<table>
<thead>
<tr>
<th>Workshop Program</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Monday, September 12, 2011</strong></td>
</tr>
<tr>
<td>8:00 AM</td>
</tr>
<tr>
<td>9:00 AM – 9:10 AM</td>
</tr>
<tr>
<td>9:10 AM – 9:30 AM</td>
</tr>
<tr>
<td>9:30 AM – 12:30 PM</td>
</tr>
<tr>
<td>12:30 PM – 2:00 PM</td>
</tr>
<tr>
<td>2:00 PM - 5:00 PM</td>
</tr>
<tr>
<td>5:30 PM</td>
</tr>
<tr>
<td><strong>Tuesday, September 13, 2011</strong></td>
</tr>
<tr>
<td>9:30 AM – 12:30 PM</td>
</tr>
<tr>
<td>12:30 PM - 2:00 PM</td>
</tr>
<tr>
<td>2:00 PM - 5:00 PM</td>
</tr>
<tr>
<td>6:30 PM</td>
</tr>
<tr>
<td><strong>Wednesday, September 14, 2011</strong></td>
</tr>
<tr>
<td>9:30 AM – 12:30 PM</td>
</tr>
<tr>
<td>12:30 PM - 2:00 PM</td>
</tr>
<tr>
<td>2:00 PM - 3:00 PM</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

REGISTRATION
http://www.dsrc.rpi.edu/NSF-RPI-WORKSHOP/registration
September 12, 2011
Breakfast (8:00am)

Welcome (9:00am-9:10am) -- Robert E. Palazzo, Provost.

Opening Remarks (9:10am-9:30am) -- Bülent Yener, Director DSRC.

Session I (9:30am-12:30pm): E-Science for Data Intensive Biomedicine

**Computational Biomedicine: A Challenge for the 21st Century**
_Peter Coveney, University College London_

Much scientific research and innovation requires a wide range of data intensive computations—including high performance computing (HPC)—to be run as part of a complex workflow, for example with different steps for data access, data acquisition, data transfer, data processing and compute-intensive simulation. To simplify the process for the user, we can orchestrate these steps using a workflow manager and provide seamless access to remote resources using audited credential delegation and the application hosting environment. This talk will outline several biomedical applications our group has been working on to enable better medical/clinical treatments which draw on the use of HPC inter alia. These include patient specific HIV drug therapy, personalized treatment of aneurysms in the brain and patient specific cancer therapy. In this talk I will describe the e-Science techniques used in each project and will make a case for an integrated computational infrastructure (data storage, networks and computational resources) to ensure the successful development of future biomedical applications. I will also provide an overview of the developments at the EU level to further computational biomedicine through the FP7 Virtual Physiological Human (VPH) Initiative. VPH projects in which we are involved include VPH-SHARE, which aims to provide robust cloud based infrastructure for translational biomedical research and clinical usage, and p-medicine, where we process large amounts of federated medical data from different sources to provide personalized clinical decision support. Such scenarios require a heterogeneous computational infrastructure, comprising secure resources from the desktop (within a hospital) to the very largest supercomputers available nationally and internationally, in new and disruptive ways, for example to deliver urgent results on demand.

**Accelerating Data-Intensive Science by Outsourcing the Mundane**
_Ian Foster, University of Chicago and Argonne National Laboratory_

Whitehead observed that "civilization advances by extending the number of important operations which we can perform without thinking of them." I propose that cloud computing can allow us to accelerate dramatically the pace of discovery by removing a range of mundane but timeconsuming research data management tasks from our consciousness. I describe the Globus Online system that we are developing to explore these possibilities, and propose milestones for evaluating progress towards smarter science.

**Developing Standards: Why standards are useful and why they are so hard to develop**
_Herbert Sauro, University of Washington_

In this presentation I will describe two standards I have been heavily involved in, the Systems Biology Mark Language (SBML) and the Synthetic Biology Open Language. SBML is now the de facto standard for model exchange in systems biology and SBOL is gaining traction both in the US and Europe synthetic biology communities. Developing a standard is akin to having a root canal done, its need to be done, no body wants to pay for it, its a painful and stressful process, and once over, everyone forgets you ever had it done. Developing standards is extremely difficult and roadblocks occur at every level. As one grant reviewer recently remarked at the NIH biodata management study section, "not anther standard".
Analysis of Molecular Networks  
Mark Gerstein, Yale University

My talk will be concerned with understanding protein function on a genomic scale. My lab approaches this through the prediction and analysis of biological networks, focusing on protein-protein interaction and transcription-factor-target ones. I will describe how these networks can be determined through integration of many genomic features and how they can be analyzed in terms of various topological statistics. In particular, I will discuss a number of recent analyses: (1) Improving the prediction of molecular networks through systematic training-set expansion; (2) Showing how the analysis of pathways across environments potentially allows them to act as biosensors; (3a) Analyzing the structure of the regulatory network indicates that it has a hierarchical layout with the "middle-managers" acting as information bottlenecks; (3b) Showing these middle managers tend be arranged in various "partnership" structures giving the hierarchy a "democratic character"; (4) Showing that most human variation occurs at the periphery of the protein interaction network; (5) Comparing the topology and variation of the regulatory network to the call graph of a computer operating system; and (5) Developing useful web-based tools for the analysis of networks (TopNet and tYNA).

Computational Insights into Population Biology  
Tanya Berger-Wolf, University of Illinois at Chicago

Computation has fundamentally changed the way we study nature. Recent breakthroughs in data collection technology, such as GPS and other mobile sensors, high definition cameras, GIS, and genotyping, are giving biologists access to data about wild populations, from genetic to social interactions, which are orders of magnitude richer than any previously collected. Such data offer the promise of answering some of the big questions in population biology: Why do animals form social groups and how do genetic ties affect this process? Which individuals are leaders and to what degree do they control the behavior of others? How do social interactions affect the survival of a species? Unfortunately, in this domain, our ability to analyze data lags substantially behind our ability to collect it. In this talk I will show how computational approaches can be part of every stage of the scientific process, from data collection (identifying individual zebras from photographs) to hypothesis formulation (by designing a novel computational framework for analysis of dynamic social networks).

Exploiting Social Networks for Disease Surveillance  
Wagner Meira Jr., Universidade Federal de Minas Gerais, Brazil

The internet has been evolving from a communication media to an environment where users talk about the most diverse topics, including their health conditions. Characterizing and understanding how the internet data may be used for assessing real phenomena, such as disease epidemics, becomes a key component of many Internet-based application and demands the development of new data mining models and techniques. Data mining in such scenarios is challenging because the data is intrinsically uncertain and multi-scale, and the patterns to be mined are complex and evolve through time, and there is a huge amount of information that need to be processed in real time. In this talk we present an active surveillance methodology that is based on four dimensions: volume, location, time and public perception. This methodology is the basis of the Dengue Observatory, a surveillance tool for real-time analysis and spatio-temporal tracking of Dengue, which is a mosquito-borne infectious disease that is a leading cause of illness and death in tropical and subtropical regions, including Brazil.
Model Reduction of Parametrized Systems  
*Thanos Antoulas, Rice University*

Dynamical systems are a principal tool in modeling and control of physical processes in engineering, economics, the natural and social sciences. In many areas, direct numerical simulation has become essential for studying the rich complexity of these phenomena and for the design process. Due to the increasing complexity of the underlying mathematical models, unmatched by the increase in computing power, model reduction has become an indispensable tool in order to facilitate or even enable simulation (in particular parameter studies and uncertainty analysis), control, and optimization of dynamical systems. Here, we will focus on parametrized models where the preservation of parameters as symbolic quantities in the reduced-order model is essential. The main tool that we will use in our investigations is the Loewner matrix.

Discovery of Principles of Nature from Matrix and Tensor Modeling of Large-Scale Molecular Biological Data  
*Orly Alter, University of Utah*

In my Genomic Signal Processing Lab, we believe that future discovery and control in biology and medicine will come from the mathematical modeling of large-scale molecular biological data, such as DNA microarray data, just as Kepler discovered the laws of planetary motion by using mathematics to describe trends in astronomical data. In this talk, I will first describe novel generalizations of the matrix and tensor computations that underlie theoretical physics, that we are developing for comparison and integration of multiple genomic datasets recording different aspects of, e.g., the cell division cycle and cancer. Second, I will describe our recent experiments that verify a computationally predicted genome-wide mode of regulation, and demonstrate that singular value decomposition (SVD) and higher-order SVD (HOSVD) modeling of DNA microarray data can be used to correctly predict previously unknown cellular mechanisms. Third, I will show that mode-1 HOSVD modeling of rRNA sequence alignments suggests a new way of looking at evolution as a composition of changes rather than a hierarchy of changes, and might be used to predict evolutionary mechanisms, i.e., evolutionary pathways and the underlying structural changes that these pathways are correlated, possibly even coordinated with. Last, I will describe the computational prognosis of brain cancers by using generalized SVD to compare global DNA copy numbers in patient-matched normal and tumor samples from the Cancer Genome Atlas.

All-at-once Optimization for Mining Higher-order Tensors  
*Evrim Acar, University of Copenhagen*

In many disciplines, we deal with data that has more than two axes of variation. For instance, multi-channel EEG (electroencephalogram) signals have spatial, spectral and temporal dimensions. We represent such data as higher-order tensors and use tensor factorizations, i.e., higher-order analogues of matrix factorizations, to extract the underlying patterns in the data. However, information extraction is often challenging due to large-scale and incomplete data sets, i.e., data with missing or unknown values. Besides, we are commonly faced with data from multiple sources and need to fuse data sets to capture the underlying structures. In order to address these issues, we introduce a gradient-based optimization framework for factorizing higher-order tensors, that can easily extend to the analysis of incomplete tensors as well as joint analysis of multiple data sets. In particular, we are interested in fitting the tensor model called CANDECOMP/PARAFAC(CP), which expresses a tensor as the sum of rank-one tensors and has been widely used in various disciplines including chemometrics, signal processing, computational neuroscience and social network analysis. The traditional CP algorithm is an alternating least squares algorithm, which is computationally very efficient but not robust in the case of overfactoring, i.e., extracting more factors than the true underlying factors. Alternative algorithms shown to be robust to overfactoring do not scale to large-scale data sets. The proposed gradient-based approach uses all-at-once optimization and solves for all CP factor matrices simultaneously. Using numerical experiments, we demonstrate that it is both accurate and scalable. We also show how the proposed approach extends to fit a CP model to incomplete data and joint analysis of multiple data sets. We further demonstrate the usefulness of the proposed framework on several biomedical applications.
Data-driven Multiscale Analysis in Musculoskeletal Biomechanics

Ahmet Erdemir, Lerner Research Institute, Cleveland Clinic

In musculoskeletal biomechanics, in vivo measurements of motion and deformation are common, particularly at the spatial scales of the body, joints, and organs. A plethora of imaging modalities, e.g., magnetic resonance imaging, computed tomography, also provide detailed anatomical descriptions of the musculoskeletal system. Unfortunately, at the spatial scales of tissues and cells, in vivo measurement capabilities on human subjects are limited. At best, such measurements are conducted in an indirect fashion as in the case of the description of cartilage contact mechanics in the knee joint using biplanar fluoroscopy. Measurements of cell level deformations are scarce; yet promising examples are emerging, e.g., in situ measurement of chondrocyte deformations in mice knees. As a result, establishing the mechanical relationship between the macroscopic and microscopic levels of the spatial scale commonly requires a modeling and simulation approach. The goal of this abstract is to provide a summary of our ongoing experiences in data-driven biomechanical analysis to identify tissue properties and to predict in situ mechanical response of tissues and cells as a function of body level mechanical loads. In gait analysis, movement of the lower extremity can be measured with a camera based motion analysis system, including ground reaction forces using force plates, and electromyography to record muscle activity. A data-driven musculoskeletal model can help estimate joint reaction forces and muscle forces. If anatomical image data of the subject's knee are available, a subject specific finite element representation of the joint can be built, with loading and boundary conditions determined by the joint reaction forces and/or muscle forces. Further simulations indicate the stress and strains at the tissue scale, i.e., for the cartilage, as a function of a daily activity. In another example, inverse finite element analysis can help process in vivo (or in situ) measurements at the organ level to obtain tissue mechanical properties. Such investigations may lead delineation of material response of healthy and diseased tissue, i.e., in the diabetic foot, and may potentially help tracking the mechanical effects of the progression of pathology. Current state of technology allows complex cadaver simulations reproducing lifelike loading scenarios. Detailed data sets obtained from such experimentation protocols, combined with multilevel finite element representations of joints, tissues, and cells has the promise to predict in situ mechanical environment of cells as a function of joint loading, dysfunction, and injury. Understanding the mechanical properties and lifelike loading environment of the tissues and cells has significant implications for establishing risk of injury, damage mechanisms, and the foundations of mechanobiological response. The sought-after information is also critical for tissue engineering as well as for prescription of surgical and rehabilitative procedures and for management of musculoskeletal disorders. Establishing this knowledgebase is challenging yet attainable through multidisciplinary collaborations. Advancements in the basic tools of the trait, e.g., finite element method, optimization; management of multimodal complex data sets; multiscale computational techniques and surrogate modeling based coupling approaches; all will significantly facilitate addressing musculoskeletal research problems. Data, model, and software sharing will likely expedite collaborative efforts, potentially providing a pathway for crowd-sourcing. The ultimate goal to translate the knowledge to the clinics for support of decision making will rely on effective processing, interpretation and summary of data, aware of cultural differences between research and clinical groups.

Concurrent Prediction of Muscle Force and Cartilage Stress during Movement

Trent M. Guess, University of Missouri – Kansas City

The absence of detailed knowledge regarding mechanical loading on knee structures inhibits our understanding of joint degeneration and injury. Information on the interrelationships between muscle activations and tissue response is crucial to the development of tissue engineered cartilage and menisci, meniscus and ligament injury and repair, and our understanding of degenerative joint disease, specifically osteoarthritis. Personalized prediction of joint and tissue level loading during ambulation has the potential to significantly enhance orthopedic medicine. In addition to providing a greater understanding of knee biomechanics and tissue function, tools with this capability would enable subject specific intervention strategies aimed at modifying gait for targeted outcomes, such as reducing articular cartilage stress. Combining dynamic multibody musculoskeletal simulations with predictions of tibio-menisco-femoral contact mechanics would be a valuable tool for studying the relationships of muscle force and tissue loading. The presented work combines muscle driven forward dynamics movement simulations with anatomical knee models that include neural network surrogates of cartilage tissue stress response. The musculoskeletal and knee models were built in the multibody framework and include both human and canines. Subject specific limb and knee geometries (tibia, femur, patella, articular cartilages, medial and lateral menisci, and ligaments) for the multibody models were obtained from magnetic resonance imaging. Ligaments crossing the knee were represented as non-linear springs and the menisci were represented as discrete elements connected by 6x6 stiffness matrices. A macro divided the medial and lateral menisci geometries, assigned deformable contacts between each meniscus element and tibio-femoral cartilages, and connected meniscus elements with stiffness matrices. A macro also divided the tibia plateau cartilage into discrete elements. Each cartilage element was rigidly attached to the tibia bone with deformable contacts assigned to the femoral cartilage and neighboring meniscus elements. The knee models also included a
deformable contact patello-femoral joint. Reaction forces at the bone-cartilage interface of the tibia plateau provided input to a neural network surrogate model that predicted cartilage von Mises stress. A tissue level finite element model was used to train the neural network nonlinear data-driven surrogates. For the musculoskeletal models, motion, ground reaction forces, and surface electromyography were measured during various ambulatory activities for both humans and canines. The gait data was used in forward dynamics simulations of the lower body that included 90 muscles, the anatomical knee models, and the cartilage surrogates.

Multi-Cell Simulations of Biological Development and Disease: Prospects and Problems
James A. Glazier, Indiana University

Embryonic development, adult homeostasis, wound healing, regeneration, and diseases like cancer, all result from a basic repertory of cell behaviors. Synthetic Biology and Tissue Engineering employ the same behaviors to build or induce biological structures. Biochemical signaling, regulatory and metabolic networks interact with the motility, force generation and shape changes of cells, extracellular signals and extracellular matrix creating additional complexity. Because of the difficulty of experiments, simulations play a crucial role in disentangling the multiple behaviors which interact in any organism or experiment. I will outline the issues we face in building such simulations, introduce the GGH approach and discuss applications including the formation of the early body plan, the development of blood vessels growth of tumors and Age-Related Macular Degeneration (an eye disease) to illustrate some of the questions this type of simulation can address and its potential application to biomedical problems. I will also discuss the special problems such models raise in the development of simulation environments, model sharing strategies and data generation and analysis.

Many-Task Computing Tools for Multi-Scale Modeling
Daniel S. Katz, University of Chicago and Argonne National Laboratory

This talk will examine the use of many-task computing tools, specifically the Swift scripting and runtime system (http://www.ci.uchicago.edu/swift/) for multi-scale modeling. It will present Swift and give some examples of how it has been used in the past, then discuss some ways it could be used for two specific multi-scale modeling applications. The first is subsurface transport, involving both pore-scale and continuum-scale models. The pore-scale model can be a parallel application, and multiple instances can also be run simultaneously, while the continuum model is a sequential application. The second application is in building coarse-grain (CG) molecular dynamics (MD) models that have a minimal number of CG sites and that match the properties of an all-atom model. This involves running a set of codes in series, starting with a small number of CG sites, to: determine the CG structure, determine the interactions between CG sites, run an MD simulation of the CG model, and determine the simulated properties of an all-atom model or experimental data wherever applicable. If the properties are not good enough, the number of CG sites is increased and the loop is repeated.
Multiparameter Computational Modeling of Tumor Invasion  
*Vittorio Cristini, University of New Mexico*

Clinical outcome prognostication in oncology is a guiding principle in therapeutic choice. A wealth of qualitative empirical evidence links disease progression with tumor morphology, histopathology, invasion, and associated molecular phenomena. However, the quantitative contribution of each of the known parameters in this progression remains elusive. Mathematical modeling can provide the capability to quantify the connection between variables governing growth, prognosis, and treatment outcome. By quantifying the link between the tumor boundary morphology and the invasive phenotype, this work provides a quantitative tool for the study of tumor progression and diagnostic/prognostic applications. This establishes a framework for monitoring system perturbation towards development of therapeutic strategies and correlation to clinical outcome for prognosis.

Multi-scale analysis of histopathological images: Experience with follicular lymphoma and neuroblastoma  
*Metin Gurcan, Ohio State University*

At the Clinical Image Analysis Lab at The, we have been developing computer-assisted image systems for the diagnosis and prognosis of cancers from histopathological slides. These slides result in rather large images with sizes around 100,000 by 100,000 pixels. We take advantage of the information at different scale to improve accuracy and efficiency of the developed algorithms. Accordingly, we have developed multi-scale processing techniques that model the underlying structures. For instance, follicles are visible at 10x magnification level, while centroblasts cells need to be detected at 40x magnification level within detected follicles. At the lowest scale, we can quickly identify areas of non-interest (e.g. background regions), while texture characteristics extracted at intermediate scales are useful to differentiate varying tissue characteristics. We have integrated this logic into our classification mechanism, so that the classifiers not only recognize the classes but also report the associated confidence levels in making a classification decision. By assigning gradually increasing levels of confidence thresholds at different scales, we let the classifiers automatically decide on the right level of information at the right scale. A critical part of this process is to allow “don’t know” type decisions into the classification mechanism. In this talk, we will give examples of this approach from our current NIH supported studies.

Inferring Transcriptional and MicroRNA-mediated Regulatory Programs in Glioblastoma  
*Christina Leslie, Sloan-Kettering Institute*

Large-scale cancer genome characterization projects are generating multiple layers of genomewide data – e.g. copy number, DNA methylation, mRNA and miRNA expression – for hundreds of tumors, but few computational approaches try to integrate these layers in a way that models the underlying relationships between them. We developed a supervised learning framework for integrating multiple molecular profiles with regulatory sequence analysis to reveal mechanisms of dysregulation of gene expression in cancer, including miRNA-mediated expression changes. In particular, we trained sparse regression models that account for differential gene expression beyond copy number changes in each tumor sample relative to normal brain reference samples, and we used a multi-task group lasso approach to learn regression models for all samples at once while encoding assignments of tumors to GBM subtypes. We applied our approach to a data set of ~400 glioblastoma tumors and identified key miRNAs and transcription factors as potential drivers of common or subtype-specific expression changes. We tested two predicted miRNA drivers for the proneural subtype, both underexpressed in proneural samples, by overexpression in neurospheres and found that they drove expression changes concordant with (i.e. opposite in direction to) those in the tumor data, while control miRNAs did not. Computationally dissecting the role of miRNAs in cancer may ultimately lead to small RNA therapeutic strategies, tailored to subtype or individual.

Panel (2:00pm-3:00pm): Medical and Life Science Informatics: Information Rich? Data Poor? Where is the Knowledge Frontier?  
-- *Chair: Peter Fox.*

Discussion & Closing Remarks  
-- *Bülent Yener, Director DSRC*