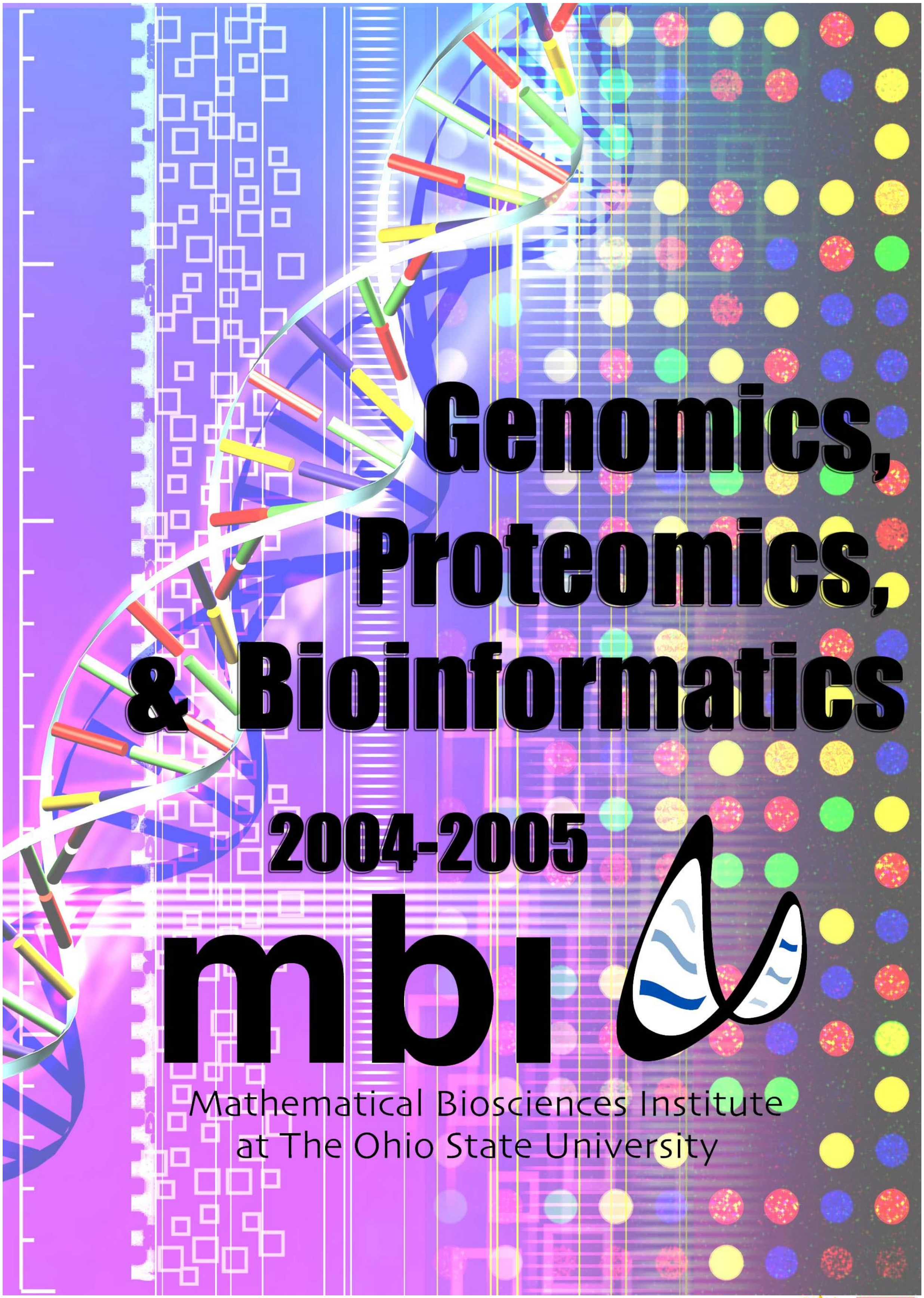


Genomics, Proteomics, & Bioinformatics

2004-2005

mbi

Mathematical Biosciences Institute
at The Ohio State University

The MBI receives major funding from the National Science Foundation Division of Mathematical Sciences and is supported by The Ohio State University.

Director's Letter

The explosion of research in the life sciences has created the need for new mathematical theories, statistical methods, and computational algorithms with which to draw knowledge from the rapidly accumulating data. This need offers a great opportunity and challenge for the mathematical sciences. But to be successful, mathematicians and statisticians must learn the scientists' language and develop a certain level of understanding of their biological problems. This can best be achieved by direct interaction between mathematical scientists and bioscientists.

The Mathematical Biosciences Institute at the Ohio State University, funded by the National Sciences Foundation (NSF), was created in 2002 to provide a national forum for mathematical biosciences that can catalyze such interactions between the biological, medical, and mathematical scientists through vigorous programs of research and education, and to nurture a nationwide community of scholars in this emerging new field. The MBI aims to reinforce and build upon existing research efforts in mathematical biosciences, and quicken intellectual growth in this area.

The MBI runs "Emphasis Year" programs, concentrating on a broad range of topics in one area of bioscience, with six to eight 1-week workshops preceded by tutorials. In the summer, the MBI runs an educational program based on tutorials and team projects led by MBI postdoctoral fellows. Occasional "Current Topics" workshops introduce mathematical scientists to new opportunities for research. The topics of the first two emphasis years were Mathematical Neurosciences and Mathematical Modeling of Cell Processes. This year was devoted to Genomics, Proteomics, and Bioinformatics.

Genomics is the study of the genes and their function in organisms on a global scale. Proteomics is the study of the proteins that are transcribed by a process which decodes genes. Genomic and Proteomics form a discipline that combines sequencing of the DNA, identifying the segments which form genes, and determining the structure and function of the proteins which are created by translation and transcription of the genes.

A major milestone in genomics was the completion of the mapping and sequencing of the human and mouse genomes in the period 2001-2003. This was followed by the sequencing of many bacterial genomes, as well as those of numerous other species of biological or medical importance, such as yeast, the roundworm, and the malaria parasite and its associated mosquito vector. This massive amount of DNA sequence data and the current study of the genes/proteins interactions brings with it the ability to make progress on the molecular mechanisms of disease, including the complex interplay of genetic and environmental factors, and to generate thousands of new biological targets for the development of drugs, vaccines, diagnostics, and therapies. Fundamental biological research is greatly aided by this wealth of data, permitting not only a genome-wide perspective in the study of particular organisms, but a greatly enhanced evolutionary perspective through the use of comparative genomics.

The MBI program began with a 2-week tutorial on microarrays and the statistical methods that have been developed to analyze them. It was followed by workshops which dealt with principles and applications of gene expression data, and regulatory networks aimed at understanding the signaling pathways at the level of genes. Subsequent workshops dealt with computational proteomics, that is, with protein structure and function; with emerging new genomics technologies, such as DNA chips and high-throughput mass spectrometry; and with statistical approaches to estimating recombination rates of different genes and determining the causes of such haplotype structures. A workshop on biomarkers for HIV and Cancer combined much of the recent developments in microarray technology with new statistical methods.

A unique feature this year was a workshop organized by the MBI postdoctorate fellows. Their workshop featured talks by several world class researchers in the mathematical biosciences. Participants included 45 young researchers from all over the country. This novel and exciting workshop included poster presentations by the young researchers, as well as group discussions. There are currently 15 postdoctorate fellows at the MBI, each having two mentors, one from the mathematical sciences and another from the biosciences. Five of them have just finished their 3-year term, and took positions in various research universities.

As we said a year ago in this annual report, despite the clear importance of biology for the future of mathematics, it is still not an easy matter for a mathematician to make the switch to working in this area. Vocabulary is different, the methods may seem strange, and the criteria by which one's work is judged can be radically different. Workshops, such as those which took place this year, play an important role; they are, in essence, role models for those mathematicians interested in broadening their research interests; they provide examples of how interdisciplinary work is done, and how to work with experimental colleagues; and, with the provision of extensive tutorials, they provide a gentle introduction to the field of genomics and proteomics.

This document provides a summary of events and talks that took place in the third year of the MBI. Further details can be found on the MBI web site http://mbi.osu.edu.

Avner Friedman
Director

Mission and Goals

The explosion of research in the life sciences has created the need for new mathematical theories, statistical methods, and computational algorithms with which to draw knowledge from the rapidly accumulating data. The Mathematical Biosciences Institute catalyzes interactions between the biological, medical, and mathematical sciences through vigorous programs of research and education and nurtures a nationwide community of scholars in this emerging new field.

The mission of the MBI is:

- To develop mathematical theories, statistical methods, and computational algorithms for the solution of fundamental problems in the biosciences;
- To involve mathematical scientists and bioscientists in the solutions of these problems; and
- To nurture a community of scholars through education and support of students and researchers in mathematical biosciences.

Participants in the First Young Researchers Workshop in Mathematical Biology smile for the camera.

Corporate Members

The MBI encourages involvement from those in private industry. The institute offers incentives to pharmaceutical and bioengineering companies interested in becoming a corporate member.

Membership benefits include:

- Regular visits by MBI Directors to identify problems and topics of interest, where mathematical sciences could be helpful;
- Follow-up to these problems by institute researchers;
- An invitation to present industrial challenges and problems to MBI audiences and to participate in MBI Programs and workshops.

Current Corporate Members:

- Pfizer
- Eli Lilly
- GlaxoSmithKline

2005 Summer Program on Microarray Gene Expression Data Analysis.

Presentation of a summer project by a student.

Institute Partners

The MBI Institute Partners Program subsidizes the travel and local expenses of IP members and faculty, postdoctoral fellows, and students to allow their participation in research and education programs at the MBI; for details see the MBI web site http://mbi.osu.edu.

Current Institute Partners

Arizona State University
Case Western Reserve University
Indiana University and Purdue University, Indianapolis
Iowa State University
Michigan State University
New Jersey Institute of Technology
Ohio University
University of California at Irvine
University of Cincinnati
University of Georgia
University of Iowa
University of Maryland, Baltimore County
University of Minnesota
Vanderbilt University

MBI Postdoctoral Fellows

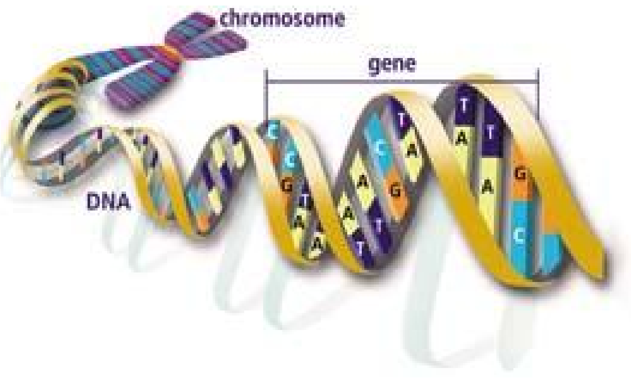
Postdoctoral fellows fall into two support categories: Supported at 100 percent by the MBI or split 50/50 percent by the MBI and another bioscience organization. Postdoctoral fellows sponsored by a specific organization spend 50 percent of their time on research suggested by the sponsor. All postdocs are provided with two mentors: one from the mathematical and statistical sciences, and another from one of the biosciences departments at The Ohio State University. Long-term visitors may also serve as mentors. More details are available in the *MBI Postdoctoral Research Program*

Martin Wechselberger and Dan Dougherty listen attentively at the First Young Researchers Workshop in Math Biology.

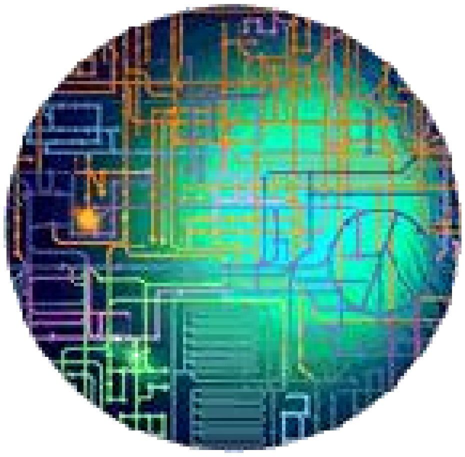
- Janet Best— Department of Mathematics, Cornell University
- Alla Borisyuk—Courant Institute of Mathematical Sciences, New York University
- Gheorghe Craciun— Department of Mathematics, The Ohio State University
- Daniel Dougherty— Department of Statistics, North Carolina State
- Pranay Goel— Department of Mathematics, University of Pittsburgh
- Sookkyung Lim— Courant Institute of Mathematical Sciences, New York University
- Diego Pol— Department of Earth and Environmental Sciences, Columbia University
- Firas Rassoul-Agha— Courant Institute of Mathematical Sciences, New York University
- Katarzyna Rejniak— Department of Mathematics, Tulane University
- Mike Stubna— Theoretical and Applied Mechanics, Cornell University
- Jianjun (Paul) Tian— Department of Mathematics, University of California, Riverside
- Zailong Wang— Department of Statistics, University of California, Davis
- Martin Wechselberger— Mathematics Department, Vienna University of Technology
- Geraldine Wright— Department of Entomology, Oxford University
- Jin Zhou— Department of Statistics, University of Georgia

Summary of the Year in Genomics, Proteomics, and Bioinformatics 2004-2005

Genomics was defined in the 1980s as the new discipline of mapping, sequencing, and analyzing genomes, that is, the study of genes and their function in organisms on a global rather than a local scale. Proteomics, the study of the PROTEin in complement to a genOME, emerged in the 1990s as the qualitative and quantitative comparison of proteomes under different conditions to further unravel biological processes. Both subject areas are at the forefront of the revolution taking place in biological and medical research, which is transforming them from data poor to data rich fields. While most biomedical research continues to be centered around single investigators or small groups of investigators -recording their experimental data in notebooks– increasing use is being made of novel technologies generating massive amounts of data, and requiring careful computational, mathematical, and statistical analyses. In this third year of the MBI, our focus was on these aspects of



DNA with features. U.S. Department of Energy Human Genome Program, http://www.ornl.gov/hgmis.



Wired Cell. U.S. Department of Energy Genomics:GTL Program, http://doegenomestolife.org.

Organizing Committee for 2004-2005

- Vineet Bafna, Department of Computer Science and Engineering, University of California, San Diego
- Victor De Gruttola, Department of Biostatistics, Harvard University
- Rick Durret, Department of Mathematics, Cornell University
- Paul Fuerst, Department of Ecology Evolution, and Organismal Biology, The Ohio State University
- Jeff Hasty, Department of Bioengineering, University of California, San Diego
- Terry Speed, Department of Statistics, University of California, Berkeley

Board of Governors

The Board consists of 12 internationally recognized mathematical scientists and bioscience researchers from academia and industry. The Board meets annually to advise the directors and The Ohio State University regarding management of the institute, to review its programs, and to suggest new programs and give advice regarding programmatic goals.

- Leah Edelstein-Keshet - Department of Mathematics, University of British Columbia
- Lisa Fauci - Department of Mathematics, Tulane University
- Louis Gross - Professor of Ecology and Evolutionary Biology, The University of Tennessee
- Kirk Jordan - IBM Computational Biology Center
- Jim Keener - Departments of Mathematics and Bioengineering, University of Utah
- Douglas Lauffenburger - Massachusetts Institute of Technology
- Gregory Mack - Vice President of Environmental Monitoring and Assessment, Battelle Memorial Institute
- Claudia Neuhauser - Professor of Ecology, Evolution, and Behavior, University of Minnesota
- Stephen Ruberg - Director, Clinical Data Technology and Services, Eli Lilly and Company
- Terry Speed - Professor of Statistics, University of California, Berkeley
- Terry Therneau - Division of Biostatistics, Mayo Clinic College of Medicine, Rochester
- Raimond L. Winslow - Center for Cardiovascular Bioinformatics and Modeling, Whitaker Biomedical Engineering Institute, Department of Biomedical Engineering, The Johns Hopkins University School of Medicine and Whiting School of Engineering

Emphasis Year Scientific Advisory Committee 2004-2005

The Emphasis Year Scientific Advisory Committee reviews the emphasis year proposal as they evolve and offers suggestions throughout the development of the emphasis year. A new committees is appointed for each emphasis year program.

- Andrew G. Clark, Cornell University
- J.J. Collins, Center for BioDynamics, Boston University
- Sandrine Dudoit, Division of Biostatistics, School of Public Health, University of California, Berkeley
- Walter M. Fitch, Ecology and Evolution, University of California, Irvine
- Pavel Pevzner, Department of Computer Science & Engineering, University of California, San Diego
- Mark Ptashne, Memorial Sloan-Kettering, Cancer Center
- Peg Riley, Osborn Memorial Laboratories, Department of Ecology & Evolutionary Biology, Yale University
- David M. Rocke, Department of Applied Science, University of California, Davis
- Michael Savageau, Department of Biomedical Engineering & Microbiology Graduate Group, University of California, Davis
- Eric Siggia, Department of Physics Center for Studies in Physics & Biology, Cornell University

Mathematics Tower and Mathematics Building at OSU.

Local Scientific Advisory Committee

The Local Scientific Advisory Committee helps identify current topics workshops, future emphasis programs and organizers, and potential mentors for postdoctoral fellows.

- Michael Beattie - Department of Neuroscience
- Laura Bohn - Department of Pharmacology and Psychiatry
- Helen Chamberlin - Department of Molecular Genetics
- Albert de la Chapelle - Human Cancer Genetics
- Meg Daly - Department of Evolution, Ecology, and Organismal Biology
- Charis Eng - Division of Human Genetics
- Martin Feinberg - Department of Chemical Engineering
- Paul Fuerst - Department of Evolution, Ecology, and Organismal Biology
- Erich Grotewold - Department of Plant Biology
- Fernand Hayot - Department of Physics
- Charles R. Hille - Department of Molecular and Cellular Biochemistry
- Daniel Janies - Department of Biomedical Informatics
- Lee Johnson - Department of Molecular Genetics
- Doug Kniss - Department of Obstetrics and Gynecology
- Stanley Lemeshow - Dean School of Public Health, Center for Biostatistics
- Gustavo Leone - Department of Molecular Virology, Immunology, and Medical Genetics
- Shili Lin - Department of Statistics
- Charles Orosz - Department of Surgery
- Dennis Pearl - Department of Statistics
- John Reeve - Department of Microbiology
- Andrej Rotter - Department of Pharmacology
- Wolfgang Sadee - Department of Pharmacology
- Joel Saltz - Department of Biomedical Informatics
- Larry S. Schlesinger - Division of Infectious Diseases and Center for Microbial Interface Biology
- Petra Schmalbrock - Department of Radiology
- Brian Smith - Department of Entomology
- David Terman - Department of Mathematics
- Deliang Wang - Department of Computer and Information Science

Program Participation

	# Partici-
Tutorial on Microarrays: September 13-17, 2004	**53**
Tutorial on Statistical Methods: September 20-24, 2004	**53**
Workshop 1: Analysis of Gene Expression Data: Principles and Applications: October 11-15, 2004	**75**
Workshop 2: Regulatory Networks: November 8-12, 2004	**89**
Miniworkshop: Quantitative Mathematical Modeling of Gene Regulatory Networks: December 2-4, 2004	**56**
Workshop 3: Computational Proteomics and Mass Spectrometry: January 11-14, 2005	**74**
Workshop 4: Emerging Genomic Technologies and Data Integration Problems: February 21-24, 2005	**69**
First Young Researchers Workshop in Mathematical Biology: March 29 - April 1, 2005	**69**
Workshop 5: Biomarkers in HIV and Cancer Research: April 18-22, 2005	**84**
Current Topics Workshop: Enzyme Dynamics and Function: May 19-21, 2005	**68**
Workshop 6: Recombination: Hotspots and Haplotype Structure: June 13-16, 2005	**51**
Summer Program	**26**
Total	**767**
Long Term Visitors	
(a) 2-3 weeks	**2**
(b) 4 weeks - 3 months	**7**
(c) 3 months - 1 year	**13**
Total	**22**

Program Details

Workshop 1: Analysis of Gene Expression Data: Principles and Applications
October 11-15, 2004

Organizers:
Terry Speed, Department of Statistics, University of California at Berkeley
Shili Lin, Department of Statistics, The Ohio State University

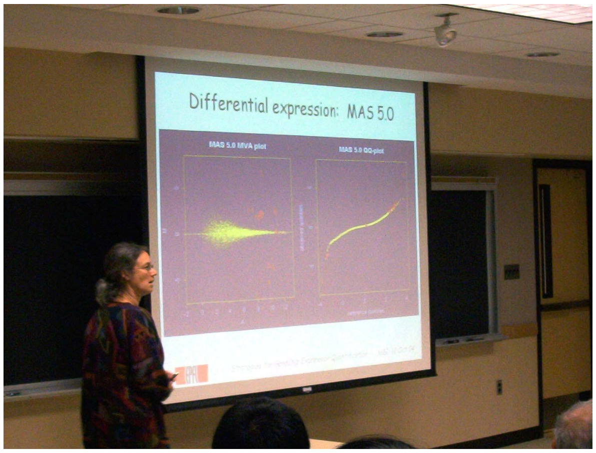
Summary of Talks

Day 1 was devoted to "low level" analysis, focusing on topics including image segmentation, expression level quantification, and normalization. Earl Hubbell (Affymetrix) discussed estimators for measuring expression intensities of transcripts that have a concentration near zero. For transcripts that are expressed at such low levels, nonspecific binding can be a significant portion of the observed probe intensity, thus it is of great importance to design estimators that can deal with such situations satisfactorily. Among the estimators discussed, an M-estimator (PLIER) and its variants were showed to provide good intensity measures with little positive bias, and they can be variance stabilized, if desired, using standard statistical tools. Kathleen Kerr (University of Washington) talked about an experiment designed to compare measurements of relative gene expression from quantitative rtPCR to measurements from Affymetrix gene chips. Her particular interest was on how different methodologies for processing Affymetrix data influence the agreement between Affymetrix and qrtPCR measurements. Her results indicated that measurements processed using MAS5, gcRMA, and dChip—pm-mm model—provided better agreements with the qrtPCR data than several other processing methods—RMA, VSN, dChip (pm only). However, how different methods -especially MAS5- performed for low-intensity genes was unclear, as the genes selected for this analysis had an overall medium intensity. The last lecture of the day was presented by David Kreil (University of Cambridge) who introduced a variable Bayesian implementation of Independent Component Analysis and reported on successes and problems they had experienced in specific microarray data analysis applications. He further discussed how the complex experimental process underlying microarray experiments affected data, and why low level analysis was still a major challenge in the field.

The morning talks of Day 2 continued on the general theme of low level analysis, but with a focus on the issue of combining data from large studies and/or multiple

sources. Darlene Goldstein (Ecole Polytechnique Federale Lausanne) discussed strategies for quantifying expression intensities for large studies. As clinical gene expression studies increase in size, quantifying expression using multiarray-based strategies, although generally perform well, is more challenging in terms of memory and time requirements. She reported on results of a study that examined properties and tradeoffs of various strategies for quantifying expression in large size studies. In particular, a partition-resampling strategy was explored and recommended as a reasonable alternative when the number of arrays is too large to be quantified as a whole. Evan Johnson (Harvard University) described an empirical Bayes method to adjust for batch effects. The Bayesian framework shrinks genewise batch adjustments by pooling information across genes within batches, adjusting for gene-batch interactions while respecting differences in expression estimates across batches. This method was compared with univariate genewise location-scale adjustments and with some other methods. The preliminary results showed that the empirical Bayes adjustments were robust and improved the consistency of within-batch fold changes for treatment effects.

The afternoon lecture was given by Raymond Carroll (Texas A&M University), who presented a semiparametric profile likelihood method for efficient estimation of main effects as well as gene-environmental interactions in case-control studies with quantitative gene information. In particular, the distribution of the gene expression level was modeled parametrically while leaving that of the environmental factors unspecified. Under this setting, a semiparametric profile likelihood was constructed and profiled over the nonparametric distribution of the environmental factors. He showed that this analysis framework led to more efficient estimation of main effects and interaction effects than a standard logistic regression approach.



Multiple testing/comparison was the main topic of Day 3. Jason Hsu (The Ohio State University) promoted the philosophy of viewing microarray experimental design and gene expression analysis as integral parts of specific decision-making processes. He described a project to design a microarray experiment with randomization, replication, and blocking that would allow for assessment of sensitivity and specificity of genetic profiling prognostic chips. The controlling for family wise error rate vs. false discovery rate when analyzing gene expression data was discussed in the context of specific applications. The principle of partition multiple testing was also given. In particular, he described the conditions under which stepwise testing becomes a valid computational shortcut to partition testing. Susmita Datta (Georgia State University) proposed an empirical Bayes approach to deal with multiple testing in microarray analysis.

Gene-Environment Case-Control Studies
Raymond J. Carroll
Department of Statistics
Faculty of Nutrition and Toxicology
Texas A&M University
http://stat.tamu.edu/~carroll
Image of rat colon displaying apoptosis (green) and cell differentiation (red)
STATISTICS

Both parametric and nonparametric versions of empirical Bayes adjustment to the p-values were considered. Through pooling evidence from all p-values across the tests, a new set of accept/reject decisions were reached for each null hypothesis using the empirical Bayes adjusted p-values. Specifically, resampling based step-down p-values were calculated with respect to a prespecified overall—familywise—type I error rate. This new procedure was shown to produce improvement in sensitivity in a number of examples. David Allison (University of Alabama at Birmingham) discussed new ways of thinking about dealing with new challenges of multiple testing issues in high dimensional biological research. The discussion focused on methods that capitalize on, rather than penalize for, the large number of tests through mixture modeling procedures. Power and sample size estimations were also considered. Furthermore, composite hypothesis testing involving both union-intersection testing and intersection-union testing were presented.

The focus of the last talk of the day by Eric Schadt (Rosetta Inpharmatics/Merck) shifted to the important issue of genetic network reconstruction. Such reconstruction has emerged as one of the primary goals in biological research as they can elucidate not only common human diseases, but also living systems more generally. Schadt presented a statistical procedure for inferring causal relationships between gene expression traits and more classic clinical traits, including complex disease traits. This procedure was then generalized to the gene network reconstruction problem, where naturally occurring genetic variations in segregating mouse populations were used as a source of perturbations to elucidate tissue-specific gene networks. Differences in the extent of genetic control between genders and among four different tissues were highlighted. Schadt also demonstrated that the networks derived from expression data in segregating mouse populations using the novel network reconstruction algorithm were able to capture causal associations between genes that resulted in increased predictive power, compared to more classically reconstructed networks derived from the same data.

The topics of Day 4 were differential expression and co-expression of genes. Kim-Anh Do (MD Anderson Cancer Center) proposed a model-based inferential procedure for differential gene expression, using a Bayesian probability model for the distribution of gene intensities under different conditions. The probability model is a variation of traditional Dirichlet process mixture models. The model includes an additional mixture corresponding to the assumption that transcription levels arise as a mixture over nondifferentially and differentially expressed genes. This full Bayesian approach overcomes some of the limitations of certain popular empirical Bayes methods, albeit with an increased, though still manageable, computational burden. Do illustrated the ease of the procedure for making joint inference about a group of genes, and elaborated on how the control of false positive rates can be automatically incorporated into this approach. Rainer Spang (Max Planck Institute for Molecular Genetics), on the other hand, addressed the problem of detecting sets of differentially co-expressed genes in two phenotypically distinct sets of expression profiles. He introduced a score for differential co-expression, and suggested a computationally efficient algorithm for finding high scoring sets of genes. The method was demonstrated in the context of simulations and on real expression data from a clinical study.

On Day 5, Harmen Bussemaker (Columbia University) addressed the challenge to extract useful information about the global regulatory network from data in functional genomics studies. He presented an integrative modeling framework that combines libraries of expression and occupancy data to define the functional targets of each transcription factor. Multivariate regression analysis was used to infer transcription factor activity levels for each condition, and the correlation between the mRNA expression profile of an individual gene and the inferred activity profile of a transcription factor is interpreted as regulatory coupling strength. The application of the method to yeast S. cerevisiae resulted in the finding that, on the average, 58% of the genes whose promoter region was bounded by a transcription factor were true regulatory targets. These results enabled them to assign directionality to transcription factors controlling divergently transcribed genes that shared the same promoter region. Hongyu Zhao (Yale University) described their efforts to develop a statistical framework to integrate diverse genomics and proteomics information to dissect transcriptional regulatory networks and signal transduction pathways. Different data sources offer deferent perspectives on the same underlying system, and thus it was anticipated that the combined information could increase one's chance of uncovering underlying biological mechanisms. The method was illustrated through applications to yeast data.

Terry Speed (University of California at Berkeley) offered concluding remarks in the final talk of the workshop, focusing on a number of open problems. The two-dimensional multiple testing problem was the first focus, with a projection that entirely satisfactory classical inference procedures, which control familywise type I error rates or estimate false discovery rates, were on the horizon. Gene set analysis was next, with an appeal for research to further develop two types of analysis: determining which known sets of genes are coregulated in a given experiment, and discovering sets not previously known to be co-ordinately regulated. Four open problems were highlighted among many others in the area of microarray time series analysis. These were: (a) flexible modeling of longitudinal data, (b) clustering methods suited to time-course profiles, (c) robust fitting procedures to deal with aberrant curves, and (d) suitable permutation-based or bootstrap analyses to assign p-values and deal with multiple testing. How to avoid confounding to identify main and interaction effects in the new wave of observational studies that include data on gene expression profiles is another open problem. The last area highlighted was about joint analysis of expression and sequence data. He believed that, although a lot of work had been done, there were still many problems needing the attention of statisticians.

Conclusion

The workshop provided an excellent forum for exchange of knowledge and ideas. The 1-hour afternoon discussion sessions were very well received and valued by the participants. The workshop covered many areas of active research in microarray analysis, re-energizing seasoned researchers with fresh ideas, and in the mean time, providing new researchers to the field with the necessary tools to tackle the many open problems. In particular, the post-doctoral fellows and visitors at the MBI, and the faculty and graduate students in the Statistics Department benefited tremendously from the lectures and discussions. Furthermore, the workshop provided an avenue for fostering collaborative research between biological/medical researchers and statisticians.

Workshop 2: Regulatory Networks
November 8-12, 2004

Organizers:
Jeff Hasty, Department of Bioengineering, University of California at San Diego
Ralf Bundschuh, Department of Physics, The Ohio State University
Fernand Hayot, Department of Physics, The Ohio State University

Summary of Talks

The workshop began with a presentation by Dan Gillespie. Nearly thirty years ago, Dan developed the algorithmic approach that is now universally used to simulate noisy genetic networks. In his lecture, he laid the groundwork for the whole workshop. He explained the microscopic justification for the algorithm named after him, and then went on to derive the higher level descriptions of genetic networks in terms of Langevin equations, and finally rate equations. In this way, he not only introduced all the modeling techniques currently used in the field, but clearly pointed out from a fundamental point of view which assumptions go into which level of description. On the way, he also presented his new tau-leaping technique that can be used to significantly speed up genetic network simulations. In the next talk, Daniel Forger (New York University) presented a specific model of a genetic network, namely of the circadian clock. He showed that on the level of modeling through reaction rate equations, the key model parameters can be identified and derived simply by fitting the model to the available experimental data. Then, he went on to stochastic simulations of the same model and found that stable oscillations in the presence of noise require further narrowing of the model parameters. Specifically, he found that duplication of some key genes, which usually has not been included in previous models of the circadian clock, is crucial for stable oscillations in the presence of noise.

In the afternoon, Tim Elston (University of North Carolina) further elaborated on the theme of noise in genetic network simulations. He was particularly interested in this issue in the context of bistable systems, i.e., genetic switches. A striking result that sparked much discussion during and after the talk was that close to the boundary of the bistable region, a difference in the concentration of a molecule by one single molecule can have drastic effects on the system's behavior in spite of the fact that the total number of molecules is on the order of one thousand. The second afternoon hour was dedicated to informal discussions. In addition to more general questions to all speakers, there was an especially heated discussion about Daniel Forger's model of the circadian clock with an essential disagreement if building more biological detail into a model—and thus introducing more parameters that have to be fitted—is beneficial or not.

Outline

I. Stochastic modeling and BioNetS
II. Binary versus Graded Response
III. The Pheromone Signaling Pathway
IV. MAP Kinase regulation
V. Protein Turnover and Oscillations

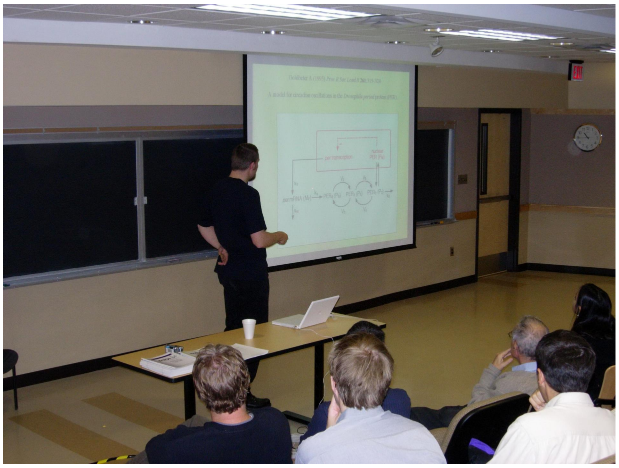
During the evening reception, posters were presented. All these posters, in one way or another, dealt with the question of genetic networks as stochastic processes. Chang Lee (University of Minnesota) presented a general master equation approach to such networks focusing on the separation of slow and fast degrees of freedom. Tomasz Lipniacki (Inst. of Fundamental Technological Research) showed an explicit model of eukaryotic transcriptional regulation—a subject that was reason for arguments several times during the workshop.

The second and third day of the workshop were dedicated to questions of modeling very specific genetic networks. Michael Savageau (University of California at Davis) gave a talk on the discovery of design principles and the construction of genetic circuits. He addressed the issues by comparing and contrasting what has been learned about design principles for gene circuits in their complex natural setting and how these have been put to use in designing, constructing, and analyzing simple synthetic gene circuits. He discussed the use of fractional kinetic equations and negative and positive modes of control, corresponding respectively to low demand and high demand for gene expression. Jean-Christophe Leloup discussed again a model for circadian clocks. In contrast to Daniel Forger's model, Leloup's model was specific for the mammalian circadian clock. His very detailed deterministic five-gene model shows behavior reminiscent of physiological disorders related to circadian rhythms in humans. Luis Serrano (European Molecular Biology Laboratory) gave a talk on engineering gene networks to emulate Drosophila embryonic pattern formation and in silico biological validation of protein interaction networks. He stressed that protein interaction networks are an important part of the post-genomic effort to integrate a parts-list view of the cell into system-level understanding. He and his

to Protein Networks
Treat & Understand Diseases
Understand in a Quantitative way a Cell and be able to re-design it
Simulation Tools

group used 11 yeast genomes to show that combining comparative genomics and secondary structure information can greatly increase consensus based prediction of SH3 targets. Their findings highlight several novel S. cerevisiae SH3 protein-interactions and the importance of selection of optimal divergence times in comparative genomics studies. Serrano also talked about pattern formation in the Drosophila embryo, where maternal morphogen gradients establish gap gene expression domain patterning along the anterior-posterior axis. He discussed an artificial transcription/translation network that generates simple patterns, crudely analogous to the Drosophila gap gene system. From a model computer simulation, several features of interest emerge. For example, the model suggests that simple diffusion may be too rapid for Drosophila-scale patterning, implying that sublocalization or 'trapping' is required. It also shows that for pattern formation to occur under the conditions of the in vitro reaction-diffusion system, the activator molecules must propagate faster than the inhibitors; furthermore, controlled protease degradation stabilizes patterns.

The first talk on Day 3 was by Ron Milo (The Weizmann Institute of Science). He started by identifying network motifs in E. Coli. Network motifs are small groups of genes with connection patterns that occur significantly more often than expected by chance. Interestingly, there is only one such motif each with three and four genes respectively. Each of these motifs has its specific function that Ron discussed. Then, he showed that the same analysis applied to yeast yields the same motifs which further underline their apparent importance in biology. John Reinitz (Stony Brook University) specializes on the developmental network of the fruitfly. He showed how a detailed understanding of the patterning of the fruitfly embryo can be achieved in interplay of systematic experiments that image the localization of proteins in the fly embryo and mathematical modeling. A new insight into fruitfly development that could also be reproduced in the quantitative model was that some of the stripes that are the precursors of future body parts actually move slightly during the development. Finally, Lingchong You (California Institute of Technology) explained his work on utilizing synthetic gene regulatory networks to control a cellular population. Experimentally, he employed a genetic sensor that can be placed in each cell. The sensor responds to the amount of a small signaling molecule that is secreted by other cells in the population. If the number of cells is large, then there is an abundance of the signaling molecule, and the sensor sends a signal that results in cellular death. Mathematically, Dr. You introduced a type of preditor-prey population model, and showed how the model could explain interesting population dynamics observed



in the experiments.

The last two days of the workshop were devoted to more generic questions in genetic network performance and synthetic circuits. On Day 4, Michael Simpson (The University of Tennessee) argued that instead of looking at noise in a genetic network as a necessary evil, one can use the noise in order to learn something about the genetic network itself. His approach, which is based on electrical engineering techniques, consists of isolating the noise from the output of a genetic circuit and calculating its frequency power spectrum. In the frequency domain, the power spectrum can be decomposed into a sequence of clearly identifiable features, each of which corresponds to a specific rate constant in the regulatory network. Thus, the elusive rate constants can be determined just by observing the noise of a genetic network. Terry Hwa (University of California at San Diego) elaborated on two different questions during his talk. In the first part, he demonstrated how complex Boolean functions can be implemented by transcriptional regulation alone, which is arguably much faster and much more resourceful than constructing complex Boolean functions from simple Boolean functions spread over several genes that regulate each other. In the second part of his talk, Terry put the idea forward that differential degradation between dimers and monomers of some protein species can be an important ingredient in generating the kind of nonlinearities that are crucial for genetic networks like switches and oscillators. In a concrete example, he showed that this mechanism can make the difference between the implementability of a switch or oscillator with “standard'“ promoters or the necessity to use extremely strong promoters in the circuit design. In the afternoon, Ron Weiss (Princeton University) showed how, by implementing synthetic genetic circuits including circuits for intercellular communication, cell populations can be forced into making predefined patterns. The climax of this effort was an in vivo implementation of the game of life.

The last day of the workshop began with a talk by Alexander von Oudenaarden (MIT). He presented a combination of experimental and modeling work on memory in genetic networks. In his experiments, he explicitly showed hysteresis in a genetic switch. Even more impressively, he demonstrated that cells remember their state in the bistable region of a switch for tens of generations. Tim Gardner spoke about a mathematical aspect of genetic networks that was not really touched upon by the previous speakers, namely the reconstruction of genetic networks from microarray data. While all models of previous speakers of the workshop were built by integrating a lot of precise biological information known about the specific networks in question, Tim presented his algorithms aimed at reconstructing a previously unknown genetic network from observations of its function by way of microarrays. Finally, Chetan Gadgil (GlaxoSmithKline) concluded the workshop by presenting his analytical approach to describing noise in genetic networks. As an application of his approach, he discussed how network topology influences the noise in the network.

Conclusion

The workshop brought together many of the leading experts in the field of genetic network modeling. As the sometimes very vivid discussions during and after the talks, as well as during the informal discussion sessions, demonstrated that this is a field that is very much in flux. Many participants explicitly commented how stimulating the relatively loose schedule was that allowed plenty of time for interactions among the participants outside of the lectures.

The speakers came from departments of Mathematics, Applied Mathematics and Statistics, Bioengineering, Electrical Engineering, Chemical Engineering, Material Sciences and Engineering, Physical Chemistry, Physics, Molecular Biology, and Biology. While commensurate with this variety, many different approaches to the question of quantitative description of genetic networks were presented. Nevertheless, there were several issues that came up recurrently throughout the workshop, such as the importance of understanding spatial structures in genetic networks, the question of how much detail is needed for successful modeling of a network, the problem of choosing and/or determining modeling parameters, and the question if noise in genetic networks is just a nuisance for cells or if it is actually positively used in biology. It is likely that not only the outside participants of this workshop, but also some of the many attendees from the Ohio State community who contributed to the discussions will grapple with these questions for quite some time to come.

Miniworkshop: Quantitative Mathematical Modeling of Gene Regulatory Networks
December 2-2, 2004

Organizers:
Erik M. Boczko, Department of Biomedical Informatics, Vanderbilt University
Tomas Gedeon, Department of Mathematical Sciences, Montana State University
Konstantin Mischaikow, Dynamical Systems and Nonlinear Studies, Georgia Institute of Technology

Overall Summary

The goal of this workshop was to bring the best mathematicians interested in the interplay between the structure of gene regulatory networks, and the best biophysicists who measure the essential quantities necessary for modeling, to jointly develop new approaches toward modeling, and understanding the dynamics of gene regulatory networks. The central biological example considered at this workshop was yeast nitrogen metabolism, although other networks were presented and discussed.

As measured by the interaction among the participants during the workshop, this venture was a success. That is, we believe we achieved the most important goal that we set for the workshop and that was to get a group of highly diverse biologists and mathematicians to effectively and actively communicate with one another about the topic of gene regulatory networks and structure theorems. During the conference, two of the experimental biologists, Terry Cooper and George Myself, remarked at how completely different this experience was for them than the scientific meetings that they normally attend. After a talk by biologist Jason Lowry, mathematician Hal Smith remarked that he was amazed at the level of detail and intricacy of the biological information presented. These comments illustrate to us the importance of this kind of workshop. Ultimately, the success of this workshop will be measured by the number of collaborations and ideas that were spawned or nurtured there.

The speakers were chosen and the talks organized around three central themes: (1) network structure and its relation to dynamics (i.e., structure theorems, symmetry, and phenotypic attractors); (2) mathematical issues surrounding models (i.e., fixed and state dependent delays, cell division and dilution, transport, and transcription and translation); and (3) the dynamics, biology, and evolution of nitrogen regulation in yeast.

Summary of Talks

The workshop began with an overview talk by Erik Boczko (Vanderbilt University) of the systems biology approach to nitrogen regulation and structure theorems taken by the organizers. The talk generated many questions and effectively set a mood of interaction that was maintained throughout the workshop. This talk was followed by Terry Cooper (University of Tennessee), who gave a more formal and complete biological introduction to the elements of nitrogen regulation in yeast by the GATA factor proteins. Professor Cooper's talk was well conceived and aimed at bringing those mathematicians not familiar with details into the loop. Nitrogen regulation was the dominant theme, and by choosing a single network as the theme, the short workshop was able to focus on the essential elements that we wanted to stress and that are the connection between network and dynamics. Had the workshop attempted to explore a large variety of networks, this would not have been possible. Perhaps the most important and exciting talk of the entire conference was given by Martin Feinberg (The Ohio State University). Professor Feinberg has pioneered the development and importance of structure theorems in chemical reaction networks. This talk, more than any other, demonstrated the power of the structure theorems approach in understanding dynamic phenomena in systems biology. The talk presented new theorems and classification of enzyme networks and bistability. The work demonstrates yet again that stochasticity, while often invoked in ignorance (not intended as a pun), is not the only extragenic mechanism at work that can account for heterogeneous behavior. This talk solidified in the minds of biologists the tools that mathematicians can provide, and it did so within a biological context easily understandable to them.

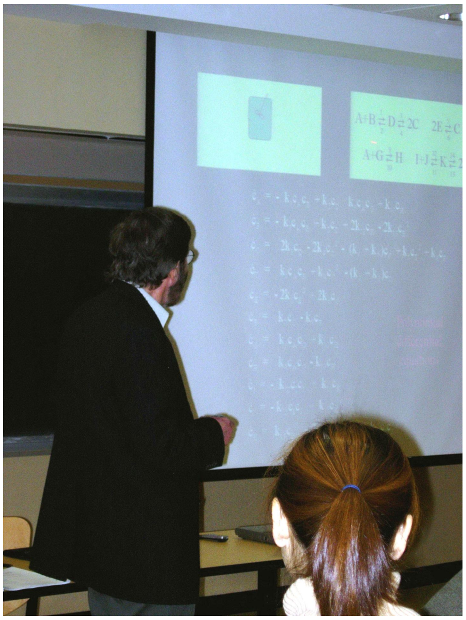
The first talk of the afternoon was presented by physicist Jon Lorsch (Johns Hopkins University School of Medicine), who detailed the molecular complexities involved in translational regulation and his attempts to model this process. This talk was well received and generated a tangible interaction and collaboration. Through interactions between Professor Lorsch, his graduate student, the organizers, and mathematician John Mallet-Paret, a quantitative model was developed that includes a state dependent delay. This model, that involves regulation by the gene GCN4 that is closely tied to nitrogen regulation, is the first of its

A model of nitrogen catabolite
repression gene regulatory circuit.

kind and its detailed mathematical analysis through this collaboration is a great indication of the success of the workshop. Day 1 ended with an important talk by biologist George Marzluf (The Ohio State University). This talk showed that nitrogen regulation in closely related fungi is handled by some of the same genes but in a very different network configuration. This illustrates perhaps the most important systems biology question that was discussed at this workshop.

NCR-Circuit
Dynamics and ?
Erik Boczko – Vanderbilt
Terrence Cooper – UT Memphis
Tomas Gedeon – Montana State
Konstantin Mischaikow – Georgia Tech
Sam Wells – Vanderbilt
Sidd Pratap

Day 2 began with a talk by Tomas Gedeon (Montana State University). This talk and a later one by Professor Sontag demonstrated that monotone systems theory is a powerful tool that allows one to prove rigorous results about complex biological systems of differential and delay differential equations. However, as shown by this talk, new extensions of the theory are required and gene regulatory networks are to be generally understood. This talk showed that extensions to periodically forced systems are essential, and it further introduced the important concept of a stackable system or property. The work of Leon Glass (McGill University), and later Professor Mahaffy, showed excellent examples of a pervasive problem in the field of gene regulatory networks so that much of current biologically oriented modeling work is ignorant of the large mathematical literature that extends back at least 30 years. The work presented dates back to a 1978 article on stable oscillations, where the authors proved an important structure theorem, and this work is now experiencing a huge revival. For instance, the recent work of Ben-Hur and Siegelman show that the model introduced by Professor Glass can be viewed as a memory bounded Turing machine and has applications in theoretical computer science. In the final morning talk, John Mallet-Paret (Brown University) noted that an intriguing mathematical observation from singular perturbation studies of state dependent delay equations is the existence of super stable solutions. His talk described his current work on state dependent delays and explored the possibility that their properties are likely to be exploited by natural systems. This idea was later explored in the workshop in the particular context of transitional control by the gene GCN4.



Eduardo Sontag (Rutgers University) began the afternoon with a talk about an important result in monotone systems called the small gain theorem and

its application to real signaling networks. The power of this approach is that with some simple steady state assumptions, an experimentally or theoretically calculated characteristic can be used to fully describe the dynamics of a given network. The speaker gave a fast paced and exciting talk. Finally, Martin Golubitsky (University of Houston) discussed a different and beautiful structure theory. How much dynamics is determined solely by the structure of the connections? The theory of coupled cell systems was described; whose central concept is group (oid) equivariance. There is some evidence to indicate that oscillatory solutions seen in a synthetic gene network arises in analogy with a rotating wave solution seen in a coupled three-cell network.

Day 3 began with a systems biology talk by Natal A.W. van Riel (Eindhoven University of Technology), who described mathematical modeling of nitrogen metabolism of continuous culture. The work is an excellent example of quantitative modeling and the interplay between experiment and theory. This is precisely the kind of work that brought together the themes of interest in this workshop. In an impromptu talk, Reka Albert (Pennsylvania State University) described recent advances at understanding pattern formation in Boolean network models of the Drosphilia segment polarity network. The result has emerged that there is a central circuit motif that appears to be essential for the proper pattern formation. Jason Lowry (University of Sydney) demonstrated once again the fascinating observation that closely related fungi use different complements of genes to regulate nitrogen metabolism. For instance, it was shown that many species of yeast and fungi do not have a DEH1 analog. This gene participates in the NCR circuit but its role in the dynamics remains a mystery.

In the afternoon, Joseph Mahaffy (San Diego State University) described half a career's worth of mathematical biology of the highest caliber, most of which has focused on genetic networks and has remained largely unrecognized by contemporary gene network modelers. Some of the talk was devoted to a discussion of the replication cycle of E. coli, and it was shown that the modeling uncovered a key component of the network that had previously been unknown. This work together with the seminal work of Mackey and Glass on cyclic neutropenia and the more recent work by Jens Timmer and colleagues on Stat5 signaling stand as examples of what can be achieved through modeling. The final talk of the workshop was one of the best. One of the most important and contentious topics in networks is the question of noise, its origins, and how to model it. Sebastian Schreiber (The College of William and Mary) described an approach where these questions can be precisely formulated and rigorously answered.

Conclusion

Finally, the organizers wish to express their appreciation to the MBI for hosting this workshop. Everything at the MBI was done to maximize the effectiveness and productivity of the event.

Workshop 3: Computational Proteomics and Mass Spectrometry January 11-14, 2005

Organizers:
Vineet Bafna, Computer Science and Engineering, University of California, San Diego
Tim Ting Chen, Departments of Biology, Computer Science, and Mathematics, University of Southern California

Overall Summary

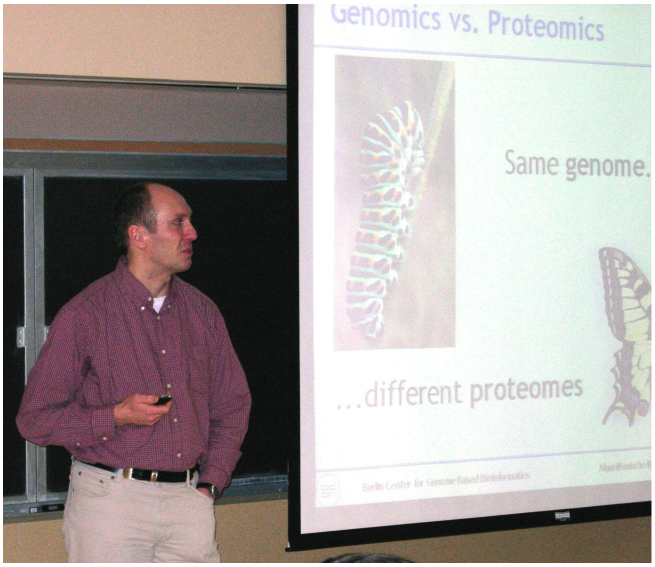
Proteomics—defined as a systematic investigation of the total protein complements of a cell—is a broad term that covers a lot of ground including, but not limited to, protein identification and quantification in specific cellular environments, structural genomics and fold recognition, identification and characterization of functional domains, and finally, the networks defining the interactions of proteins with biomolecules (proteins, DNA, etc.). This broad definition ensures an abundance of meetings and workshops devoted to this theme. Within this larger context, the Ohio workshop stood out by emphasizing a few well chosen themes and excellent presentations by experts from industry and Academia.

The workshop focused on computational analysis of mass spectrometry data and its applicability to broader proteomic analysis. Simply speaking, a mass spectrum is a collection of masses and (relative) intensities of charged molecules. The spectrum of mass fragments of a protein (or peptide) sequence form a fingerprint that can be used for identification and relative quantification. Post translational modifications can be measured using characteristic shifts in the spectrum. Various computational issues arise in the analysis of mass spectrometry data for protein identification and quantification. Until recently, the algorithms for such analysis were tied in to the specifics of the instrumentation, and the data generated was not generally available to computational scientists: these made it difficult to abstract and formulate problems and exchange algorithmic ideas for data analysis. The situation has changed dramatically in the last few years with a large number of data sets now freely available. Not surprisingly, it has led to a number of publications with novel algorithmic improvements. The Ohio workshop brought together many of these researchers.

Summary of Talks and Presentations

The first day of the conference was devoted to introductory talks on the use of mass spectrometry and related technologies for data analysis. The first talk was given by Scott Patterson (Amgen), who is a pioneer in the area. He started with a general overview of the area and differentiated it from the measurement of gene expression. As he pointed out, Parallel protein measurements, a.k.a. proteomics, have the potential to provide information on biological systems in isolation as cell culture systems, tissues, or in an organism. Whereas parallel measures of transcript (mRNA) abundance can be multiplexed more easily through microarray analysis of even small quantities of sample following amplification using PCR, parallel measures of protein abundance are more difficult due to the heterogeneity of protein properties compared with nucleic acids, and the inability to amplify the signal. Therefore, while much useful data can be generated, the (computational) interpretation of such data sets is challenging. His talk laid out challenges for the computational community, and set the tone for the rest of the workshop.

Peter de B. Harrington (Ohio University) focused on the issues of experimental design in analyzing MS data. He started with the application of MS technologies in identifying biomarkers and making predictions from noninvasive samples. Next, he described the use of fuzzy classification and rule based systems in identifying protein biomarkers. Spectra from studies of amniotic fluids from women who had normal, normal with inflamed uteri, and premature delivery were used for building classification models. The fuzzy classification technology, coupled with the Latin-partition method based experimental design helped obtain precise bounds. His talk was followed by his collaborator, Alfred Yergey (NIH), who also demonstrated the power of a rational experimental design applied to preliminary experiments directed towards discovery of biomarkers in amniotic fluid. He showed that combining analysis of variance with principal component analysis (ANOVA/PCA) provides a powerful tool for the discovery of biomarkers in chemical measurements of biological systems. The last talk of the day was given by Benno Schwikowski (Institut Pasteur). He

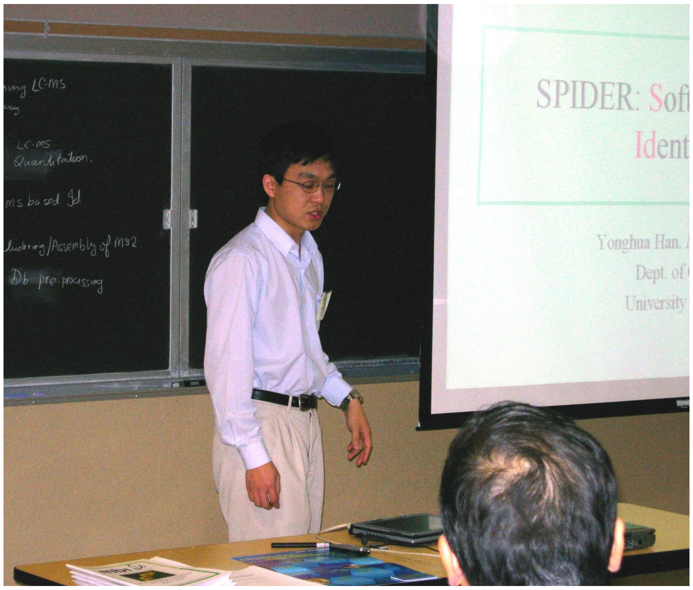


talked about algorithms for protein quantification using MS data, a collaboration with Amol Prakash, and the research groups of Professors Aebersold and Pavlovitch in Seattle. In his approach, all data acquired across a whole experiment are first aligned into an *n+1*-dimensional space, where *n* is the number of dimensions used for the LC separation. This condenses all peaks generated by the same protein fragment throughout the experiment into a single dense signal, which allows a much better separation of signal and noise. He presented algorithms that addressed the main computational challenge, which is to compensate for fluctuations in the separation process. Taken together, the talks on the first day describe the various applications of MS data to proteomics. The next 2 days were dedicated to a discussion of the underlying algorithms.

The first two talks on Day 2 of the workshop were devoted to exploration of MS data for peptide quantification. Oliver Kohlbacher (Tuebingen) and Knut Reinert (Freie Univ., Berlin) described the algorithms underlying their open source software library for Mass Spectrometry called OpenMS. Even as a black box, OpenMS is a tremendous asset for researchers as it allows for interpretation and visualization of data from different instruments. However, the real value of the software is the clear description of the underlying algorithms for MS signal processing and benchmark comparisons with instrument software. This separation of the instrument and data will make it possible to combine results from different instruments and greatly spur research in the area. Professor Reinert also presented a case study in the use of OpenMS in development of a standard protocol. The use of MS data in peptide quantification is a nascent but important research area, and OpenMS will greatly spur research in the area. Next, Nathan Edwards (Univ. Maryland) presented on a topic that straddles Mass Spectrometry and traditional Bioinformatics. He described an algorithm for removing peptide redundancy in a sequence database, as a means for improving MS2 throughput. The construction, based on a de Bruijn graph representation of sequence overlaps is elegant and provides a practically useful tool.

Continuing with the theme of nontraditional MS analysis, Nuno Bandeira (UCSD) presented his research on “shotgun protein sequencing”, the name being derived as a tribute to the successful strategy for sequencing genomes. All current techniques for protein identification rely upon identification of peptides. Instead, Nuno presented a framework for sequencing the entire protein by starting with tandem MS from a nonspecific digestion of the protein and subsequent clustering, overlapping, and assembly of spectra. His research provides a viable alternative to sequencing proteins when the corresponding genomic sequence is not known. More important, it provides a systematic algorithmic treatment by abstracting the problem into different modules. This opens up a new area for algorithm development. Also, the modules for spectral clustering are likely to be of independent interest.

Finally, Alexey Nesvizhskii (Institute for Systems Biology, Seattle) spoke on the importance of post-processing output from tandem MS database search and on various statistical measures and approaches for estimating the confidence level of peptide identifications, including p-values, expectation values, reverse database searching, and the Bayesian classification. He compared these approaches to methods developed for the analysis of other types of data such as microarray gene expression. Next, he addressed the problem of inferring proteins the original sample, based on peptide identifications. He explained a statistical model for assembling peptides into proteins paying special attention to the problem of nonrandom grouping of peptides according to their corresponding proteins ('single hit' identification problem). He concluded with a description of a new project involving the annotation of genomic sequence with mass spectra.



Day 3 of the workshop was devoted to algorithms for MS2 based peptide identification. This is the single most researched area in computational mass spectrometry. However, algorithms for identifying modifications and mutated peptides are still under development. The last 2 years have seen rapid development in the area, and also the amalgamation of two, previously distinct, lines of research in MS2 identification: de novo sequencing and database search. This workshop summarized the most interesting lines of research in this area.

The first talk was offered by Ari Frank (UC San Diego) on de novo peptide sequencing. Like other researchers, he started with the notion of a spectral graph. This is a structure in which the traversal on any path constitutes a peptide interpretation of the spectrum. However, Ari greatly improved upon the state of the art by describing novel approaches to score paths so that the correct interpretation is the top scoring one. His approach was based on a Bayesian network model of fragmentation, which was trained using a large data-set describing fragmentation propensity. His results demonstrated superiority over existing tools and showed that de novo sequencing can often compete with the database search approaches for high quality spectra.

The talk was followed by Vineet Bafna (UC San Diego), who showed that de novo sequencing and database search, when combined, provide an effective tool for identifying post-translational modifications. Such modifications are a key component of cellular processes, and while mass spectrometry has the potential to identify such modifications, this has been difficult to realize in practice, often because of a computational bottleneck in exploring all combinatorial possibilities. This talk presented an approach in which de novo sequence analysis was used to identify tags that can be used to efficiently filter databases, giving more time for a combinatorial exploration of modifications in the remaining database. The talks by Bin Ma (Univ. Western Ontario) explored a related theme, in which only a modified/mutated form of the true peptide is available in the database. The goal is to find a peptide that matches both the spectrum, and some database peptide with appropriate mutations. His tool, Spider, has been used to identify a number of novel modifications.

Brian Searle (Proteome Software Inc.) addressed the interesting notion that different search engines often identify different peptides, and the problem of how to best combine the search results, so as to maximize peptide identifications. His approach, based on a novel application of Peptide Prophet (Nesvizhskii) ideas, showed that a surprisingly large fraction of spectra can be reliably identified using this idea.

The last two talks of the day were devoted to improving peptide identifications based on scoring. Tim Chen modeled MS2 spectra using an HMM and showed that incorporation of various features greatly improved peptide identification. Rovshan Sadygov (ThermoElectron Corp.) also described the pitfalls in scoring for low intensity spectra. He described a probability model for two of the parameters that affect the quality of peptide identification the most: the number of product ion matches and the sum of the product ion abundances. The probabilities obtained from each model are correlated and normalized to derive a single score: significance of peptide identification.

On the last day, Frederic Schutz's (Walter and Eliza Hall Institute of Medical Research) talk continued the theme of scoring. He described results on comparing various database search tools on a large data-set. The results are interesting, and in line with common knowledge: the intersection of different results is a very large subset, but the union is significantly more sensitive than any single search result. In general, Mascot was found to be more specific, while Sequest was the more sensitive tool. In his own research, he also described a statistical model for modeling mass spectra and showed improvements in scoring.

The other two talks of the day were offered by Fengzhu Sun (USC) and Sebastian

Bocker (Bielfeld), and covered novel themes. Dr. Sun's talk was based on the idea of combining large amounts of biological data (including, but not limited to mass spectrometry data), and its applicability to studying protein interaction networks. They combine features from these data sets using Markov random fields, and further study the relationship between gene lethality, protein interaction networks, and protein function annotation.

Bocker's talk was on mass spectrometry of DNA. Starting with compositional data from MS experiments, he addresses the problem of inferring DNA sequence. Thus, the problem leads to the study of weighted strings and compomers: A string's compomer is an integer vector specifying the number of occurrences of each character. He described algorithms for determining all or some compomers with a given mass, the number of such compomers, and related questions.

Conclusion

The workshop emphasized specific themes, which led to a very high level of interaction during, and between the talks. An informal poll showed that while the researchers were familiar with each other's work, they came from very different communities and had not met in person. Therefore, the workshop enabled them to initiate discussions and collaborations, and there was great enthusiasm for meeting on a regular basis. The MBI is to be commended for providing excellent facilities, and for enabling a relatively relaxed schedule that encouraged discussion and audience participation. Finally, the staff at the MBI worked very hard to cater to various requests, and making the organizers task simple.

Workshop 4: Emerging Genomic Technologies and Data Integration Problems

February 21-24, 2005

Organizers:
Terry Speed, Department of Genetics and Bioinformatics, University of California , Berkeley
Hongyu Zhao, Department of Epidemiology and Public Health, Yale University School of Medicine

Overall Summary

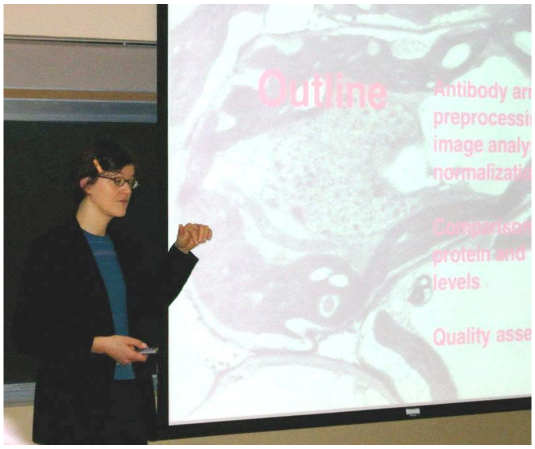
The purpose of Workshop 4 was, loosely speaking, to address some of the mathematical, statistical, and computational problems that arise on genomic technologies not discussed elsewhere in the year-long program, particularly promising novel ones, and the questions arising in integrating data from different technologies.

Summary of Talks

Day 1: Data Integration and Transcriptional Regulation:

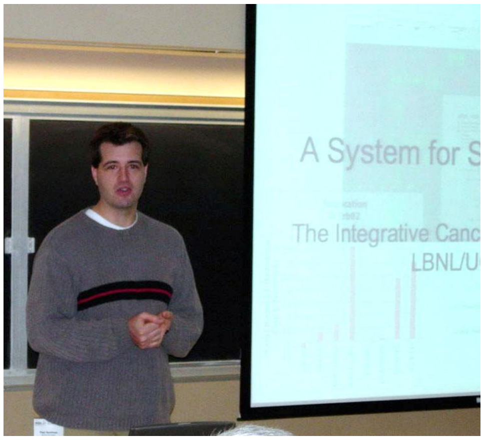
The conference started with two talks on data integration. It was opened with an exciting talk by Eric Schadt of Rosetta Inpharmatics (a subsidiary of Merck). Researchers at Rosetta have been leaders in the creative use of microarray technologies, and Eric spoke on the initial analyses of one of their more ambitious projects: the combining of genotypic and gene expression data in segregating mouse populations. Here, mice from a traditional F2 intercross between two inbred strains are genotyped at a dense set of markers, permitting the linkage mapping of any quantitative trait whose distribution differs between the parent lines. Data on the expression levels of thousands of genes is then obtained for each of the F2 mice by hybridising mRNA from suitable tissues (here their livers) to microarrays. The novel idea is that each gene expression level may then be regarded as a quantitative trait and mapped to one of more genomic locations, giving rise to thousands of so-called eQTLs; e for (gene) expression. Many of these eQTL are in cis relative to the gene whose expression levels are being measured, but many are not. Many such loci clusters and the wealth of data offer a wide range of analytical challenges. He demonstrated conclusively that this integration of microarray, genetic, and clinical data were a very powerful approach for directly identifying genes underlying QTLs.

The second talk of the conference was by Ning Sun (Yale University). She works with a research team that is developing methods to integrate gene expression data, in vivo DNA-protein data, protein-protein interaction data, and genomic sequence data. This is an ambitious goal and, in her talk, Dr. Sun presented a measurement error model, which permitted the integration of DNA-protein binding data with gene expression data form microarrays. The model was described and its performance with simulated was data discussed. Finally, it was applied to the well-known yeast cell-cycle data set of P. Spellman and colleagues.



The afternoon session was devoted to two closely related talks on transcriptional regulation. There are a number of inter-related problems here. One is to identify the DNA sequence motifs—or clusters of motifs—to which transcription factors (TFs) bind. Another is to identify the genes regulated by a given TF. A third is to understand the transcriptional regulation of a given gene: which factors bind, and under what conditions, to promote its expression.

The afternoon began with a talk by Xiaole (Shirley) Liu (Harvard University). She outlined the analytical problems that arise when we are presented with genome-wide information on the binding of a given TF obtained from so-called ChIP-chip data, that is, data from Chromatin ImmunoPrecipitation followed by microarray (chip) experiments. Such data permit the unbiased mapping of TF binding sites, and are bringing in a new era in our understanding of transcriptional regulation. The microarrays, which permit such data to be collected, are not the familiar gene expression microarrays, but novel, not yet widely available tiling arrays, which have hybridisation probes uniformly and densely covering entire genomes. Professor Liu and colleagues developed a fast method to identify putative binding sites from data on such genome tiling arrays, which had at its core a hidden Markov model. She described the results of applying her method to recent published and unpublished data sets, and summarized some data validating her findings.



Ramana Davuluri (The Ohio State University) continued the same theme, though he had different ChIP-chip data, and his focus was different. His data was from a microarray whose probes were from ~9000 GC-rich regions previously shown to be preferentially located at the 5'-end of genes (and so near regulatory regions), and his aim was to identify genes directly or indirectly regulated by the TF known as the Estrogen Receptor a (ERa). He also made use of data from mouse, performing what has come to be known as a comparative genomic analysis. What Professor Davaluri sought was an algorithm that would predict direct and indirect targets of ERa, and to do this he utilized the method known as CART (Classification and Regression Trees) developed by

Breiman and others. He described his results and the follow-up validation.

Day 2: Protein interactions and SNPs and Chips:

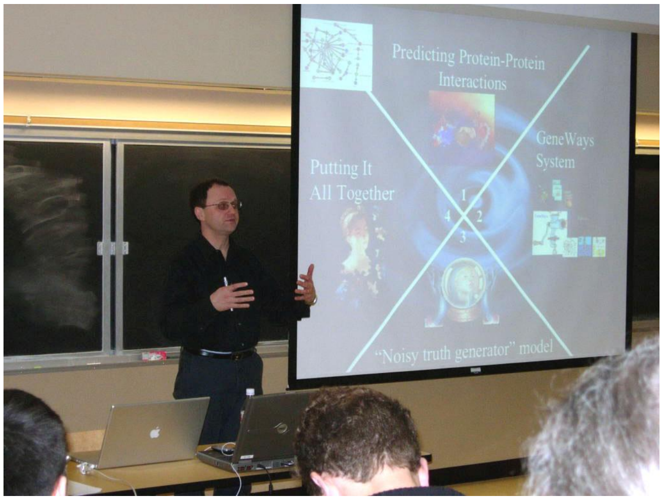
Prior to his joining Johns Hopkins University, Joel Bader worked in the research group at the company CuraGen Inc., which produced a genome-wide protein interaction map of the proteome of the fly Drosophila melanogaster. This impressive map was based on the relatively recently developed yeast two-hybrid assay. Professor Bader began by explaining how the assay worked and then took us through some of the statistical modeling he carried out on the data. He described two different levels of organization, and how the network recapitulated known pathways, extended pathways, and uncovered previously unknown pathway components. Finally, he discussed how maps such as the one he worked on are a starting point for systems biology modeling of multicellular organisms.

Andre Rzhetsky (Columbia University) continued the theme of protein interactions, but with a difference: his approach was through searching the biomedical literature. His system GeneWays automates the selection of articles in molecular biology, and the extraction of and visualization of information, aimed at achieving a consensus view of molecular networks. We were shown an impressive variety of models for these tasks: some highly creative representations of knowledge, and some serious testing of his system.

The final morning talk on protein interactions was by Amy Keating (MIT). She described special protein microarrays designed to identify the interaction specificities of what are known as bZIP transcription factors. Her analysis of the microarray data was combined with machine learning techniques (especially support vector machines) to predict bZIP interaction preferences, and she foreshadowed future research combining these analyses with those involving physical modeling of protein structure.

The afternoon of Day 2 saw analysis challenges from two scientists working in companies producing high-throughput genotyping assays. Fiona Hyland (AppliedBiosystems) presented an overview of her company's platform, and then presented data in which samples of 45 individuals from each 4 different populations were genotyped, to provide information for the selection of so-called tagging Single

Nucleotide Polymorphisms (tSNPs), these being SNPs, which help characterize haplotypes efficiently. The challenge was to select tSNPs, which can be effective across different populations.


Predicting Protein-Protein Interactions
GeneWays System
Putting It All Together
1
2
3
4
"Noisy truth generator" model

Earl Hubbell (Affymetrix) began by outlining the whole genome sampling assay used by his company to generate data for 10K or 100K SNPs. He then reviewed three algorithms that have been used by Affymetrix to call genotypes: first MPAM, then DM, and most recently his own AAFYnity model. The context in which the models were developed was then reviewed, their performance described, and future needs outlined.

Day 3: Novel High-Throughput technologies:

Julia Brettschneider (UC Berkeley) began the 2 days devoted to analysis issues arising with novel technologies. She described her experience with a form of protein array (manufactured by BD Biosciences) in which several hundred antibodies are attached to a glass slide in an array format, and probed with proteins extracted from cell samples. In her case, the proteins were from post-mortem human brain tissue, and she also had Affymetrix microarray measurements on mRNA expression levels from the same samples. After describing the analysis methods she devised for the antibody array, Dr. Brettschneider compared measured protein expression levels with the corresponding mRNA measurements from the microarrays. The agreement was very poor, and she carefully considered a number of possible explanations for this observation.

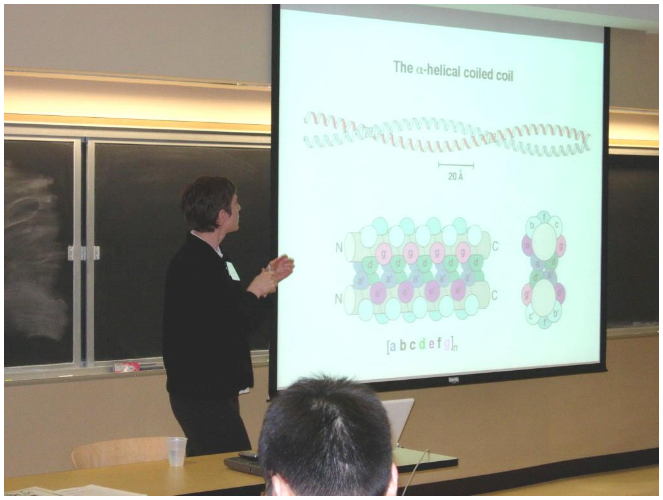
In the second talk, Martha Bulyk (Harvard University) described a novel technology termed protein-bending microarrays (PBMs), which permit high-throughput characterization of the in vitro DNA binding site sequence specificities of transcription factors. She compared her results with those from in vivo experiments, and noted several substantial differences, undoubtedly deriving from the fact that in vivo experiments can only sample a limited number of the contexts in which a TF is active. She concluded that PBMs show promise in elucidating transcriptional regulatory networks.

Joakim Lundeberg comes from the Department of Biotechnology of the Royal Institute of Stockholm, which is one of the pioneering proteomics groups in Europe. He outlined the work of their Center, and then described their affinity proteomics strategy for profiling gene products in human tissues and the associated analysis challenges. Along the way, Dr. Lundeberg described a special project in which he was involved, concerning the evolutionary history of the Australian dingo.

Michael Uhler (University of Michigan) has developed a novel microarray-based transfection method, which permits high-throughput experimental verification of potential transcriptional regulatory mechanisms, such as those determined by bioinformatic studies. This method, termed STEP (Surface Transfection and Expression Protocol) uses engineered recombinant proteins spotted onto microscope slides. After explaining his method, Professor Uhler summarized several analytical challenges which arise with his arrays.

Day 4: Novel High-Throughput Technologies (cont.):
In recent years, much interest has been paid to proteomic profiling seeking to identify differences between diseased and healthy tissue samples, typically blood. Keith Baggerly (MD Anderson Cancer Center) gave us a fascinating survey of the problems and pitfalls associated with these studies using MALDI and SELDI technologies. His basic conclusions were simple and universal: design matters, dealing adequately with uncontrollable variation are important, and validation is essential.

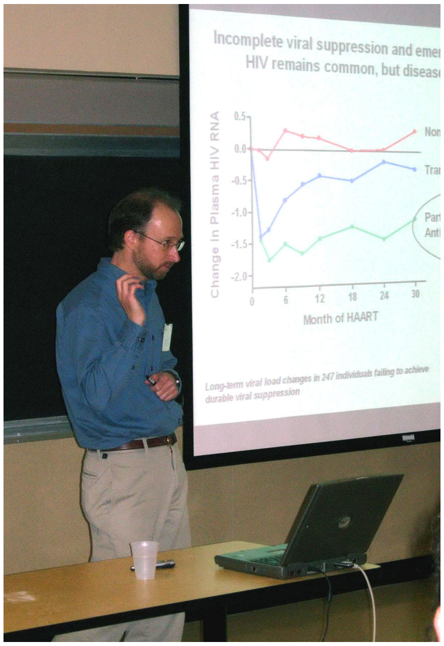
Paul Spellman (Lawrence Berkeley National Laboratory) gave us a clear introduction to efforts there to develop a systems approach to understanding breast cancer using cell lines. The group there plans to use a variety of novel high-throughput genomic and proteomic technologies, including Affymetrix High-Throughput Assays (HTA), these being 96 Affymetrix chips in one assay, reverse-phase protein lysate assays for measuring absolute protein abundance, DNA copy number analysis using Comparative Genomic Hybridisation (CGH), SNP genotyping, and DNA sequencing. The challenges in integrating these data in a systems model were presented, and we were told that the MSRI in Berkeley will be holding an annual meeting to discuss these data and the methods for their analysis.



The last talk in the workshop was by Professor Steve Horvath of UCLA, who discussed tissue microarrays (TMAs). These are new, moderately high-throughput tools for the study of protein expression patterns, and are increasingly used to evaluate the diagnostic and prognostic importance of tumor biomarkers. He described the use of tree-based methods such as CART and random forests for relating immunohistochemistry data obtained on the arrayed samples to survival. In passing, he described a good number of analytical challenges arising in the routine use of tissue microarrays.

Conclusion

On two occasions during the workshop we held panel question/answer sessions, which were well attended and productive. As there were experts presenting a wide variety of technologies at the workshop, there were many participants who were wholly unaware of some, and hence a lot of opportunity for meeting new people and learning about novel technologies, and the analytical challenges they pose. Many participants commented on how valuable they found this aspect of the workshop. The workshop facilities were excellent and everything went very smoothly, so the MBI staff is very much to be thanked for their excellent organization and good spirit.


Incomplete viral suppression and emer
HIV remains common, but diseas
Change in Plasma HIV RNA
0.5
0.0
-0.5
-1.0
-1.5
-2.0
0
6
12
18
24
30
Month of HAART
Non
Tra
Part
Anti
Long-term viral load changes in 247 individuals failing to achieve durable viral suppression

Workshop 5 - Part 1: Biomarkers in HIV
April 18-19, 2005

Organizers:
Victor De Grutolla, Department of Biostatistics, Harvard School of Public Health
Mark Segal, Department of Biostatistics, University of California, San Francisco
Alan Perelson, Los Alamos National Laboratory

Overall Summary

The first part of the workshop addressed the medical application of new scientific technologies, such as PCR and genetic sequencing to research on, and treatment of HIV infection. The medical applications included use of measurements based on these technologies as biomarkers for assessing disease progression and effects of antiviral treatment as well as for treatment selection. As an example of how these technologies are influencing medical practice, several participants noted that HIV gene sequencing is now used to evaluate drug susceptibility and select treatment regimens for drug-experienced patients. PCR technology has made it possible to count HIV-RNA particles in body compartments, which allows evaluation of drug efficacy in suppressing the virus in plasma or in genital secretions. Presenters also noted that modeling of HIV dynamics, made possible because of the accuracy of PCR measurements, provide insight into the mechanisms of drug action. In addition to viral genomics, human genomics is also a developing area of research. In particular, there is interest in determining whether polymorphisms in specific host genes explain patient variability in treatment response, toxicity, and pharmacokinetics of antiretroviral drugs.

The sessions included methods for relating HIV genotype to resistance phenotype; methods for modeling the accumulation of HIV resistance mutations; and relationship of host genomics to treatment response, toxicity, and pharmacokinetics of ARV therapy. This workshop also highlighted the statistical challenges involved in the areas of HIV and medical practice, presented statistical research in progress, and provided a forum for discussing current answers to the statistical challenges and future directions.

Summary of Talks

The first day of the conference was devoted to HIV dynamics and modeling. Steven Deeks (University of California San Francisco Medical School) started the meeting by presenting an overview of HIV infection and treatment, emphasizing problems of drug resistant virus. Dr. Deeks noted that many patients treated with combination antiretroviral therapy fail to achieve complete viral suppression, but treatment may nonetheless provide clinical benefit in such patients. Optimizing individual treatment strategies is challenging, however, in part because it requires an understanding of the complex relationship between replication of a drug-resistant virus and the host response. In particular, the distinction between persistent drug activity, alterations in replicative capacity ("fitness"), and the ability of a newly emergent variant to cause disease ("virulence") may prove to be important in designing long-term therapeutic strategies. These issues will likely become even more relevant with entry inhibitors, where drug-pressure may select for X4 variants that may be less fit but more virulent. To address these issues, Dr. Deeks and colleagues have performed a series of studies focusing on the determinants of disease outcome in patients with drug-resistant viremia, and have observed the following: (1) HIV is often constrained in its ability to develop high-level drug resistance while maintaining replicative capacity; (2) immune activation is reduced in patients with drug-resistant HIV (after controlling for the level of viremia); and (3) patients who durably control HIV replication despite the presence of drug-resistance exhibit immunologic characteristics comparable to that observed in long-term nonprogressors (e.g., low levels of T-cell proliferation and activation, and preserved HIV-specific IL-2 and gamma-interferon-high producing CD4+ T-cells). Dr. Deeks went on to describe the interventional studies he has designed to investigate the hypothe-

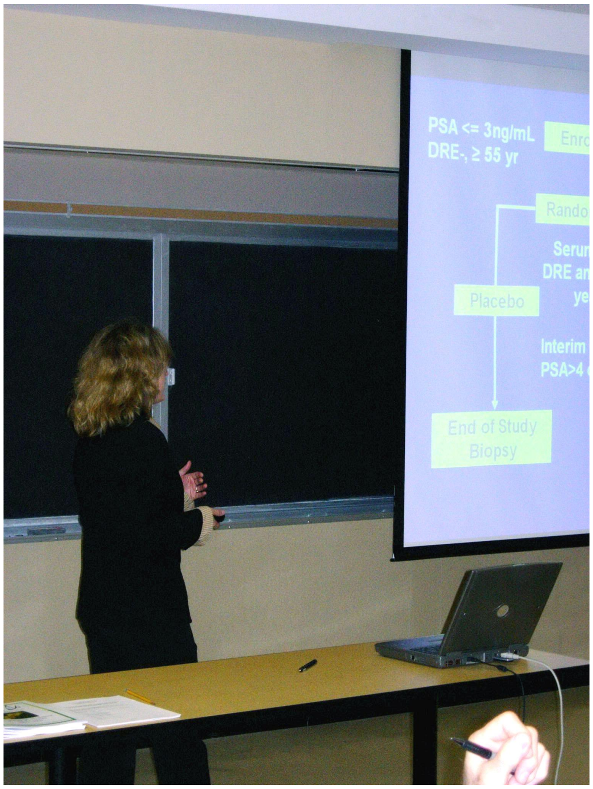
sis that drug-mediated alterations in HIV fitness/virulence may be clinically useful in patients with limited therapeutic options.

Alan Perelson (Los Alamos National Laboratory) a pioneer in mathematically modeling the dynamics of HIV infection, reviewed the current state of modeling HIV with emphasis on a new class of models that incorporate pharmacokinetic and pharmacodynamic information. Previous HIV models simply assumed that drug was present and has a given, fixed efficacy, often 100%. In these new models, drug pharmacokinetics is used to establish how the concentration of drug changes with time after administration and then the models relate drug concentration to antiviral efficacy. These new models were then used to interpret data from HIV-HCV co-infected patients treated with pegylated interferon, in which the plasma drug level was measured along with viral load very frequently for the first few weeks of treatment. The new models could explain both drops and rebounds in viral levels, with the rebounds occurring when plasma drug levels fell. Issues of the best ways to estimate parameters in these more complex models were discussed.

Dr. Perelson's talk was followed by a presentation from Hulin Wu (University of Rochester), who continued on the theme of incorporating pharmacokinetic (PK) and pharmacodynamic information into HIV models. Professor Wu discussed a large clinical trial where viral load, CD4 cell count, drug adherence, drug resistance, and drug pharmacokinetic parameters were measured for each patient. Using an Emax model for drug efficacy that included adherence information, he showed how one could use a hierarchical Bayesian modeling approach to analyze the data and extract model parameters. No single PK parameter was significantly related to a virological response, but by including drug susceptibility (IC50), or IC50 and adherence together, C_trough, C_12h, C_max, and AUC_0-12h were each significantly correlated to long-term virologic response. Adherence measured by pill counts and multiple trough drug concentrations did not provide additional information for virologic response presumably due to the data quality and noise problems. He concluded that HIV dynamic modeling is a powerful tool to establish a PD relationship and correlate other factors such as adherence and drug susceptibility to long-term virologic response, since it can appropriately capture the complicated nonlinear relationships and interactions among multiple covariates.

Mark van der Laan (UC Berkeley) then spoke about methods to interpreting HIV mutations and using gene sequences to predict response to antiretroviral therapy, using the deletion/substitution/addition (DSA) algorithm for the estimation of direct causal effects. The goal of his talk was to estimate the causal effect of mutations detected in the HIV strains infecting a patient on clinical virologic response to specific antiretroviral drugs and drug combinations. He considered the following data structure: (1) viral genotype, summarized as the presence or absence of each viral mutation considered by the Stanford HIV Database as likely to have some effect on virologic response to antiretroviral therapy; (2) drug regimen initiated following assessment of viral genotype (the regimen may involve changing some or all of the drugs in a patient's previous regimen); and (3) change in plasma HIV-RNA level (viral load) over baseline at 12 and 24 weeks after starting this regimen. The effects of a set of mutations on virologic response are heavily confounded by past treatment. In addition, viral mutation profiles are often used by physicians to make treatment choices. This confounding needed to be addressed, because he was interested in the direct causal effect of mutations on virologic outcome, not mediated by choice of other drugs in a patient's regimen. Finally, the need to consider multiple mutations and treatment history variables, as well as multiway interactions among these variables, results in a high-dimensional modeling problem. This application thus requires data-adaptive estimation of the direct causal effect of a set of mutations on viral load under a particular drug, controlling for confounding and blocking the effect the mutations have on the assignment of other drugs. The algorithm developed by Dr. van der Laan was based on a mix of the direct effect causal inference framework and the data adaptive regression deletion/substitution/addition (DSA) algorithm.

Rodolphe Thiébaut (Université Victor Segalen) spoke about issues in longitudinal modeling of HIV markers using mixed models. As noted in the introduction, plasma HIV-RNA and T-lymphocytes CD4+ count are major biomarkers used to decide when to start, change, or stop a treatment as well as to evaluate treatment efficacy in HIV-infected patients. Thus, repeated measurements of those biomarkers are commonly collected in HIV studies. Those data may be analyzed by using models for longitudinal data such as mixed models. However, the statistical analysis is complicated by several methodological difficulties. Three of them are of particular importance: (1) left-censoring of HIV-RNA due to a lower quantification limit; (2) correlation between CD4+ T lymphocytes and plasma HIV RNA; and (3) missing data due to informative dropout or disease progression. Dr. Thiebaut presents a unified approach to deal with those issues by jointly modeling longitudinal measurement data and event history data. Likelihood inference was used to estimate the parameters of such a model. He illustrated the model by studying HIV markers response to antiretroviral treatment in randomized clinical trials and observational cohort studies. This approach might help in studying the change in markers, their prognostic value, and their surrogacy.



Victor De Gruttola (Harvard School of Public Health) spoke about joint modeling of progression of HIV resistance mutations measured with uncertainty and time to virological failure. Development of HIV resistance mutations is a major cause for failure of antiretroviral treatment. He and colleague Chengcheng Hu proposed a method for jointly modeling the processes of viral genetic changes and treatment failure. Because the viral genome is measured with uncertainty, a hidden Markov model was used to fit the viral genetic process. The uncertain viral genotype was included as a time-dependent covariate in a Cox model for failure time, and an EM algorithm is used to estimate the model parameters. This model allowed simultaneous evaluation of the sequencing uncertainty and the effect of resistance mutation on the risk of virological failure. The method was then applied to data collected in three phase II clinical trials testing antiretroviral treatments containing the drug efavirenz. Various model checking tests are provided to assess the appropriateness of the model.

Day 2 began with two talks regarding biomarker issues in vaccine development and evaluation. First, Mark Segal (University of California at San Francisco) discussed prediction of HIV-1 epitopes using amino acid sequences of MHC binding peptides. Dr. Segal began by reviewing some basic biology of HIV infection. He noted that following infection, HIV-1 proteins are digested into short peptides that bind to major histocompatibility complex (MHC) molecules. Subsequently, these bound complexes are displayed by antigen presenting cells. T-cells with receptors that recognize the complexes are activated, triggering an immune response. Peptides with this ability to induce T-cell response are called T-cell epitopes; prediction of which peptides are epitopes is therefore important for vaccine development. Sung and Simon (JCB, 2004) start with compilations of peptide sequences that either do or do not bind to specific MHC molecules. Using biophysical properties of the constituent amino acids, they develop a classifier. Properties are used because of the inability of select classifiers to effectively handle amino acid sequence itself. Tree-structured methods are not so limited (Segal et al., Biometrics, 2001). Dr. Segal applied these methods, along with their ensemble extensions (bagging, boosting, and random forests), and showed they provide improved accuracy. Both additional properties (QSAR derived) and classifiers (SVMs, ANNs) are also investigated. HIV-1 genomewide comparisons with respect to predicted / conserved epitopes were also presented.

Betz Halloran (Emory University) switched the focus from HIV infection to influenza in a presentation regarding the use of validation sets for outcomes with time-to-event data in vaccine studies. In many vaccine studies, confirmatory diagnosis of a suspected case is made by doing a culture to confirm that the infectious agent of interest is present. However, often such cultures are too expensive or difficult to collect, so that an operational case definition, such as "any respiratory illness", is used. This leads to many misclassified cases and serious attenuation of efficacy and effectiveness estimates. Dr. Halloran discussed the use of a validation sample can be used to improve the attenuated estimates. She proposed a new method of analysis for validation sets with time-to-event in vaccine studies when the baseline hazards of both the illness of interest and similar, nonspecific illnesses are changing. She analyzed data from an influenza vaccine field study with these methods and showed that they could have a major impact on the estimated vaccine efficacy.

Joe Hogan (Brown University) gave the final presentation on biomarker evaluation and analysis in a causal framework. Dr. Hogan reiterated that biomarkers can be used for several purposes, such as surrogate markers of treatment effect or as inputs to a diagnostic algorithm. His presentation described applications of causal modeling and inference for both settings, and highlighted the role of potential outcomes for understanding properties of a biomarker. First, he illustrated the use of instrumental variables and associated sensitivity analysis for estimating causal treatment effects of HAART from observational cohort studies. His focus was on transparent representation of underlying assumptions, and on the role of coherent sensitivity analyses to understand the effects of departures from those assumptions. He also described the role of potential outcomes for assessing diagnostic utility of a continuous biomarker. An important measure of diagnostic utility is area under the ROC curve. The area represents $P(X>Y)$, where X and Y are, respectively, randomly-drawn marker values from the 'case' and 'non-case' populations. In some observational studies, the 'case' and 'non-case' populations are systematically different, and bias can be introduced by confounders. He proposed a new definition for area under the ROC curve that is written in terms of potential outcomes, and appeals to a causal interpretation of diagnostic utility. Standard methods for causal inference can be used to estimate the area under the curve; The ideas were illustrated by examining the diagnostic utility of viral load and CD4 as markers for HIV-related mortality, using inverse probability weighting to adjust for potential confounders. He also made qualitative and quantitative comparisons to standard methods.

Workshop 5 - Part 2: Biomarkers in Cancer Research
April 20-22, 2005

Organizers:
Steven Skates, Massachusetts General Hospital and Harvard Medical School
Jeremy Taylor, Biostatistics - School of Public Health, University of Michigan

Overall Summary

There were approximately 65 attendees for some or all of the workshop sessions. A strength of the workshop was the breadth of experience and backgrounds of the attendees. They included graduate students, post-docs, a large contingent of junior level faculty, and some established researchers. The majority of them had a background in statistics or biostatistics, but there were also a significant number who had a biological science background. They came from universities, the government, industry, and medical research centers.

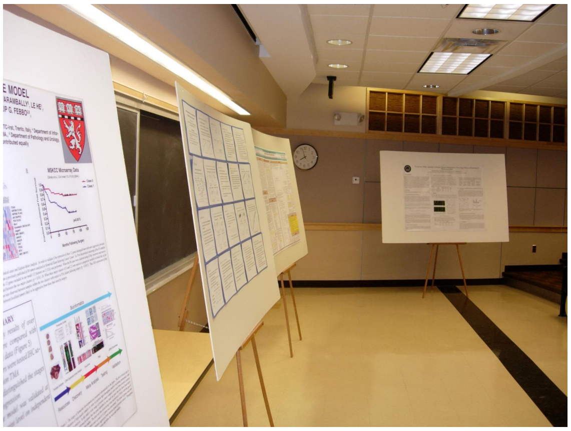
There were five sessions: three of the sessions had designated speakers and discussants, one session had designated speakers only, and one session consisted of poster viewing followed by minipresentations. There was considerable time allowed for general discussion after each talk and at the end of each session.

Summary of Talks

The opening session was a general overview session. The speakers were a cancer pathologist, an epidemiologist, a statistician, and a NCI representative. The session highlighted the importance of biomarkers in cancer research, the substantial numbers of areas of application, and the complexities that can arise. The speakers were Sudhir Srivastava (National Cancer Institute), Mark Rubin (Brigham and Women's Hospital), Bruce Trock (Johns Hopkins School of Medicine), and Jeremy Taylor. The discussants were Colin Begg (Memorial Sloan-Kettering Cancer Center), and Kevin Coombes (MD Anderson Cancer Center).

The Thursday morning session was on the use of biomarkers in detection of cancer. The talks highlighted the use of sophisticated statistical methods to extract the information from biomarker data. The speakers were Steven Skates, Donna Pauler Ankerst (Fred Hutchinson Cancer Research Center), and Steve Horvath (David Geffen School of Medicine). The discussants were Elizabeth Slate (MUSC) and Alexander Tsodikov (UC Davis).

The Thursday afternoon session was on identification of biomarkers, with particular emphasis on proteomic methods. The session highlighted the technology aspects of many of the proteomic assays and the potential for considerable bias that can arise in studies without proper experimental design methodology. The speakers were John Semmes (Eastern Virginia Medical School), Eleftherios Diamandis (University of Toronto), Rick Higgs (Eli Lilly), and Kerry Bemis (Indiana Centers for Applied Protein Sciences), and the discussants were Keith Baggerly (MD Anderson Cancer Center) and Zhen Zhang (Johns Hopkins Medical Institutions).



The Friday morning session was a poster session, followed by mini presentations of selected posters. The posters included a range of topics including gene expression, prostate cancer biomarkers, proteomics, regression trees, and DNA adducts. Posters were presented by Dan Normolle (University of Michigan Medical Center), Annette Molinaro (National Cancer Institute – NIH HHS), Sally Thurston (University of Rochester Medical Center), Ronglai Shen (University of Michigan), Natasha Rajicic (Massachusetts General Hospital), Jeff Morris (MD Anderson Cancer Center), and Francesca Demichelis (Brigham and Women's Hospital).

The Friday afternoon session was on genetics, and included talks on gene expression data and CGH array data. The talks highlighted the need and power of incorporating the biological context and genetic knowledge into the analysis methods to extract the most information from the data. The speakers were Debashis Ghosh (University of Michigan), Adam Olshen (Memorial Sloan-Kettering Cancer Center), and Jane Fridlyand (UCSF).

Conclusion

The feedback from the attendees of the conference was very positive. The mixed backgrounds of the attendees made for many interesting interactions. As an educational workshop, it certainly broadened many people's understanding of the area of cancer biomarkers. An underlying theme of the whole conference was the need for statisticians to incorporate the scientific context of cancer biomarkers into the methods they proposed. Throughout the workshop, many fine examples of this were presented.

The workshop ran very smoothly. The attendees appreciated the logistical support of the MBI Director and staff.

First Young Researchers Workshop in Mathematical Biology

March 29-April 1, 2005

Organizers:
The MBI Postdoctoral Fellows

Overall Summary

The principal aim of this workshop was to bring together over 60 young researchers in Mathematical Biology, to broaden their scientific perspective, and to develop connections that will be important for their future careers.

This workshop provided opportunities for the young researchers (postdoctoral researchers, tenure-track faculty, and advanced graduate students) to start collaborations with each other, to find out what is going on in the field of Mathematical Biology at other universities and research institutions, and to have very open discussions. After all, they are going to be colleagues for a long time.

The workshop included six plenary presentations from biology and mathematical biology. The plenary speakers were leading researchers in Mathematical Biosciences.

The young researchers made poster presentations and gave short talks at the end of each day. This ensured that all the participants had the opportunity to present their work. Posters were displayed for a whole day, to provide ample time for interactions between the young researchers. The interactions during the poster sessions were very lively and generated animated discussions that continued into the evening.

Additionally, there were working group discussions on broad scientific issues on the impact of science, career development, research funding, education, mathematical biology in industry, challenges of a developing field.

Summary of Talks

Day 1: The workshop began with Charles Peskin (New York University) who gave an overview of his work on constructing a combined electrical, mechanical, and fluid-mechanical model of the heart. He discussed the mathematical principles governing cardiac fiber architecture. In addition, he showed how the immersed boundary paradigm can be effectively used in conjunction with a description of the fiber architecture to study both fluid-structure interaction and electrophysiology of the heart.

This was followed by a first group of short talks by some of the young participants (for a list of speakers see [1] below). The goal of this session was to give a preview of the posters exhibited on this day. This also gave a chance to participants to connect a topic with a particular person. The topics ranged from models of zebrafish development, to cell-based models of growing tumors, to computational models of cell motility.

In the afternoon James Keener (University of Utah) talked about how cells make measurements and then make behavioral decisions in response to these measurements. He suggested the following principle: the rate of molecular diffusion contains quantifiable information that can be transduced by biochemical feedback to give control over physical structures. His models of colony size regulation in P. aeruginosa and flagellum growth in salmonella illustrate this principle.

Day 2: Kirk Jordan (IBM) and Frank Tobin (GlaxoSmithKline) started the day with a presentation on Mathematical Biology in Industry. They discussed the differences between skills necessary for a successful mathematical biology career in academia and industry. They also presented examples of recent problems that their groups had worked on. This was followed by a panel discussion.

The morning session continued with a second group of short talks by some of the young participants, introducing their posters (for a list of speakers see [2] below). The topics ranged from differential equations in chemical kinetics on graphs to macroscale and microscale modeling of osmotic effects.

In the afternoon, Alex Mogilner (UC Davis) talked about self-organizing molecular machines in cell division. He addressed the problem of the self-organization of cytoplasmic fibers during mitosis, considering experiments and models of fragments of fish melanophore cells that aggregate pigment granules coated with dynein molecular motors at the center. Also, he described experiments and models of the “search and capture” process and demonstrated that cells use chemical gradients to bias and optimize microtubule dynamics for fast division.

This was followed by another plenary talk by MBI Director Avner Friedman. He presented some new mathematical problems arising from models of tumor growth. He concentrated on free-boundary PDE models that can be used to address the following questions: What are the shapes of dormant tumors? Are these shapes stable? The answers to these questions involve Liapounov-Schmidt and Hopf bifurcations for free boundary problems.

Day 3: The third day started with another group of short talks by young participants, introducing their posters (for a list of speakers see [3] below). The topics ranged from foraging dispersal strategies in rainforest canopies to genetic algorithms in phylogenetics.

This was followed by working group sessions on the following topics: "Women and Minorities in Mathematical Biology", "Establishing Successful Mathematical/Biological Collaborations", "Mathematics and the Biosciences: Philosophies and Shifting Paradigms", and "Ensuring that your Models get used: Routes to Successful Dissemination". These topics were chosen by the organizers (MBI postdoctoral researchers), summarizing suggestions made by applicants for the workshop.

Each topic attracted a number of participants. They held round-table discussions on the respective topic and summarized conclusions. The afternoon session started with each working group presenting a short report to all workshop participants. The rest of the audience actively participated by asking questions and further extending the discussion.

The afternoon continued with a plenary talk by Lou Gross (University of Tennessee). He talked about natural resource management as a spatial control problem. He summarized a variety of mathematical and computational approaches that are available to address this problem. He argued that this field presents new opportunities for mathematicians to collaborate with computational scientists, natural resource managers, and geographers, to develop a science of spatial control of natural systems.

Day 4: The final day of the workshop began with a plenary talk by Claudia Neuhauser (University of Minnesota). She presented her recent work on spatial effects of trophic interactions. This was studied by considering a spatially explicit, stochastic model that investigates the role of explicit space and host-specificity in multispecies host-symbiont interactions. It was found that, surprisingly, pathogens can significantly alter the spatial structure of plant communities, promoting co-existence,

whereas mutualists appear to have only a limited effect.

The morning session continued with a fourth group of short talks by some of the young participants, introducing their posters (for a list of speakers see [4] below). The topics ranged from modeling the 2001 foot-and-mouth epidemic to mechanisms of synaptic facilitation.

Next, working group sessions were held on the following topics: "Mathematical Biology Curriculum at the Graduate and Undergraduate Levels", "The Future of Mathematical Biology", "Funding Interdisciplinary Research In Mathematical Biology", and "How to Sell Mathematical Biology to Mathematicians and Biologists". As in the previous day, each topic was discussed in a round-table format, and the conclusions were presented in the afternoon for further discussion by all workshop participants.

Conclusion

This was the first workshop held at the MBI specifically for the benefit of the young researchers in mathematical biology, as well as the first workshop organized by MBI postdoctoral researchers. Judging from the very large number of applications and the extremely positive feedback from participants, the workshop was successful in achieving its goals. The young researchers had an opportunity to meet each other, to exchange scientific ideas, and to think about broader career and research issues in a friendly and productive environment. They also benefited from the experience of plenary speakers. The plenary speakers were some of the most influential researchers in the field; they truly cared about interacting with the young researchers and provided mentoring, guidance, and advice.

The Mathematical Biosciences Institute decided to make this Young Researchers Workshop an annual event.

References

[1] Yong-Tao Zhang, Zhijun Wu, Joshua S. Weitz, Yulia Timofeeva, Patrick De Leenheer, Jun Feng (Jeff) Sun, Magdalena Stolarska, Eunha Shim, Lixin Shen, Evelyn Sander, Antonio Politi, Brad Peercy, Katarzyna Rejniak.

[2] Andrew L. Nevai, Maya Mincheva, Anastasios Matzavinos, Paul Atzberger, Hugh R. MacMillan, Manuel Lladser, Steven H. Kleinstein, Petro Babak, Christine E. Heitsch, Robert Guy, Cengiz Gunay, Paula B. Grajdeanu.

[3] Javier Garcia-Perez Gamarra, John Fricks, German A. Enciso, Louis Tao, Kim Cuddington, Huseyin Coskun, Nick Cogan, Ariel Cintron-Arias, Diego Pol.

[4] Gerardo Chowell, Mark Byrne, Khalid Boushaba, John Matthew Beggs, Ruth Baker, Yongsam Kim, Victor Matveev, Robyn Araujo, Alexei Medovikov, Gheorghe Craciun.

Current Topics Workshop: Enzyme Dynamics and Function May 19-21, 2005

Organizers:
Russ Hille, Molecular and Cellular Biochemistry, The Ohio State University
Ming-Daw Tsai, Department of Chemistry, The Ohio State University

Overall Summary

Over the past several years, it has become increasingly appreciated that the dynamic properties of enzymes can play a significant role in modulating their catalytic properties. The motions involved can range from the vibration of individual chemical bonds or groups of bonds (taking place on the femtosecond timescale and involving distances of less than 1 Å) to large domain motions (taking place on a timescale of milliseconds to seconds and involving distances as great as 10 Å or more). With the accumulating experimental evidence attesting to the importance of these motions in catalysis, it has become important to develop appropriate mathematical models for enzyme behavior that provide a conceptual framework within which to understand the influence of this dynamic behavior.

This workshop brought together leaders in this emerging field to present their recent work and to participate in discussion groups that provided a forum for both mathematicians and enzymologists to consider the fundamentals relevant to the field. Eight recognized leaders in the field spoke on their most recent research, and two others led discussion round tables that focused on various areas of interest. The invited speakers were selected on the basis of their prominence in the field, with a conscious effort to hear from individuals with diverse backgrounds. These included theoreticians trained in mathematics, physical chemistry and physics, and experimentalists with backgrounds in biochemistry, organic chemistry, and nuclear magnetic resonance methods. Total registration for the workshop was over 50, and included a number of individuals from other institutions as well as faculty, graduate students, and postdoctoral scholars at the Ohio State University.

Mornings were devoted to research presentations by the invited speakers, and the afternoons to round tables at which various topics were more broadly discussed. Presentations were deliberately kept informal, and there was lively discussion during, as well as after, the talks. A diverse range of perspectives on how to approach the role of molecular dynamics in enzyme function were presented, and at the same time a broad consensus emerged as to the nature of the key questions to be asked and the manner (both theoretical and experimental) in which they were to be addressed. Immediately after the meeting, a paper was prepared and broadly distributed that summarized the overall discussion and topics of the round tables (below).

What should the textbooks say about how enzymes work?

- The reaction is different in the enzyme than in solution (notion of enzymes as a specialized solvent; number of involved atoms and functional groups; different and more complex chemical mechanisms; more complex and non-static potential energy surfaces; different/appropriate energy scales; conformational sampling/NAC’s)
- A portion of catalytic effectiveness is paid for in the synthesis of the polypeptide (or polynucleotide) and in binding.
- Transition state analog and catalytic antibody approaches have limitations (imperfect mimics of transition state; poor template and/or dynamic characteristics).
- Different factors may be important to different degrees in going from one enzyme to another (the “specifics” of specific enzyme-catalyzed reactions).
- There are distinct differences between “bio-organic” and “biophysical” approaches to understanding rate acceleration (but these are not necessarily mutually exclusive).
- The dynamic properties of enzymes (on a wide range of distance and time scales) play important roles in catalysis.

What are the important remaining questions in enzymology?

- Are “starting states” for computational analysis appropriate to the calculation time periods?
- Why are enzymes so big?
- How do mutational, temperature, and kinetic isotope effects on enzyme behavior correlate?
- What methods can be developed to span the differences in time scales of calculations (ps-ns) and actual/experimental catalysis (µs-ms)?
- How can we develop more, better, faster spectroscopic methods?
- How can the behavior of individual molecules best be extrapolated to the behavior of ensembles molecules?
- How can computation and experiment be better reconciled?
- How do we probe the role of exchangeable protons and other (larger) group transfer reactions?
- How can we better compare results between different groups (sharing of coordinates, programs, and so on)?

Where do you want to be in two years?

- Studying cell-crowding effects on enzyme-enzyme interactions and catalysis.
- Making more accurate $\Delta S^{\ddagger}$ predictions; getting better B-factors and higher-resolution crystal structures.
- Identifying and examining synergistic elements of protein structure.
- Reminding the larger biochemical community that understanding the basis of catalysis in detail is a central issue for cell function, inhibitor (drug) design and that it is important to answer questions of catalysis correctly.
- Making better “predictions” rather than “retrodictions”.
- Designing enzymes *de novo*.
- Better correlating NMR properties with molecular dynamics.
- Hawaii.

Workshop 6: Recombination: Hotspots and Haplotype Structure
June 13-16, 2005

Organizers:
Rick Durrett, Department of Mathematics, Cornell University
Paul Fuerst, Department of Ecology, Evolution, and Organismal Biology, The Ohio

State University

Overall Summary

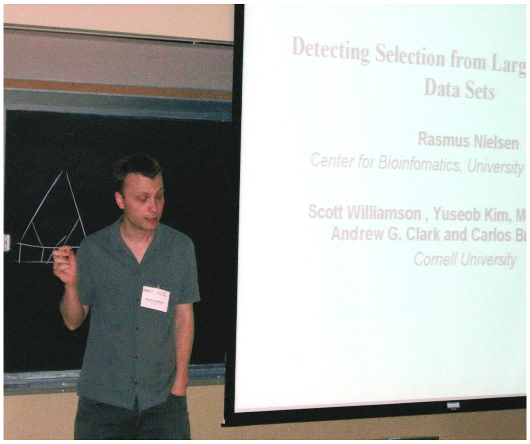
The workshop highlighted new approaches to understanding the nature and causes of linkage disequilibrium in the genomes of higher organisms. Emphasis was placed upon the role of using new genomic information, especially the availability of high density SNPs (single nucleotide polymorphisms) to elucidate the mapping of complex disease loci through association studies. Data following from the sequencing of the human genome suggested that an intricate haplotype structure exists. This suggestion led to the HapMap project whose goal is to understand the patterns of DNA sequence variation, and whose results are now being reported. Several of the papers presented in this workshop examined preliminary data from the HapMap project. In a parallel development, recent studies have shown that much recombination within the genome occurs at hot spots, and is not uniformly distributed across chromosomes. This workshop concentrated on mathematical, statistical, and computational approaches to estimating local and global recombination rates, and determining the causes of haplotype structure in humans and other species.

Efficient Simulation Methods for a Class of Nonneutral Population Genetics Models
Paul Joyce
Alan Genz
Abstract
Any assessment of the forces that generate and maintain genetic diversity must include the possibility of selection. Computationally intensive methods for approximating likelihood functions and generating samples for a class of nonneutral models were proposed by Donnelly, Nordborg, and Joyce (DNJ 2001). In this article, we develop algorithms for likelihood analysis that are substantially more efficient than those in DNJ (2001). The DNJ (2001) method relies on simulating data from the neutral model and then uses acceptance sampling to generate data under the model with selection. For example, in some cases DNJ (2001) requires on average 1 × 10^9 rejections before a sample is accepted. Our method simulates samples directly from the distribution

Summary of Talks

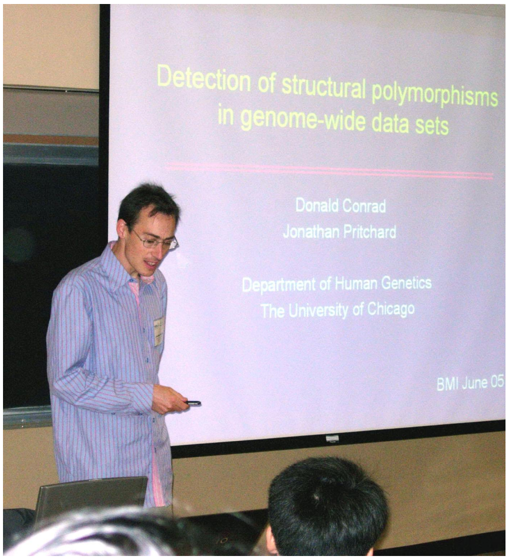
The first talk was given by Paul Fearnhead (Lancaster University) on “Likelihood-based methods for detecting recombination Hotspots from Population Data.” He focused on two methods for detecting recombinational hotspots based on analyzing subregions of the data. The approach calculates likelihoods for sub-regions (e.g., six consecutive sites), assuming constant recombination. Although the methods can detect hotspots, they are computationally slow. Consideration is being given to ways to improve speed of the methods, as well as increasing the ability to distinguish conversion from crossing-over. Next, Noah Rosenberg (University of Michigan) spoke about “Population structure and homozygosity-based measures of linkage disequilibrium.” He adapted an approach first proposed by Tomoko Ohta to identify excess homozygosity in haplotypes in order to recognize regions having significant linkage disequilibrium. The effects of population structure were specifically investigated because structured populations contain more multilocus homozygosity than predicted from single locus considerations.

In the afternoon, Fengzhu Sun (University of Southern California) presented a consideration of “Haplotype block partition and tag SNP selection and their applications to association studies.” He argued that a small fraction of SNPs (tag SNPs) in any genomic segment are sufficient to capture most of the haplotype structure of the human genome. By using simulations based on a coalescent model, he examined the problems of identifying tag SNPs, and the equivalence of power to detect associations from the use of tag SNPs compared to using uniformly spaced marker SNPs. Susan Ptak (Max Planck Institute for Evolutionary Genetics) presented the results of a study on comparative genomics and comparative recombination. The presentation was entitled "Fine-scale recombination patterns differ between chimpanzees and humans." This analysis suggested that less than 10% of recombinational hotspots are conserved between humans and chimpanzees. The analysis further indicates that average recombination rates in homologous regions are only weakly correlated. Taken together, the results indicate that recombination rates are dynamically changing during evolution, and that the recombination landscape has changed dramatically when comparing humans and their most similar evolutionary cousin.


Detecting Selection from Larg
Data Sets
Rasmus Nielsen
Center for Bioinformatics, University
Scott Williamson , Yuseob Kim, M
Andrew G. Clark and Carlos B
Cornell University

The second day of the workshop was started by Paul Joyce (University of Idaho). His talk, "Efficient Simulation Methods for a Class of Nonneutral Population Genetics Models", considered problems inherent in dealing with more complex evolutionary models, especially when incorporating selection. Likelihood methods, which had previously been proposed, are computationally inefficient. Central to previous approaches, is the need to calculate the constant of integration for the "K" allele model with selection. He presents a new method for likelihood analysis that is substantially more efficient, using numerical analysis techniques, including fast Fourier transforms to calculate the intractable constant of integration. New algorithms are presented and examined, which substantially improve the performance of likelihood simulations of non-neutral scenarios. The new methods make likelihood analysis practicable for a wider set of parameters. In particular, if the selection intensity is much greater than the mutation rate, previous methods become increasingly inefficient. However, this is the case where one has the best hope of drawing meaningful (more precise) inferences about evolutionary processes. Rasmus Nielsen (Univ. Copenhagen Bioinformatics Center, Denmark) finished the morning with a talk on "Analysis of ascertained SNP data." This paper dealt with the statistical complications arising from the non-random nature of SNP discovery. The usual pattern of SNP identification in small samples, followed by subsequent analysis of large samples, has implications for statistical properties of the data including linkage disequilibrium, frequency spectrum, and levels of population differentiation. The ascertainment bias also implies that standard population genetic analyses are not applicable to the vast majority of human SNP data. A composite likelihood method is proposed that can be implemented to allow valid population genetic inferences.

The afternoon of the second day included the presentation of a number of short talks. Among these was a presentation by Vincent Plagnol (University of Southern California) on “Demographic inference of human population’s history.” He investigated the problem of fitting models for the history of two human populations: European and African-American, using data from the Seattle SNPS's database. By ancestral inference, the problem of causation of differences in the pattern of polymorphism in the two populations was considered. He also considered the issue of ascertainment bias when dealing with the HapMap dataset.


Detection of structural polymorphisms
in genome-wide data sets
Donald Conrad
Jonathan Pritchard
Department of Human Genetics
The University of Chicago
BMI June 05

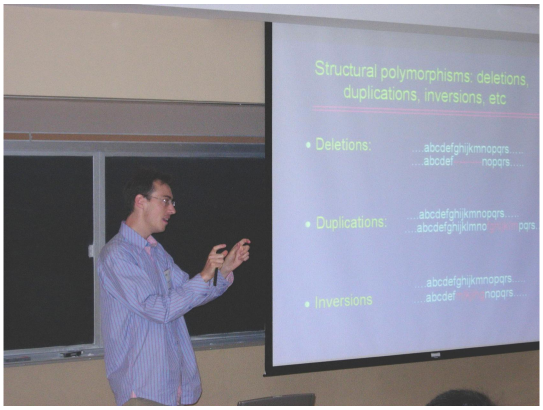
Continuing in the short paper presentations, Yuguo Chen (Duke University) discussed “Stopping-Time Resampling for Sequential Monte Carlo Methods with Applications to Population Genetics.” Resampling is important to sequential Monte Carlo methods used in statistical genetics. However, existing resampling techniques do not work well for coalescent-based inference problems in population genetics. A new method called “stopping-time resampling” was presented, which allows the comparison of partially simulated samples at different stages of the coalescence process to terminate unpromising partial samples and allow the early identification and multiplication of promising partial samples.

Graham M. Coop (University of Chicago) presented a short talk -co-authored with Simon R. Myers (Department of Statistics, University of Oxford)- with the provocative title “Live hot, die young: transmission distortion in recombination hotspots.” This paper dealt with the issue of the conservation of recombinational hotspots in the genome. Transient occurrences of recombinational hotspots are believed to be the result of biased gene conversion in favor of alleles that locally disrupt hotspots. The results indicate that a lack of sharing intense hotspots between species is to be expected even if there are few sites where hotspot-disrupting alleles arise. Effective population size plays a significant role in the fate of hotspots. Alleles that reduce the intensity of a hotspot leave little trace of their presence in the patterns found in population data.

Kui Zhang (University of Alabama at Birmingham) talked on “Haplotype Inference for Tightly Linked SNPs in General Pedigrees.” He presented results from the use of an efficient computer program, HAPLORE, for haplotype frequency estimation and reconstruction in general pedigrees with tightly linked SNP markers. The paper compared the performance of this program with two others in its ability to efficiently estimate haplotype frequencies and accurately infer haplotype configurations in general pedigrees with a large number of tightly linked SNPs, especially in the presence of missing data. Relative performance of the programs was also evaluated.

The last short talk was given by Scott Williamson (Cornell University) who discussed "Searching for evidence of balancing selection by comparison of the gene frequency spectrum." This analysis utilized the Perlgen SNP data set to search for an excess of alleles with intermediate frequency. A composite Likelihood ration test was proposed to identify the number and location of such alleles in the data.

On the third day of the workshop, Jonathan Pritchard (University of Chicago) talked about "Detecting partial selective sweeps from SNP data." He outlined approaches to detect the effects of selective sweeps in haplotype data, such as the HapMap and Perlegen data sets, emphasizing the examination of patterns of long-range Linkage Disequilibrium. He presented a new approach that extends the PAC-likelihood model of Li and Stephens (GENETICS 165: 2213, 2003) in order to test long range LD signals in an approximate likelihood framework. The new test controls for local recombination rate heterogeneity, which may confound simpler approaches. Marcy Uyenoyama (Duke University) presented the "Likelihoods from summary statistics." The talk emphasized analysis of summary statistics in population genetics when we have no knowledge of the gene genealogy. She showed the development of an importance sampling (IS) approximation to the time-consuming computation of exact likelihoods during the computations of a maximum-likelihood estimate of the rate of recombination between a neutral marker locus and the target of strong balancing selection to which it shows nearly completely linkage. She also presented examples of the use of this approach for the analysis of data on linkage and balancing selection is *Drosophila*. Finally she considered some difficulties presented by aspects of the biology of an organism, such as the genomic location of a target of selection, with respect to areas of the genome having low recombination, and problems of the process of introgression of genes.



Structural polymorphisms: deletions, duplications, inversions, etc

- Deletions: ....abcdefghijkmnopqrs..... ....abcdef nopqrs.....
- Duplications: ....abcdefghijkmnopqrs..... ....abcdefghijklmno pqrs.
- Inversions ....abcdefghijkmnopqrs..... ....abcdef nopqrs.....

Eran Halperin (International Computer Science Institute, Berkeley, CA) began the afternoon with a talk about "Estimating haplotype frequencies efficiently." This presentation illustrated the use of the program HAPLOFREQ to estimate haplotype frequencies over a short genomic region given the genotypes or haplotypes with missing data and/or sequencing errors. The likelihood function, which forms the basis of this approach, is guaranteed to efficiently converge to its global optimum. The relationship between haplotype frequency estimation and tag SNP selection was also considered. The day ended with a presentation by Rick Durrett (Cornell University), who talked about "The impact of spatial structure on genetic data." He reviewed recent results about migration models in population genetics. Specifically he considered the stepping stone model, and the fact that this model has a much different impact on genetic data than the often used island model. Theoretical results for coalescence times were presented, as well as simulation results concerning the site frequency spectrum and decay of linkage disequilibrium along a chromosome.

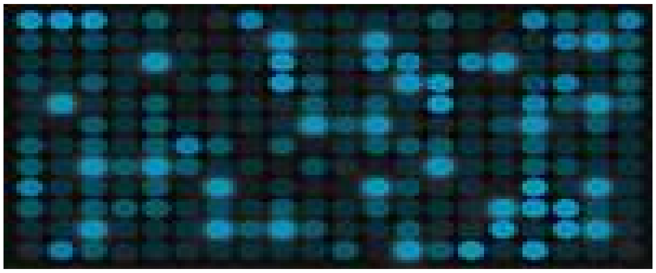
The final day of the workshop included two talks. Ranajit Chakraborty (University of Cincinnati College of Medicine) presented a talk on “Effects of Mutation and Population Demography on the Dynamics of Linkage Disequilibria and their Relevance for Mapping Complex Disease Genes.” He reviewed the problems inherent in the analysis of complex disease traits and the use of linkage disequilibrium for the identification of genes contributing to these traits. Some properties of genome-wide background LD were examined through a coalescence-based simulation study. When microsatellite loci are used as genomic markers for disease-gene association studies, the expectation of the weighted normalized LD between two loci decreases with recombination distance between loci. However, the extent and trend of such decay is dependent on the rate and pattern of mutations as well as on the demographic history of populations. In a growing population, the power of detecting LD is substantially reduced, being comparable to that expected in a constant population of the largest size reached by the population. The presence of multiple alleles at microsatellite loci makes such markers more powerful to detect LD than single nucleotide polymorphism sites (SNPs) residing at the same recombination distance. Carlos Bustamante (Cornell University) ended the conference by discussing “Inferring the distribution of selective effects among mutation, SNPs, and fixed differences using Polymorphism and Divergence Data.” He discussed both Baysian and frequentist approaches to the problem of inferring the distribution of selection coefficients on newly arising mutations. Methods were considered that use whole-genome SNP frequency data, polymorphism and divergence across protein-coding gene data, and combined SNP frequency, invariant, and divergence data. Simulations with selection and recombination were used to gauge the sensitivity, robustness, and accuracy of the models. Lastly, we apply the method to human polymorphism and divergence data to estimate the proportion of mutations, SNPs, and nucleotide substitutions in the human genome that are deleterious, neutral, and adaptive. The fact that mutations undergoing negative selection can interfere with the processes affecting alleles undergoing positive selection are considered in obtaining the joint frequency spectrum of alleles.

Tutori-

Tutorial on Microarrays

Organizers and Speakers:
Sashwati Roy and Chanden Sen - Department of Surgery, The Ohio State University

DNA microarrays are becoming a standard tool more molecular biology research and clinical diagnostics by providing a simple and natural means for surveying the genome in a very systematic and comprehensive manner. Microarrays are miniature arrays of gene fragments immobilized in a dense order on a solid substrate. Because thousands or tens of thousands of gene fragments can be present on a single microarray, data for an entire genome can be acquired in a single experiment. The power of DNA microarrays lies in the ability to simultaneously score the hybridization signals, which represent global gene expression patterns of biological processes, and their dynamic variations. The tutorial introduced all key aspects of microarray, array design, probe selection, array fabrication, biological questions, experiment design, target preparation as well as labeling, visualizing and analysis of microarray images, and basic data analysis. The lectures included general principles underlying microarray printing, enzymatic labeling, signal amplification, clustering methods, and electronic resources.



Microarrays. U.S. Department of Energy Genomics: GTL Program, http://doegenomestolife.org.

Tutorial on Statistical Methods and Software for the Analysis of Microarray Experiments September 20-24, 2004

Organizers and Speakers:
Nick Jewell and Sandrine Dudoit - Division of Biostatistics, UC Berekeley

DNA microarray and other high-throughput genomic experiments generate complex high-dimensional datasets of multiple types. Extracting meaningful and reliable biological information from the analysis of these data presents new statistical and computational challenges. The tutorial discussed statistical design and inference methods for microarray experiments. Topics covered included: pre-processing (image analysis and normalization); multiple testing procedures for the identification of differentially expressed genes; hierarchical and partitioning cluster analysis; prediction; and model selection.

The statistical methods discussed could apply to a broad range of problems beyond the analysis of microarray data, such as the genetic mapping of complex traits using single nucleotide polymorphisms (SNPs) and the identification of transcription factor-binding sites in ChiP-Chip experiments.

The tutorial included computer lab sessions to allow participants to explore statistical software resources for the analysis of genomic data, with emphasis on R packages developed as part of the Bioconductor Project.

Summer Program

Microarray Gene Expression Data Analysis

August 1-19, 2005

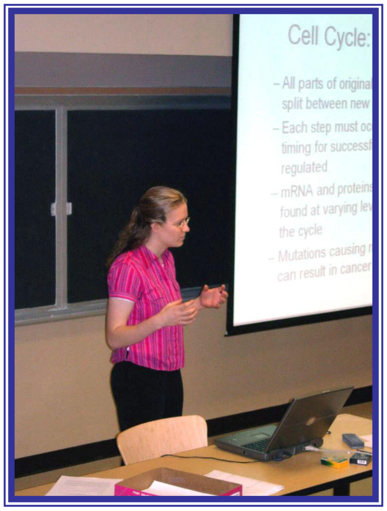
The program included 26 participants, most of them graduate students from departments of mathematics, statistics, biology, and computer science from the U.S.; a few came from Europe. The first week included tutorial lectures in statistics needed for analysis of microarray gene expression.

Shili Lin presented one lecture on exploratory data analysis which discussed a number of simple, yet very useful numerical and graphical summary methods for microarray data. She also gave three lectures on statistical analysis of both Affymetrix and cDNA microarrays. Topics discussed included data preprocessing (image analysis and normalization), identification of differentially expressed genes, and class discovery and class prediction problems.

Joe Verducci lectured on the framework of statistical inference and gave an introductory talk on how to implement statistical methods using the Bioconductor software based on the programming language R. He followed this with two computer lab sessions in which participants analyzed gene expression data from the NHLBI Program for Genomic Applications.

There were also several introductory talks in molecular biology. Greg Singer described the structure of the cell, the structure of chromosomes and DNA; he explained the transcription (from DNA to RNA) and translation (from RNA to protein), and gene expression regulation. The program included visits to several labs.

Following the tutorials, the participants were divided into five groups, and each worked on one project as described below. Toward the end of the two weeks, each group made a presentation of the results they obtained; these presentations are available on the



Miniconference: Group Projects Report August 18-19, 2005

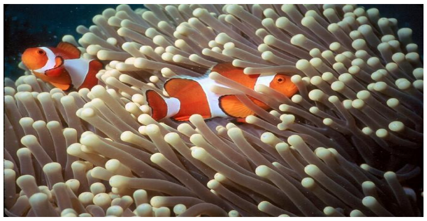
Project 1: Image analysis and normalization (cDNA array data)
Project Leader:
Bertram Zinner

Project 2: Identification of differently expressed genes
Project Leader:
Zailong Wang

Project 3: Cluster analysis of gene expression data
Project Leader: Jin Zhou

Project 4: Class discovery and prediction of tumor subtypes
Project Leader:
Nusrat Rabbee

Project 5: Use of ChIP-on-chip to interrogate cancer epigenome
Project Leaders: Victor Jin and Alfred Cheng



Future Programs

Ecology and Evolution

September 2005 - August 2006

Ecology and evolutionary biology have historically been two of the areas of biology which have most benefited from, and made use of, mathematical methods. Many distinguished mathematical biologists have contributed to these areas, and their efforts have illuminated much of ecological and evolutionary theory over the past century. An objective of this special year is to focus on specialized areas that offer particularly challenging mathematical problems, which are relatively unexplored and are of potentially great interest to observational biologists. Thus, an underlying goal of the proposed activities is to maintain direct connections to observable biology.

One thread of connection between the various proposed activities concerns spatial aspects of natural systems. Central questions about the history and structure of biological systems are affected by spatial variation. Additionally, numerous problems, which have great public impact, necessarily involve the spatial heterogeneity of biological systems, both those occurring through natural processes and those deriving from human actions. Conservation biology, biodiversity, harvest planning, invasive species control, and wildlife management are just a few of the applications that utilize mathematical methods to address major public policy issues. These applied areas rely greatly upon general ecological and evolutionary genetics theory. Determining how natural systems are affected by interactions of space and time leads to problems that require mathematical approaches. Although a large body of mathematical literature has developed over the past several decades dealing with spatiotemporal interactions, there are still many biologically important questions that require new mathematical approaches and would benefit from close collaborations between ecologists, evolutionary biologists, and mathematicians.

Beyond emphasizing the spatiotemporal nature of natural systems and the mathematical approaches that are used to address them, the special year is intended to foster interactions between individuals working on problems at different spatial/temporal scales. While the underlying biological questions may operate on quite different scales, the necessary mathematical approaches may be similar. Another theme for the year is linking between scales, for example, how might evolutionary models that account for the dynamics of spatial structure relate to ecological models, which operate on shorter time periods? How might genomic information that is rapidly becoming available assist in developing a theory for whole organism interactions with environment and the functioning of populations, communities, and ecosystems? What new mathematical approaches might contribute to better models for natural system response across the genome/organism/population interfaces? The proposed set of activities will enhance our ability to address these questions and hopefully lead to new collaborations between mathematicians and biologists that are beneficial to both fields.

Tutorials

Tutorial on Tree Reconstruction and Coalescence Theory
September 7-9 and 12-13, 2005

Organizers:
Dennis Pearl - Department of Statistics, The Ohio State University
Paul Fuerst - Department of Evolution, Ecology, and Organismal Biology, The Ohio State University

Tutorial on Reaction-Diffusion Models
March 9-10, 2006

Organizer:
Chris Cosner - Department of Mathematics, University of Miami

Workshops

Phylogeography and Phylogenetics
September 26-30, 2005

Organizers:
Craig Moritz - Department of Integrative Biology, University of California, Berkeley
Michael Hickerson - Department of Integrative Biology, University of California, Berkeley
Dennis Pearl - Department of Statistics, The Ohio State University

Aspects of Self-Organization in Evolution
November 14-18, 2005

Organizers:
Chris Adami - Keck Graduate Institute, California Institute of Technology
Claus O. Wilke - Keck Graduate Institute, California Institute of Technology

The Problems of Phylogenetic Analysis of Large Datasets
December 1-2, 2005

Organizers:
Daniel Janies - Department of Biomedical Informatics, The Ohio State University
Diego Pol - Mathematical Biosciences Institute, The Ohio State University
John Wenzel - Materials Science and Engineering, Rutgers University
Dennis Pearl - Department of Statistics, The Ohio State University
Ward Wheeler - Division of Invertebrate Zoology, American Museum of Natural History

Spatial Heterogeneity in Biotic and Abiotic Environment: Effects on Spe-

cies Ranges, Co-evolution, and Speciation
February 6-10, 2006

Organizers:
Sergey Gavrilets - Department of Ecology and Evolutionary Biology, The Institute for Environmental Modeling; Department of Mathematics, University of Tennessee
Mark Kirkpatrick - Section of Integrative Biology, University of Texas at Austin
John Thompson - Department of Mathematics, University of Florida

Spatial Ecology
March 13-17, 2006

Organizers:
Lou Gross - Department of Ecology and Evolutionary Biology, The Institute for Environmental Modeling; Department of Mathematics, University of Tennessee
Claudia Neuhauser - Department of Ecology, Evolution and Behavior, University of Minnesota
Chris Cosner - Department of Mathematics, University of Miami
Mark Kot - Department of Applied Mathematics, University of Washington

Second Young Researchers Workshop in Mathematical Biology
March 27-30, 2006

Organizers:
MBI Postdoctoral Fellows

Uncertainty in Ecological Analysis
April 3-7, 2006

Kate Calder - Department of Statistics, The Ohio State University
Jim Clark - Department of Electrical and Computer Engineering, McGill University
Noel Cressie - Department of Statistics, The Ohio State University
Jay Ver Hoef - Alaska Department of Fish and Game
Chris Wikle - Department of Statistics, University of Missouri

Symposium on Drug Safety and Public Policy
April 20-22, 2006

Organizers:
Rajesh Balkrishnan - College of Pharmacy, The Ohio State University
Avner Friedman - Mathematical Biosciences Institute, The Ohio State University
Michael Grever - Department of Internal Medicine, The Ohio State University

Cross Cutting Minisymposium on Theoretical and Empirical Perspectives on Speciation Dynamics

April 24-26, 2006

Organizers:
Craig Moritz - Department of Integrative Biology, University of California, Berkeley
Sergey Gavrilets - Department of Ecology and Evolutionary Biology, The Institute for Environmental Modeling; Department of Mathematics, University of Tennessee

Microbial Ecology
May 15-19, 2006

Organizers:
Frede Thingstad - Department of Microbiology, University of Bergen, Norway
George Jackson - Department of Oceanography, Texas A&M University

Global Ecology
June 26-30, 2006

Organizers:
John Pastor - Department of Biology, Center for Water and the Environment, University of Minnesota, Duluth
John Harte - Energy and Resources Group and the Ecosystem Sciences Division of the College of Natural Resources, University of California, Berkeley
David Schimel - Terrestrial Sciences Section, National Center for Atmospheric Research

Summer Program 2006

Ecology and Evolution
July 17 - August 4, 2006

Program Leaders:
Kate Calder - Department of Statistics, The Ohio State University
Yuan Lou - Department of Mathematics, The Ohio State University

Systems Physiology
September 2006 - August 2007

Much of the biological investigation of the past can be described as a compilation and categorization of the list of parts, whether as the delineation of genomic sequences, genes, proteins, or species. The past decade, for example, has uncovered the genetic basis for many diseases. A remaining and larger challenge is to provide an understanding of how the interactions of these biological entities across spatial and temporal scales lead to observable behavior and function. This is what systems biology is concerned with. Two important organizing principles need emphasis: (1) An integrated understanding of systems requires mathematics and the development of theory, supplemented by simulations; and (2) Theory cannot be relevant if it is not driven and inspired by experimental data. Thus the development of system biology requires collaborative work by theoreticians and experimentalists.

The goal of systems physiology is to understand how various human organs and tissues are organized and regulated to produce their normal function and pathologies. This year at the MBI will examine features of several human organ and tissue systems, including the cardiac system, the respiratory system, the microcirculatory system, the renal system, the visual processing system, the endocrine system, and the auditory system. Although these are at first glance quite different, the underlying theme is how cellular level behavior participates in the function of the whole and how feedback from the function of the whole contributes to the regulation of the cellular level behavior. Understanding of these processes may lead to new insights into the causes of diseases and how they can be treated.

Tutorials

Tutorial on Heart and Lung
September 18-21, 2006

Organizers:
Jim Keener - Departments of Mathematics and Bioengineering, University of Utah
Rai Winslow - Department of Biomedical Engineering, Johns Hopkins University School of Medicine
Andrew McCulloch - Department of Bioengineering, Whitaker Institute for Biomedical Engineering, University of California, San Diego
Ken Lutchen - Department of Biomedical Engineering, Boston University

Workshops

Cardiac Electrophysiology and Arrhythmia

Organizers:

Jim Keener - Departments of Mathematics and Bioengineering, University of Utah
Rai Winslow - Department of Biomedical Engineering, Johns Hopkins University School of Medicine

Cardiac Mechanics and Remodeling

Organizers:
Jim Keener - Departments of Mathematics and Bioengineering, University of Utah
Andrew McCulloch - Department of Bioengineering, Whitaker Institute for Biomedical Engineering, University of California, San Diego

The Lung and the Respiratory (Structure, Oxygen, Transport)

Organizers:
Ken Lutchen - Department of Biomedical Engineering, Boston University
Jason Bates - College of Medicine, The University of Vermont

Blood Flow in the Microcirculation: Function, Regulation, and Adaptation

Organizers:
Tim Secomb - Department of Physiology, The University of Arizona Health Sciences Center
Daniel A. Beard - Department of Bioengineering, University of Washington

Renal System

Organizers:
Harold Layton - Department of Mathematics, Duke University
Leon Moore - Department of Entomology, University of Arkansas
S. Randall Thomas - Necker Faculty of Medicine
Alan Weinstein - Department of Mathematics, University of California at Berkeley

Information Processing in the Visual System

Organizers:
Paul C. Bressloff - Department of Mathematics, University of Utah
Alessandra Angelucci - John A. Moran Eye Center, University of Utah

Endocrine Physiology: Type 2 Diabetes, Metabolism, and Obesity

Organizers:
Artie Sherman - NIH-NIDDK-MRB
Richard Bertram - Department of Mathematics, Florida State University

The Auditory System

Organizers:
James Sneyd - Department of Mathematics, University of Auckland, New Zealand
David Mountain - Department of Information Science, City University

Publications

Technical Report No. 25

Authors: Baltazar D. Aguda, Gheorghe Craciun, and Rengul Cetin-Atalay
Title: Data sources and computational approaches for generating models of gene regulatory networks
Date of Publication: September 2004

Technical Report No. 26

Authors: Talia Konkle, Ning Jiang, Jie Zhang, Fatma Gurel, Christopher Scheper, and Gheorghe Craciun
Title: Image segmentation using neural oscillators
Date of Publication: September 2004

Technical Report No. 27

Authors: Daniel P. Dougherty, Dorjsuren Badamorj, Michelle Carlton, Magdalena Musielak, Laura Wherity, and Alice Yew
Title: A mathematical model of the spiking behavior in olfactory receptor neurons
Date of Publication: October 2004

Technical Report No. 28

Authors: Avner Friedman and Bei Hu
Title: Bifurcation from stability to instability for a free boundary problem arising in a tumor model, I
Date of Publication: November 2004

Technical Report No. 29

Author: Katarzyna A. Rejniak
Title: A single cell approach in modeling the dynamics of tumor microregions
Date of Publication: December 2004

Technical Report No. 30

Authors: Katarzyna A. Rejniak, Adrienne Frostholm, Julie Besco, Magdalena Popesco, and Andrej Rotter
Title: *MmSAGEClass* - software manual: An online database for the functional classification of mouse SAGE tags
Date of Publication: December 2004

Technical Report No. 31

Authors: Chandan K. Sen, Joseph S. Verducci, Vicent F. Melfi, Savita Khanna, Cata-

lin Barbacioru, and Sashwati Roy
Title: Post-reperfusion healing of the heart: Focus on oxygen-sensitive genes and DNA microarray as a tool
Date of Publication: January 2005

Technical Report No. 32

Authors: Janet Best, Alla Borisyuk, Jonathan Rubin, David Terman, and Martin Wechselberger
Title: The dynamic range of bursting in a network of synaptically coupled square-wave bursting respiratory pacemaker cells
Date of Publication: February 2005

Technical Report No. 33

Author: Yuan Lou
Title: On the effects of migration and spatial heterogeneity on single and multiple species
Date of Publication: March 2005

Technical Report No. 34

Author: Zhijun Wu
Title: Linear algebra in biomolecular modeling
Date of Publication: April 2005

Technical Report No. 35

Authors: Gheorghe Craciun and Martin Feinberg
Title: Multiple equilibria in complex chemical reaction networks: II. The species-reactions graph
Date of Publication: July 2005

Technical Report No. 36

Authors: Gheorghe Craciun and Martin Feinberg
Title: Multiple equilibria in complex chemical reaction networks: III. Extensions to entrapped species models
Date of Publication: July 2005

Technical Report No. 37

Author: Sookkyung Lim
Title: Whirling instability of a rotating elastic filament based on a bacterial flagellar structure

Date of Publication: July 2005

Technical Report No. 38

Authors: Pranay Goel and Klaus Röbenack
Title: Observing the current input in neutrons
Date of Publication: August 2005

Technical Report No. 39

Authors: Avner Friedman, Jianjun Paul Tian, Giulia Fulci, E. Antonio Chiocca, and Jin Wang
Title: Glioma virotherapy: The effects of innate immune suppression and increased viral replication capacity
Date of Publication: August 2005

Technical Report No. 40

Author: Jianjun Paul Tian
Title: Algebraic structure of non-Menelian inheritance
Date of Publication: August 2005

Technical Report No. 41

Authors: Jianjun Paul Tian and Zhenqiu Liu
Title: Coalescent random walks on graphs
Date of Publication: August 2005

Technical Report No. 42

Author: Jianjun Paul Tian
Title: Evolution algebras and their applications
Date of Publication: August 2005

MBI Volumes on Tutorials in Mathematical Biosciences
Published by Springer-Verlag

Volume I: Mathematical Neuroscience (2004)
Volume II: Mathematical Modeling of Calcium Dynamics and Signal Transduction (2005)
Volume III: Cell Cycle, Proliferation, and Cancer (in press)

Directors

Avner Friedman, Director
Mathematical Biosciences Institute
afriedman@mbi.osu.edu

Peter March, Associate Director
Department of Mathematics
march@mbi.osu.edu

Dennis Pearl, Associate Director
Department of Statistics
dpearl@mbi.osu.edu

Andrej Rotter, Associate Director
Department of Pharmacology
arotter@mbi.osu.edu

Tony Nance, Assistant Director
Mathematical Biosciences Institute
tony@mbi.osu.edu

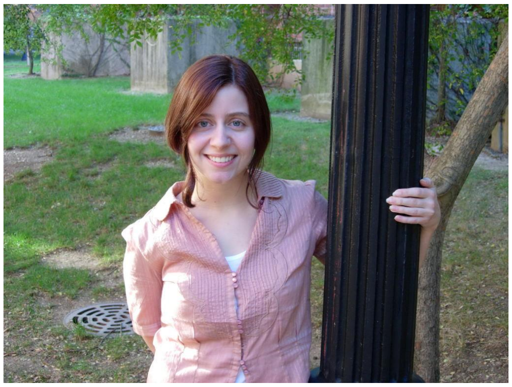
Staff

Kimberly Holle
Program Specialist

Michael Siroskey
Systems Manager

Matt Thompson
Program Assistant

Stella Cornett
Program Assistant



Rebecca Martin
Office Associate

Postdocs

Alla Borisyuk
Courant Institute of Mathematical Sciences,
New York University

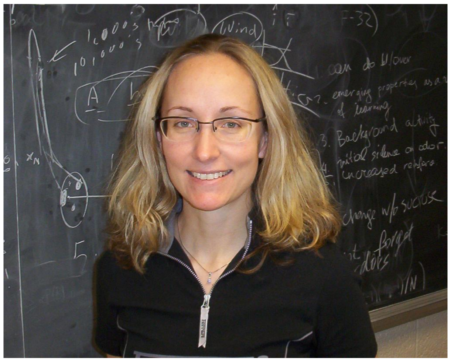
Gheorghe Craciun
Department of Mathematics
The Ohio State University

Daniel Dougherty
Department of Statistics
North Carolina State

Katarzyna Rejniak
Department of Mathematics
Tulane University

Martin Wechselberger
Mathematics Department
Vienna University of Technology

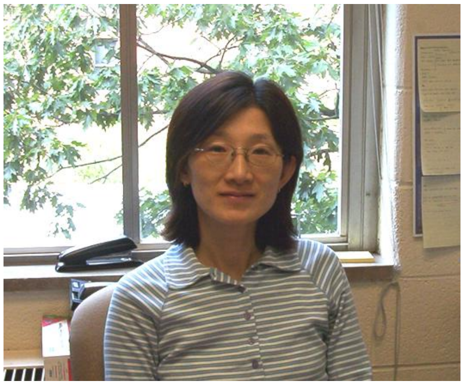
Postdocs



Geraldine Wright
Department of Entomology
Oxford University

Janet Best
Department of Mathematics
Cornell University

Pranay Goel
Department of Mathematics
University of Pittsburgh



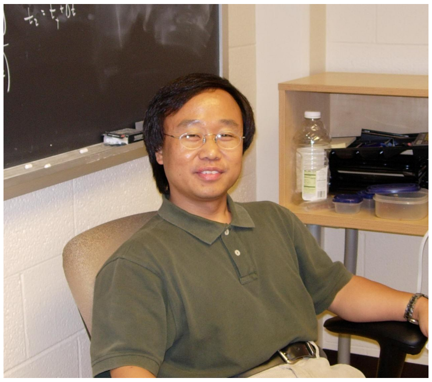
Sookkung Lim
Courant Institute of Mathematical Sciences
New York University

Diego Pol
Department of Earth & Environmental Sciences
Columbia University

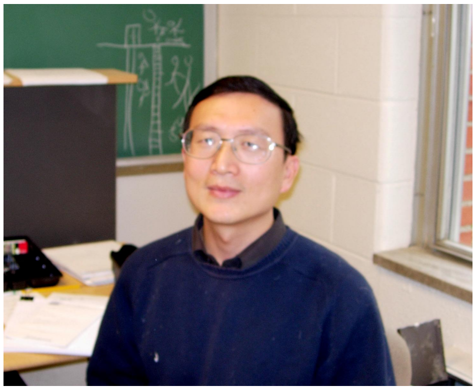
Postdocs

Firas Rassoul-Agha
Courant Institute of Mathematical Sciences
New York University

Mike Stubna
Theoretical and Applied Mechanics
Cornell University



Jianjun (Paul) Tian
Mathematics Department
University of California, Riverside



Zailong Wang
Department of Statistics
University of California, Davis

Jin Zhou
Department of Statistics
University of Georgia