peroxide, and hydroxyl radicals can be produced by a wide variety of cellular processes. The production of reactive oxygen species has in particular been co-opted by animals (Shepherd, 1986) and plants (Sutherland, 1991) as a defense response against microbial infection. In animals, macrophages and neutrophils use an oxidative burst to destroy engulfed microbes. In plants, the oxidative burst can destroy microbes directly, as well as create signal molecules to trigger additional defense responses. As a consequence, most plant pathogens and many animal pathogens (especially those such as Mycobacterium and Histoplasma that colonize macrophages), must have active mechanisms to protect against oxidative damage as an essential component of their pathogenicity machinery (Ismail et al., 2002). In the case of the malaria parasite *Plasmodium*, oxidative stress is also created during infection of red blood cells by the degradation of hemoglobin and the release of free heme (Robert et al., 2002).

Pathogens of humans, animals, and plants are found in many different kingdoms of life. Some kingdoms, such as eubacteria, plants, animals, and fungi contain well-studied model organisms. Information these model organisms greatly aids the understanding of related pathogens. Many other important pathogens are found in kingdoms that do not include model organisms. These kingdoms include alveolates (malaria and other apicomplexan parasites), diplomonads (e.g. *Giardia*), euglenoids (e.g. trypanosomes, *Leishmania*), and stramenopiles (e.g. *Blastocystis*). Others include plant pathogens such as *Phytophthora*. Because these organisms are so divergent evolutionarily from model organisms such as yeast and *Arabidopsis*, it is significantly more difficult to infer the functions of their genes simply on the basis of sequence similarity.

The dual goals of this project are: (i) to characterize the role of oxidative stress protection in the pathogenicity of the plant pathogen *Phytophthora sojae* and the human malaria pathogen *Plasmodium falciparum*; and (ii) in the process, explore and validate novel bioinformatic approaches for transferring functional information from model organisms across kingdom boundaries. The oxidative stress response is well-suited to the exploration and validation of new approaches as it is well-studied in several model organisms including yeast and *Arabidopsis*, and many of the components of the response are well-conserved.

As a first step in this project, we have used Affymetrix GeneChips to measure the transcriptional responses of cells of *Saccharomyces cerevisiae* (a fungus), *Arabidopsis thaliana* (a plant) and *Phytophthora sojae* (a stramenopile) to cumene hydroperoxide. The cells of each species were grown under as closely analogous conditions as possible, and were each subjected to the maximum concentration of cumene hydroperoxide that did not produce a reduction in

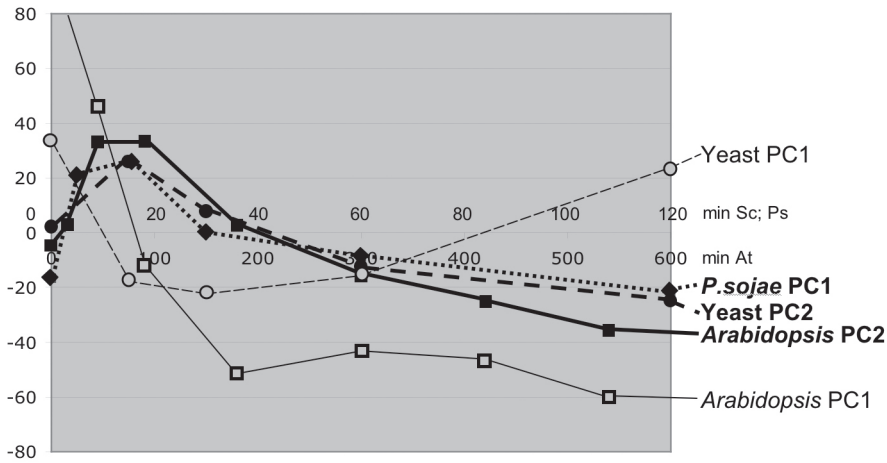


Figure 2. Cross-kingdom Comparison of the Oxidative Stress Response using Principal Components Analysis. Exponentially growing cells of yeast (*S. cerevisiae*), *Phytophthora sojae* and *Arabidopsis thaliana* were exposed an IC_5 of cumene hydroperoxide (CuOOH) or a matching concentration of cumene alcohol (CuOH) for the periods of 5 to 540 minutes. Each experiment was repeated three times, with two internal replications. Transcriptional profiles were determined from each culture using Affymetrix GeneChips. Expression levels were normalized by RMA analysis and and the $\log_2$ratios between the CuOOH and CuOH expression levels subjected to principal components analysis. The principal components cumulatively accounting for >90% of the variation in each case were plotted against treatment time. The time scale for *Arabidopsis* was compressed five-fold to highlight the similarity of PC2 to PC2 of yeast and PC1 of *P. sojae*. The vertical scale is in arbitrary units and was not adjusted for any of the species.

growth. The *Arabidopsis* cells were grown in suspension culture. For each organism, the transcriptional response was measured at five timepoints ranging from 5 min to 9 hours, with six replicates. In order to identify common and specific responses in the three organisms, Principal Components Analysis was used to identify major sets of genes with correlated expression profiles. As shown in Figure 2, a common response was identified in all three species that consisted of rapid induction of a set of genes followed by a rapid return to normal (PC2 in yeast and Arabidopsis and PC1 in *Phytophthora*). The yeast and *Phytophthora* responses occurred on the same time scale while the *Arabidopsis* response occurred on a time-scale that was approximately five-fold longer. We are using Hopfield networks (Karaoz et al., 2004) to integrate functional information from the yeast and *Arabidopsis* genes involved in the responses to predict the full set of oxidative protection genes in *Phytophthora sojae*, including both stress-induced and constitutively expressed genes.

References

Burnham KD, Dorrance AE, Francis DM, Fioritto RJ, and St-Martin SK (2003) Rps8, A New Locus in Soybean for Resistance to *Phytophthora sojae*. *Crop Sci* **43**: 101–105

Chamnanpunt J, Shan W-X, and Tyler BM (2001) High frequency mitotic gene conversion in genetic hybrids of the oomycete *Phytophthora sojae*. *Proc Natl Acad Sci USA* **98**: 14530–14535

Chang JH, Goel AK, Grant SR, and 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 and Ribiero OK (1996) *Phytophthora Diseases Worldwide*. APS Press, St. Paul, Minnesota

Fast NM, Kissinger JC, Roos DS and Keeling PJ (2001) Nuclear-encoded, plastid-targeted genes suggest a single common origin for apicomplexan and dinoflagellae plastids. *Mol Biol Evol* **18**: 418–426

Hiller NL, Bhattacharjee S, van-Ooij C, Liolios K, Harrison T, Lopez-Estraño C, and Haldar K (2004) A Host-Targeting Signal in Virulence Proteins Reveals a Secretome in Malarial Infection. *Science* **306**: 1934–1937

Ismail N, Olano JP, Feng H-M, and Walker DH (2002) MiniReview: Current status of immune mechanisms of killing of intracellular microorganisms. *FEMS Microbiol Lett* **207**: 111–120

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

Luo M-C, Thomas C, You FM, Hsiao J, Ouyang S, Buell CR, Malandro M, McGuire PE, Anderson OD, and Dvorak J (2003) High-throughput fingerprinting of bacterial artificial chromosomes using the SNaPshot labeling kit and sizing of restriction fragments by capillary electrophoresis. *Genomics* **82**: 378–389

Marti M, Good RT, Rug M, Knuepfer E, and 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, and Beynon JL (2005) Differential recognition of highly divergent downy mildew avirulence gene alleles by RPP1 genes from two Arabidopsis lines. *Plant Cell* accepted

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

Robert A, Dechy-Cabaret O, Cazelles J, and Meunier B (2002) From Mechanistic Studies on Artemisinin Derivatives to New Modular Antimalarial Drugs. *Acc Chem Res* **35**: 167–174

Shan W, Cao M, Leung D, and 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

Shepherd VL (1986) The role of the respiratory burst of phagocytes in host defense. *Semin Respir Infect* **1**: 99–106

Soderlund C, Longden I, and Mott R (1997) FPC: a system for building contigs from restriction fingerprinted clones. *Bioinformatics* **13**: 523–535

Sutherland MW (1991) The generation of oxygen radicals during host/plant responses to infection. *Physiol Mol Plant Pathol* **39**: 79–93

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. *Annu Rev Phytopath* **40**: 137–167

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

Publications

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

2005 Research Reports From the Virginia Bioinformatics Institute's Faculty Fellows

Epi-fluorescent Image Modeling and Denoising for Viral Infection Analysis

Amy E. Bell
Faculty Fellow, VBI
Associate Professor of Electrical and Computer Engineering, Virginia Tech
abell@vt.edu

Introduction

Quantitative fluorescence microscopy is increasingly being used for studying localization, dynamics, and interaction of organelles, proteins, and pathogens (Lippincott-Schwartz et al., 2001). These applications demand precise measurements in both the spatial and temporal domains. Furthermore, multiple protein markers and an image sequence over time are now commonly employed to understand complex biological processes (Gerlich and Ellenberg, 2003; Gerlich et al., 2003). The changing intensity values over a sequence of images indicate how much protein is present, where it is located (i.e., spatial information), and how it is changing over the course of an experiment (i.e., time information). For example, an image sequence of a viral protein marker and its location, combined with an examination of viral plaques (kill zones), contributed to the development of a phenotypic characterization method for virus-host interactions, based on an evolving dynamic relationship in vitro (Duca et al., 2001). However, in that study, only localization measurements were possible because the images suffered from shading artifacts that prevented accurate intensity measurements.

At low magnification (such as in our application: 4X and 10X), wide-field fluorescence microscopy (WFM) is plagued by several types of distortion that may preclude meaningful quantitative analysis of the captured image sequence. The distortions in the captured image are inherent to the image collection process; they arise from various sources, including: the spectral characteristics of the markers; impurities in the glass or plastic substrate to which the cells are attached; thickness variation in the sample or substrate; fluctuations in the illumination source; and optical characteristics of the imaging system (Swedlow 2003).

This paper presents a model and methods to rectify problems associated with montaged images of thin cells generated with WFM. While methods have been developed to deal with these distortions individually, the literature fails to comprehensively address these distortions and their correction (Sled et al., 1998; Tomazevic et al., 2002). Our fluorescence intensity image model for WFM includes four common types of distortion: random noise, background shading artifact, intensity normalization, and spectral overlap of the markers. We also describe an automated, retrospective correction procedure for each distortion, resulting in a denoised image sequence. Finally, we present a method for obtaining the spatial profile of the relative fluorescent intensity, called the immunofluorescent intensity signal (IIS), for each denoised image. The IISs provide an initial quantitative profile of the nature of virus-host interactions.

Fluorescent Image Model and Denoising

In order to capture the entire specimen at one time point, several individual images (30 in our case) taken at low magnification (4X) are combined. The result is one montage image (MI) comprised of several sub-images (SI) arranged in a matrix (see Figure 3A). One MI represents one time point in the experiment; thus, an experiment is comprised of a sequence of MIs.

We model the noisy image corresponding to the j^{th} SI of an experiment as

$$F_n(x,y,j) = I_0(j)\left[Kc(x,y,j) + B(x,y,j) + S(x,y,j)\right] \quad (1)$$

where $F_n(x,y,j)$ is the captured, *i.e.*, measured, noisy fluorescent intensity corresponding to pixel (x,y) in the j^{th} SI; $c(x,y,j)$ represents the true signal intensity; $B(x,y,j)$ is the additive non-uniform background illumination profile from optical features associated with WFM; $S(x,y,j)$ is the "ghost" image from spectral overlap from the second channel corresponding to the second fluorophore; $I_0(j)$ is the true maximum brightness of the illumination source when the j^{th} SI was acquired; and, K reflects a combination of other parameters that remain constant throughout the experiment. Random noise is not included in our model; instead, it is treated as a separate, spurious distortion.

Figure 1 depicts our correction procedure for a noisy fluorescent image modeled by $F_n(x,y,j)$. Our algorithm accounts for the four types of noise: (1) random noise; (2) background shading, $B(x,y,j)$; (3) intensity normalization, $I_0(j)$; and (4) spectral overlap, $S(x,y,j)$. The result is a denoised fluorescent image that estimates the true fluorescent intensity, $c(x,y,j)$. The correction process depends on whether the collected noisy images are corrupted by spectral overlap (SO). In our experiments, the viral proteins are tagged with FITC (green channel), and the host proteins with CY3 (red channel). The images corresponding to host proteins are corrupted by spectral leakage from the viral channel; however, the images corresponding to viral proteins are relatively SO free. Finally, the denoised fluorescent images are used to provide quantitative spatio-temporal information about host-virus interactions. With a denoised image, an immunofluorescent intensity signal (IIS) may be generated at a given time point and compared to others in a series.

Denoising Viral Images (Spectral Overlap is Absent)

$S(x,y,j)$, the distortion due to spectral overlap, is negligible in our viral images. Consequently, (1) reduces to

$$F_n(x,y,j) = I_0(j)\left[Kc(x,y,j) + B(x,y,j)\right] \quad (2)$$

As shown in Figure 1, a viral image undergoes three denoising steps (in order): random noise removal, background shading correction (BSC), and intensity normalization.

Random noise removal: The random noise in $F_n(x,y,j)$ is removed by replacing the corrupted pixels with the same pixels in a similar SI (within the same MI) that is not corrupted by random noise. The resulting image is $F_r(x,y,j)$; it corresponds to II in Figure 1.

Background shading correction: Here the goal is to estimate the background, $\hat{I}_0(j)\hat{B}(x,y,j)$, in order to remove it. We derive an adaptive block size method, based on the SI and MI variances, that estimates the background illumination for each SI. This method is based on several assumptions inherent to the noisy viral images. The primary assumption is that, throughout the experiment, the viral MIs are primarily comprised of background; only 10-30 percent of the SIs exhibit any signal. Hence, the median variance corresponds to background and not signal. Our iterative process that considers minimum values within increasingly larger blocks ensures that our final estimate reflects the background and not the signal. The background estimate is computed in the following four steps.

Step 1: (a) Compute the variance of every SI within one MI.
(b) Find the median variance for the MI.

Step 2: Perform this step for each SI.

(a) Begin with a smallest block size (*e.g.*, 4x4) and divide the SI into blocks.

(b) Compute the minimum value within each block.

(c) Create a background estimate for the SI: the pixel values in each background block are the minimum value from the corresponding SI block in (2b).

(d) Compute the variance of the background estimate.

(e) Increase the block size and repeat steps (2b) through (2d) until the background variance is within 10% of the median variance in (1b).

Step 3: Arrange all of the SI background estimates obtained in (2) into one MI background estimate, $\hat{I}_0(j)\hat{B}(x,y,j)$.

Step 4: Subtract the background estimate from $F_r(x,y,j)$ to obtain F_b (Figure 1, V).

$$F_r(x,y,j) - \hat{I}_0(j)\hat{B}(x,y,j) = \hat{I}_0(j)\hat{K}\hat{c}(x,y,j) = F_b(x,y,j) \quad (3)$$

Intensity normalization: Here, the goal is to estimate an SI normalization factor, *NF*, in order to compensate for varying source brightness across the entire set of SIs. The main idea is to use the background estimate in order to normalize the SIs and place them on the same intensity scale. *NF(j)* is computed for each SI in the following four steps.

Step 1: Identify the brightest pixel in the SI background estimate $\hat{I}_0(j)\hat{B}(x,y,j)$

Step 2: Identify the brightest pixel among all of the candidates in Step 1, call it $\hat{I}_0(m)\hat{B}(x,y,m)$; the m^{th} SI has the brightest background pixel for the experiment.

Step 3: Compute $NF(j) = \dfrac{\hat{I}_0(m)}{\hat{I}_0(j)}$. Note that $\hat{B}(x,y,m) \approx \hat{B}(x,y,j)$ since the background shape is constant throughout an experiment.

Thus, $NF(j) \approx \dfrac{\hat{I}_0(m)}{\hat{I}_0(j)}$.

Step 4: Apply *NF(j)* to $F_b(x,y,j)$ to obtain F_c (Figure 1, VI):

$$NF(j) \times F_b(x,y,j) = \frac{\hat{I}_0(m)}{\hat{I}_0(j)} F_b(x,y,j) \approx \hat{I}_0(m)\hat{K}\hat{c}(x,y,j) = F_c(x,y,j) \quad (4)$$

Denoising Host Images (Spectral Overlap is Present)

Unlike viral images, host images suffer from distortion due to spectral overlap of the two fluorophores. Consequently, the host image model is indicated by (1)

$$F_n(x,y,j) = I_0(j)\left[Kc(x,y,j) + B(x,y,j) + S(x,y,j)\right]$$

As shown in Figure 1, a host image undergoes four denoising steps (in order) that are done in a different sequence than the corresponding viral images: random noise removal, intensity normalization, background shading correction, and spectral overlap correction. The algorithms for the first three noise types are modified versions of those algorithms when SO is absent (Rout et al., 2004).

***Spectral overlap correction*:** This is the final step in denoising host images. Spectral overlap correction is based on the assumption that the dead zone, *e.g.*, region R1 in Figures 2A and B, in the host MI is virtually devoid of any host

cells. Thus, any signal inside R1 corresponds to spectral leakage from the viral channel. For cytolytic viruses, including VSV, this is a very good approximation and introduces little error. The DZ, being the initial viral reservoir, contains mostly debris from the initial innoculum or released by locally lysed cells. Spectral overlap correction is accomplished in four steps.

Step 1: (a) The DZ (R1) is calculated from $F_b(x, y, j)$, the partly denoised host MI. Beginning at the center of infection, identify the radius at which the circular region first touches the "live" host cell region.
(b) Use the corresponding, denoised viral MI to calculate R2, the radius that indicates the outer edge of the viral infection spread.

Step 2: Compute the correlation coefficient (ρ) in R1 for a host/viral MI pair:

$$\rho = \frac{\sum_{i=1}^{n} x_i y_i - \mu_x \mu_y}{n \sigma_x \sigma_y} \quad (5)$$

where x_i, y_i are the host and viral pixel intensities; μ_x, μ_y, and σ_x, σ_y correspond to the means and standard deviations of host and viral images in R1.

Step 3: (a) For the R1 circular region, obtain a candidate host MI by subtracting a scaled version of the viral MI from the original host MI. Use a small initial scaling factor.
(b) Compute ρ for the candidate host MI and viral MI.
(c) Repeat steps 3a-b to find the scaling factor that results in $\rho = 0$ in R1.

Step 4: Use the final scaling factor to subtract the scaled viral MI from the original host MI within the circular region R2; F_c is the result (Figure 1, VI).

$$F_b(x, y, j) - \hat{I}_0(m)\hat{S}(x, y, j) = \hat{I}_0(m)\hat{K}\hat{c}(x, y, j) = F_c(x, y, j) \quad (6)$$

Generation of Immunofluorescent Intensity Signals

An immunofluorescent intensity signal (IIS) is generated from a denoised montaged image (MI). It reflects the relative amount of protein present at a given radial distance from the point of initial infection. A collection of IISs provides a phenotypic profile of host-virus dynamics by revealing changes in relative protein concentrations, as a function of radial distance, over time. A denoised MI is divided into concentric radial bands and the average fluorescent intensity for each band is obtained. However, this calculation is complicated by uneven source illumination that cannot be entirely eliminated by our methods, especially for the host images. In addition to generating noise in an SI, uneven source illumination also results in signal loss; insufficient illumination produces weak fluorescence that may not be captured in the recorded image. We detect those pixels suffering from signal loss and exclude them from the average estimates (Rout, S., et al., 2004).

Results and Discussion

The objective of the experiment is to test whether the spatial and temporal behavior of certain proteins in the host and/or virus, revealed by the IISs, might serve as reporters or markers of the induction of an antiviral state. Using immunocytochemistry, we label the viral protein VSV surface glycoprotein (VSV-G) as an indicator of propagation and two host enzymes (NOS-1, NOS-3) associated with host defensive response. VSV-G is identified with the FITC fluorescent marker and NOS-1 and NOS-3 are identified with the Cy3 fluorescent

marker. The images were denoised and IISs were subsequently generated using the methods described above. IIS data for 10 time points ranging from 6hrs-48hrs are obtained from denoised fluorescent images.

Direct comparison of the noisy (Figures 3A and 3C) and corrected (Figures 3B and 3D) images clearly illustrates the power of the method for extracting the true fluorescent signal from the noise, thereby enabling quantification. The noisy viral and host images exhibit a strong montage effect at the SI boundaries. It arises from a mismatch of intensity profiles in adjoining SIs: the mismatch occurs due to the BSA and time varying source illumination. The montage effect is completely eliminated in the viral images (Figure 3B) and is significantly minimized in the case of host images (Figure 3D). While the remaining artifact in the case of host images is undesirable, it is mitigated further during the subsequent computation of the IIS.

A smooth curve is fitted to the IIS generated from the denoised images (Figures 4A-H). Figures 4A-D depict noisy and corrected VSV-G viral and NOS1 host proteins at three time points (12 hours, 21 hours, and 36 hours). Similarly, Figures 4E-H correspond to noisy and corrected VSV-G viral and NOS3 host proteins at 9hr, 30hr, and 48hr. IIS for the host proteins are shown as a ratio of IIS derived from infected images to those derived from uninfected (mock) images.

A comparison of the noisy and denoised viral IISs illustrates the impact of the background shading correction on the viral fluorescent intensities. The oscillation associated with the montage effect in the noisy images is readily observed in Figures 4A and 4E (particularly at larger radial distances (> 4mm) where the image is mostly background); the oscillations are not present in Figures 4B and 4F. Furthermore, intensity normalization allows the denoised IISs to be compared across time points. The 48 hour IIS in Figure 4E exhibits a higher intensity peak than the earlier time points; however, the 48 hour IIS in Figure 4F has the lowest intensity peak.

A comparison of the noisy and denoised host IISs illustrates the value of the background shading and spectral overlap corrections on the host fluorescent intensities. Background shading correction increases contrast and allows a comparison of peak intensities and protein distributions relative to the advancing viral front across time. Figure 4D illustrates distinct upregulation (IIS ratio > 1) of NOS-1 in response to the virus at 12 and 21 hrs. Spectral overlap correction permits a better understanding of the host IIS at the smaller radial distances. Figure 3H indicates complete loss of NOS-3 at radial distances > 3mm between the 9 and 30 hour time points where the cells died and lifted off the well during fixation; this is not evident in Figure 4G. Figures 4G and 4H indicate that NOS-3 is not upregulated during infection (IIS ratio < 1). This result suggests it is unlikely that NOS-3 plays a role in the infection process. However, Figures 4C and 4D depict that NOS-1 is upregulated proximal to the advancing viral front at 12 hrs post infection, more upregulated at 21 hrs before viral arrest, and back below baseline at 36 hr after propagation has stopped. As previously reported (Komatsu et al., 1999), this suggests a possible role for NOS-1.

Conclusion

Low magnification montaged images obtained from fluorescent microscopy are denoised to remove inherent distortions associated with the measurement process; this is necessary for meaningful quantitative analysis. In this paper, we present a fluorescent image model and denoising process that corrects four types of distortion common in fluorescent microscopy. Our denoised images show significant improvement in image quality; the final corrected images are free from montage artifacts and spurious intensity variations. Quantitative analysis of the IIS derived from the denoised images provides accurate measurements of relative protein concentrations and distributions not possible with the IIS obtained from the noisy images.

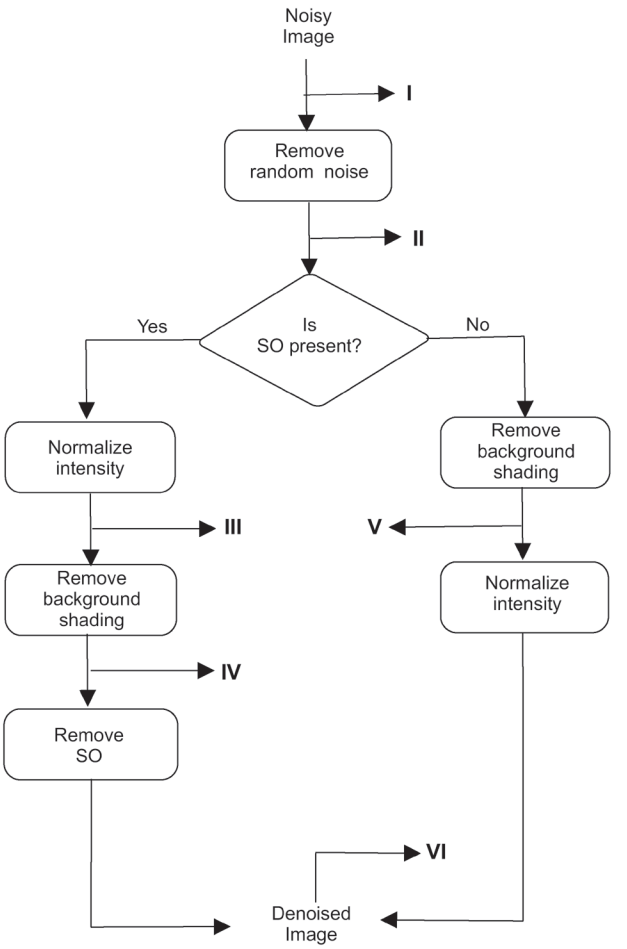


Figure 1. Flowchart of the correction procedure for double-labeled fluorescent images. I represents the original, noisy image measured by the microscope. II is the image obtained after random noise removal. In the presence of spectral overlap distortion, III corresponds to the image normalized for intensity variations and IV is the result after background shading correction. When spectral overlap is absent, V is the result after background shading correction. VI is the final, denoised image in all cases.

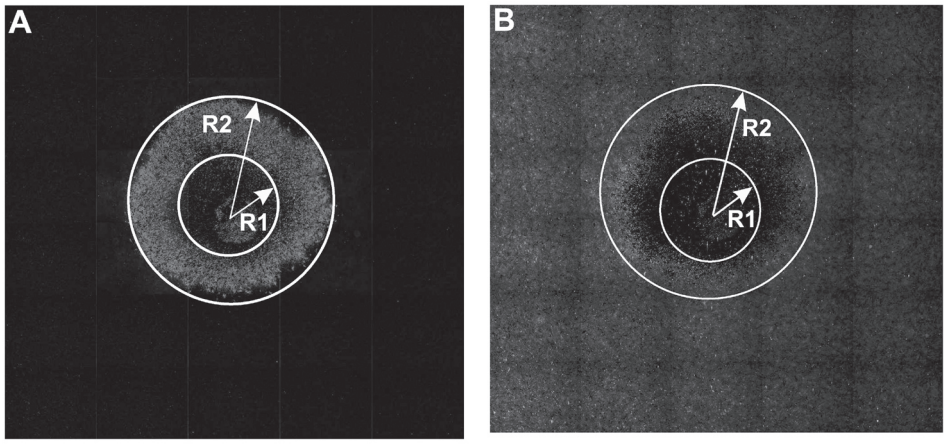


Figure 2. (A) is the denoised viral image (VSV-G) at 36 hours; (B) is the host image (NOS1) at 36 hours. Due to the overlap of the fluorescent markers, the virus signal leaks into the host image. A correlation-based method uses the dead zone (R1) and the extent of the viral spread (R2) to generate a spectral overlap corrected host image.

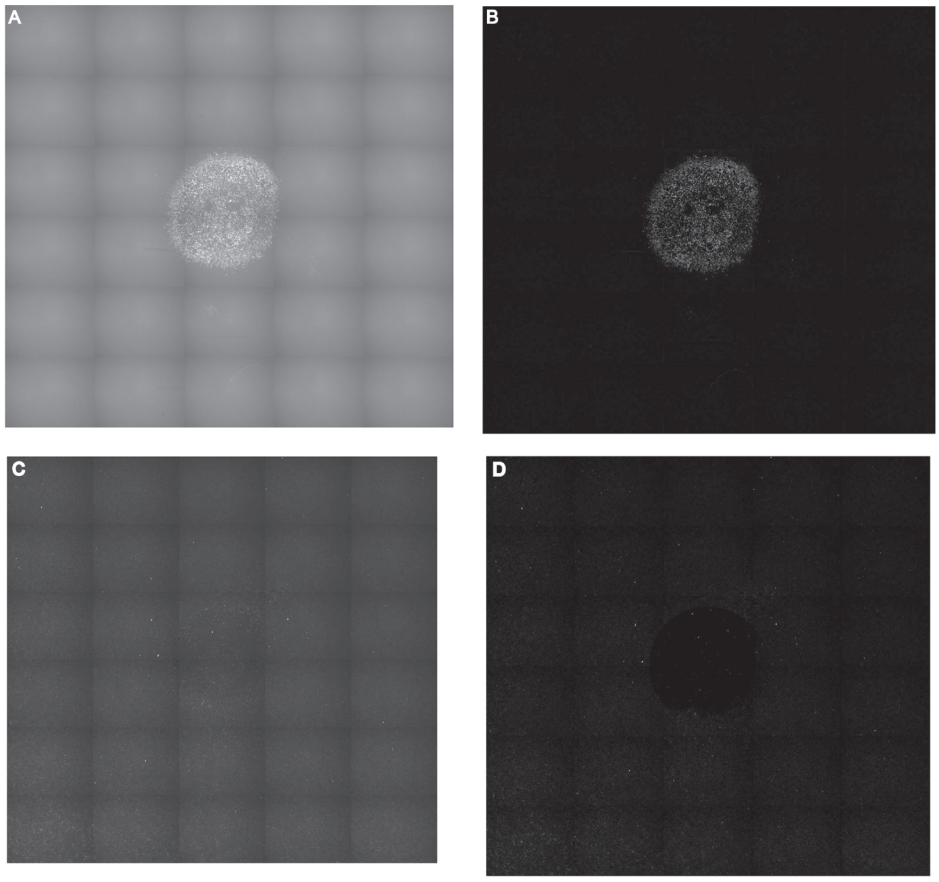


Figure 3. Fluorescent micrographs (A) - (D) show the noisy and corrected viral (VSV-G) and host (NOS3) images at 18hrs. (A) Noisy viral image (B) Corrected viral image (C) Noisy host image (D) Corrected host image. Note that the non-uniform background shading artifact in each SI as well as the montage artifact in the MI present in (A) and (C) have been removed by the denoising process.

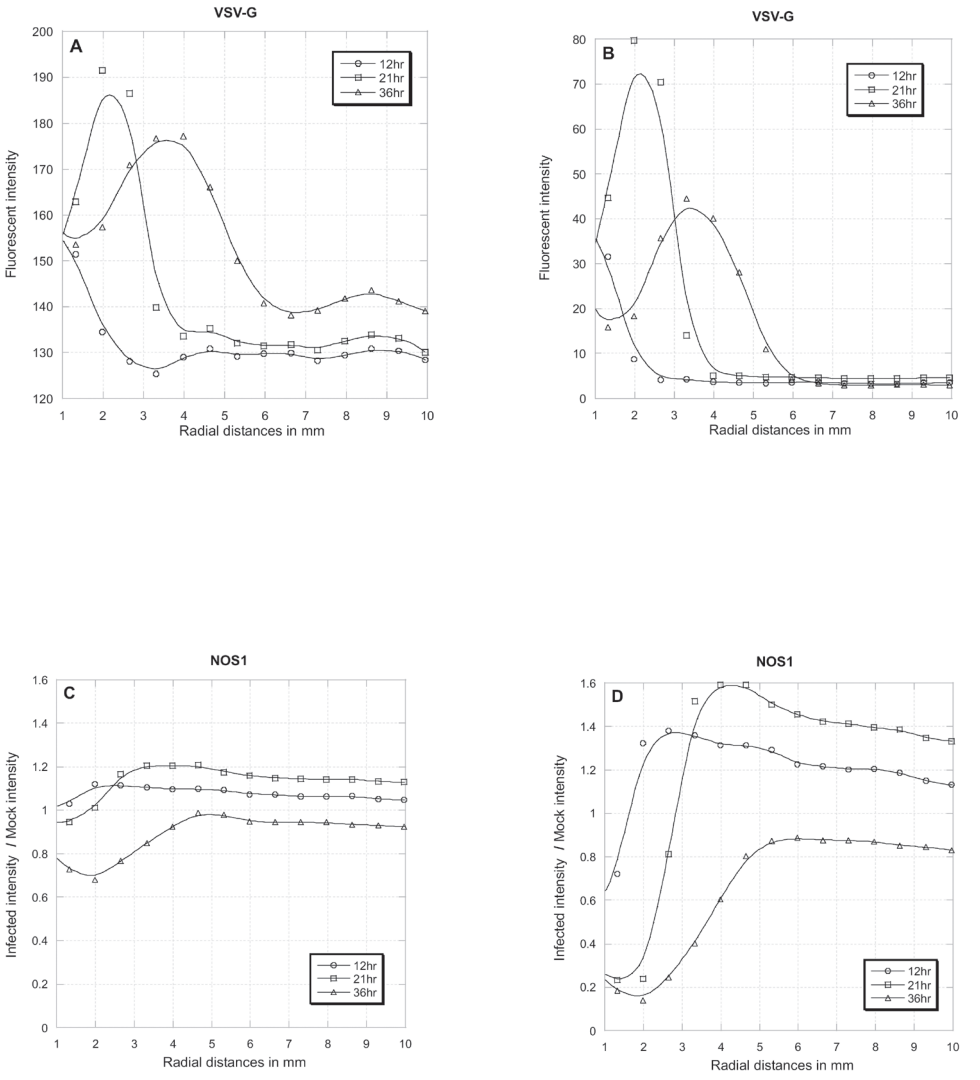


Figure 4. Continued on the next page.

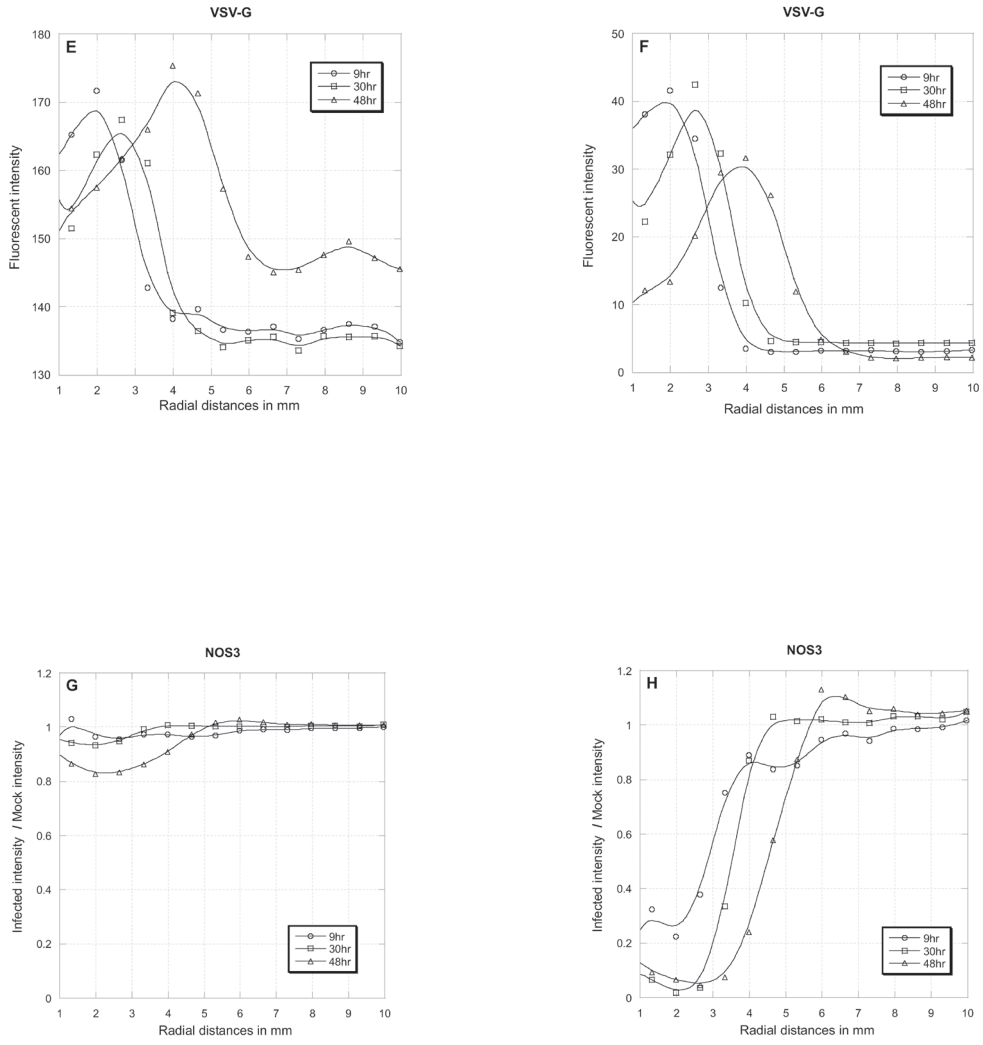


Figure 4. IIS for time points (12hr, 21hr and 36hr): (A) noisy VSV-G (B) corrected VSV-G IIS (C) noisy NOS1. (D) corrected NOS1. IIS for time points (9hr, 30hr and 48hr): (E) noisy VSV-G (F) corrected VSV-G IIS (G) noisy NOS3. (H) corrected NOS3.

References

Duca KA, Lam V, Keren I, Endler EE, Letchworth G J, Novella I S, and Yin J (2001) Quantifying Viral Propagation *in vitro*: Toward a Method for Characterization of Complex Phenotypes. *Biotechnol Prog* **17**:1156–1165

Gerlich D and Ellenberg J (2003) 4D Imaging to Assay Complex Dynamics in Live Specimen. *Nat Cell Biol Supplement* **5**: S14–19

Gerlich D, Mattes J, and Eils R (2003) Quantitative Motion Analysis and Visualization of Cellular Structures. *Methods* **29**: 3–13

Komatsu T, et al (1999) Mechanisms of Cytokine-mediated Inhibition of Viral Replication. *Virology* **259**: 334–341

Lippincott-Schwartz J, Snapp E, and Kenworthy A (2001) Stuyding Protein Dynamics in Living Cells. *Nat Rev Mol Cell Biol* **2**: 444–456

Rout S, Lam V, Bell AE, and Duca KA (2004) Epi-fluorescent Image Modeling for Viral Infection Analysis. Proceedings of the 38th Asilomar Conference on Signals, Systems and Computers. 1841–1846

Sled JG, Zijdenbos AP, and Evans AC (1998) A Nonparametric Method for Automatic Correction of Intensity Nonuniformity in MRI Data. *IEEE Trans Med Imag* **17**: 87–97

Swedlow J R (2003) Quantitative Fluorescence Microscopy and Image Deconvolution. Methods Cell Biol **72**: 349–367

Tomazevic D, Likar B, and Pernus F (2002) Comparative Evaluation of Retrospective Shading Correction Methods. *J Microsc* **208**: 212–223

Publications

Bell AE, Anderson-Cook C, and Spencer SJ (2004) Stereotype Threat in the Engineering Classroom. *Proceedings of the ASEE Annual Conference and Exposition*, Salt Lake City, Utah

Barua S, Carletta JE, Kotteri KA, and Bell AE (2005) An Efficient Architecture for Lifting-based Two-Dimensional Discrete Wavelet Transforms. *Integration: the VLSI Journal* **38**: 341–352

Barua S, Kotteri KA, Bell AE, and Carletta JE (2004) Optimal, Quantized Lifting Coefficients for the Biorthogonal 9/7 Wavelet. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* pp. 193–196. Montreal, Canada

Barua S, Kotteri KA, Carletta JE, and Bell AE (2004) An Efficient Architecture for Lifting-based Two-dimensional Discrete Wavelet Transforms. *Proceedings of the Great Lakes Symposium on VLSI*, Boston, MA

Kotteri KA, Barua S, Bell AE, and Carletta JE (2004) Quantized FIR Filter Design: A Collaborative Project for Digital Signal Processing and Digital Design Courses. *Proceedings of the ASEE Annual Conference and Exposition*, Salt Lake City, Utah

Kotteri KA, Bell AE, and Carletta JE (2004) Polyphase Structures for Multiplierless Biorthogonal Filter Banks. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* pp. 197–200 Montreal, Canada

Lyons R and Bell AE (2004) The Swiss Army Knife of Digital Networks. *IEEE Signal Processing Magazine* **21**: 90–100

Rout S and Bell AE (2004) Narrowing the Performance Gap Between Biorthogonal and Orthogonal Wavelets. *Proceedings of the 38th Asilomar Conference on Signals, Systems and Computers* pp. 1757–1761. Pacific Grove, California

Rout S, Lam V, Bell AE, and Duca KA A Comprehensive Correction Method for Low Magnification, Montaged Images: Application to Host-Virus Interactions. *Accepted for publication in Cytometry*, date pending

Rout S, Lam V, Bell AE, and Duca KA (2004) Epi-fluorescent Microscope Image Modeling for Viral Propagation Analysis. *Proceedings of the 38th Asilomar Conference on Signals, Systems and Computers* pp. 1841–1846. Pacific Grove, California

Varma K and Bell AE (2004) Improving JPEG2000's Perceptual Performance with Weights Based on Contrast Sensitivity and Standard Deviation. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* pp. 665–668. Montreal, Canada

Varma K and Bell AE (2004) JPEG2000—Choices and Tradeoffs for Encoders. *IEEE Signal Processing Magazine* **21**: 70–75

Robust and Scalable Comparative Whole-Genome Functional Annotation Systems

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

Corban Rivera, Rob St. Clair, and Shomir Wilson

Introduction

The genomes of more than 150 organisms have been fully sequenced (Bernal et al., 2001). About 25 percent of the genes in these genomes have no known function since they share no sequence or structural similarity to any annotated genes, while another 50 percent have poor annotations (Enright et al., 2003). For example, 60 percent of the genes in *Plasmodium falciparum* are "hypothetical" (Gardner et al., 2002). Discovering the functions of these genes will provide critical insights into the biology of many organisms. These insights can lead to an improved understanding of the causes of human diseases; accelerate the discovery of drugs that target human, animal, or plant pathogens; explain how plants adapt to environmental stresses; and shed light on how some micro-organisms live in extreme conditions such as heat, high acidity, and radiation. The focus of our research is the development of systematic algorithms that can integrate diverse types of functional genomic data to produce testable, quantified predictions of the functions of poorly understood genes.

In the absence of sequence or structural similarity to genes of known function, a promising approach for predicting gene function identifies functional associations of genes of unknown function with those of known function. We can deduce such associations from diverse sources of evidence. For instance, evolutionary pressure dictates that pairs of genes that function in concert often co-evolve, appear together, or exhibit spatial proximity in the genome (Marcotte, 2000). The complete genome sequences of more than 150 organisms are a rich resource for these types of functional linkages. In contrast, genome-scale functional genomics experiments such as gene or protein expression profiling (Brown and Botstein, 1999; MacBeath, 2002) and screens for detecting interactions between proteins (Fields and Song, 1989), proteins and DNA (Lee et al., 2002), and proteins and metabolites provide species-specific data for such association approaches.

A *functional linkage network* (FLN) (Karaoz et al., 2004; Marcotte et al., 1999) is a powerful framework for representing and analysing such functional relationships between genes. An FLN is a graph in which each node corresponds to a gene, and an edge connects two genes if some experimental or computational procedure suggests that these genes might share the same function. Each edge in the network has a real-valued weight between 0 and 1. The larger the weight, the more evidence we have that the two genes share the same function. The edge usually does not provide information on which specific functional annotation the genes share.

Many existing databases have assembled large collections of functional links between genes by curating the literature or by combining multiple experimental and computational procedures (Bader et al., 2003; Bowers et al., 2004; Stark and Tyers, 2003; von Mering et al., 2005). Other authors have proposed novel techniques for constructing FLNs that integrate multiple sources of data (Lee et al., 2004) or construct FLNs that are based on gene expression

data analysed across multiple species (Bergmann et al., 2003; Stuart et al., 2003; von Mering et al., 2005). Although these databases and algorithms are valuable sources of functional associations, they do not appear to address the issue that an edge in an FLN does not indicate *which* function the connected genes share. Therefore, in order to harness the power of FLNs for functional annotation, we need to develop techniques that provide 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.

We have developed a computational technique called "Gene Annotation using Integrated Networks" (GAIN) (Karaoz et al., 2004) that addresses many of these issues. GAIN automatically and robustly suggests putative functions for genes of unknown function. 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. One of the attractive features of this algorithm is that we can represent the flow of information in the FLN as a directed graph and provide visualizations of this graph to biologists. 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) and in *Drosophila melanogaster* (results not yet published).

Over the last year, the main focus of our research has been to extend GAIN to harness information flowing from multiple species simultaneously. Collaborations with VBI scientists aim to provide precise functional annotations for many organisms including *S. cerevisiae*, *Arabidopsis thaliana*, the malaria-causing pathogen *P. falciparum*, and plant pathogens such as *Phytophthora sojae*. The statistical significance and cross-validation results that accompany our predictions will help biologists to formulate precise and prioritized experiments that will validate our predictions. In turn, these new experiments will provide new datasets for integration and analysis by GAIN. These mutually reinforcing cycles of experiments and computational analysis promise to result in the validated functional annotation of a large number of microbial genes.

Description of the GAIN Algorithm

We represent the FLN for an organism as an undirected graph $G = (V, E)$, where V is the set of nodes and E is the set of edges. Each node in V corresponds to a gene in the FLN and each edge in E corresponds to a functional association between two genes in the FLN. For an edge connecting the nodes u and v in V, we use w_{uv} to denote the weight of the edge. All edge weights lie in the range [0, 1]. Our algorithm is reminiscent of discrete-state Hopfield networks (Hopfield and Tank, 1986) and Sequential Dynamical Systems (Mortveit and Reidys, 2001). The input to our algorithm is the graph G and a function f a function of interest, for instance, a function in the Gene Ontology (Ashburner et al., 2000). We repeatedly invoke our algorithm for each function f of interest. With each node v in V, the algorithm associates a *state* variable, s_v, defined as follows: s_v=1, if v is annotated with f, s_v = -1 if v is annotated with a function other than f, and s_v = 0, otherwise. The nodes with state 0 correspond to the genes of unknown function. The goal of our procedure is to assign a state of 1 or –1 to each node with state 0. The output of the procedure is set of nodes with initial state 0 to which it assigns a state of 1. We predict that the corresponding genes have the function f.

Intuitively, we would like to assign states to these nodes in such a way that the two nodes connected by an edge receive the same functional assignment; we say that such an edge

is *consistent*. Since it is not always possible to ensure that every edge is consistent, it is desirable to compute a maximally consistent state assignment (one in which the weighted sum of consistent edges is as large as possible). We can achieve such a maximally consistent assignment by minimizing the following *energy* function: $E=-1/2\sum_{(u,v)\in E} w_{uv}s_u s_v$. In this equation, a consistent edge makes a positive contribution to E, while an inconsistent edge makes a negative contribution to E. Therefore, minimising E maximizes the weighted sum of consistent edges. The problem of computing maximally consistent assignments is computationally intractable (Kasif et al., 1993). Given this difficulty, we employ the following local search procedure. We do not change the state of the nodes in initial state 1 or -1; these nodes correspond to genes with known annotations. We iteratively process the nodes in initial state 0. In each iteration, for each node v, the algorithm sets $s_v = sgn(\sum_{u\in N_v} w_{uv}s_u - \theta)$ where N_v is the set of neighbors of v in G and q is an *activation threshold*. The right hand side of this equation computes the weighted sum of the states of the neighbors of node v and compares this sum to q; if the sum is greater than q, then node v's state is set to 1, otherwise to -1. Iterative application of this rule achieves a more globally consistent state assignment for all the nodes in the network than a single application of this rule. It can be shown that using this update function monotonically changes the value of E, guaranteeing the convergence of the procedure. Moreover, this final value of E is guaranteed to be at least half the optimal solution (Kasif et al., 1993). In practice, we have noted that our networks converge within two or three iterations over all the nodes.

Ongoing Research

We now describe the collaborative projects we have been working on since September 2004. The success of these projects hinges upon the construction of FLNs that integrate diverse types of biological clues about shared function. We exploit two primary sources of information about shared function: (i) information on co-evolution of non-homologous genes that is encoded in genomic context across all sequenced genomes, and (ii) species-specific functional genomic information such as gene expression measurements, protein-protein interaction networks, or transcriptional regulatory networks. By combining different types of data from multiple species and supplementing with species-specific information, when available, we are able to generate more accurate predictions of gene function than information from a single species or a single type of data.

Functional Annotation of All Sequenced Microbes

In collaboration with Allan Dickerman and Brett Tyler, we are exploring the use of GAIN to simultaneously annotate the genomes of all sequenced microbes of interest. Each node in the FLN is a gene in one of the genomes. Edges connect genes that have similar sequences, as well as non-homologous genes. We have developed a novel technique for constructing cross-microbe FLNs. This method defines clusters of genes with similar sequences and connects genes in different clusters using evidence based on co-evolution. We start by constructing a similarity graph where a gene in an organism is connected to the most similar gene in each of the other organisms. We use a greedy algorithm to extract dense subgraphs from the similarity graph; each dense subgraph has the property that each gene in it is connected to at least half the other genes in the subgraph. We include all the dense subgraphs in the final FLN. Next, we connect pairs of genes in different subgraphs by asking if the genomic neighbors of these genes are similar. We map this question to an appropriate measure of the similarities of the dense subgraphs that the neighbours belong to. This approach naturally subsumes chromosomal proximity and gene fusion. We are currently applying GAIN to this FLN. We will post the results of this study on a publicly-available website for the biological community to evaluate.

Study of Oxidative Stress Response in Model Microbes

In collaboration with Allan Dickerman, Dharmendar Rathore, and Brett Tyler, we are using GAIN to annotate model organisms such as *A. thaliana* and *S. cerevisiae* using gene expression profiles collected by Tyler's lab on exposing these organisms to the oxidizing agent cumene hydroperoxide. We construct separate FLNs for each organism from this gene expression data and molecular interaction networks. We interconnect the FLNs for each organism using gene sequence and protein structure similarity data. Our system operates on all FLNs simultaneously to provide putative annotations for genes in all these organisms. An advantage of constructing FLNs for each organism, as opposed to simply transferring information from one organism to another, is that the FLN in the target organism can reinforce the evidence propagated from the FLN in the source organism.

It is natural to model a protein-protein interaction or a protein-DNA interaction as an edge in the FLN. We construct FLNs from gene expression profiles using the following methods: (i) We connect two genes by an edge if the similarity of their expression profiles is greater than a threshold. We employ permutation-based random sampling techniques to estimate statistically significant thresholds for these similarity measures. (ii) We also construct edges based on bi-clustering algorithms. In a set of molecular profiles, a bi-cluster is a set of genes (or proteins or metabolites) that cluster only under a subset of the conditions (Tanay et al., 2004). Thus bi-clusters capture associations in a subset of the data that standard clustering algorithms are prone to miss. By disregarding irrelevant conditions, bi-clusters naturally contain genes with higher correlation and potentially higher statistical significance. We use the xMotif software tool developed by our group for finding bi-clusters (Murali and Kasif, 2003; Wu et al., 2004). An edge connects two genes if they belong to the same bi-cluster; the weight of the edge is the correlation between the expression profiles of the two genes, computed only for the conditions that belong to the bi-cluster. We will use our experience with these well-studied organisms to extend this approach to *P. falciparum* and *P. sojae*. Oxidative stress is the mechanism by which many anti-malarial drugs attack *P. falciparum*. For *P. falciparum*, we will use publicly-available gene expression and proteomic data sets. We will construct an FLN for *P. sojae f*rom DNA microarray data collected during infection of soybean plants.

In addition to these specific case studies, we have made the GAIN system freely available to all researchers (http://bioinformatics.cs.vt.edu/~murali/software/gain). Life scientists can use GAIN to analyze the genomes of various organisms using functional genomic data collected in their laboratories, obtain predicted annotations for these genes, and prioritise the predictions for experimental validation. We have also created "Functional Annotation Search Tool" (FAST), a prototype web resource that enables biologists to access the predictions, assess the plausibility of the results, and design targeted experiments to validate interesting predictions. We are currently testing this web resource with our collaborators.

Future Research

Our collaborations with life scientists to use GAIN to annotate a large number of genomes have pointed out several interesting directions for extending GAIN. The FLNs we construct can contain 10,000-100,000 nodes and edges numbering in the millions, especially when we study a large number of genomes. We are examining three approaches for developing efficient algorithms that can analyse such large FLNs: (i) Exploit the hierarchical structure of the Gene Ontology; the relationships between the functions in GO form a directed acyclic graph (DAG). By definition, if a function *f* annotates a gene, then all the ancestors of *f* in GO also annotate the gene. Therefore, we may be able to use the results obtained by applying GAIN to the FLN of a particular function as a starting point for applying the GAIN to the FLN of a child of that function. (ii) Construct a sparse FLN that contains a considerably smaller

number of nodes and edges than the original FLN. The sparse FLN should have the property that we can use the results of applying GAIN on it as an estimate of the results of applying GAIN to the original FLN. (iii) GAIN currently constructs a separate FLN for each function. We are examining extensions to GAIN that can potentially improve the quality of the annotations by incorporating correlations and dependencies between functions into account.

For a biologist to assess the plausibility of the predictions made by GAIN, we must provide multiple measures of confidence in the predictions. The GAIN software already includes the functionality to perform leave-one-out and *k*-fold cross-validation. However, the current system treats GAIN's performance for a function as being independent from GAIN's performance for the parent of the function. We are devising techniques that address this issue. An important aspect of functional annotation is assessing the statistical significance of putative functional assignments. The value measures the possibility that we may make a particular functional assignment to a particular gene even when presented with a "random'' functional linkage network. We will endow our system with the ability to compute a *p*-value for each putative functional annotation by running GAIN on several randomised versions of the FLN.

We have also identified several intriguing organisms to apply GAIN to. In collaboration with Biswarup Mukhopadhyay, we will study the closely related organisms *Methanocaldococcus jannaschii*, *Methanothermoccoccus lithotrophicus*, and *Methanococcus maripaludis* that grow optimally at 85°C, 65°C, and 37°C, respectively. GAIN will provide numerous insights into the comparative environmental biology of these organisms and suggest mechanisms by which they adapt to extreme conditions. We are also in discussions with the PATRIC group at VBI on the possibility of including an implementation of GAIN in PATRIC's computational pipelines.

In the long term, we envision that our research on cross-organism and cross-kingdom functional annotation also points to meta-websites that use rapidly developing web service and GRID technology to query organism-specific data sources and generate functional predictions by integrating this information. In particular, we believe that researchers will be able to use GAIN to functionally annotate a genome as soon as it is sequenced. Finally, our research may serve as a suitable model for the more daunting task of annotating the functions of complex eukaryotic genomes, such as the human genome.

Acknowledgments

We thank Dianjing Guo and Eric Nordberg for their significant contributions to data analysis and algorithm development and implementation in our collaborative projects.

References

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, and Sherlock G (2000) Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet* **25**: 25–29

Bader GD, Betel D, and Hogue CWV (2003) BIND: the biomolecular interaction network database. *Nucleic Acids Res* **31**: 248–250

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

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

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

Bowers P, Pellegrini M, Thompson M, Fierro J, Yeates T, and Eisenberg D (2004) Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol* **5**: R35

Brown PO and Botstein D (1999) Exploring the new world of the genome with dna microarrays. *Nat Genet* **21**: 33–37

Enright A, Kunin V, and Ouzounis C (2003) Protein families and TRIBES in genome sequence space. *Nucleic Acids Res* **31**: 4632–8

Fields, S and Song, O (1989). A novel genetic system to detect protein-protein interactions. *Nature* **340**: 245–246

Gardner MJ et al (2002) Genome sequence of the human malaria parasite plasmodium falciparum. *Nature* **419**: 498–511

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

Karaoz U, Murali TM, Letovsky S, Zheng Y, Ding C, Cantor CR, and Kasif S (2004) Whole genome annotation using evidence integration in functional linkagenetworks pp. 2888–2893. *Proceedings of the National Academy of Sciences*

Kasif S, Banerjee S, Delcher A L, and Sullivan G (1993) Some results on the complexity of symmetric connectionist networks. *Ann Math Artif Intell* **9**: 327–344

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

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne J-B, Volkert TL, Fraenkel E, Gifford DK, and Young RA (2002) Transcriptional Regulatory Networks in Saccharomyces cerevisiae. *Science* 298(5594): 799–804

MacBeath G (2002) Protein microarrays and proteomics. *Nat Genet* **32** Suppl: 526–532

Marcotte EM (2000) Computational genetics: finding protein function by nonhomology methods. *Curr Opin Struct Biol* **10**: 359–365

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

Mortveit HS and Reidys CM (2001) Discrete, sequential dynamical systems. *Discrete Math* **226**: 281–295

Murali TM and Kasif S (2003) Extracting conserved gene expression motifs from gene expression data pp. 77–88. In *Proceedings of the Pacific Symposium on Biocomputing*

Stark C and Tyers M (2003) The GRID: the general repository for interaction datasets. *Genome Biol* **4**: R23

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

Tanay A, Sharan R, and Shamir R (2004) *Handbook of Bioinformatics*, chapter Biclustering Algorithms: A Survey.

von Mering C, Jensen L, Snel B, Hooper S, Krupp M, Foglierini M, Jouffre N, Huynen M, and Bork P (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res* **33** Database Issue:D433–7

Wu C-J, Fu Y, Murali TM, and Kasif S (2004) Gene expression module discovery using gibbs sampling. *Genome Informatics*, **15**: 239–248

Bio-Microfluidics Modeling

Joseph Wang
Faculty Fellow, VBI
Associate Professor of Aerospace and Ocean Engineering, Virginia Tech
jowang@vt.edu

Yong Cao

Motivations and Objectives

There have been significant research efforts in recent years to develop bio-analytical microfluidic devices for sorting, separation, or manipulations of particles of biological origin, such as cells, virus, proteins, and DNA. The most critical issue in the design of a bio-analytical microfluidic device is the control of bio-particle motion in fluids. However, understanding biofluidics in micro-scale devices represents a significant challenge. In micro-scale flow regimes, electrokinetic forces, energy dissipation, and surface interaction often dominate the fluid dynamics aspect. The bio-particles are of irregular shape and typically have a characteristic length ranging from nano-meter to micro-meter. The physical process governing the dynamics of the bio-particles is typically determined by the interactions between particles, fluids, solid surfaces, and external electric fields. Due to the intrinsic complex nature of the problem and a lack of effective experimental means to diagnose the detailed processes in microchips, questions on many fundamental issues related to bio-microfluidics still remain unanswered. As a result, the design process often lacks a clear, physics based guidance. The objective of this research is to develop a first-principle based numerical model to understand the dynamic of bio-particles in electrofluidics.

Introduction

The research reported here concerns a particular type of bio-analytical device, the dielectrophoresis (DEP) based device. Dielectrophoresis is the motion of neutral matter caused by polarization effects in a non-uniform electric field (Pohl, 1978). In recent years, DEP has been broadly used in experiments for separations or manipulations of micrometer-sized to nanometer-sized bio-particles. Some recent experimental studies include the following: Green and Morgan (1997) showed for the first time that it is possible to separate a population of nanoparticles (93 nm diameter latex beads) into two subpopulations solely on the basis of their dielectric properties by using nanofabricated electrode arrays; Gascoyne and Wang (1997) were able to separate human breast cancer cells from blood using AC electric field DEP separation; Yang et al (2000) were able to separate the four major leukocyte subtypes by their sizes, densities, and membrane properties using dielectrophoretic field-flow-fractionation (DEP-FFF); Washizu (2003) demonstrated that the electrostatic force can be used for stretching DNA and for manipulation, dissection and acquisition of targeted location of DNA; Zhang et al (2004) used dielectrophoresis to position the DNA and proteins at well-defined positions on a chip; Chou et al (2002) used electrodeless DEP to trap and concentrate single- and double-stranded DNA and showed strong dielectrophoretic response of the DNA in the audio frequency range. However, quantitative understanding of bio-particle handling and manipulations still significantly lag behind the experimental studies.

In a bio-electrofluidics system, the fluid produces a drag force, as well as forces due to buoyancy on the particles (and vice versa). The external electric field induces the dielectrophretic (DEP) force and electrophoretic (EP) force on the bio-particles. These forces combined with inter-atom interaction force determine translational

and rotational motion, as well as deformation of the particle. The motion and deformation of bio-particles in turn also affect the dynamics of the fluid and the distributions of the electric double layer, which in turn affect the forces experienced by bio-particles. It is through this interaction that the continuum fluid and discrete molecule exchange their momentum and energy.

Some recent modeling studies of DEP include Kadaksham et al.(2004); Schnelle et al (1999); Castellanos et al (2003); Nedelcu et al (2004); Snyder et al (2001); and Green et al (2002). All previous modeling studies make the following two common simplifications. First, the particles were always modeled as hard spherical particles, which is obviously an over-simplification for bio-particles. A direct consequence from this assumption is that only translational motions under the DEP force were included. Second, the drag force from the fluid to particle was calculated for that of a continuum flow at the particle-fluid interface. This assumption is typically valid for micrometer sized particles, but will become invalid for nano-particles.

Accomplishments

A new hybrid fluid-particle simulation model has been developed. This simulation model combines computational fluid dynamics with computer particle simulations and solves the Navie-Stokes equation for fluid dynamics, Newton's second law for particle dynamics, and the Poisson's equation for the electric field. The simulation model also advances the current state-of-the-art in the following two aspects: a) bio-particles are modeled as elliptical particles rather than spherical particles; and b) the calculation of the fluid drag force on particles is extended to include both the continuum and slip condition at fluid-particle interface.

Due to page limits, this report only discusses the effects of particle shape on particle dynamics in a DEP device. Results on particle-fluid interactions are not included here. A more complete description of this research can be found in (Cao and Wang, 2005).

Model Outline

A bio-particle in a DEP device may experience the following forces: hydrodynamic force (drag force), buoyancy force, particle-particle interaction force, and dielectrophretic force. In this paper, we only consider the DEP force. The numerical simulation model solves the Poisson's equation for the electric field, and Newton's second law for particle dynamics.

The translational particle motion is solved from:

$$\dot{\vec{\mathbf{v}}}_0 = \vec{\mathbf{F}} / m$$

If (x, y, z) is the reference frame centered on the particle center of mass, the equations for rotational motion may be written as:

$$\sum \vec{\mathbf{T}}_c = \left(\dot{\vec{\mathbf{H}}}_c\right)_{xyz} + \vec{\omega} \times \vec{\mathbf{H}}_c$$

where $\left(\dot{\vec{\mathbf{H}}}\right)_{xyz}$ is the time rate of change of angular momentum, measured from the (x, y, z) reference. The particle dynamics are solved using the discrete element method (DEM) (Tsuji et al, 1993; Xu et al, 1997; Limtrakul et al, 2003; and Cao 2003).

When a homogeneous ellipsoid with isotropic dielectric permittivity ε_2 is immersed in a dielectric fluid of permittivity ε_1, its effective moment can be expressed as (Jones, 1995):

$$\mathbf{p}_{\text{eff}} = \frac{4\pi abc}{3}(\varepsilon_2 - \varepsilon_1)\mathbf{E}_{\text{e}}^*$$

$$\mathbf{E}_{\text{e}}^* = \mathbf{i}_{\text{x}} \frac{E_{e,x}}{1 + (\frac{\varepsilon_2 - \varepsilon_1}{\varepsilon_1})L_x} + \mathbf{i}_{\text{y}} \frac{E_{e,y}}{1 + (\frac{\varepsilon_2 - \varepsilon_1}{\varepsilon_1})L_y} + \mathbf{i}_{\text{z}} \frac{E_{e,z}}{1 + (\frac{\varepsilon_2 - \varepsilon_1}{\varepsilon_1})L_z}$$

where L_x is defined by an elliptical integral:

$$L_x = \frac{abc}{2}\int_0^\infty \frac{ds}{(s+a^2)R_s}$$

and $R_s \equiv \sqrt{(s+a^2)(s+b^2)(s+c^2)}$. Similar expressions may be obtained for L_y and L_z by appropriate substitutions of y (or z) for x and b (or c) for a, respectively. For an ideal elliptical particle with no energy dissipation, the dielectrophoretic (DEP) force is:

$$\mathbf{F}_{\mathrm{DEP}} = \frac{4\pi abc}{3}(\varepsilon_2 - \varepsilon_1)\mathbf{E}_e^* \nabla \mathbf{E}_e$$

and the torque on the elliptical particle is:

$$\mathbf{T} = \frac{4\pi abc}{3}(\varepsilon_2 - \varepsilon_1)\mathbf{E}_e^* \times \mathbf{E}_e$$

A variety of energy dissipation mechanisms such as conducting and dielectric relaxation may influence the behavior of real dielectric particles (i.e., a lossy particle) in an electric field. For lossy dielectric ellipsoids, the effective moment will be modified by the Clausius-Mossotti function.

The electric field is obtained by solving the electric potential ϕ from

$$\nabla^2\phi = -\frac{\rho_e}{\varepsilon}$$

The finite element method is used to solve the electric potential.

Figure 1 shows the model setup for the device along with boundary conditions for the electric field. We consider a 2-dimensional problem where the device is assumed to be infinitively long in the z direction (vertical to the paper). A 140μm×100μm rectangular domain is simulated. The length of electrode is 100μm. The electrode is has an applied potential of 10V. Figure 1 also shows the distribution of potential and the vector **E** field.

Results and Discussions

In the simulations presented here, the relative permittivity of particle and the suspending medium is taken to be 2.55 and 78, respectively. The mass density of particle is taken to be 1050kg/m^3. The lateral boundary is periodic boundary condition for particles. We consider only the DEP force acting on particle. The particle surface is described by $x^2 / a^2 + y^2 / b^2 + z^2 / c^2 = 1$, with a =2μm, b = 1μm, and c = 1μm.

The first set of simulation results concerns the effect of particle orientation on particle trajectories under the DEP force. Figure 2 shows the trajectories of a set of particles with the same initial position but different initial orientation. In Figure 2, the initial velocity of particle is taken to be zero, the initial position is taken to be at x=65 μm, y=50 μm, and the initial angle between the major axis of the ellipse and the x axis, α, is taken to be $i\pi/12$, i= 0, 1, …., 11. The second set of simulation results concerns the effect of particle orientation on particle trajectories under the DEP force. In Figures 3 and 4, we compare the trajectories of elliptical particles with that of a spherical particle for two different initial positions. The results show that the dynamics

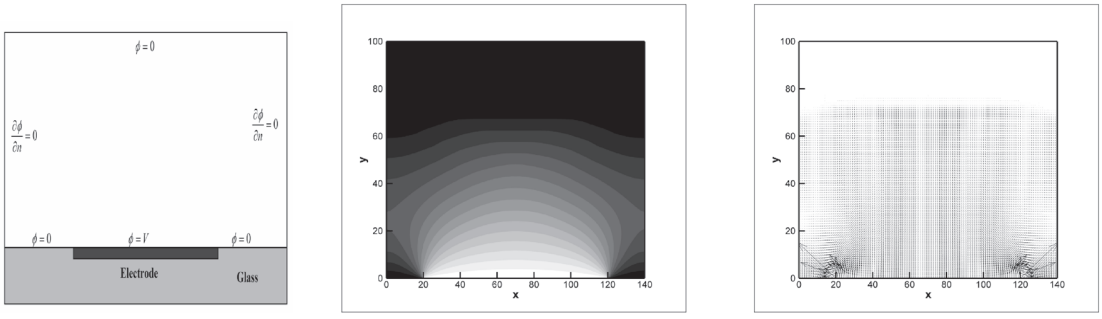


Figure 1. Model setup (left), initial electric potential (middle), and initial electric field (right).

of an elliptical particle can be very different from that of a spherical particle in a DEP device as particle trajectories are sensitively dependent upon particle orientation, and particle shape.

Conclusions and Future Plan

An improved DEP simulation model, which uses elliptical shaped particles to better represent bio-particles, is presented here for bio-microfluidics applications. Initial simulation results suggest that the DEP force on a particle sensitively depends upon the particle's shape and initial orientation. A hybrid fluid-particle simulation model has also been developed which combines computational fluid dynamics solutions with the particle dynamics solutions described in this paper. Detailed bio-microfluidics simulations will be performed to support the design of micro-fluidics devices at VBI.

Acknowledgements

This research is supported by the Virginia Bioinformatics Institute under the VBI College of Engineering Faculty Fellows program.

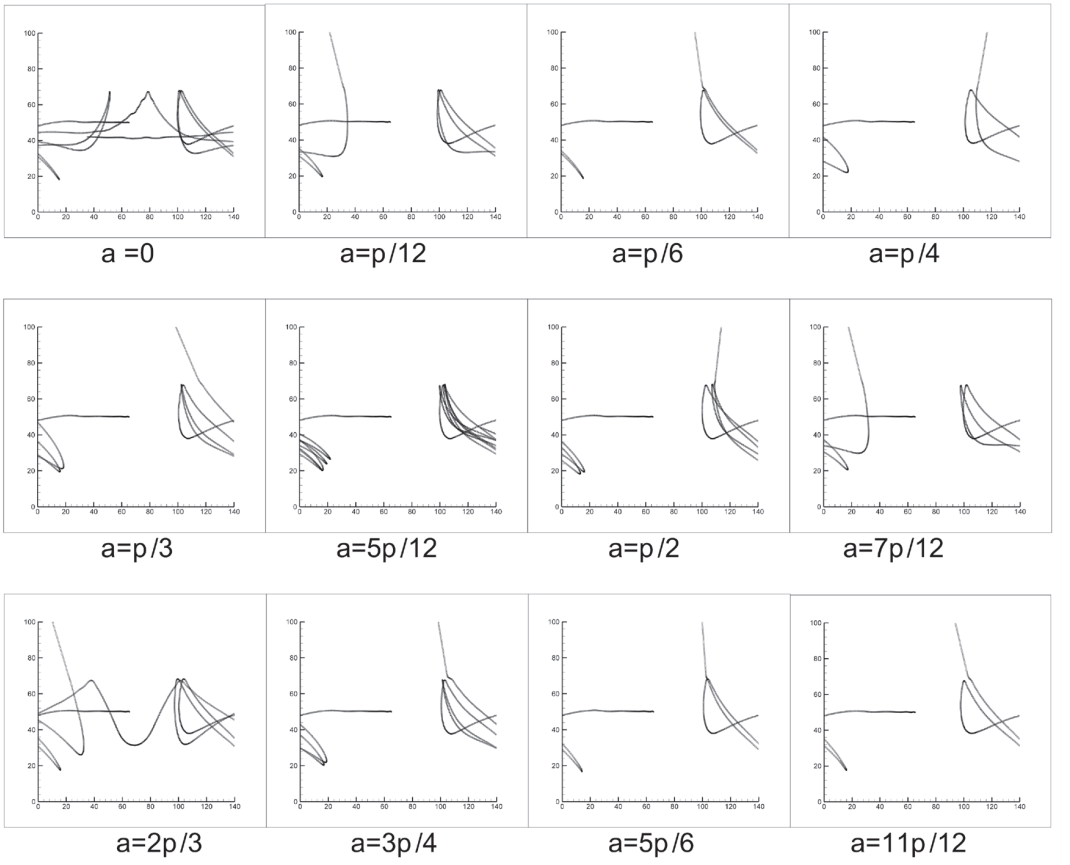


Figure 2. Particle trajectories under DEP force. Initial position at (65, 50)

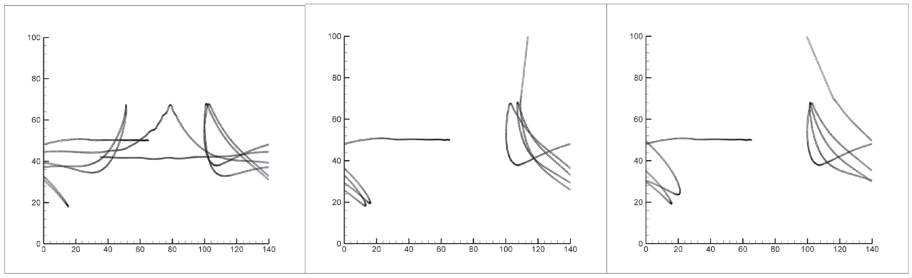


Figure 3. Particle trajectories under DEP force. Initial position at (65, 50)

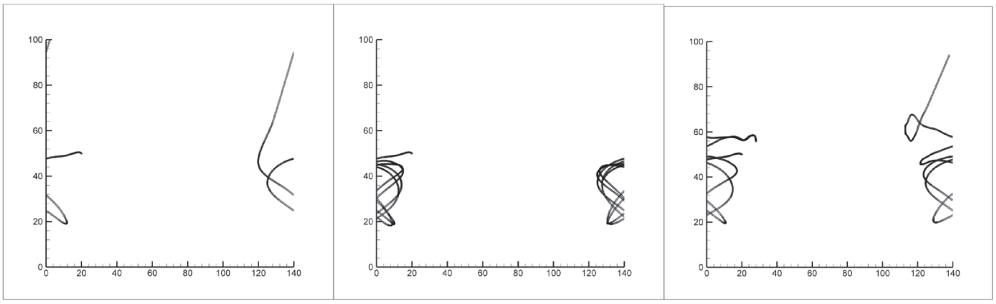


Figure 4. Particle trajectories under DEP force. Initial position at (20, 50)

References

Castellanos A, Ramos A, Gonzalez A, Green NG, and Moprgan H (2003) Electrohydrodynamics and dielectrophoresis in Microsystems: scaling laws. *J Phys D: Appl Phys* **36**: 2584–2597

Cao Y and Wang J (2005) Modeling micro- and nano-scale particle dynamics in micro-channel flows. submitted to *Dynamics of Continuous, Discrete, and Impulsive Systems*

Chou C, Tegenfeldt JO, Bakajin O, Chan SS, and Cox EC (2002) Electrodeless Dielectrophoresis of Single- and Double-Stranded DNA. *Biophys J* **83**: 2170–2179

Gascoyne P and Wang X (1997) Dielectrophoretic separation of cancer cells from blood. *IEEE Trans Indus Appl* **33**: 670–678

Green N, and Morgan H (1997) Dielectrophoretic separation of nano-particles. *J Phys D: Appl Phys* **30**: L41–L48

Green NG, Ramos A, and Morgan H (2002) Numerical solution of the dielectrophoretic and traveling wave forces for interdigitated electrode arrays using the finite element method. *J Electrostat* **56**: 235–254

Jones TB (1995) Electromechanics of Particles. Cambridge University Press, New York City, NY

Kadaksham J, Singh P, and Aubry N (2004) Dynamics of electrorheological suspensions subjected to spatially nonuniform electric fields. *J Fluid Eng, Trans ASME* **126**: 170–179

Limtrakul S, Chalermwattanatai A, Unggurawirote K, Tsji Y, Kawaguchi T, Tanthapanichakoon W (2003) Discrete particle simulation of solids motion in a gas-solid fluidized bed. *Chem Eng Sci* **58**: 915–921

Nedelcu S and Watson JHP (2004) Size separation of DNA molecules by pulsed electric field dielectrophoresis. *J Phys D: Appl Phys* **37**: 2197–2204

Pohl H (1978) Dielectrophoresis. Cambridge Unversity Press, Cambridge

Schnelle T, Muller T, Fiedler S, and Fuhr G (1999) The influence of higher moments on particle behaviour in dielectrophoretic field cages. *J Electrostat* **46**: 13–28

Synder TJ, Schneider JB, and Chung JN (2001) Bieletrophoresis with application to boiling heat transfer in microgravity. I. Numerical analysis. *J Appl Phys* **37**: 4076–4083

Tsuji Y, Kawaguchi T, and Tanaka T (1993) Discrete particle simulation of two-dimensionalized fluidized bed. *Powder Technol* **77**: 79–87

Washizu M (2003) DNA Manipulation in Electrostatic Fields. 7th international Conference on Miniaturized Chemical and Biochemical Analysts Systems pp 869–873. Squaw Valley, California USA

Xu BH and Yu AB (1997) Numerical simulation of the gas-solid flow in a fluidized bed by combing discrete particle method with computational fluid dynamics. *Chem Eng Sci* **52**: 2785–2809

Yang J, Huang Y, Wang XB, Becker FF, and Gascoyne PRC (2000) Differential analysis of human leukocytes by dielectrophoretic field-flow-fractionation. *Biophys J* **78**: 2680–2689

Zhang L, Brody JP, and Burke PJ (2004) Electronic manipulation of DNA, proteins, and nanoparticles for potential circuit assembly. *Biosens Bioelectron* **20**: 606–619