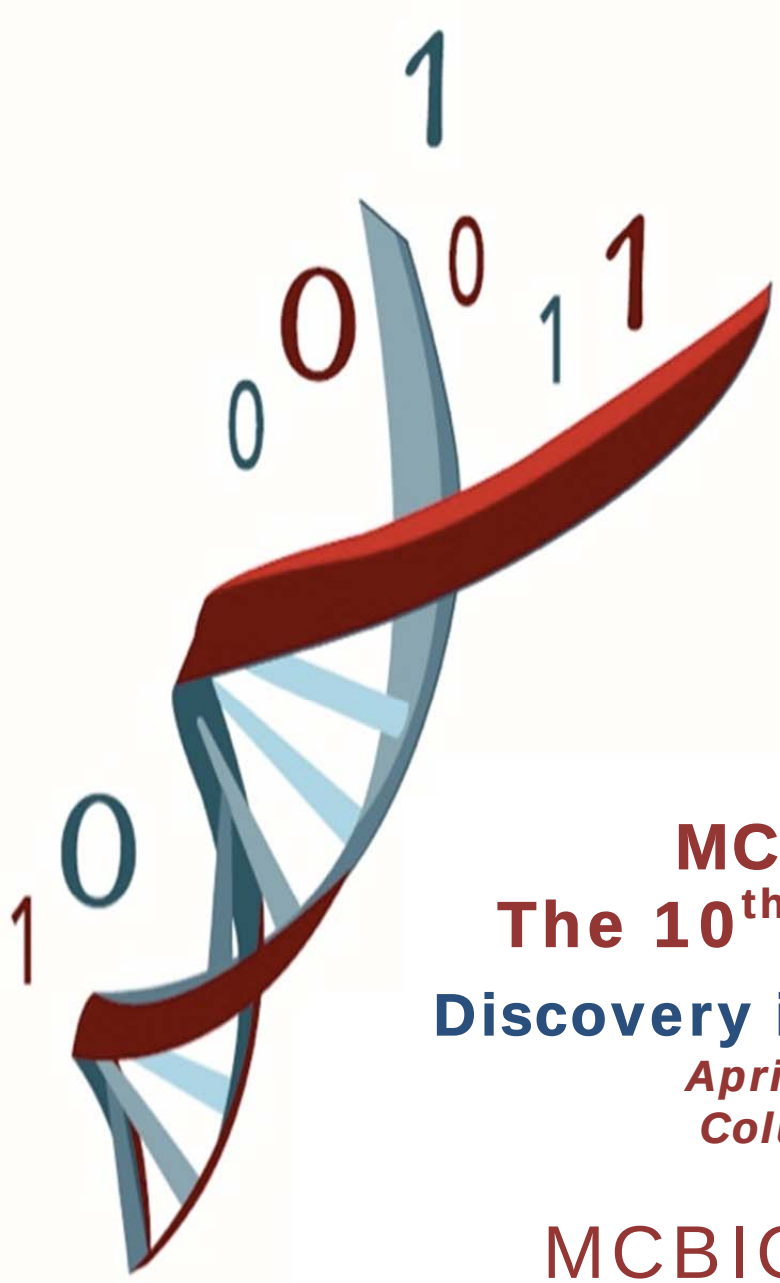


MCBIOS X:
The 10th Anniversary

Discovery in A Sea of Data

April 5-6, 2013
Columbia. MO

MCBIOS.ORG

MIDSOUTH COMPUTATIONAL BIOLOGY & BIOINFORMATICS SOCIETY

Conference Program

MCBIOS Mission Statement

The mission of the MidSouth Computational Biology and Bioinformatics Society (MCBIOS) is to foster networking and collaboration, and promote the professional development of our members.

As stated in the bylaws, we seek to advance the understanding of bioinformatics and computational biology, bring together scientists of various backgrounds and disciplines, facilitate the collaboration of researchers with similar or complementary backgrounds to solve biological, health, and/or medical problems, promote education, inform the general public on the results and implications of current research, and promote other activities that will contribute to the development of bioinformatics and computational biology.

We have a strong orientation toward supporting our student members.

Sponsors:

"Funding for this conference was made possible, in part, by the Food and Drug Administration through grant 1R13FD004229-01. The views expressed in written Conference materials or publications and by speakers and moderators do not necessarily reflect the official policies of the Department of Health and Human Services; nor does any mention of trade names, commercial practices, or organization imply endorsement by the United States Government."

Mid-South Computational Biology and Bioinformatics Society (MCBIOS)

Tenth Annual Conference

Stoney Creek Inn

University of Missouri, Columbia Mo

April 5-6, 2013

Officers

President: Edward Perkins, ERDC, Army Corp of Engineers

Past President: Doris Kupfer, CAMI FAA

President-Elect: Andy Perkins, Mississippi State University

Treasurer: Dennis Burian, CAMI FAA

Secretary: Debra Knisley, East Tennessee State University

Board Members

Gordon K Springer, University of Missouri

Weida Tong, National Center for Toxicological Research, FDA

Joshua Yuan, Texas A&M University

Debra Knisley, East Tennessee State University

Chaoyang (Joe) Zhang, University of Southern Mississippi

2013 Conference Committee

Gordon K Springer, University of Missouri

Ed Perkins, ERDC, Army Corp of Engineers

Doris Kupfer, CAMI FAA

Dennis Burian, CAMI FAA

Joshua Yuan, Texas A&M University

Jian-Lin Cheng, University of Missouri

Dmitry Korkin, University of Missouri

Table of Contents

MCBIOS 2013 CONFERENCE PROGRAM 9

Keynote Speaker John Quackenbush, PhD 18

Keynote Speaker Veronica J. Vieland, PhD 19

Invited Speaker William Slikker, PhD 20

NCBI Database Tools Webinar Facilitator Peter Cooper, PhD 21

MCBIOS Timberland Rattlesnake Genome Project 22

Soybean Knowledge Base (SoyKB) Workshop 24

MCBIOS 2013 Conference Proceedings 25

Oral Presentation Abstracts 27

Systematic classification of common disease-associated regulatory DNA variations by their epigenomic relationship 28

A Bioinformatics Tool for Cross-mapping Microarray-based Genes to Next-generation RNA-Sequencing 29

A Time Series Analysis Method for Gene Regulatory Network Reconstruction 30

Application of topic model to toxicogenomics: clustering gene expression profiles by topic distributions 31

Extensive modulation of circadian transcription cycle by microRNAs 32

SASD: the Synthetic Alternative Splicing Database for Identifying Novel Isoform from Proteomics 33

Combining Probabilistic Graphical Model-based and Knowledge-based Methods for Automatic Construction of Metabolic Pathways 34

Contacts-Assisted Protein Structure Prediction 35

PAGED v2.0: an update to enable molecular phenotype discoveries through gene-set regulatory networks 36

Structural Variants Discovery in the Water Flea Daphnia pulex by Genome Re-Sequencing 37

A Cross Disciplinary Study of Link Decay and the Effectiveness of Mitigation Techniques 38

A molecular dynamics approach for the analysis of the binding of vitamin E analogs to the ?-tocopherol transfer protein 39

The impact of CpG islands in promoters on evolution of gene expression in mammals 40

PREDICTING ALTERNATIVE INDICATIONS OF THE MARKETED DRUGS WITH LATENT DIRICHLET ALLOCATION 41

Effect of treatment effect and sample size on reproducibility in a toxicogenomics study 42

Quantitative Structure-Activity Relationship Models for Predicting Drug-Induced Liver Injury Based on FDA-Approved Drug Labeling Annotation 43

The Properties of Human Genome Conformation and Spatial Gene Interaction and Regulation Networks 44

Topic Modeling on LiverTox: Discovering the Hidden Knowledge for Predicting Potential to Cause Acute Liver Failure of a Drug 45

Structure-based engineering to produce high affinity anti-methamphetamine antibodies 46

Homology Modeling, Molecular Docking, Molecular Dynamics Simulations and Free Energy Calculations Elucidated Rat α-Fetoprotein-Ligand Binding Modes 47

Inference of Genetic Regulatory Networks from Perturbation Data – Application to Melanoma cell lines 48

Bioinformatics Analysis Pipeline for Characterization of Intra-strain Variability in Food-borne Bacteria 49

Statistical analysis of molecular pathways 50

Consensus Spectral Techniques via Vertex Weighted Graphs in the Analysis of Genomic and Proteomic Data 51

Model quality assessment of proteins using group-based redundancy 52

Inference of Temporally rewiring genetic networks using time-varying differential equations 53

Comparative Analyses Demonstrated Reliability of 1000 Genome Project and Confirmed Usefulness of GWAS Findings Based on SNP-Arrays 54

A New Hidden Markov Model for the Compatibility between Protein Sequence and Structure 55

A study and extension of DNcon: a method for protein residue-residue contact prediction using deep networks 56

Determining beneficial and detrimental effects of non-synonymous SNPs on protein-protein interactions using supervised and semi-supervised learning 57

Choosing the right coverage depth and read length for an RNA-seq experiment 58

Towards the integration, annotation and association of historical microarray experiments with RNA-seq 59

Drug Activity Prediction Using Multi-Instance Learning via Joint Instance and Feature Selection 60

BinAligner: a heuristic method to align biological networks 61

Discrimination of fungal secretory (effector) proteins from plant host secretory proteins 62

Several quartet-based methods to reconstruct phylogenetic networks 63

Towards Human-Computer Synergistic Analysis of Large-Scale and Complex Biological Data 64

Pathway mergeability enables the evaluation of merging signaling pathways with protein interaction data 65

Identification of A-to-I RNA Editing Sites in Honey Bee (Apis mellifera) Using RNA-Seq Data 66

Rule Based Regression and Feature Selection for Biological Data 67

Topological Data Analysis of Triple Negative Breast Cancer Transcriptome and Proteome 68

Identification of genes specific to the lineage of the Hymenopteran insect, honey bee (Apis mellifera) 69

What can we learn from phenotypes? 70

Predicting Multi-target Protein Subcellular Localization combining homology and Machine Learning

Approaches **71**

Bioinformatics Approaches to Deciphering Host-Pathogen Protein Interaction Networks (PINs) **72**

NeedleFinder: A Data Analysis Tool for LCMS-Based Metabolomics **73**

MULTICOM – RNA-Seq Data Analysis of Mouse Transcriptomes Perturbed by Botanicals **74**

Local and global quality assessment by MULTICOM during CASP10 ..... **75**

Paternal Influence on Transcriptomic Landscapes of *in vitro* derived bovine embryos **76**

A Scalable and Deterministic Finite State Automaton-based Model to Determine Ancestor-Descendant Relationships in Directed Acyclic Graphs **77**

Contacts-Assisted Protein Structure Prediction **78**

Constructing Three-Dimensional Structures of Human Chromosomes from Chromosomal Contact Data **79**

Poster Abstracts **80**

Analysis of co-localization of 5-hydroxymethylcytosine and G-quadruplex-forming regions in gene transcriptional start sites: might hydroxymethlation affect the C-rich i-motif structure? **81**

Comparison of Data Mining Methods on Microarray Gene Expression Data on Cancer **82**

Comparing transcriptome response to amphetamine and environment induced hyperthermia in rat brains and blood **83**

Iterative reconstruction of three-dimensional model of human genome from chromosomal contact data **84**

Bridging the gap between soybean translational genomics and breeding with Soybean Knowledge Base (SoyKB) **85**

SoyKB can be publicly accessed at http://soykb.org.Clustering Gene Expression Data using Probabilistic Non-negative Matrix Factorization **85**

Predicting Protein Model Quality from Sequence Alignment by Support Vector Machines **87**

A Bioinformatics Study on Evolutionary Diversification of Multiple Inositol Polyphosphate Phosphatase1 as an Aid to Understanding its Functional Significance in Mammalian Systems **88**

GMOL: A Tool for 3D Genome Structure Visualization **89**

Molecular Dynamic Simulation of β-amyloid peptide **90**

A Bioinformatics Study on Evolutionary Diversification of Multiple Inositol Polyphosphate Phosphatase1 as an Aid to Understanding its Functional Significance in Mammalian Systems **91**

A Constrained Importance Sampling Approach for Inference of Time-Varying Gene Regulatory Networks **92**

Challenges in Inferring Large-Scale Networks **93**

NETS TOOL: A computational Approach to predict potential drug candidates **94**

Inference of Temporally rewiring genetic networks using time-varying differential equations **95**

In-silico evaluation of chemical interactions between active chemical constituents from ayurvedic medicinal plants and carrier protein ligands **96**

Uncovering Protein-Protein Interactions in Brassica napus Using Integrative Methods **97**

Intergenic Region Analysis Pipeline for Bacteria (BIRAP): A tool for analyzing expression profile of intergenic regions generated by RNA-seq....98

A "core genome" approach to decoding host-pathogen dual RNA-Seq data....99

Proteogenomic mapping of bovine respiratory disease pathogens....100

PCA Based Analysis for Classification of Multiple Myeloma Microarray Data....101

The Genome of Reniform Nematode, Rotylenchulus reniformis....102

Probabilistic Methods for Accurate Mapping of Metatranscriptomic Sequence Data....103

A Distributed CPU-GPU Framework for Pairwise Alignments on Large-Scale Sequence Datasets....104

Performance of Hadoop Based REMD on Cloud Computing Platform....105

Transcriptional and epigenetic variation: two integrative ways to explain the soybean innate immunity triggered by PAMPs....106

Analysis of correlation between b and y ion in tandem mass spectra library....107

Protein interaction binding site prediction for comparative models by combining sequential and structural properties....108

Transgenerational Effects of Chronic Low-Dose Radiation Exposure in Medaka Fish model....109

A language-based approach to mining host-pathogen interactions from biomedical literature....110

TOPOLOGICAL NETWORK ALIGNMENT BASED ON GRAPHLET DEGREE SIGNATURE....111

NeedleFinder, a Server for Binning and Statistical Analysis of Mass Spectra Data....112

Coherence in Evolution: An Influenza Story....113

New method for efficiently sampling protein 3D conformations....114

Unexpected evolutionary recursive patterns in influenza A proteins....115

PHDcleav: A SVM-based method for predicting human Dicer cleavage sites using sequence and secondary structure of miRNA precursors....116

A Comparative Study of Linear and Nonlinear Dimensionality Reduction Methods Using Gene Expression Data....117

Computational analysis of 2D protein gel images for identification of differentially expressed proteins associated with resistance to aflatoxin accumulation in maize....118

Classification and Feature Selection Using Hybrid Top Pairs on Microarray Data....119

Top-Down Transversing the Gene Ontology to Extract Biological Knowledge for Biomedical Models 120

Identification of protein coding genes in the plant-parasitic nematode Rotylenchulus reniformis through comparative proteogenomic mapping with Caenorhabditis elegans....121

Circles within circles: Crosstalk between protein Ser/Thr/Tyr-phosphorylation and methionine oxidation....122

MOLECULAR MODELING OF THE INHIBITORY EFFECTS HUMAN CYTOCHROME CYP3A4 BY DILLAPIOL DERIVATIVES....123

Combining Next-generation Sequencing and Comparative Genomics to Identify Novel microRNA. ...124

Quantitative RT-PCR data analysis of RNA transport pathway genes associated with resistance to

aflatoxin accumulation in maize.....125

De novo Transcriptome Assembly of the Plant-Parasitic Nematode Rotylenchulus reniformis.....126

Identification of oncogenic pathways of breast cancer in a genome wide association study.....127

Analysis of Military Unique Chemical-Induced Neurotoxicity through Gene Regulatory Network Reconstruction.....128

An evolution-inspired computational framework for computer-aided molecular design 129

Differential Reconstructed Pathways for Deriving Toxicity Thresholds in Chemical Risk Assessment .130

Utilizing RNASeq in Eukaryotic Genome Annotation and Genome Databases.....131

Role of Global and Targeted hypomethylation in cancer 132

PLpred: a bioinformatics system for the identification and classification of plastid type proteins.....133

Constructing Three-Dimensional Structures of Human Chromosomes from Chromosomal Contact Data 134

Statistical Methods for Ambiguous Sequence Mappings 135

Evaluation and application of microarray data: normalization to gene network inference. 136

A Graphical Processing Unit Supported Neuroimaging Software in JAVA.....137

Fractal approach for automatic malignancy determination in dermoscopy images.....138

Transcriptome Analysis of an Oleaginous Filamentous Fungus for Novel Biomass Consolidated Bioprocessing Model.....139

Armband tracker and reminder for patients with Dementia and Alzheimer's disease.....140

MCBIOS 2013 CONFERENCE PROGRAM

"Discovery in a Sea of Data"

Thursday, April 4, 2013

2:00 PM - 5:00 PM	Workshop on Array track and More - Weida Tong – Columbia Room
5:00 PM - 7:00 PM	Registration Desk Open Poster Setup in Rooms: Pines, Timberlands, & Meadows NOTE: Posters will be on display for entire Conference

Friday, April 5, 2013

7:45 AM - 8:30 AM	Registration and coffee, Poster set-up Poster Setup in Rooms: Pines, Timberlands, & Meadows
8:30 AM - 10:00 AM	**General Session – Salons A & B**
8:30 - 8:45 AM	Welcome and Introductory Remarks Ed Perkins, MCBIOS President Gordon Springer, MCBIOS X Conference Chair University of Missouri – Welcome - Gary Allen, UM VP for Information Technology and CIO
8:45 AM - 9:45 AM	**Keynote Address 1** John Quackenbush, Director, Center for Cancer Computational Biology (CCCB) Dana Farber Cancer Institute, Harvard University **"The Road to Genomic Medicine is Paved with Data"**
9:45 AM - 10:00 AM	**Conference Break – Salon C**

10:00 AM - 12: 00 Noon Friday	
Lewis & Clark Room	NCBI Database Tools Webinar Peter Cooper, Facilitator

10:00 AM - 11: 00 AM **Friday Morning Parallel Sessions**	
Track 1 **Computation I** **Salon A**	J. Wang Model quality assessment of proteins using group-based redundancy S. Liu Rule based regression and feature selection for biological data J. Yang Several quartet-based methods to reconstruct phylogenetic networks Steven Embry NeedleFinder: A Data Analysis Tool for LCMS-Based Metabolomics
Track 2 **Biology I** **Salon B**	A.K. Bennett Identification of genes specific to the lineage of the Hymenopteran insect, honey bee (Apis mellifera) A. Markovets Topological data analysis of triple negative breast cancer transcriptome and proteome S. Tao Identification of A-to-I RNA editing sites in honey bee (Apis mellifera) using RNA-seq data Robyn Kelley Bioinformatics Approaches to Deciphering Host-Pathogen Protein Interaction Networks (PINs)

11:00 AM - 11:15 AM Conference Break – Salon C

11:15 AM - 12: 00 PM Friday Morning Parallel Sessions	
Track 1 **Computation I** **Salon A**	Zhiquan He A new hidden Markov Model for the compatibility between protein sequence and structure J.S. Yuan Statistical analysis of molecular pathways Awantika Singh A molecular dynamics approach for the analysis of the binding of vitamin E analogs to the α-tocopherol transfer protein
Track 2 **Biology I** **Salon B**	Rahul Singh Towards Human-Computer Synergistic Analysis of Large-Scale and Complex Biological Data Mihir Jaiswal Bioinformatics analysis pipeline for characterization of intra-strain variability in food-borne bacteria Yugang Ban The impact of CpG islands in promoters on evolution of gene expression in mammals

12:00 PM -1:20 PM Lunch and Business Meeting

12:00 PM-12:45 PM **Lunch - Salon C**

12:45 PM- 1:20 PM Announcements
Business Meeting, Ed Perkins, President
Treasurer's Report, Dennis Burian
Election of Board (All Members) - Doris Kupfer

1:30 PM - 2: 30 PM Friday Afternoon Parallel Sessions	
Track 1 **Computation II** **Salon A**	Jason Hennessey A cross disciplinary study of link decay and the effectiveness of mitigation techniques Chayaporn Suphavilai PAGED v2.0: an update to enable molecular phenotype discoveries through gene-set regulatory networks Jilong Li MULTICOM – RNA-Seq Data Analysis of Mouse Transcriptomes Perturbed by Botanicals Renzhi Cao Local and global quality assessment by MULTICOM during CASP10
Track 2 **Biology II** **Salon B**	Mohammed Rasheed Inference of genetic regulatory networks from perturbation data – application to Melanoma cell lines Sule Dogan Paternal Influence on Transcriptomic Landscapes of in vitro derived bovine embryos Bhavsar-Jog Yogini Analysis of co-localization of 5-hydroxymethylcytosine and G-quadruplex-forming regions in gene transcriptional start sites: might hydroxymethlation affect the C-rich i-motif structure? Andrew Overton A Scalable and Deterministic Finite State Automaton-based Model to Determine Ancestor-Descendant Relationships in Directed Acyclic Graphs

2:30 PM - 2:45 PM Conference Break – Salon C

2:45 PM - 3:45 PM	**Friday Afternoon Parallel Sessions**
Track 1 **Computation III** **Salon A**	Debswapna Bhattacharya Improving Structural Accuracy of Protein Models by a Decoy-Based Iterative Fragment Exchange Greedy Sampling and Selection Strategy Annick Dongmo Inference of Temporally rewiring genetic networks using time-varying differential equations Badri Adhikari Contacts-Assisted Protein Structure Prediction Zhenqiang Su A Bioinformatics Tool for Cross-mapping Microarray-based Genes to Next-generation RNA-Sequencing
Track 2 **Biology III** **Salon B**	Tuan Trieu Constructing Three-Dimensional Structures of Human Chromosomes from Chromosomal Contact Data Shweta Chavan Towards the integration, annotation and association of historical microarray experiments with RNA-seq Shraddha Thakkar Structure-based engineering to produce high affinity anti-methamphetamine antibodies Mikhail Dozmorov Systematic classification of common disease-associated regulatory DNA variations by their epigenomic relationship

3:45 PM - 4:00 PM **Conference Break**

4:00 PM - 6:00 PM **Poster Session, Pines, Timberlands, & Meadows**

Poster Presenters Must be Present for Judging

Friday Evening April 4, 2013

6:15 PM - 7:15 PM **General Session**
Salon A&B

Keynote Address II
Veronica J. Vieland, Vice President for Computational Research and Director of the Battelle Center for Mathematical Medicine at The Research Institute at Nationwide Children's Hospital, and Professor in the Department of Pediatrics at The Ohio State University College of Medicine and the Department of Statistics.
"Is the Universe Made of Information?"

7:15 PM - 8:30 PM **Dinner Session - Salon C**

8:15PM - 8:30PM William Slikker - "10 Year Celebration of MCBIOS"
MCBIOS Board 2014 Election Results
Introduction to MCBIOS 2014 - Rakesh Kaundal

Saturday, April 6, 2013

8:00 AM – 9:00 AM Breakfast Business Meeting (Board Members Only)
Lewis & Clark Room

9:00 AM – 10:00 AM **Special Interest Sessions**

Salon A Distributed Rattlesnake Annotation Project – Ed Perkins

Salon B MCBIOS Student Council Meeting – Shraddha Thakkar

10:00 AM - 10:30 AM **Conference Break - Salon C**

Posters remain on display in Pines, Timberlands, & Meadows

10:30 AM - 11:30 AM **Saturday Morning Parallel Sessions**	
Track 1 **Computation IV** **Salon A**	Mikyung Lee Application of topic model to toxicogenomics: clustering gene expression profiles by topic distributions Sitanshu Sahu Predicting Multi-target Protein Subcellular Localization combining homology and Machine Learning Approaches Jialiang Yang BinAligner: a heuristic method to align biological networks
Track 2 **Biology IV** **Salon B**	Ruchi Verma Discrimination of fungal secretory (effector) proteins from plant host secretory proteins Jesse Eickholt A study and extension of DNcon: a method for protein residue-residue contact prediction using deep network

11:30 AM – 12:45 PM **Conference Lunch – Salon C**

1:00 PM - 3: 30 PM **Saturday - Soybean Knowledge Base (SoyKB) Workshop**	
Lewis & Clark Room	**Soybean Knowledge Base (SoyKB) : A Powerhouse for Soybean Research and Breeding** Trupti Joshi & Dong Xu, University of Missouri

12:45 PM - 2:00 PM **Saturday Afternoon Parallel Sessions – Part 1**	
Track 1 **Computation V** **Salon A**	Zhendong Zhao Drug activity prediction using multiple-instance learning via joint instance and feature select Nan Zhao Determining beneficial and detrimental effects of non-synonymous SNPs on protein-protein interactions using supervised and semi-supervised learning Halil Bisgin Predicting alternative indications of the marketed drugs with latent Dirichilet allocation

Track 1 (continued) **Computation V** **Salon A**	Qi Qi Combining probabilistic graphical model-based and knowledge-based methods for automatic construction of metabolic pathways Xiaogang Wu Pathway mergeability enables the evaluation of merging signaling pathways with protein interaction data
Track 2 **Biology V** **Salon B**	Binsheng Gong Effect of treatment effect and sample size on reproducibility in a toxicogenomics study Zheng Wang The properties of human genome conformation and spatial gene interaction and regulation networks Wenqian Zhang Comparative analyses demonstrated reliability of 1000 genome project and confirmed usefulness of GWAS findings based on SNP-arrays Jie Shen Homology modeling, molecular docking, energy calculations elucidated rat-fetoprotein- molecular dynamics simulations and free ligand binding modes Ke Yu Topic Modeling on LiverTox: Discovering the Hidden Knowledge for Predicting Potential to Cause Acute Liver Failure of a Drug

2:00 PM - 2:15 PM **Conference Break - Salon C**

2:15 PM - 3:30 PM **Saturday Afternoon Parallel Sessions**	
Track 1 **Computation VI** **Salon A**	Andrey Ptitsyn Extensive modulation of circadian transcription cycle by microRNAs Zhenqiang Su A Bioinformatics Tool for Cross-mapping Microarray-based Genes to Next-generation RNA-Sequencing Fan Zhang SASD: the Synthetic Alternative Splicing Database for Identifying Novel Isoform from Proteomics

Track 1 (continued) **Computation VI** **Salon A**	Minjun Chen Quantitative Structure-Activity Relationship Models for Predicting Drug-Induced Liver Injury Based on FDA-Approved Drug Labeling Annotation
Track 2 **Biology VI** **Salon B**	Hua Li Choosing the right coverage depth and read length for an RNA-seq experiment Toni Kazic What can we learn from phenotypes? Debra Knisley Consensus Spectral Techniques via Vertex Weighted Graphs in the Analysis of Genomic and Proteomic Data Chaoyang Zhang A Time Series Analysis Method for Gene Regulatory Network Reconstruction Ping Gong Structural Variants Discovery in the Water Flea Daphnia pulex by Genome Re-Sequencing

3:30 PM - 4:15 PM **Awards Ceremony & Photo Session – Salons A & B**
Award winners announced and photos of all recipients

4:15 PM **Conclusion**

Please don't forget to turn in your evaluations!

Keynote Speaker John Quackenbush, PhD

John Quackenbush, Dana-Farber Cancer Institute and the Harvard School of Public Health

John Quackenbush received his PhD in 1990 in theoretical physics from UCLA working on string theory models. Following two years as a postdoctoral fellow in physics, Dr. Quackenbush applied for and received a Special Emphasis Research Career Award from the National Center for Human Genome Research to work on the Human Genome Project. He spent two years at the Salk Institute and two years at Stanford University working at the interface of genomics and computational biology. In 1997 he joined the faculty of The Institute for Genomic Research (TIGR) where his focus began to shift to understanding what was encoded within the human genome. Since joining the faculties of the Dana-Farber Cancer Institute and the Harvard School of Public Health in 2005, his work has focused on the analysis of human cancer and expanded to embrace systems-based approaches to understanding and modeling biological problems.

Title: The Road to Genomic Medicine is Paved with Data and Information

Abstract:
Since the introduction of second-generation DNS sequencing technologies in 2007, the cost of genome sequencing has been consistently by 33% per quarter, with the $1000 genome arriving in 2012 and the $100 genome not far off. As DNA sequencing increasingly becomes a commodity, biomedical research is rapidly evolving from a purely laboratory science to an information science in which the winners in the race to cure disease are likely to be those best able to collect, manage, analyze, and interpret data. Here I will provide an overview of the approach we have been developing to deal with the challenge of personal genomic data, including integrative approaches to data analysis and the creation of data portals focused on addressing the most common use cases presented by different user constituencies. By effectively collecting genomic and clinical data and linking information available in the public domain, we have made significant advances in uncovering the cellular networks and pathways that underlie human disease and building predictive models of those networks that may help to direct therapies.

Keynote Speaker Veronica J. Vieland, PhD

Veronica J. Vieland, PhD

Veronica J. Vieland, Ph.D., is Vice President for Computational Research and the Director of the Battelle Center for Mathematical Medicine at The Research Institute at Nationwide Children's Hospital. She is a Professor in the Department of Pediatrics at The Ohio State University College of Medicine and in OSU's Department of Statistics. Dr. Vieland's research focuses on statistical and computational methods for discovery of genetic influences on human disease. She is also interested more generally in the measurement of evidence in biomedical research.

Title: Is the universe made of information?

Abstract

Previous work has suggested deep connections between statistical mechanics and certain aspects of both information theory and statistical inference, based primarily on the shared concept of entropy. In this talk I go beyond familiar information theoretic treatments of entropy to develop purely information-based interpretations of both the 1st and 2nd laws of thermodynamics. This allows us to ask and answer a question that has gone begging until now: What is the analogue of temperature (T) on the information/inferential side? I argue that the physical quantity T has a familiar, but surprising, interpretation as *statistical evidence*. Moreover, this formulation provides a template for measuring evidence on an absolute (Kelvin) scale for the first time. This has far reaching implications for bioinformatics, since the measurement and interpretation of statistical evidence is a critical element of how we make scientific use of bioinformatic results. In a more speculative vein, this work also raises the question of whether our physical theories require us to posit the existence of matter. If fundamental physical laws can be interpreted in purely informational terms, perhaps it is mathematically cogent to say that the universe is in fact made of information.

Invited Speaker William Slikker, PhD

Title: 10 Year Celebration of MCBIOS

William Slikker, Jr., Ph.D.
Director, National Center for Toxicological Research Jefferson, AR

Dr. William Slikker, Jr. is the Director of the Food and Drug Administration's (FDA) National Center for Toxicological Research (NCTR). He received his Ph.D. in Pharmacology and Toxicology from the University of California at Davis in 1978. Dr. Slikker holds Adjunct Professorships in the Departments of Pediatrics, and Pharmacology and Toxicology at the University of Arkansas for Medical Sciences. Dr. Slikker is currently Vice President of the Society of Toxicology (SOT) and Associate Editor for *NeuroToxicology* and *Toxicological Sciences*. He was past Treasurer, SOT and past President of The Academy of Toxicological Sciences. He held committee chairmanships or elected offices in several other scientific societies including the Teratology Society (serving as President); the American Society for Pharmacology and Experimental Therapeutics (Chair, Developmental Pharmacology Section and member of the Program Committee); and co-founder and past President of the MidSouth Computational Biology and Bioinformatics Society. Dr. Slikker has authored or co-authored over 300 publications in the areas of transplacental pharmacokinetics, developmental neurotoxicology, neuroprotection, systems biology, and risk assessment. He has also served on several national/international advisory panels for HESI/ILSI, CIIT Centers for Health Research, EPA, NIEHS, NAS, NIH, and WHO

NCBI Database Tools Webinar Facilitator Peter Cooper, PhD

Peter Cook, PhD
Director Scientific Outreach and Training, NCBI, Washington DC

Peter Cooper directs the scientific outreach and training program for the National Center for Biotechnology Information at the National Library of Medicine. Peter has conducted and developed training courses for biologists in the use of NBCI molecular databases and has provided scientific user support for the NCBI since 1998.

Prior to joining the NCBI, Peter pursued diverse biological research interests including peptide neurochemistry, marine environmental toxicology, and taught biology and chemistry. Peter earned a BS from Virginia Tech, a MA in chemistry from the Johns Hopkins University and a Ph.D. in Marine Science from the College of William and Mary, School of Marine Science in 1996.

Advanced data access and analysis at the National Center for Biotechnology Information

After a review of the basics of using the Entrez system and BLAST search tool, you will explore advanced ways to access biomedical data at NCBI as well as some of the specialized analysis tools provided by the NCBI. You will learn about the Entrez Assembly, Genome and BioProject databases as useful entry points and how to find and download large datasets through FTP and Aspera protocols. NCBI provides application programming interfaces to both the Entrez system (E-utilities) and the BLAST system (URL API). You will learn the basics of using these to search and download records. The standalone BLAST package also can access the web BLAST service using the remote option. You will learn how use BLAST+ programs as clients to run searches at NCBI. Genome Workbench is a standalone package that also provides client access to Entrez and BLAST services. You will see how to set-up Genome Workbench and use it to download and manipulate and analyze sequences from the NCBI site.

Those with laptop computers can follow the live examples on the NCBI website. You will need to download and install NCBI software and data to follow some of the examples.

Useful software:
-BLAST: ftp://ftp.ncbi.nih.gov/blast/executables/LATEST/
-Genome Workbench: http://www.ncbi.nlm.nih.gov/tools/gbench/

MCBIOS Timberland Rattlesnake Genome Project

Co-Facilitators Edward J. Perkins, PhD
and William "Shane" Sanders, PhD

Edward J. Perkins
Senior Research Scientist (ST)
Environmental Networks and Genetic Toxicology
Environmental Laboratory
ERDC, U.S. Army Corps of
Engineers Vicksburg, MS

Dr. Edward J. Perkins currently serves as Senior Research Scientist (ST) in Environmental Networks and Genetic Toxicology in the U.S. Army ERDC, Environmental Laboratory. Prior to joining ERDC, Dr Perkins worked in development of transgenic plants for phytoremediation and molecular measures of soil quality. Dr. Perkins joined the ERDC Environmental Laboratory in 1996 where he established a genetics research lab. His research focuses on using biological networks and systems biology to understand chemical impacts on animals, bioinspired approaches for novel materials, the use of gene expression to monitor adverse environmental impacts, the use of environmental DNA to monitor invasive species, and the effect of military activities on genetic viability of threatened and endangered species on Department of Defense lands.

Co-Facilitator:
William S. "Shane" Sanders, Post-Doctoral Fellow,
Institute for Genomics, Biocomputing & Biotechnology
Mississippi State University
Starkville, MS

MCBIOS Timberland Rattlesnake Genome Project

Saturday, April 6, 2013 – Columbia, MO
9:00-9:50 AM
Project Home: http://rattles.mcbios.org/

Facilitators: Edward J. Perkins and William S. Sanders

9:00am – 9:10am: Introduction & Background

- Project History
- Why Rattlesnake?

9:10am – 9:20am: Project Overview

- Available Data
- Available Resources
- Current Progress

9:20am - 9:40am: Participation & Working Groups

- Assembly
- Annotation
- Comparative Genomics
- Data Management

9:40am – 9:50am: Project Timeline, Scheduling Regular Progress Meetings

Soybean Knowledge Base (SoyKB) Workshop

Saturday, April 6, 2013 – Columbia, MO
1:00-3:30 PM

Project Home: http://soykb.org

Facilitators: Trupti Joshi and Dong Xu

SoyKB: A Powerhouse for Soybean Research and Breeding

Soybean Knowledge Base (SoyKB) is a comprehensive all-inclusive web resource for soybean translational genomics. SoyKB is designed to handle the management and integration of soybean genomics, transcriptomics, proteomics and metabolomics data along with annotation of gene function and biological pathway. It provides a wide range of tools including pathway viewer, genome browser, breeding program, 3D protein structure viewer, etc. for visualization and integration of data for breeding and multi-omics analysis.

During the workshop, we will introduce and provide hands-on computational demonstrations featuring various functionalities and tools available in SoyKB. Participants will be provided detailed tutorial material to follow along. Participants are encouraged to bring along their laptops.

Topics covered:

- SoyKB resources for gene, miRNA, metabolite and SNP searches.
- Discover available tools in SoyKB.
- Explore Missouri Breeding Program tools and Chromosome Visualizer for PI, QTL and Trait searches.
- Build modules in Cyber Studio system integrating multi-omics data.

Please RSVP at: http://soykb.org/MCBIOSworkshop

MCBIOS 2013 Conference Proceedings

MCBIOS 2013 presenters who had their poster or platform abstracts accepted for presentation are eligible to submit a full paper on the work they presented to be considered for formal, peer-reviewed publication in the conference proceedings. The proceedings will appear in a special issue of *BMC Bioinformatics*. **Past MCBIOS Proceedings have yielded an average impact factor of 5.68** over the 5 years we have data for, which speaks strongly of the impact of MCBIOS and its participants in bioinformatics. The deadline for submission of these papers is **Monday, May 6th, 2013.**

BMC Bioinformatics is an open access, peer-reviewed journal that considers articles on all aspects of the development, testing and novel application of computational and statistical methods for the modeling and analysis of all kinds of biological data, as well as other areas of computational biology. Submissions must be within this scope of interest and represent original work. **Important note:** This year, unlike previous years, BMC has requested that we restrict acceptance to the top 15 papers.

Specific formatting instructions for Proceedings papers can be found on their website (http://www.biomedcentral.com/info/authors/instprepdoc). Note that this is a different web address than the one for their regular papers. Authors of accepted papers will be asked to **pay an article processing charge of £615** (about $1,000 US at current exchange rates), an amount discounted for this event from the normal $1,960 charge. Because this is a special issue, fee waivers and institutional discounts do not apply.

If you intend to submit a paper, please send your tentative title/abstract to the Senior Editor, (Jonathan-Wren@OMRF.org) as soon as possible to enable us to better plan for reviews, paper handling, etc. Submitted papers also need to be sent to the Senior Editor.

Timeline:

May 6, 2013	– Manuscripts should be submitted to Jonathan Wren
June 7, 2013	– Reviewers return comments to editors
June 30, 2013	– Revisions due back from authors
July 14, 2013	– Final decisions made on submitted papers by reviewers
July 21, 2013	– Editors notify authors of acceptability of papers
Aug 14, 2013	– All final manuscript revisions due to editors along with payment of article processing charges due to MCBIOS

MCBIOS XI

Theme: From Genome to Phenome - Connecting the Dots

Location: Oklahoma State University, Stillwater

The Wes Watkins Conference Center

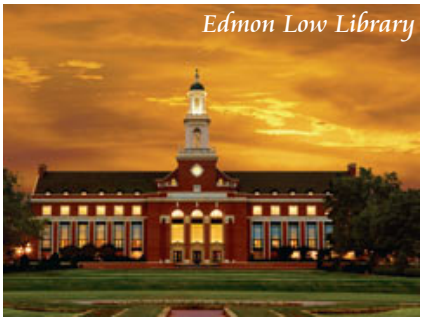
Dates: March 27-29, 2014

ORGANIZING COMMITTEE:

Conference Chair: Rakesh Kaundal

Program Chairs: Ulrich Melcher
Doris Kupfer

Organizing Committee: Members of the OSU *i*CREST Center for Bioinformatics, and a cast of many volunteers



Stillwater: Home of the COWBOYS!

Oral Presentation Abstracts

Systematic classification of common disease-associated regulatory DNA variations by their epigenomic relationship

Mikhail G. Dozmorov, OMRF
Cory B. Giles, OMRF
Kristi Koelsch, OMRF
Jonathan D. Wren, OMRF

Background: The success of genome-wide association studies (GWASs) in finding causative genomic variants for Mendelian phenotypes is nonetheless unable to elucidate complex patterns and biological roles of mutations underlying non-Mendelian inheritance. Our motivation was to use disease-associated SNP correlations with the functional genome annotation data for systematic classification of diseases by their (epi)genomic background.

Methods: Human disease-specific sets of SNPs were extracted from the UCSC GWAS catalog and tested for statistically significant associations with other UCSC genome annotation data. All SNPs from the GWAS catalog were used as a background for testing disease-specific sets of SNPs for statistically significant associations with genome annotations using Fisher's exact test.

Results: 212 disease and 363 trait/phenotype associated sets of SNPs were tested for associations with >1,000 genome annotation features. We identified that diseases/traits of similar origin tend to act within similar genome annotation context, implying similar effect upon (epi)genomic elements. Our results suggest that there may be common factors within disease etiology, such that different disease-associated mutations may lead to equivalent outcomes in disease pathology. Identifying these themes will empower us to interpret GWAS data in terms of unifying mechanisms rather than on a gene-by-gene basis. .

Conclusion: The vast and growing amount of genome annotation data contains enormous potential to interpret sets of disease-associated mutations within a common, unifying theme of functional (epi)genomic elements. Conversely, similarities and differences in (epi)genomic context of disease- and trait-associated SNPs provide a new means to classify diseases and understand their common (epi)genomic denominators.

A Bioinformatics Tool for Cross-mapping Microarray-based Genes to Next-generation RNA-Sequencing

Zhenqiang Su, National Center for Toxicological Research, US Food and Drug Administration
Hong Fang, National Center for Toxicological Research, US Food and Drug Administration
Huixiao Hong, National Center for Toxicological Research, US Food and Drug Administration
Don Ding, National Center for Toxicological Research, US Food and Drug Administration
Binsheng Gong, National Center for Toxicological Research, US Food and Drug Administration
Joshua Xu, National Center for Toxicological Research, US Food and Drug Administration
Weigong Ge, National Center for Toxicological Research, US Food and Drug Administration
Roger Perkins, National Center for Toxicological Research, US Food and Drug Administration
Weida Tong, National Center for Toxicological Research, US Food and Drug Administration

Microarrays have been used to profile global gene expressions for more than fifteen years. Recently, next-generation RNA sequencing (RNA-Seq) has emerged as an alternative technology for transcriptome analysis with promising capabilities in the discovery of splicing junctions, transcripts, and un-annotated genes. Both technologies have been demonstrated to be comparable and complementary to each other in a variety of real genomics applications. Therefore, the integration of information obtained from microarrays and from RNA-Seq will provide a more comprehensive view of real biology, which in turn increases our understanding of human illnesses and diseases and would help in developing more effective treatments. However, it is difficult to directly combine data generated with microarrays and RNA-Seq due to the complexity of transcriptome analysis as well as the complications in microarray probe design and annotation. Here we present a bioinformatics tool to facilitate the integration of microarrays and RNA-Seq data based on either gene ID mapping or probe sequence chromosome location mapping. It comes with a user-friendly interface and is open for public use. Currently, the tool supports gene mapping from Affymetrix, Illumina, and Agilent microarrays to RNA-Seq genes which can be either known genes from RefSeq, UCSC, Ensembl, or novel genes from de novo transcript assembly (such as those generated by Cufflinks or Scripture). In conclusion, our mapping tool provides a missing link from microarray to RNA-Seq for speeding up the use of new biotechnology and knowledge discovery from both RNA-seq and microarray.

A Time Series Analysis Method for Gene Regulatory Network Reconstruction

Chaoyang Zhang
Xi Wu
Nan Wang
Ping Gong

Gene regulatory networks (GRNs) are commonly viewed as dynamic systems, consisting of a set of components whose properties change in response to internal interactions and external signals. There is an increasing demand for modeling and inferring GRNs from gene expression data to better understand the functioning of cellular organisms, address why complicated response patterns to stressors are observed, and generate hypotheses for experimental verification. The complexity of the eukaryotic transcriptional regulation machinery makes GRN reconstruction from time series gene expression data a very challenging task. These challenges become exacerbated when analysis of a set of time series experiments representing various environmental conditions or treatment is desired. The existing GRN reconstruction approaches such as probabilistic Boolean networks and dynamic Bayesian networks have various limitations and have relatively low accuracy or significant computational overhead. In this work, we introduced a times series analysis method to reduce the number of parameters in a simplified linear model (SLM) for improving the accuracy of GRN reconstruction. A total of 20 synthetic GRNs were generated by GeneNetWeaver. These networks were used to evaluate the performance of SLM and to compare it with state space model (SSM). Our results show that SLM performs better than SSM and SLM is more stable with a different length of hidden variables. Moreover, SLM can be applied to a smaller dataset than SSM since it has less number of parameters.

Application of topic model to toxicogenomics: clustering gene expression profiles by topic distributions

Mikyung Lee
Zhichao Liu
Hong Fang
Weida Tong

Background

Over the last few decades, toxicogenomics has been transforming traditional phenotype observation-based toxicology into a predictive science with the development of a variety of algorithms and models, leading to the systematic understanding of toxicological mechanism. The application of appropriate computational method can help uncover underlying patterns in the toxicogenomics dataset. The topic modeling is widely used technique in the field of text mining, however it was not actively applied to toxicogenomics in spite of their similar nature of data structure. In this study, we used topic distributions to cluster 131 compound treated samples, demonstrating the applicability of topic modeling in this field.

Method

We used a combined approach of topic modeling and network analysis in order to cluster gene expression profiling of TGP (Toxicogenomics Project in Japan), which is composed of an *in-vitro* test set with three time points and three dose levels, and an *in-vivo* single and repeated treatment with four time points and three dose levels for 131 compounds. The topic model was constructed from documents to record DEG (Differentially Expressed Gene) information differentiated by their fold change for every sample. Then, the network was generated by associating sample pairings with highly similar topic distributions. Finally, we used the Cytoscape Plugin MCODE to identify strongly connected sub-clusters in our network.

Result

This study demonstrated that topic distributions from topic modeling can be useful to discriminative features for sample clustering. As a consequence of our analysis, a number of strongly connected sub-clusters were identified with being significantly enriched in a particular therapeutic category. One of the intriguing sub-clusters was overrepresented in PPARα agonist, namely WY-14643, clofibrate, gemfibrozil, and fenofibrate in both the *in-vitro* and *in-vivo* rat tests indicating that both assays can catch the similar mode of action in this therapeutic category.

Conclusion

This study suggests that our approach could be a powerful tool for clustering samples and provide very meaningful biological insight for further studies.

Extensive modulation of circadian transcription cycle by microRNAs

Andrey Ptitsyn University of Florida Whitney Lab

In recent years it has been established that microRNAs are important modulators of gene expression. There is also anecdotal evidence that the abundance of some microRNAs varies in circadian or approximately daily rhythm. Studies in which the abundance of both mRNA and microRNA is estimated with regular sampling over sufficiently long time are rare, but have recently became available in public databases. In this study we attempt a systematic analysis of co-expression between microRNAs and their prospective mRNA targets. We have selected a set of data from where mRNA and microRNA have been sampled every 4 hours from mouse liver in a continuous timeline for 48 hours. Microarray analysis has been performed with 3 biological replicates, which allows reconstruction of up to 6 complete circadian periods. The questions we are attempting try to answer are: if a microRNA is expressed in a rhythm, do the target genes of this microRNA also oscillate in the same rhythm, and, if a microRNA abundance peaks at a certain time of the day, how does it affect the expression pattern of its multiple targets?

An advanced analysis of periodicity with the application of digital filters in phase continuum revealed a baseline rhythmic oscillation on over 80% of both mRNA and microRNA populations. From this data we have selected only the top 26% of confidently rhythmic transcripts with highest signal to noise ratio as estimated by a Pt-test. For each transcript and microRNA we have also established the most likely phase of oscillation and circadian amplitude. Next, for each microRNA in the study we identified all predicted targets using both Miranda and TargetScan software. The combined list of microRNAs and corresponding targets has been clustered to eliminate cases of cross-modulation of the same mRNAs by multiple microRNAs expressed in a different pattern over time. For the resulting list of microRNAs we tallied the occurrence of all possible phases of oscillation among target mRNA transcripts. The resulting empirical distribution of phases was tested for bias using chi-square criterion.

In spite of all imperfections in the quantitative estimation of gene expression and incorporated assumptions on multiple stages of analysis, we can confidently (with $p=0.05$ cutoff) claim that 78% of tested microRNAs modulate expression of predicted targets. This computational observation adds evidence to the theory that microRNAs play an active role in modulation of rhythmic expression as a general rule rather than special exemption.

SASD: the Synthetic Alternative Splicing Database for Identifying Novel Isoform from Proteomics

Fan Zhang UNTHSC
Renee Drabier UNTHSC

Alternative splicing is an important widespread mechanism for generating protein diversity and regulating protein expression. In human cells, about 40-60% of the genes are known to exhibit alternative splicing. Recent methodological advances, including EST sequencing, exon array, exon-exon junction array, and next-generation sequencing of all mRNA transcripts, have made it possible to perform high-throughput alternative splicing analysis. However, high-throughput identification and analysis of alternative splicing in the protein level has several advantages. For example, mRNA abundance in a cell often correlates poorly with the amount of protein synthesized, and proteins rather than mRNA transcripts are the major effector molecules in the cell. The combination of alternative splicing database and tandem mass spectrometry provides a powerful technique for identification, analysis and characterization of potential novel alternative splicing protein isoforms from proteomics.

Therefore, we used a three steps pipeline to create an synthetic alternative splicing database(SASD) for tandem mass spectrometry data analysis. First we derived exons and introns from UCSC Genome Database, then we analyzed six types of combinations of exons and introns for the transcription of artificial splicing gene (exon_exon_normal, exon_exon_skipping, intron_exon, exon_intron, single exon, and single intron), and lastly we performed the translation of the artificial transcripts.

In addition, we built a web interface for users to browse 1) by genes/proteins, 2) by biological process, 3) by signaling and metabolic pathway, 4) by disease, 5) by drug, and 6) organ.

Lastly, we presented two case studies: 1)in breast cancer and 2) in liver cancer, to demonstrate that the SASD can enable users to analyze, characterize, and understand the impact of alternative splicing on genes involved in drug, disease, pathway, function, and organ-specificity.

The SASD provides the scientific community with an efficient means to identify and characterize novel Exon Skipping, Intron Retention, and alternative 3' splice site and 5' splice site protein isoforms from mass spectrometry data. We believe that it will be useful in annotating genome structures using rapidly accumulating proteomics data and assist scientific research on signal transduction pathways regulating pre-mRNA, clinical therapy, disease prevention, and drug development.

Combining Probabilistic Graphical Model-based and Knowledge-based Methods for Automatic Construction of Metabolic Pathways

Qi Qi, University of Missouri
Jianlin Cheng, University of Missouri

Automatic reconstruction of metabolic pathways from genome data has been a challenging problem for many years. Traditionally a reference pathway can be mapped into an organism-specific one based on its genome annotation. However, the prediction for unknown interactions is the deficiency of this mapping-based method. Also its network reconstructions tend to contain gaps that caused by gene products missed from genome annotation of the target organism. In contrast, computational methods can predict networks of gene relationships through data reverse engineering. But it lacks reliability as opposed to the mapping-based method.
The motivation of this study was to create such a system for automatic reconstruction of metabolic networks that would mix in the merits of those two types of methods. Specifically, we built a knowledge base including protein interactions and metabolic reactions from reference pathways in KEGG database. The knowledge then would be served as constraints for Bayesian networks learning methods to predict metabolic pathways.

We tested the knowledge-based approach to predict for a set of ground-truth metabolic networks comprised of 62 yeast pathway maps in KEGG database. The comparison between predicted and ground truth pathways was based on their underlying protein-protein relationships. The experimental results showed favorably over the knowledge-based approach as against the mapping-based method.

Contacts-Assisted Protein Structure Prediction

Badri Adhikari, Department of Computer Science, University of Missouri-Columbia
Xin Deng, Department of Computer Science, University of Missouri-Columbia
Jilong Li, Department of Computer Science, University of Missouri-Columbia
Debswapna Bhattacharya, Department of Computer Science, University of Missouri-Columbia
Jianlin Cheng, Department of Computer Science, University of Missouri-Columbia

One recent approach for protein tertiary structure prediction from its residue sequence is to predict which residues are close to each other first, and then build a complete structure solely from this contacts information. These methods use a threshold distance to define this closeness or residue-residue contact. Instead of building a structure purely from these contacts, we summarize a contact-assisted structure prediction approach that uses only a few known contacts to improve the quality of already predicted models. Assuming that we already have some predicted structures and some known contacts, we designed and implemented an automated pipeline that starts with a predicted structure and improves the structure using the inputted contacts as constraints. The system also handles cases when some non-contact information (i.e., knowledge that two residues are not in contact) is provided as input along with or instead of contact information. Our approach for contact assisted structure prediction is a model selection and improvement process comprising of three major steps. First, we select models from a predicted model pool using a scoring scheme. We then refine these selected models using existing protein refinement tools. Finally, we improve the structure of these refined models with given residue-residue contacts information as distance restraints. Our experiment during the 10th Critical Assessment of Techniques for Protein Structure Prediction (CASP10) in 2012 shows that in most cases the quality of predicted structures is improved. The server for contact-assisted protein structure prediction is available at: http://protein.rnet.missouri.edu/contact_assisted/index.html

PAGED v2.0: an update to enable molecular phenotype discoveries through gene-set regulatory networks

Chayaporn Suphavilai, Purdue University - Indianapolis
Xiaogang Wu, Indiana University - Indianapolis
Hui Huang, Indiana University - Indianapolis
Jake Y. Chen, Indiana University - Indianapolis

To discover molecular phenotypes specific for a human disease, one crucial question is which pathways and gene signatures (both can be considered as gene sets) are associated with this disease, but the more important question is how these gene sets are functionally coordinated. Constructing pathway crosstalk networks based on similarities and protein interactions has been introduced in 2008. Recently, a comprehensive approach to construct multi-edge gene-set networks based on co-memberships, protein interactions, and co-enrichment has also been proposed. Nevertheless, the high-level directional relationships between gene sets (gene-set regulation, instead of gene regulation) have never been revealed. To meet this challenge, we developed an update version of the Pathway And Gene Enrichment Database (PAGED v2.0). First, we retrieved experimentally-validated directional information from NCI-Nature (PID) and computationally-predicted directional information from String 9.0, and integrated them into the PAGED. Second, we evaluated disease-gene association data sources and developed a scoring system to prioritize disease-associated genes based on OMIM/GAD and additional gene mutation information retrieved mainly from NGS Catalog and GWAS Catalog. Third, we enlarged the gene set pool by integrating pathway data from ConsensusPathDB. Finally, we enhanced visualization functions, including interactive text cloud for disease-associated genes and interactive network layout for gene-set networks. We used non-small cell lung cancer (NSCLC) as case studies to demonstrate how to build gene-set regulatory networks (directed graphs) for a specific disease. We also evaluate each directional relationship between gene sets, and used two NSCLC-related microarrays from GEO/ArrayExpress to validate these directional relationships as well as NSCLC-associated genes.

Structural Variants Discovery in the Water Flea Daphnia pulex by Genome Re-Sequencing

Yan Peng, School of Computing, University of Southern Mississippi
Natalie Barker, SpecPro Inc.
Jenny Laird, ERDC Environmental Lab
Chris Lounds, ERDC Environmental Lab
Al Kennedy, ERDC Environmental Lab
Chaoyang Zhang, School of Computing, University of Southern Mississippi
Nan Wang, School of Computing, University of Southern Mississippi

Genomic structural variation (SV) is defined as the variation in DNA sequence structure within an organism's chromosome. There are two types of structural variations: balanced rearrangements including inversion and translocation, and copy number variation including insertions, deletions and duplications. They play an important role in phenotypic variation and plasticity, and have been associated with many kinds of human diseases. This study aims to identify genome-wide structural variations in *Daphnia pulex* from genome re-sequencing data. Findings from this study will help establish relationship between genotype (SV) and phenotype (sensitivity to toxic chemicals) in *D. pulex*. To achieve this goal, we chose eight different *D. pulex* populations collected from different geographical regions, which possessed variations in such phenotypes as chemical sensitivity, reproductive output and reproductive strategy. Five clonal individuals from each population were re-sequenced using Illumina MiSeq Genome Analyzer (one individual per paired-end 500-cycle run). A pipeline was designed to analyze the acquired sequencing data to identify SVs. The pipeline consists of four steps: data pre-processing, sequence alignment, SV detection, and consensus SV calling. Through the pipeline, candidate structural variants (i.e., consensus SV call sets) were identified for each population and a selected subset of them were validated by PCR assays. Finally, we annotated the putative functions of confirmed SVs and further related them to phenotypic variations (e.g., chemical sensitivity) of the eight *D. pulex* populations.

A Cross Disciplinary Study of Link Decay and the Effectiveness of Mitigation Techniques

Xijin Ge, South Dakota State University, Department of Mathematics and Statistics

Motivation: Many pieces of published academic research consist of or depend on online resources. Many of these resources, however, disappear within a few years after publication of the related papers, taking with them online databases, tools and references. To address this problem, some solutions have been implemented, including the Internet Archive and WebCite.

Results: Previous research is complemented by performing a cross-discipline characterization of the problem and testing the effectiveness of the existing solutions. We accessed 14,489 web pages found in the abstracts within Thomson Reuters' Web of Science citation index that were published between 1996 and 2010 and found that the median lifespan of these web pages was 9.3 years with 62% of them being archived. Survival analysis and logistic regression were used to find significant predictors of URL lifespan. The availability of a web page is most dependent on the time it is published and the top-level domain names. Similar statistical analysis revealed biases in current solutions: the Internet Archive favors web pages with less layers in the Universal Resource Locator (URL) while WebCite is significantly influenced by the source of publication. We also prototyped a process to submit web pages available but not archived to the archives and increased the coverage of the Internet Archive and WebCite by about 22% and 255% respectively.

Conclusion: Our results show that link decay continues to be a problem across different disciplines and that current solutions for static web pages are helping and can be incrementally improved.

A molecular dynamics approach for the analysis of the binding of vitamin E analogs to the ?-tocopherol transfer protein

Awantika Singh, UAMS/UALR Joint Bioinformatics Program, Department of Pharmaceutical Sciences, University of Arkansas for Medical Sciences
Phillip J Breen, Department of Pharmaceutical Sciences, University of Arkansas for Medical Sciences
Martin Hauer-Jensen, Department of Pharmaceutical Sciences, University of Arkansas for Medical Sciences, Surgical Service, Central Arkansas Veterans Healthcare System, Little Rock, AR
Cesar M Compadre, Department of Pharmaceutical Sciences, University of Arkansas for Medical Sciences

Analyzing and predicting the binding mode of ligands to proteins are some of the most common tasks in computational chemistry and biology. Currently, there are a number of robust computational approaches to perform these tasks. Unfortunately, the current approaches have difficulty to successfully predict the binding of ligands to buried protein pockets. In this research we have used molecular dynamics simulations studies to analyze the binding of vitamin E analogs a-tocopherol transfer protein (ATTP). ATTP is responsible for maintaining the plasma levels of vitamin E, which is a mixture of a-, b-, g- and d-tocopherols and the corresponding tocotrienols. Available high-resolution crystal structures of ATTP complexed with a-tocopherol show that the ligand is bind in a buried cavity inside the protein. Conventional docking methods predict binding modes inconsistent with experimental data. In this study, *in vacuo* molecular dynamics simulations were performed on vitamin E analogs with a dielectric constant of 4 using an NVT ensemble. For the analysis 20 ns production simulation were conducted. A total of 100,000 conformations for each molecule were generated and their structural, dynamics, and energetic properties were analyzed. The results of the analysis correlated well with the experimental results and also allowed us predicted binding of tocoflexol, a vitamin E analog with modified side chain.

Acknowledgement

This research was supported by funds from the National Center for Research Resources, award IULlRR029884, Defense Threat Reduction Agency grants number HDTRAI-07-0028 to Martin Hauer-Jensen, and the IDeA Networks of Biomedical Research Excellence (INBRE)

The impact of CpG islands in promoters on evolution of gene expression in mammals

Yuguang Ban, South Dakota State University
Xijin Ge, South Dakota State University

Comparative analysis is important in modern biology. Evolution of gene function is well-documented in comparative genomics through the study of protein sequence. But the evolution of gene expression is not well-understood. Changes in gene expression may affect significantly the fate of a morphological structure. To gain insight into the genetic mechanisms underlining the evolution of gene expression, we used species-specific whole-gene microarrays to analyze 9 corresponding tissues from human, mouse, and rat. We found that gene expression diverges rapidly but conservation can be observed in 0selected genes. Using linear regression and canonical correlation analysis, we found that gene expression conservation is not linked to promoter sequence conservation; rather, we identified the enrichment of CpG dinucleotides in promoter sequences most associated with conserved gene expression comparing to other genetic constraints. Genes containing CpG islands (CGIs) showed lower divergence in both housekeeping and tissue-specific genes. Evolutionary changes of CGIs affect more divergence of gene expression in non-housekeeping genes than housekeeping genes. It is possible that the conservation of gene expression is a result of low mutation rates of CGIs rather than strong stabilizing selection. Our results suggest the possible interplay of epigenetics and transcriptional mechanisms for preserving the expressions of essential genes.

PREDICTING ALTERNATIVE INDICATIONS OF THE MARKETED DRUGS WITH LATENT DIRICHLET ALLOCATION

Halil Bisgin, Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, US Food and Drug Administration
Zhichao Liu, Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, US Food and Drug Administration
Hong Fang, Office of Scientific Coordination, National Center for Toxicological Research, US Food and Drug Administration
Xiaowei Xu, Department of Information Science, University of Arkansas at Little Rock; Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, US Food and Drug Administration
Weida Tong, Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, US Food and Drug Administration

Drug repositioning, a process of identifying new uses of existing drugs, has gained acceptance as an alternative to *de novo* drug discovery due to its manageable risk and low cost. Phenotypic data in conjunction with probabilistic modeling such as Bayesian Network (BN) could be used for this application via developing a generative model that can infer new relationships between indications and drugs. In this study, 811 drugs with reported side effects and indications from the Side Effect Resource (SIDER) were employed. The combined drug and phenotypic information was used to build a Latent Dirichlet Allocation (LDA) model, which is a BN. The model contains total of 5748 known drug-indication pairs. A leave-one-out method was developed, where a known pair was assumed to be unknown and its probability as a pair was subsequently predicted by the model. This process was repeated for all the known pairs, which resulted in 3172 predicted pairs that are statistically significant by comparing with random chance. If we consider only these pairs with the highest probability (justified by the number of indications the drugs have) as the reliable repositioning opportunities, 2155 known pairs are successfully fished, which gives a success rate of 68%. Using the same approach, we also identified 2365 unknown drug-indication pairs, indicating that some alternative indications of a drug are not recorded in SIDER. We confirmed 9 pairs through various sources. The results demonstrated that the proposed method using LDA could be a promising approach to explore new uses of existing drugs.

Effect of treatment effect and sample size on reproducibility in a toxicogenomics study

Joshua Xu
Weida Tong

Toxicogenomics has become a powerful tool for compound classification, mechanistic elucidation, toxicity biomarker discovery and prediction. At the core of many applications of toxicogenomics lie two issues that may have a large influence on the outcome. Both are related to the gene signatures with the first being their reproducibility for repeated experiments and the second being the similarity (overlapping ratio) of gene signatures between treatments grouped together by a chosen common factor such as chemical structure similarity, a shared mechanism, or a toxicity endpoint. Here we take an existing microarray dataset to examine these two issues regarding gene signatures. 120 rats in total were treated 12 biological replicates in each treatment and control group, by tolcapone and entacapone in 3 days with high dose and 28 days with high, medium and low dose. The preliminary results show that the above mentioned reproducibility and similarity ratio are linearly correlated with the treatment effect. The stronger the treatment effect is, the higher reproducibility and similarity ratio are reached and the smaller sample size is needed. Additionally, more variance has been observed for the low-expressed genes and thus increases the false positive rate in DEG identification among those genes. Our findings could be useful for toxicogenomics study design and guiding data analysis plans.

Quantitative Structure-Activity Relationship Models for Predicting Drug-Induced Liver Injury Based on FDA-Approved Drug Labeling Annotation

Minjun Chen
Huixiao Hong
Hong Fang
Reagan Kelly
Guanxuan Zhou
Jürgen Borlak
Weida Tong

Drug-induced liver injury (DILI) is a leading cause of aborted drug development program. Consequently, identifying risk of DILI for drug candidates during the early stages of the development process would greatly reduce drug attrition rate in the pharmaceutical industry, but requires the implementation of new research and development strategies. In this regard, several *in silico* models had been proposed as alternative means in prioritizing drug candidates. Since the accuracy and utility of a predictive model rests largely on how to annotate the potential of a drug to cause DILI in a reliable and consistent way, we annotated 197 drugs using the FDA-approved drug labeling and developed a quantitative structure-activity relationship (QSAR) model built using our in-house Decision Forest based on molecular descriptors obtained from Mold^2. The performance of the model was assessed with cross-validations and permutation analysis, and was challenged by three independent validation datasets for drugs and yielded prediction accuracies of 61.6%, 63.1%, and 68.4%. Of note, the QSAR model's performance varied for drugs with different therapeutic uses; it achieved a better estimated accuracy (73.6%) as well as negative predictive value (77.0%) when focusing only on high confidence therapeutic subgroups without significantly sacrificing the positive predictive value (69.1%), as compared to those reported for other published QSAR models. We concluded that the developed QSAR model has the potential to hallmark compounds with risk for DILI in humans, particularly for those falling into high confidence therapeutic subgroups like analgesics, antibacterial agents, and antihistamines.

The Properties of Human Genome Conformation and Spatial Gene Interaction and Regulation Networks

Zheng Wang, University of Missouri
Renzhi Cao, University of Missouri
Kristen Taylor, University of Missouri
Aaron Briley, University of Missouri
Charles Caldwell, University of Missouri
Jianlin Cheng, University of Missouri

The spatial conformation of a genome plays an important role in the long-range regulation of genome-wide gene expression and methylation, but has not been extensively studied due to lack of genome conformation data. The recently developed chromosome conformation capturing techniques such as the Hi-C method empowered by next generation sequencing can generate unbiased, large-scale, high-resolution chromosomal interaction (contact) data, providing an unprecedented opportunity to investigate the spatial structure of a genome and its applications in gene regulation, genomics, epigenetics, and cell biology. In this work, we conducted a comprehensive, large-scale computational analysis of this new stream of genome conformation data generated for three different human leukemia cells or cell lines by the Hi-C technique. We developed and applied a set of bioinformatics methods to reliably generate spatial chromosomal contacts from high-throughput sequencing data and to effectively use them to study the properties of the genome structures in one-dimension (1D) and two-dimension (2D). Our analysis demonstrates that Hi-C data can be effectively applied to study tissue-specific genome conformation, chromosome-chromosome interaction, chromosomal translocations, and spatial gene-gene interaction and regulation in a three-dimensional genome of primary tumor cells. Particularly, for the first time, we constructed genome-scale spatial gene-gene interaction network, transcription factor binding site (TFBS) – TFBS interaction network, and TFBS-gene interaction network from chromosomal contact information. Remarkably, all these networks possess the properties of scale-free modular networks.

Topic Modeling on LiverTox: Discovering the Hidden Knowledge for Predicting Potential to Cause Acute Liver Failure of a Drug

Ke Yu (Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, US Food and Drug Administration, 3900 NCTR Road, Jefferson, AR 72079, USA)
Jie Zhang (Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, US Food and Drug Administration, 3900 NCTR Road, Jefferson, AR 72079, USA)
Minjun Chen (Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, US Food and Drug Administration, 3900 NCTR Road, Jefferson, AR 72079, USA)
Yijun Ding (Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, US Food and Drug Administration, 3900 NCTR Road, Jefferson, AR 72079, USA)
Jürgen Borlak (Center of Pharmacology and Toxicology, Hannover Medical School, Hannover, 30625, Germany)
Hong Fang (Office of Scientific Coordination under NCTR, US Food and Drug Administration, 3900 NCTR Road, Jefferson, AR 72079, USA)
Xiaowei Xu (Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, US Food and Drug Administration, 3900 NCTR Road, Jefferson, AR 72079, USA; Department of Information Science, University of Arkansas at Little Rock, 2801 S. University Ave., Little Rock, AR 72204, USA)
Weida Tong (Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, US Food and Drug Administration, 3900 NCTR Road, Jefferson, AR 72079, USA)

Topic modeling can be used to discover the latent topics, which provide an abstract view to the underlying corpus. The NIH LiverTox database provides a comprehensive and up-to-date summary of clinical information on drug-induce liver injury (DILI), including acute liver failure (ALF) with textual drug records. ALF is of great concern in clinics, and more than half of ALF cases in the United States are caused by drugs. In this study, we applied topic modeling on the drug records in the LiverTox database to discover the latent topics relevant to ALF. Each drug record contains several sections, of which the hepatotoxicity section was used for the topic modeling. A fishing dataset comprised of 30 well-known ALF drugs was applied to identify the diagnostic topic to indicate ALF drugs. This diagnostic topic identified 94 additional ALF drugs, 31 of which could not be identified using text search. More interestingly, the relevant words to interpret the diagnostic topic were highly related to the most predictive rule of ALF in clinics, i.e., Hy's law, stating that hepatocellular injury accompanied with jaundice will result in 10% to 50% fatality caused by ALF. The significant relationship between hepatic jaundice observed in the clinical trials and ALF cases observed in post-marketing for a drug was confirmed in a follow-up survey using the FDA's approved drug labels. In conclusion, topic modeling can be used not only for identifying drugs with ALF potential from the LiverTox database, but also for discovering the hidden knowledge for predicting serious hepatotoxicity.

Structure-based engineering to produce high affinity anti-methamphetamine antibodies

Shraddha Thakkar, Bioinformatics Graduate Program,
University of Arkansas for Little Rock, Little Rock AR 72205
Nisha Nanaware-Kharade, Dept. of Pharmacology and Toxicology,
University of Arkansas for Medical Sciences, Little Rock, AR 72205
Guillermo Gonzalez III, Dept. of Pharmacology and Toxicology,
University of Arkansas for Medical Sciences, Little Rock, AR 72205
Reha Celikel, Dept. of Physiology and Biophysics,
University of Arkansas for Medical Sciences, Little Rock, AR 72205
Eric C. Peterson, Dept. of Pharmacology and Toxicology,
University of Arkansas for Medical Sciences, Little Rock, AR 72205
Kottayil I. Varughese, Dept. of Physiology and Biophysics,
University of Arkansas for Medical Sciences

Methamphetamine (METH) abuse is a threat in the US and worldwide, with no FDA approved treatments available. Anti-METH IgGs and smaller single chain fragments (scFvs) have shown efficacy in preclinical studies. Here we describe our efforts to enhance their efficacy through designing higher affinity antibodies. The crystal structure of the scFv6H4, in complex with METH elucidated important binding interactions. Using the structural data, we made single amino acid mutations in the scFv-6H4 binding pocket in order to increase the affinity for METH. The new mutants; scFv-S93T, -I37M, and -Y34M were cloned, expressed in yeast and tested for affinity by saturation binding techniques. Two of the mutant scFv showed enhanced binding affinity for METH; scFv-I37M by 25% and scFv-S93T by 150%. We have now determined the crystal structure of the highest-affinity mutant scFvS93T in complex with METH, revealing a more hydrophobic binding pocket in comparison with the wild type. ScFv6H4 binding pocket has METH along with two water molecules buried deep in the pocket, interacting with METH hydrocarbon. The substitution of a Ser residue by a Thr in scFvS93T caused expulsion of a water molecule from the cavity, relieving some unfavorable molecular interactions of water with METH. Thus, this mutation has increased the hydrophobicity of the pocket, and dramatically enhanced affinity for METH.

Homology Modeling, Molecular Docking, Molecular Dynamics Simulations and Free Energy Calculations Elucidated Rat α-Fetoprotein-Ligand Binding Modes

Jie Shen, Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. Food and Drug Administration
Wenqian Zhang, Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. Food and Drug Administration
Hong Fang, Office of Scientific Coordination, National Center for Toxicological Research, U.S. Food and Drug Administration
Roger Perkins, Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. Food and Drug Administration
Weida Tong, Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. Food and Drug Administration
Huixiao Hong, Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. Food and Drug Administration

Endocrine activity is a complex process and can occur in multiple mechanisms. One important mechanism is that chemicals interact with hormone receptors such as estrogen receptor (ER) in target cells. To interact with receptors, chemicals have to enter cells via transport proteins. α-Fetoprotein (AFP) is a major transport protein in rodent serum that can bind with estrogens and thus change a chemical's availability for entrance into the target cell. Sequestration of an estrogen in the serum can alter the chemical's potential for disrupting estrogen receptor-mediated responses. Recently, we have reported the rat AFP binding affinities of a large set of structurally diverse chemicals, including 53 binders and 72 non-binders. These data are useful in understanding binding interactions and mechanisms between chemicals and rat AFP. However, lack of bound AFP crystal structure data hinders further understanding of AFP binding interactions. Therefore, AFP homology modeling, docking simulation and molecular dynamic simulations were conducted to elucidate rat AFP structure and rat AFP-ligand binding modes which can be used for estimation of rat AFP binding affinity of chemicals. More specifically, the AFP tertiary structure was first estimated using homology analysis. Then, 12 structurally representative rat AFP binders were selected from our recently reported 53 binders and were docked into the binding pockets of rat AFP using molecular docking. Lastly, molecular dynamics simulations and free energy calculations were performed to compare ligand binding modes and estimate relative binding affinity. We have constructed the first rat AFP tertiary structure and elucidated the rat AFP-ligand binding modes that have been demonstrated reliable for estimating rat AFP binding affinity of chemicals.

Inference of Genetic Regulatory Networks from Perturbation Data – Application to Melanoma cell lines

Mohammed Mohammed-Rasheed
Nidhal Bouaynaya
Hassan Fathallah-Shaykh

In this study, we investigate the underlying genetic interaction structure and dynamics of five melanoma cell lines (A375, A2058, G361, UACC647 and UACC903) from gene perturbation experiments. We derive the interactions attribute (activation, inhibition or no interaction) among the four genes (HADHB, PIRIN, Wnt5a and MART1) in the five cell lines. We model the gene expression dynamics using a linear differential equation. We estimate the genetic regulatory network interactions using a 95% t-test followed by a constrained least-squares optimization. We successfully address the differences in cells behavior and response to external stimuli by studying the dynamics of the interactions and the stability of the cell. Further mathematical analysis is conducted to find the major interactions and proteins causing the instability of the system. Finally, the analytical results in this study are under validation process through biological experiments to analyze the main molecules causing the change in the attractor cycle of the cell.

Bioinformatics Analysis Pipeline for Characterization of Intra-strain Variability in Food-borne Bacteria

Mihir Jaiswal: University of Arkansas at Little Rock and University of Arkansas for Medical Sciences Bioinformatics Graduate Program, Little Rock, AR, USA
Wen Zou: Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, US Food and Drug Administration, 3900 NCTR Road, Jefferson, AR 72079, USA
Harry Hull: University of Arkansas at Little Rock Undergraduate Program in Computer Science, Little Rock, AR 72204, USA
Udensi K. Udensi: University of Arkansas at Little Rock Postbaccalaureate Program, Little Rock, AR 72204, USA
Justin Woo: University of Arkansas at Little Rock Undergraduate Program in Physics, Little Rock, AR 72204, USA
Rinku Saha: University of Arkansas at Little Rock and University of Arkansas for Medical Sciences Bioinformatics Graduate Program, Little Rock, AR, USA
Steven F. Jennings (corresponding author: sfjennings@ualr.edu): Department of Information Science, University of Arkansas at Little Rock, 2801 S. University Ave., Little Rock, AR, 72204, USA

Salmonella enterica subsp. *enterica* serovar Montevideo is one of the top ten serovars most commonly associated with food contamination. However, its intra-strain genetic variability remains misunderstood. A bioinformatics analysis pipeline was constructed to characterize the variations across Montevideo strains. Antigenic genes such as *rfb* cluster encoding O-antigen, *fliC* and *fljB* genes encoding H antigen and *viaB* encoding V antigen were analyzed. Genome contigs for 47 Montevideo strains were retrieved from NCBI, followed by a BLAST against reference genes. Genomes of *Salmonella enterica* subsp. *enterica* serotypes Pomona and Poona and have been used as reference genomes for *rfb* cluster while *Typhi* was used as reference genome for the others. The best-aligned sequences for each of the 47 contig genomes were then extracted for the *rfb* region, *fliC* and *fljB* genes. Multiple sequence alignment was performed using ClustalW and a HMMER profile was generated. Each sequence was then matched with the HMMER profile to view the differences among the homologous sequences of the 47 strains for the antigenic genes. One sequence appeared to be different for each *rfb* region gene cluster, *fljB* and *fliC* genes. This pipeline proved useful in the identification of differences encoded in the genome contigs of 47 Montevideo strains used in the study and will allow additional strains to be analyzed using the same pipeline in the future.

Statistical analysis of molecular pathways

Joshua S. Yuan ;Synthetic and Systems Biology Innovation Hub, Texas A&M University
Hu Chen ; Department of Plant pathology and microbiology , Texas A&M University

Quantitative pathway analysis is essential for understanding the molecular mechanisms for biological processes and identifying key biomarkers for clinical applications. However, quantitative analysis of molecular pathways is complicated by the multidimensional feature of the problem, where defined statistical modeling and interpretation are challenging. Unlike the single gene analysis, the pathway analysis outcomes are often biased by various modeling assumptions. In order to address these challenges, we hereby proposed, evaluated, and implemented an alternative approach to use multivariate analysis for molecular pathway quantification, comparison, and analysis. The approach clearly defines the variables, p value, and parameter estimations that can be interpreted to classify differentially regulated pathways and identify key genes involved in the classification. The approach is implemented by Python and R into a software package named as Statistical Analysis of Molecular Pathways (SAMP) for the three-step analysis. First, gene expression data from RNA-seq was processed and mapped to specific molecular pathways. Our statistical model here specified the features of genes, i.e. gene expression, as independent variable, and the pathway outcome or classification as dependent variables. Second, MANOVA test was carried out to screen the molecular pathways for differential regulation. Third, principal component analysis (PCA) was carried out as the exemplary multivariate analysis to weigh each genes in variation contributing pathways. Using the new approach, we analyzed several breast cancer test datasets. The novel approach overcomes the limitations of assumptions in previous approaches and provided an effective approach for quantitative pathway analysis with both scientific and clinical application potentials.

Consensus Spectral Techniques via Vertex Weighted Graphs in the Analysis of Genomic and Proteomic Data

Debra Knisley, East Tenn State Univ

There is a long and fruitful history in the use of spectral methods in proteomics and genomics, especially when such techniques have been used to produce consensus spectra and consensus models for homological families of residue chains. Typically, these methods have focused on either periodicity – such as the fact that coding regions in DNA have 3-base periodicity – or on the use of biophysical measures such as the Electron-Ion Interaction Potential (EIIP) and its relationship to "hot spots" in amino acid chains. Given the rich history and sophisticated signal processing theory underpinning such methods, there may be value in extending consensus methods into more general machine learning approaches. We present one possible generalization of these methods, one that can use structural measures such as graphical invariants of network models of residue chains. Moreover, a major component of this approach is a resampling technique that allows not only the construction of a consensus model, but also a means of estimating confidence intervals from the resulting empirical distribution. We apply this generalization both to DNA microarray data as a means of finding differentially expressed genes and to a homological family of proteins.

Model quality assessment of proteins using group-based redundancy

Juexin Wang, Jilin University and University of Missouri
Yanchun Liang, Jilin University

The model quality assessment of protein (MQAP) is an essential challenging issue in protein structure prediction. Among all MQAP methods, consensus method incorporating all the model information together could obtain much more accurate results. In reality, choosing only parts of models in consensus method by removing the redundancy of these candidate models could be useful in selecting best models. In this paper, we propose a novel method using group-based criteria to define the redundancy and use META methods to incorporate point-wise and group-wise similarity together. In CASP10, this method rank first in the random choosing 20 models quality category. Experiments based on CASP7/8/9 dataset show that our method could obtain good results in Pearson's Correlation and the performance in choosing best model.

Inference of Temporally rewiring genetic networks using time-varying differential equations

Annick Dongmo UALR

Collections of complex time-dependent genetic data from dynamic biological processes such as cancer progression, immune response, and developmental processes, are expanding given the improved data collection of new technologies. Clearly, the “static” view provided by the current time-invariant network topologies is obsolete for these

dynamical systems and we must develop methods and algorithms that reflect the dynamic or rewiring nature of genetic interactions. We model time-varying gene networks using time-varying coefficients differential equations with an additive noise term representing the errors in measurements and the model. Model networks resulting from differential equations are directed, allow for feedback loops, and categorize stimulations and inhibitions. The ability to recognize whether a gene interaction is stimulative or inhibitory is critical in understanding transcriptional regulatory interactions and designing appropriate drug targets. The inference problem is to estimate the time-varying network interactions. Given the limited number of observations at each time point, this problem is severely under-determined. We first transform the time-varying problem into a time-invariant one by decomposing the time-varying parameters in a suitable basis function. We then estimate the basis coefficients using a robust Singular Value Decomposition (SVD) approach that constraints the topology of the network to be sparse and hence reduces the number of required measurements for estimation. Our simulation results show the performance of the proposed approach for different basis functions.

Comparative Analyses Demonstrated Reliability of 1000 Genome Project and Confirmed Usefulness of GWAS Findings Based on SNP-Arrays

Wenqian Zhang, Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. Food and Drug Administration, 3900 NCTR Road, Jefferson, AR 72079
Jie Shen, Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. Food and Drug Administration, 3900 NCTR Road, Jefferson, AR 72079
Hong Fang, Office of Scientific Coordination, National Center for Toxicological Research, U.S. Food and Drug Administration, 3900 NCTR Road, Jefferson, AR 72079
Roger Perkins, Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. Food and Drug Administration, 3900 NCTR Road, Jefferson, AR 72079
Weida Tong, Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. Food and Drug Administration, 3900 NCTR Road, Jefferson, AR 72079
Huixiao Hong, Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. Food and Drug Administration, 3900 NCTR Road, Jefferson, AR 72079

High-density genotyping single-nucleotide polymorphism (SNP) arrays have enabled genome-wide association studies (GWAS) that successfully identified common genetic variants associated with a variety of human hereditary traits. However, next generation sequencing (NGS) has become the preferred technology in current genetic studies because it can identify not only common genetic variants but also novel and rare genetic variants. The 1000 Genome Project Consortium (1000GPC) recently identified 38 million SNPs, providing a rich resource of reference genetic variants for GWAS. The 1000GPC data's availability enabled a comparative evaluation of reported GWAS findings based on SNP-arrays for application in personalized medicine. Our comparative analyses used data from 1000GPC, HapMap, and our recently published SNP-array data. Comparing the genotypes of the same SNPs for the same subjects between 1000GPC data and HapMap data yielded high concordance, demonstrating the high accuracy of the genotypes determined by 1000GPC, and providing confidence that the reference genetic variants provided by 1000GPC are reliable for GWAS. Comparison of 1000GPC data and our previously published SNP-array data also showed a high concordance, demonstrating that genotypes of common SNPs determined using SNP-arrays are trustable, that GWAS findings based on SNP-arrays reported in the literature are useful. The result lends confidence that GWAS findings based on SNP-array data can be reliably applied in translational studies for personalized medicine.

A New Hidden Markov Model for the Compatibility between Protein Sequence and Structure

Zhiquan He, Department of Computer Science, University of Missouri, Missouri
Wenji Ma, Department of Computer Science, City University of Hong Kong, Hong Kong
Jingfen Zhang, Department of Computer Science, University of Missouri, Missouri
Donbo Bu, Institute of Computing Technology, Chinese Academy of Science, China
Dong Xu, Department of Computer Science, University of Missouri, Missouri

The relationship between protein sequence and structure plays an important role in protein function analysis and protein structure prediction. In this work, we developed a novel Hidden Markov Model (HMM) to model the compatibility of protein sequence and structure for capturing their complex relationship. More specifically, the emission of the HMM consists of protein local structures in angular space, secondary structures, solvent accessibilities and sequence profile information. The optimal HMM model was determined by the Bayes Information Criterion (BIC). This model has two capabilities: (1) encoding local structure for each position by consideration of its sequence and structure information together, and (2) assigning a global score to estimate the overall quality of a predicted structure, as well as local scores to assess the quality of specific regions of a structure separately to provide useful guidance for further refinement. Computational results show that structures encoded by our model have less structural information loss, than other encoding schemes. For structure quality assessment, our model can achieve better ranking performance than most current single structure quality assessment methods. Most importantly, our method provides useful local structure assessment for predicted protein structures.

A study and extension of DNcon: a method for protein residue-residue contact prediction using deep networks

Jesse Eickholt, University of Missouri
Jianlin Cheng, University of Missouri

Predicted residue-residue contact information has applications in protein tertiary structure modeling, model quality assessment and drug design. In spite of the usefulness of this data and a prolonged effort on behalf of the community, progress in the area of predicted residue-residue contacts has been slow, particularly for hard targets (i.e., those proteins for which template/structural data is scarce). In an effort to increase the performance of residue-residue contact prediction we developed DNcon, a sequence based contact predictor built from boosted ensembles of deep networks. The approach represents one of the first applications of deep learning to Bioinformatics and was successfully benchmarked in the 10th Critical Assessment of protein Structure Prediction (CASP). Here, we present an analysis of DNcon's performance in CASP as well as outline further studies and extensions that we have made to the method. In particular, we report on the robustness of the deep networks used in DNcon and contact propagation (i.e., using shorter range contact predictions to fulfill longer range contact prediction).

The original method, DNcon, is available at http://iris.rnet.missouri.edu/dncon/ .

Determining beneficial and detrimental effects of non-synonymous SNPs on protein-protein interactions using supervised and semi-supervised learning

Nan Zhao, Informatics Institute and Department of Computer Science, University of Missouri
Chi-Ren Shyu, Informatics Institute and Department of Computer Science, University of Missouri
Dmitry Korkin, Informatics Institute and Department of Computer Science, University of Missouri

Functional changes by disease-causing nonsynonymous single nucleotide polymorphisms (nsSNPs) are often associated with changes in interactions involving protein that carry those nsSNPs. In spite of large amount of structural data available for protein-protein interactions (PPIs), only a few computational methods exist to analyze how a single amino acid mutation caused by a nsSNP affects a PPI. An accurate computational method that determines whether a nsSNP alters binding affinity is yet to be developed.

Here, we address this problem by developing nsSNP interaction effect classifiers that leverage supervised and semi-supervised machine learning methods. Both methods were trained based on a dataset of 2,079 single point mutations from 151 PPI complexes with experimentally determined binding affinity. The change of binding energies between mutant protein and its wild type was used to label mutations into three types: beneficial, detrimental, and neutral. Each complex involving a mutant protein was modeled by FoldX, and a set of 33 features describing energy changes were calculated. Finally, three classifiers were trained in total: (1) supervised classifier of the three classes (2) supervised classifier of beneficial and detrimental classes only, (3) semi-supervised classifier of beneficial and detrimental classes where the neutral mutations were used as the unlabeled data. The cross-validation for the most-difficult 3-class classification revealed a weighted average F-measure of 68.2%, while the 2-class classification could be done with F-measure of 87%. The near-perfect ability of the classifiers to determine nsSNP with the detrimental effects may provide useful insights in identifying novel disease-associated non-synonymous SNPs.

Choosing the right coverage depth and read length for an RNA-seq experiment

Madelaine Gogol, Stowers Institute for Medical Research
Malcolm Cook, Stowers Institute for Medical Research
Ron Yu, , Stowers Institute for Medical Research
Hua Li, , Stowers Institute for Medical Research

RNA-Seq is an approach that uses deep-sequencing technologies to profile a transcriptome. Studies show that it can provide precise measurements of expression level of transcripts and their isoforms. When planning a RNA-seq experiment, choosing the right coverage depth and read length is important to ensure complete coverage of the genome and accurate measures of gene expression. We use a RNA-seq dataset studying the genes involved in mouse olfactory development to investigate how the coverage depth and read length affect the gene expression values (mean, variance) and the detection of differentially expression genes. Preliminary results show that doubling the coverage (from 10 million reads (one lane) to 20 million reads (two lanes)) won't necessarily detect more expressed genes, but will produce more accurate measurements. In terms of differential expression analysis, at least 70% of the genes identified by the analysis of a single lane of data are also identified by analysis that uses four lanes of data. Only a subset of genes appear to be affected by read length (42/ 60 bases). We investigate the effects of these choices in detail and provide guidelines for planning future experiments.

Towards the integration, annotation and association of historical microarray experiments with RNA-seq

Shweta S. Chavan, Myeloma Institute for Research and Therapy,
University of Arkansas for Medical Sciences
Michael A. Bauer, Myeloma Institute for Research and Therapy,
University of Arkansas for Medical Sciences
Erich A. Peterson, Myeloma Institute for Research and Therapy,
University of Arkansas for Medical Sciences
Donald J. Johann, Jr., Myeloma Institute for Research and Therapy,
University of Arkansas for Medical Sciences

Introduction

Transcriptome analysis by microarrays has produced important advances in biomedicine. For instance in multiple myeloma (MM), microarray approaches led to the development of an effective disease subtyping via cluster assignment, and a 70 gene risk score. Both enabled an improved molecular understanding of MM, and have provided prognostic information for the purposes of clinical management. Many researchers are now transitioning to Next Generation Sequencing (NGS) approaches and RNA-seq in particular, due to its discovery-based nature, improved sensitivity, and dynamic range. Additionally, RNA-seq allows for the analysis of gene isoforms, splice variants, novel gene fusions, etc. There is now a need to associate and integrate microarray and NGS data via advanced bioinformatic approaches.

Methods

Custom software was developed following a model-view-controller (MVC) approach to integrate Affymetix probe set-IDs, and gene annotation information from a variety of sources. The tool/approach employees a variety of strategies to integrate, annotate, and associate microarray and RNA-seq datasets.

Results

Output from the Cufflinks tool (from the Tuxedo suite) can be directly integrated, and/or associated with Affymetrix probe set data, as well as necessary gene identifiers and/or symbols from a variety of sources. Strategies are employed to maximize the integration/annotation process. Novel gene sets (eg, MM 70 risk score) can be specified, and the tool can be directly interfaced to the RNA-seq pipeline.

Conclusion

A novel bioinformatic approach to aid in the facilitation of both annotation and association of historic microarray data in conjunction with richer RNA-seq data is now assisting with the study of MM cancer biology.

Drug Activity Prediction Using Multi-Instance Learning via Joint Instance and Feature Selection

Zhendong Zhao, Department of Computer and Information Science, School of Engineering, University of Mississippi, University, 38677, USA
Gang Fu, Department of Medicinal Chemistry, School of Pharmacy, University of Mississippi, University, 38677, USA
Sheng Liu, Department of Computer and Information Science, School of Engineering, University of Mississippi, University, 38677, USA
Khaled M. Elokely, Department of Medicinal Chemistry, School of Pharmacy, University of Mississippi, University, 38677, USA
Robert J. Doerksen, Department of Medicinal Chemistry, School of Pharmacy, University of Mississippi, University, 38677, USA, Research Institute of Pharmaceutical Sciences, School of Pharmacy, University of Mississippi, University, 38677, USA
Yixin Chen, Department of Computer and Information Science, School of Engineering, University of Mississippi, University, 38677, USA
Dawn E. Wilkins, Department of Computer and Information Science, School of Engineering, University of Mississippi, University, 38677, USA

In drug discovery and development, it is crucial to determine which conformers (instances) of a given molecule are responsible for its observed biological activity and at the same time to recognize the most representative subset of features (molecular descriptors). Due to experimental difficulty in obtaining the bioactive conformers, computational approaches such as machine learning techniques are much needed. Multi-Instance Learning (MIL) is a machine learning method capable of tackling this type of problem. In the MIL framework, each instance is represented as a feature vector, which usually resides in a high-dimensional feature space. The high dimensionality may provide significant information for learning tasks, but at the same time it may also include a large number of irrelevant or redundant features that might negatively affect learning performance. Reducing the dimensionality of data will hence facilitate the classification task and improve the interpretability of the model. In this work we propose a novel approach, named multi-instance learning via joint instance and feature selection. The iterative joint instance and feature selection is achieved using an instance-based feature mapping and 1-norm regularized optimization. The proposed approach was tested on four biological activity datasets. The empirical results demonstrate the selected instances (prototype conformers) and features (pharmacophore fingerprints) have competitive discriminative power and the convergence of the selection process is also fast.

BinAligner: a heuristic method to align biological networks

Jialiang Yang
Jun Li
Stefan Grunewald
Xiu-Feng Wan

Alignments of molecular networks across species will help detect orthologs and conserved functional modules and provide insights on the evolutionary relationships of the compared species. However, such analyses are not trivial. We develop a global network algorithm “BinAligner”. Based on the hypotheses that the information from the edges and the structures of subnetworks can be more informative than vertices alone, two scoring schema, 1-neighborhood subnetwork and graphlet, were introduced to derive the scoring matrices between networks, besides the commonly used scoring scheme from vertices. Then the alignment problem is formulated as an assignment problem, which is solved by the combinatorial optimization algorithm, such as Hungarian method. The proposed algorithm was applied and validated in aligning the protein protein interaction network of Kaposi’s sarcoma associated herpesvirus (KSHV) and that of varicella zoster virus (VZV). Interestingly, we identified several putative functional orthologous proteins with similar functions but very low sequence similarity between the two viruses. For example, KSHV open reading frame 56 (Orf56) and VZV Orf55 are helicase-primase subunits with sequence similarity 14.6%, and KSHV Orf75 and VZV Orf44 are tegument proteins with sequence similarity 15.3%. These functional pairs can not be identified if one restricts the alignment into orthologous protein pairs. In addition, BinAligner managed to identify a conserved pathway between two viruses, which consists of 7 orthologous protein pairs and these proteins are connected by conserved links. This pathway might be crucial for virus packing and infection. BinAligner is available at http://sysbio.cvm.msstate.edu/BinAligner.php.

Discrimination of fungal secretory (effector) proteins from plant host secretory proteins

Ruchi Verma, Dept of Biochemistry and Molecular biology,
Oklahoma State University, Stillwater, OK
Ulrich Melcher, Dept of Biochemistry and Molecular biology,
Oklahoma State University, Stillwater, OK

Distinguishing plant host proteins from fungal pathogen proteins computationally long has been a challenge. Both are eukaryotic in origin and the proteins share many properties. Effector proteins of fungi are well-classified proteins that are secreted by pathogens and help them to survive in their plant hosts. It has been shown that identifying and using conserved or signature sequences often allows efficient distinctions. Signals present in the putative effector proteins include RXLX[EDQ] and RXLR, also called as Host targeting or VTS signals. These help *Plasmodium* and *Phytophthora* species to survive in their respective hosts. We took the same approach to identify conserved sequences in true fungal plant pathogens using proteins from the fungal secretome database. We took 20 fungal species; averaging 2319 proteins in each species and 10 plant host species; averaging 9300 proteins in each species to determine conserved sequences using online motif searching tools. Their ability to identify the fungal proteins in a background of plant host proteins with high accuracy was also tested.

Several quartet-based methods to reconstruct phylogenetic networks

Jialiang Yang
Yifei Xu
Xiu-Feng Wan

As a generalization of phylogenetic trees, phylogenetic networks explicitly model reticulate evolutionary events such as hybridization, recombination, and horizontal gene transfer. However, reconstructing such networks is not trivial. Popular character-based methods are computationally inefficient, while distance-based methods cannot guarantee reconstruction accuracy since pairwise genetic distances only reflect partial information about a reticulate phylogeny. Recently, we published a quartet-based method Quartet-Net (Molecular Biology and Evolution, in press) to solve such dilemma. Quartets contain more information than distances since they reflect the relationships among 4 taxa, and thus quartet-based methods could be more accurate in reconstructing phylogenetic histories than distance-based methods. However, it is shown that Quartet-Net sometimes tends to reconstruct too few splits because it is a parsimony method. To solve such a problem, here, we introduce three new quartet-based methods, namely QuartetS, QuartetA and QuartetM. By applying the methods into simulated data sets, we demonstrate that these methods have the potential to be more accurate in reconstructing some phylogenetic networks than popular phylogenetic reconstruction methods like Neighbor-Joining, Split-Decomposition, Neighbor-Net, QNet and Quartet-Net. In addition, we theoretically prove that QuartetS has the same nice property as Quartet-Net, that is, it is consistent on 2-weakly compatible split systems, broader than the split systems that other methods can accurately reconstruct.

Towards Human-Computer Synergistic Analysis of Large-Scale and Complex Biological Data

Rahul Singh, Computer Science, San Francisco State University and Center for Discovery and Innovation in Parasitic Diseases, University of California, San Francisco,
Hui Yang, Computer Science, San Francisco State University,
Ben Dalziel, Computer Science, San Francisco State University,
David Foote, Open University Program, San Francisco State University,
Daniel Asarnow, Computer Science, San Francisco State University,
William Murad, Computer Science, San Francisco State University,
Matthew Gromley, Obstetrics, Gynecology, and Reproductive Sciences, U. of California, San Francisco,
Jonathan Stillman, Department of Biology, San Francisco State University
Susan Fisher, Department of Anatomy and Pharmaceutical Chemistry, U. of California, San Francisco

Background : Advances in technology have led to the generation of massive amounts of complex and multifarious biological data in areas from genomics to structural biology. The volume and complexity of such data leads to significant challenges in terms of its analysis, especially when one seeks to generate hypotheses or explore the underlying biological processes. At the state-of-the-art, the predominant paradigm for analysing biological data involves the application of automated algorithms followed by perusal and analysis of the results by an expert. This paradigm works well in many problem domains. However, it also is limiting, since domain experts are forced to apply their instincts and expertise such as contextual reasoning, hypothesis formulation, and exploratory analysis *after* the algorithm has produced its results. In many areas where exploratory analysis is crucial, what is needed is to integrate domain expertise *during* the data analysis process and use it to drive the analysis itself.

Results: We propose a design approach that combines information visualization and human-computer interaction with algorithms for exploratory analysis of large-scale and complex biological data. In the proposed approach, emphasis is laid on: (1) allowing users to directly visualize, interact, experience, and explore the data through interoperable computing components, (2) supporting unified query and presentation spaces to facilitate experimentation and exploration, (3) providing external contextual information by assimilating relevant supplementary data, and (4) encouraging experiential data exploration, hypotheses formulation, and information visualization. To demonstrate the efficacy of the proposed design paradigm, we discuss and evaluate two prototype systems. The first, called XMAS (Experiential Microarray Analysis System), is for analysis of time-series transcriptional data. The second system, called PSPACE (Protein Space Explorer) is designed for holistic analysis of structural and structure-function relationships using interactive displays of the protein structure space. Both systems promote and facilitate human-computer synergy, where cognitive elements such as domain knowledge, contextual reasoning, and purpose-driven exploration, are integrated with a host of powerful algorithmic operations for large-scale data analysis, multifaceted data visualization, and multi-source information integration.

Conclusions: The proposed design philosophy combines algorithmic components and cognitive expertise into a seamless processing-analysis-exploration framework. Using XMAS, we present case studies that explore transcriptional data from two highly complex domains: gene expression in the placenta during human pregnancy and reaction of marine organisms to heat stress. With PSPACE, we demonstrate how complex structure-function relationships can be explored. These results demonstrate the novelty, advantages, and distinctions of the proposed paradigm. Furthermore, the results highlight how biological insights can be combined with algorithms to discover meaningful knowledge and formulate evidence-based hypotheses. Finally, user studies against comparable systems indicate that both XMAS and PSPACE deliver results with high interpretability and efficacy. XMAS is available at: http://tintin.sfsu.edu:8080/xmas. PSPACE is available at: http://pspace.info/.

Pathway mergeability enables the evaluation of merging signaling pathways with protein interaction data

Xiaogang Wu, Indiana University
Jake Chen, Indiana University

It has been challenging to develop enhanced pathway tools that could enlarge the coverage and improve the quality of existing annotated human pathway data. Computational pathway discovery (i.e. identifying potential pathways) can be conducted in three ways: 1) inferring unknown pathways from microarray data, 2) merging known pathways having common molecular entities together, and 3) expanding known pathways with molecular interaction data. In this work, we will focus on the latter two ways. We aim to quantitatively evaluate the processes of merging similar or functionally-related signaling pathways together by linking them with protein-protein interaction (PPI) data. We presented a concept of pathway mergeability to examine the merging potential between two different pathways. We analyzed the mergeability for existing pathways in the Pathway And Gene Enrichment Database (PAGED) and mergeability variation for potential pathways expanded from the Human Annotated and Predicted Protein Interaction (HAPPI) database with different confidence scores. By comparing the mergeability variation between expanding existing pathways in the HAPPI and expanding existing pathways in randomly-permuted PPI networks, we revealed a quantitative relationship (a high degree of coherence) between signaling pathway data and high-quality PPI data. This relationship will further guide pathway merging processes, and the concept of pathway mergeability will also enable the next-generation pathway tool development. Finally, we used colorectal cancer (CRC) as an example to demonstrate that how to determine appropriate thresholds for pathway mergeability to identify mergeable CRC-specific pathways which focus on the same biological processes or even have the same name from different data sources.

Identification of A-to-I RNA Editing Sites in Honey Bee (Apis mellifera) Using RNA-Seq Data

Shu Tao; Department of Biology, Georgetown University, Washington, DC 20057; Division of Animal Sciences, University of Missouri, Columbia, MO 65211
Christine G. Elsik; Department of Biology, Georgetown University, Washington, DC 20057; Division of Animal Sciences, University of Missouri, Columbia, MO 65211; Division of Plant Science, University of Missouri, Columbia, MO 65211; MU Informatics Institute, University of Missouri, Columbia, MO 65211

Honey bee (*Apis mellifera*) is one of the most studied insects, due in part to its importance as a model to study molecular mechanisms related to social behavior. We performed computational identification of potential A-to-I RNA editing sites in the honey bee brain using the latest genome assembly and Illumina-generated RNASeq data from ten honey bee individuals (five nurses and five foragers, 2-3 replicates per individual). We chose the honey bee brain as the focus of this study because: *i*) it is important in controlling behaviors, and *ii*) studies in human and *Drosophila* have shown that A-to-I RNA editing is prevalent in brain. This is the first systematic investigation of A-to-I RNA editing in the honey bee brain. It will enhance the annotation of genes expressed in the honey bee brain, and provide candidate A-to-I RNA editing sites for further investigation.

Our RNASeq-based pipeline to identify A-to-I RNA editing will be presented. Some of the candidate A-to-I editing sites in *A. mellifera* were found in genes with one-to-one orthologs in *Drosophila melanogaster* that were already shown to have A-to-I RNA editing sites. For example, we found two conserved A-to-I RNA editing sites in the gene orthologous to *quiver* (*qvr*) in *Drosophila melanogaster*. The first conserved A-to-I editing site changes codon from AGU to GGU, causing the conversion of serine to glycine, and the second conserved site recodes histidine (CAC) to arginine (CGC). Conservation of these specific amino acid conversions across 300 million years of evolutionary divergence indicates important biological functions.

Rule Based Regression and Feature Selection for Biological Data

Sheng Liu
Shamitha Dissanayake
Sanjay Patel
Todd Mlsna
Xin Dang
Yixin Chen
Dawn Wilkins

Regression is widely utilized in a variety of biological problems involving continuous outcomes. There are a number of methods for building regression models ranging from linear models to more complex nonlinear ones. While linear regression techniques can identify linear correlations between input and output, in many practical applications, the relations are nonlinear. These relations can be modeled by nonlinear regression techniques effectively. However, in general, models built with nonlinear techniques are relatively harder for humans to interpret, which is crucial in many problems.

We propose a rule based regression algorithm that uses 1-norm regularized random forests. The proposed approach simultaneously extracts a small number of rules from generated random forests and eliminates unimportant features. We tested the approach on a seacoast chemical sensors dataset, a Stockori flowering time dataset, and two datasets from UCI repository. The proposed approach is able to construct a significantly smaller set of regression rules using a subset of attributes while achieving prediction performance comparable to that of random forests regression. It demonstrates high potential in aiding prediction and interpretation of nonlinear relationships of subject being studied.

Topological Data Analysis of Triple Negative Breast Cancer Transcriptome and Proteome

Aleksandra Markovets, UALR/UAMS Bioinformatics Program
Damir Herman, University of Arkansas for Medical Sciences

The latest genomics and proteomics technologies are capable of generating vast quantities of high dimensional data. We used next generation sequencing and MS/MS mass spectrometry to create large molecular data sets of Triple Negative Breast Cancer (TNBC), a form of very aggressive breast cancer that preferentially targets young African American women. We performed extensive quality control and primary bioinformatics analysis of raw reads and proteomic spectra. In the next step we investigated transcriptional and translational regulation of cancer relative to breast cancer cells and pathologically cancer-free surrounding tissue. Using molecular and clinical information for each patient and cell line, we built a vast network that we analyzed using Topological Data Analysis (TDA). TDA was developed at the Department of Mathematics at Stanford in response to the Defense Advanced Research Programs Agency's (DARPA) request to find application of results in pure mathematics to practical problems. TDA has been very successfully applied to find new insights in a 10 year old breast cancer data set that was used for validation of MammaPrint – the first FDA-approved breast cancer molecular test. We will also show results of the TDA analyses of large sequencing efforts such as the 1000 Genome Project and the Cancer Genome Project. While TDA works on a wide range of problems spanning national security, oil and gas exploration and transportation, in this presentation we will be mainly focused on the topological insights from the aforementioned publicly available genomic data sets and our TNBC data.

Identification of genes specific to the lineage of the Hymenopteran insect, honey bee (Apis mellifera)

Anna K. Bennett, Department of Biology, Georgetown University, Washington, DC 20057; Division of Animal Sciences, University of Missouri, Columbia, MO 65211
Christine G. Elsik, Department of Biology, Georgetown University, Washington, DC 20057; Division of Animal Sciences, University of Missouri, Columbia, MO 65211; Division of Plant Sciences, University of Missouri, Columbia, MO 65211; MU Informatics Institute, University of Missouri, Columbia, MO 65211

Honey bees are important as key agricultural pollinators and models for social behavior and the evolution of eusociality. As social insects, honey bees live in close proximity to each other within hives and cooperate as a society with division of labor. The honey bee genome published in 2006 by the Honey Bee Genome Sequencing Consortium, had fewer gene predictions than expected, partially due to lack gene evidence in the form of transcriptome data and protein homologs from closely related species. As part of ongoing consortium efforts, we produced an improved official gene set (OGSv3.2) for honey bee with ~5000 more protein-coding genes than the first set. In search of genomic differences that contributed to diversification of the honey bee from other insects, we identified genes specific to *Apis mellifera* and to lineages within the insect order Hymenoptera. With numerous recently sequenced arthropod genomes, the Hymenopteran clade is now well-sampled, providing the genomic and transcriptomic data necessary to allow for fine-scale, comparative analyses to detect lineage specific genes.

We will present our approach, which relied largely on Next Generation Sequencing, to detect previously unknown genes in *A. mellifera,* and to determine which genes were specific to the *A. mellifera* lineage. We identified genes that were specific to the *A. mellifera* species, *Apis* genus, Apidae family, and Hymenoptera order, including those with tentative roles in brood care, immunity and other processes important to hive health, and thus critical to the emergence of eusociality.

What can we learn from phenotypes?

Toni Kazic, Avimanyou Vatsa, Derek Kelly, Wade Mayham, Leonard Hearne

Many phenotypes --- organismal development, disease resistance, and agronomic productivity --- are complex. Biologically, complexity has two meanings: phenotypes are a composition of many distinct but interacting phenomenological features; and phenotypes are caused by a network of processes and their constituent molecular machinery that interact with the environment. Mathematically, this complexity is naturally expressed as points in high-dimensional domains. Understanding the mechanisms of complex phenotypes begins with their close observation and characterization; and using appropriate mathematical techniques to discover the relationships among phenotypic dimensions, and between phenotypes and genotypes. In this talk, we describe a set of related complex phenotypes in maize; what they can teach us about their underlying causal processes; and some principled techniques for exploring high dimensional domains.

Predicting Multi-target Protein Subcellular Localization combining homology and Machine Learning Approaches

Sitanshu S Sahu, National Institute for Microbial Forensics & Food and Agricultural Biosecurity (NIMFFAB), Department of Biochemistry & Molecular Biology, Oklahoma State University, Stillwater (OK), 74074 USA
Rakesh Kaundal, National Institute for Microbial Forensics & Food and Agricultural Biosecurity (NIMFFAB), Department of Biochemistry & Molecular Biology, Oklahoma State University, Stillwater (OK), 74074 USA

Computational prediction of subcellular localization is an important issue in protein science as it provides relevant information in determining the functionality, rational drug design, gene product annotation etc. Despite several efforts over the past few years, accurate localization prediction still remains a grand challenge. Most of the recent studies have primarily focused on single compartment localization, and a very little attention has been paid towards the dual and multi-target localizations. Here we develop a machine learned framework to predict a range of single, dual and multi-target proteins on a genome-wide scale viz. (1) Cell membrane, (2) Cell wall, (3) Plastid, (4) Cytoplasm, (5) Endoplasmic reticulum, (6) Extracellular, (7) Golgi apparatus, (8) Mitochondrion, (9) Nucleus, (10) Peroxisome, (11) Vacuole, (12) Cytoplasm-Golgi apparatus, (13) Cytoplasm-Nucleus, (14) Mitochondrion-Cytoplasm, (15) Other Dual and (16) Multi-target (>2 locations). Based on various protein features such as the compositional information, sequence order and length effects, diverse Support Vector Machine (SVM) modules have been developed with efficiency assessed through a 5-fold cross-validation test. The simple amino acid composition achieves an overall sensitivity, specificity and Matthews correlation coefficient of 81%, 89% and 0.62 respectively outperforming its counterpart features. Ongoing work includes exploring additional features such as the similarity-based PSI-BLAST, evolutionary information-based Position Specific Scoring Matrix (PSSM), autocorrelation, and a range of hybrid combinations to increase the prediction performance. These modules will be tested on independent datasets and performance compared with the existing tools. The best performing modules will be implemented as a web server for public use.

Bioinformatics Approaches to Deciphering Host-Pathogen Protein Interaction Networks (PINs)

Rakesh Kaundal Oklahoma State University

Protein-Protein Interactions (PPIs) are an important aspect of initiating pathogenesis and maintaining infection. These PPIs control interchanges between plant host-pathogen systems on the molecular level, which play a vital role in the success of the plant's defense or of the pathogenesis. Here we report the first comprehensive study to predicting genome-wide host-pathogen networks in any plant-pathogen interaction system. Experimentally proven genomic data and protein features from the Arabidopsis-*Pseudomonas syringae* pv tomato strain DC3000 model interaction system are combined into a Support Vector Machine (SVM), to develop computational models for predicting genome-wide inter-species PINs. Several models are constructed based on similarity-search, machine learning, and a series of combinatorial approaches using diverse sequence features such as amino acid and dipeptide composition, physicochemical properties, domain information. Under five-fold cross-validation training/testing, the best performance is achieved with physicochemical model with a sensitivity of 100%, specificity 99.14% and Matthews correlation coefficient of .99. These models have been further tested on independent datasets from the Host-Pathogen Interaction Database (HPIDB) that consists of 58 different hosts (plant and animal) and 416 pathogens. Independent test results achieve 100% accuracy in predicting positive Arabidopsis-*P. syringae* interacting sequences, with a false positive rate of as low as 12.92% on the negative sets. A web-based prediction system has also been developed for querying unknown host-pathogen interactions, and is freely available at http://bioinfo.okstate.edu/AP-iNET/. We believe this system could be widely used to study agriculturally relevant crop hosts and their interacting pathogens, thus benefiting the plant research community in developing better resistance cultivars.

NeedleFinder: A Data Analysis Tool for LCMS-Based Metabolomics

Stephen Embry[1], Heng Luo[1], Stephen C. Grace[2]

[1]Bioinformatics Program, Department of Information Science, [2]Department of Biology, University of Arkansas at Little Rock, Little Rock, AR 72204

There is growing interest in untargeted metabolomics in life and health sciences to identify certain conditions or phenotypes. LCMS is the most common analytical platform in metabolomics research since it provides the broadest coverage of the metabolome and requires relatively simple sample preparation. However, mining LCMS for important features presents a major bottleneck to metabolomics research. Here we introduce a novel platform called NeedleFinder to filter and extract significant features from LCMS datasets rapidly and efficiently. NeedleFinder is a data analysis tool being developed in Perl, along with an intuitive graphical interface powered by Php and a database in MySql, that can streamline the analysis of metabolomic data from full scan LCMS-based chromatograms output cdf files. The core function of NeedleFinder is to produce a graphical representation of a dataset akin to a "barcode" consisting of all masses parsed into user-defined bins that can be analyzed by exploratory methods such as PCA. Users can also search for target masses and obtain extracted ion chromatograms listing time and intensity values. Graphical representation of the outputs (bin intensity, EIC, etc) is provided, along with comma separated files, so that the user can export the results to other software for further analysis or customized graphing. A major advantage of NeedleFinder over other LCMS processing platforms is that it provides an unbiased representation of the data since peaks are not assigned to specific masses by retention time indices. However, peaks of interest can be recovered as latent variables using S/N analysis of individual bins.

MULTICOM – RNA-Seq Data Analysis of Mouse Transcriptomes Perturbed by Botanicals

Jilong Li [1,2,] Jordan Wilkins [3], Mingzhu Zhu [2], Kevin Fritsche [1,4,] Michael Greenlief [1,5,] William Folk[1,3], Dennis Lubhan [1,4,] Mark Hannink [1,3,] Jianlin Cheng[1,2]*

[1] MU Botanical Center, [2] Department of Computer Science, [3] Biochemistry, [4] Animal Sciences, [5] Chemistry, University of Missouri, Columbia, MO 65211, USA.

*chengji@missouri.edu

Abstract

We have developed MULTICOM - a bioinformatics pipeline - for RNA-Seq transcription data analysis using five steps, including: i) mapping reads to a reference genome; ii) normalizing read counts into gene expression values; iii) identifying differentially expressed genes; iv) predicting gene functions; v) constructing biological pathways. We applied MULTICOM to the RNA-Seq data generated with two mouse fibroblast cell lines treated by botanical extracts of Sutherlandia and Elderberry and the Nrf2 activator CDDO (2-cyano-3,12-dioxooleana-1,9-dien-28-oic acid). Initial analysis identifies ~430 differentially expressed coding and non-coding genes, whose functions can be largely assigned to ten biological processes; genes differentially expressed by all the botanicals can be clustered into functional groups; their hypothetical gene regulatory pathways were also constructed.

Local and global quality assessment by MULTICOM during CASP10

Renzhi Cao [1]; Jianlin Cheng [1,2,3]

[1]Computer Science Department; [2]Informatics Institute; [3] C. Bond Life Science Center, University of Missouri, Columbia, MO 65211, USA.

Corresponding author: Jianlin Cheng (chengji@missouri.edu)

Abstract

We developed and tested four model quality assessment (QA) servers: MULTICOM-REFINE, MULTICOM-CLUSTER, MULTICOM-NOVEL, MULTICOM-CONSTRUCT. For all of the four QA servers, they can generate both global quality scores and local quality scores. For MULTICOM-REFINE, the global quality score is generated by the pair-wise method, which calculates the pair-wise GDT-TS score of a pool of models and then evaluate the quality of each model; the local quality score is generated by the refined local scores. MULTICOM-CLUSTER and MULTICOM-NOVEL are new, single-model, support vector machine (SVM)-based method. The SVM was trained to predict the local quality score of each residue and the global quality scores were generated from the local quality scores. The input features for MULTICOM-CLUSTER to the SVM includes amino acids encoded by 20-digit vector of 0 and 1, the difference between secondary structure and solvent accessibility predicted by SCRATCH from protein sequence and that of a model parsed by DSSP, and predicted contact probabilities. MULTICOM-NOVEL is the same as MULTICOM-CLUSTER except that amino acid sequence features were replaced with the sequence profile features. The global quality scores of MULTICOM-CONSTRUCT are generated by refined pair-wise model comparison method. The local quality scores are all generated by using SVM method. MULTICOM-CONSTRUCT uses secondary structure difference, solvent accessibility, profiles, and SOV score as the features for SVM.

Paternal Influence on Transcriptomic Landscapes of *in vitro* derived bovine embryos

Sule Dogan[1*], Pablo Ross[2], Abdullah Kaya[3], and Erdogan Memili[1]

[1]Animal and Dairy Sciences, Mississippi State University, MS 39762
[2]Department of Animal Science, University of California-Davis
[3]Alta Genetics, Watertown, WI 53094

Embryo transfer (ET) still relies on morphological assessment of embryo quality in assisted reproductive technologies (ARTs), and the male factor is generally being neglected. Paternal influence on embryo quality is not well understood. The objective of this study was to determine global transcriptomics of single blastocyst stage embryos-derived using sperm from bulls with extreme fertility phenotypes [Low *vs*. High fertility]. We accomplished our objectives by 1) performing *in vitro* fertilization (IVF) to develop blastocysts-derived using spermatozoa from Low *vs*. High fertility bulls 2) performing RNA sequencing to determine global transcriptomics of those *in vitro*-derived blastocysts, 2) computing the RNA-Sequencing data via DAVID and IPA to illustrate the potential pathways and networks wherein the unique transcripts involve, 3) validating the RNA-Sequencing data via qRT-PCR. Our results showed that although there was no difference between IVF results, a total of 909 genes were found to be differentially expressed between two groups by RNA-seq (FDR p ≤0.05 and fold change ≥2). In addition, functional annotation clustering by DAVID showed that epidermal growth factor (EGF) signaling cluster was firstly and differently enriched in embryos of low *vs.* high fertility groups (1.69 *vs.* 2.47, respectively; p <0.05). RNA-seq data were validated by qRT-PCR using a panel of genes. In conclusion, we determined transcriptome profiling of single blastocysts-derived from spermatozoa of low *vs*. high fertility bulls, and analyzed the data using computational biology approaches. The results are significant because they help us better understand early mammalian embryos and to improve efficiency of mammalian reproduction.

Key words: RNA sequencing, Transcriptomics, male fertility, systems biology.

*Presenter: suledogann@gmail.com , sd565@msstate.edu

A Scalable and Deterministic Finite State Automaton-based Model to Determine Ancestor-Descendant Relationships in Directed Acyclic Graphs

Andrew J. Overton[1*], Natarajan Meghanathan[1], Raphael Isokpehi[2]

[*] Student Author
andrew.j.overton@students.jsums.edu

[1] Mailbox 18839
Department of Computer Science
Jackson State University
1400 J R Lynch Street
Jackson, MS 39217

[2] Mailbox 18540
Department of Biology
Jackson State University
1400 J R Lynch Street
Jackson, MS 39217

We propose the use of a binary code-based finite state machine (FSM) model to efficiently represent and verify ancestor-descendant relationships between any two nodes in a directed acyclic graph in polynomial time. The FSM of a child node, with single inheritance, is constructed by adding a unique combination of transitions to the terminal state (a state with no further transitions) of the parent's FSM corresponding to the suffix added to a parent's binary code for its children. The FSM model captures multiple inheritance of a child node by joining together the FSMs of the parent nodes which likely creates a non-deterministic finite automaton (NFA). This NFA is then minimized into a deterministic finite automaton (of smaller size and complexity) by identifying common states and transitions, leading to a single terminal state. The resulting single DFA for a child node comprehensively captures the inheritance information from all the parents and passes this along to its own children. This negates the code explosion problem (compared to the previous approach, wherein the binary code for a node is the concatenation of the code for its parent node and the unique binary representation of the immediate children of the parent node). To test for an ancestor-descendant relationship, a binary code is generated by walking through the DFA of a prospective ancestor node to one of the terminal states in the automaton. If the resulting binary code can be completely walked through the DFA of a prospective descendant node, then the two nodes are related.

Contacts-Assisted Protein Structure Prediction

Badri Adhikari, Xin Deng, Jilong Li, Debswapna Bhattacharya, and Jianlin Cheng
Department of Computer Science, University of Missouri, Columbia, MO 65211 USA
chengji@missouri.edu

One recent approach for protein tertiary structure prediction from its residue sequence is to predict which residues are close to each other first, and then build a complete structure solely from this contacts information. These methods use a threshold distance to define this closeness or residue-residue contact. Instead of building a structure purely from these contacts, we summarize a contact-assisted structure prediction approach that uses only a few known contacts to improve the quality of already predicted models. Assuming that we already have some predicted structures and some known contacts, we designed and implemented an automated pipeline that starts with a predicted structure and improves the structure using the inputted contacts as constraints. The system also handles cases when some non-contact information (i.e., knowledge that two residues are not in contact) is provided as input along with or instead of contact information. Our approach for contact assisted structure prediction is a model selection and improvement process comprising of three major steps. First, we select models from a predicted model pool using a scoring scheme. We then refine these selected models using existing protein refinement tools. Finally, we improve the structure of these refined models with given residue-residue contacts information as distance restraints. Our experiment during the 10th Critical Assessment of Techniques for Protein Structure Prediction (CASP10) in 2012 shows that in most cases the quality of predicted structures is improved. The server for contact-assisted protein structure prediction is available at:
http://protein.rnet.missouri.edu/contact_assisted/index.html

Constructing Three-Dimensional Structures of Human Chromosomes from Chromosomal Contact Data

Tuan Trieu[1] and Jianlin Cheng[1,2,3]

[1] Email: tatrz9@mail.missouri.edu, Department of Computer Science, [2]Informatics Institute and [3]C. Bond Life Science Center, University of Missouri, Columbia, MO 65211, USA

Chromosomes are not positioned randomly within the nucleus, but they tend to adopt preferred conformations. To study the topologies and how chromosomes fold, here we describe a method to generate three-dimensional structure of individual chromosomes from contact maps. An objective function was derived for structures and this objective function was then optimized to generate structures that satisfy as much as possible the corresponding contact maps. We used contact maps from a Hi-C dataset to generate three-dimensional structures for chromosomes at resolution of one megabase. Parameters used in our method were derived from a FISH data to construct structures as close to the true structure as possible. Depending on chromosome length, our generated structures satisfy from seventy to one hundred percentages of contact and non-contacts from the corresponding contact maps. We compared our structures with the FISH data and found that the graph of spatial distance as a function of genomic distance is similar to the FISH data's. We also verified our structures with the two compartments feature of chromosome and found that our structures agree with this feature.

Poster Presentation Abstracts

Analysis of co-localization of 5-hydroxymethylcytosine and G-quadruplex-forming regions in gene transcriptional start sites: might hydroxymethlation affect the C-rich i-motif structure?

Yogini P.Bhavsar-Jog The University of Mississippi
Eric Van Dornshuld The Univeristy of Mississippi
Randy M. Wadkins The Univeristy of Mississippi

5-hydroxymethylation of cytosine bases (5hmC) in genomic DNA of stem cells is an epigenetic modification occurring at the transcriptional start sites (TSS) of numerous genes. The 5hmC is known to be associated with gene enhancers, suggesting a potential function of the modified cytidine in gene regulation. Additionally, G-quadruplexes (in G-rich strands) and i-motifs (in C-rich strands) are topological variants (secondary structures) of DNA that are also known to be clustered in close proximity to TSS and regulate transcription. Genome analysis of G-rich sequences with a greater potential to form quadruplex structures suggests that these structures are present in higher abundance in proto-oncogenes as opposed to other classes of genes. Hence, both 5hmC and quadruplex-forming sequences are regulatory elements that occur at TSS of many genes. However, it is still unknown whether 5hmC enriched regions of a particular gene overlaps with, or is related to, the probability of occurrence of number of G-quadruplexes in its vicinity. Here, we present statistical analyses of the co-localization of these two elements in an effort to explore the hypothesis that enrichment of 5hmC occurs in the C-rich strands complimentary to G-rich strands with quadruplex potential in and around a TSS.. Our preliminary results indicate that conversion of 5mC to 5hmC eliminates the negative correlation between 5mC occurrence and i-motif co-localization.

Comparison of Data Mining Methods on Microarray Gene Expression Data on Cancer

Shen Lu, Soft Challenge LLC
Richard Segall, Arkanas State University at Jonesboro
Thomas Hahn, University of Arkanas at Little Rock, and University of Arkanas Medical Science

The use of pattern recognition software is not new in the field of bioinformatics, but is not as developed as it could be. With the growing amounts of data that are being produced by various Microarray technologies and other devices on the one hand, and an appreciation of the fact that there is more to the vast amount of 'non-coding' DNA than meets the eye certainly for H. sapiens on the other, we survey microarray experimental data to see possibilities and problems to control microarray expression data. We use both variable measure and attribute measure to visualize microarray expression data. According to the data structure of the attribute, we use control charts to visualize fold change and t-test attributes in order to find the root causes of the Microarray data quality. After comparison of several data mining algorithms, such as KNN, Random Forest, Self Organizing Map (SOM), and Multi-pass LVQ, we improved SOM algorithm by dividing the one map into several smaller maps, generated the model with Multiple Layer SOM, and performed data analysis with Microarray gene expression data for liver cancer with 179 samples and 772 genes after pre-processing (19536 genes before pre-processing). Our experimental results show that the precision of Multiple Layer SOM is 10.58% greater than SOM, and its recall is 11.07% greater than SOM. In future, we will use different Microarray gene expression data to test this process in order to further improve it.

Comparing transcriptome response to amphetamine and environment induced hyperthermia in rat brains and blood

John Bowyer, National Center for Toxicology Research

Rats were given amphetamines to induce hyperthermia. These rats were compared to a group of rats with hyperthermia due to a hot environment. The rats were sampled for blood, and four brain regions: choroid plexus, striatum, meninges and associated vasculature (MAV), and the parietal cortex. RNA was isolated from the samples and gene expression micro array experiments were performed. Differential gene expression was found for genes related to angiogenesis, vaso-regulation, neurotoxic damage, and immunity.

Iterative reconstruction of three-dimensional model of human genome from chromosomal contact data

Sharif Ahmed; Department of Computer Science, University of Missouri, Columbia, MO 65211
Jianlin Cheng; Department of Computer Science, Informatics Institute, C. Bond Life Science Center, University of Missouri, Columbia, MO 65211

We developed an iterative two-step approach to reconstruct three-dimensional (3D) structural model of the human genome from chromosomal contact data. In the first growth step, the method builds the initial structure of a chromosome by sequentially positioning one chromosomal region unit (e.g., 1M bp) at a time. The position of a unit is chosen probabilistically from a set of randomly sampled positions, where the positions that have more positive chromosomal contacts have a higher probability to be selected. In the second adaptation step, the method randomly picks a chromosomal unit and tries to adjust its position randomly in order to increase the satisfaction of the observed chromosomal contacts in the structural model. The fitness score that determines the probability of accepting a position adjustment (i.e., move) includes a contact satisfying score, a non-contact satisfying score, distance constraints between adjacent /contact units. The adaptation process generally runs for many cycles generating a large ensemble of chromosomal conformations. The conformations with relatively high fitness scores are chosen as candidate models. Our preliminary experiment suggests that the method can generate 3D chromosomal models satisfying a large portion of chromosomal contacts rather quickly.

Bridging the gap between soybean translational genomics and breeding with Soybean Knowledge Base (SoyKB)

Joshi T Department of Computer Science, Christopher S. Bond Life Sciences Center, National Center for Soybean Biotechnology, Informatics Institute, University of Missouri, Columbia, MO 65211, USA
Fitzpatrick MR Department of Computer Science, University of Missouri, Columbia, MO 65211, USA
Chen SY Department of Computer Science, University of Missouri, Columbia, MO 65211, USA
Liu Y Informatics Institute, University of Missouri, Columbia, MO 65211, USA
Endacott RZ Department of Computer Science, University of Missouri, Columbia, MO 65211, USA
Zhang H Department of Computer Science, University of Missouri, Columbia, MO 65211, USA
Gaudiello EC Department of Computer Science, University of Missouri, Columbia, MO 65211, USA
Cheng J Department of Computer Science, Christopher S. Bond Life Sciences Center, National Center for Soybean Biotechnology, Informatics Institute, University of Missouri, Columbia, MO 65211, USA
Stacey G Division of Plant Sciences, Christopher S. Bond Life Sciences Center, National Center for Soybean Biotechnology, Informatics Institute, University of Missouri, Columbia, MO 65211, USA
Nguyen HT Division of Plant Sciences, Christopher S. Bond Life Sciences Center, National Center for Soybean Biotechnology, Informatics Institute, University of Missouri, Columbia, MO 65211, USA
Xu D Department of Computer Science, Christopher S. Bond Life Sciences Center, National Center for Soybean Biotechnology, Informatics Institute, University of Missouri, Columbia, MO 65211, USA

Many genome-scale data are available in soybean including genomic sequence, transcriptomics (microarray, RNA-seq), proteomics and metabolomics datasets, together with growing knowledge of soybean in gene, microRNAs, pathways, and phenotypes. This represents rich and resourceful information which can provide valuable insights, if mined in an innovative and integrative manner and thus, the need for informatics resources to achieve that.

Towards this we have developed Soybean Knowledge Base (SoyKB), a comprehensive all-inclusive web resource for soybean translational genomics and breeding. SoyKB handles the management and integration of soybean genomics and multi-omics data along with gene function annotations, biological pathway and trait information. It has many useful tools including Affymetrix probeID search, gene family search, multiple gene/metabolite analysis, motif analysis tool, protein 3D structure viewer and download/upload capacity for experimental data and annotations. It has a user-friendly web interface together with genome browser and pathway viewer, which display data in an intuitive manner to the soybean researchers, breeders and consumers.

SoyKB has new innovative tools for soybean breeding including a graphical chromosome visualizer targeted towards ease of navigation for breeders. It integrates QTLs, traits, germplasm information along with genomic variation data such as single nucleotide polymorphisms (SNPs) and genome-wide association studies (GWAS) data from multiple genotypes, cultivars and *G.soja*. QTLs for multiple traits can be queried and visualized in the chromosome visualizer simultaneously and overlaid on top of the genes and other molecular markers as well as multi-omics experimental data for meaningful inferences.

SoyKB can be publicly accessed at http://soykb.org.

Clustering Gene Expression Data using Probabilistic Non-negative Matrix Factorization

Belhassen Bayar University of Arkansas at Little Rock
Nidhal Bouaynaya University of Arkansas at Little Rock
Roman Shterenberg University of Alabama at Birmingham

Non-negative matrix factorization (NMF) has proven to be a useful decomposition for multivariate data, where the non-negativity constraint is necessary to have a meaningful physical interpretation. The NMF algorithm, however, assumes a deterministic framework. In particular, the effect of the data noise on the stability of the factorization and the convergence of the algorithm are unknown. Collected data, on the other hand, is stochastic in nature due to measurement noise and sometimes inherent variability in the physical process.

We present a new theoretical and applied developments to the problem of non-negative matrix factorization. First, we extend the NMF framework to the probabilistic case (PNMF). We show that the Maximum A Posteriori estimate of the non-negative factors is the solution to a weighted regularized non-negative matrix factorization problem. We subsequently derive update rules that converge towards an optimal solution.

Finally, we apply the PNMF to cluster DNA microarrays data. The proposed PNMF is shown to outperform the deterministic NMF algorithm in clustering stability accuracy.

Predicting Protein Model Quality from Sequence Alignment by Support Vector Machines

Jianlin Cheng

Protein sequence alignment is essential for homology-based protein structure modeling. Here, we developed a SVM (Support Vector Machine) alignment-based model selection method to predict the quality score (GDT-TS score) of a protein structure model from the features extracted from the query-template alignment used to generate the model. The input features fed into the SVM predictor include the normalized e-value of the given query-template alignment, the percentage of identical residue pairs of alignment positions, the percentage of residues of the query aligned with one in the template, and the sum of the BLOSUM scores of all aligned residues divided by the length of the aligned positions. The four input features were extracted from 482 pairwise sequence alignments generated from the CASP9 (Critical Assessment of Techniques for Protein Structure Prediction) datasets by PSI-BLAST along with the real GDT-TS scores calculated by the TM-Score program. In the training process, a SVM regression predictor was trained based on the above data to predict the GDT-TS scores of the models from the input features. Three parameters (the epsilon width of the regression tube w; the margin option c; the gamma in the RBF kernel g) were tuned during training the SVM with Gaussian radial basis kernel (RBF) regression model. The root mean square error (RMSE) and the absolute mean error (ABS) between predicted and real GDT-TS scores were calculated to assess the performance. A five-fold cross validation was adopted to select the best parameter sets based on the average RMSE and ABS on the five folds. The RMSE and ABS of the SVM trained with the best parameter values were 0.085 and 0.06, respectively. We compared the SVM alignment-based predictor with the pure e-value based method. The better performance of SVM predictor indicates that integrating sequence alignment features with a SVM can improve model selection over the pure e-value based method.

A Bioinformatics Study on Evolutionary Diversification of Multiple Inositol Polyphosphate Phosphatase1 as an Aid to Understanding its Functional Significance in Mammalian Systems

Surya P. Kilaparty
Awantika Singh
Nawab Ali

Multiple Inositol Polyphosphate Phosphatase1 (Minpp1) in higher organisms dephosphorylates the most abundant inositol hexakis-phosphate (InP_6) as well as more phosphorylated ($InsP_7$ or $InsP_8$) or less phosphorylated ($InsP_5$ and $InsP_4$) than $InsP_6$. Changes in cellular levels of these InsPs have been implicated in the regulation of diverse cellular functions. Minpp1 is classified as a member of the histidine phosphatase super family of proteins and resembles phytases found in lower organisms in its function. Besides InsPs, Minpp1 also dephosphorylates other compounds and shows mutase activity. In mammalian systems, normally, it is localized inside endoplasmic reticulum as soluble luminal protein without any demonstrated access to InsPs, whereas in other organisms it may be membrane bound or secreted extracellularly. In order to understand physiological relevance of Minpp1 in higher organisms, we used a bioinformatics approach to explore the extent of evolutionary diversification in its structure and function. Human Minpp1 amino acid sequence was BLAST searched in NCBI and EMBL-EBI databases. Phylogenetic analysis revealed that MInpp1 is widely distributed and there exists a systematic pattern in sequence identity from lower to higher organisms. Multiple search tools (TASSEL, PROSITE, Pfam, SMART, PANTHER) were used to identify key functional motifs in Minpp1 which were further analyzed by ClustalX2 for comparison among species. Certain motifs that were predominant in higher organism were absent in lower organisms and vice-versa. This study demonstrates a diversification of the motifs and sheds light on evolutionary adaptability of Minpp1 function from lower to higher life forms.

Travel Supported by grants from NCRR (5P20RR016460-11) and NIGMS (8 P20 GM103429-11) at NIH

GMOL: A Tool for 3D Genome Structure Visualization

Chenfeng He (Informatics Institute, University of Missouri)
Avery O. Wells (Department of Computer Science, University of Missouri)
Jianlin Cheng (Informatics Institute, University of Missouri and Department of Computer Science, University of Missouri)

We developed a tool, GMOL, to visualize genome tertiary structure. GMOL is based upon Jmol, an open-source JAVA program for visualizing smaller molecular structures like proteins. However, when it comes to large-scale structures, such as the human genome, which consists of nearly three billion DNA base pairs, Jmol is unsatisfactory. To solve this, we developed a multi-scale strategy to meet the requirements for genome structure visualization. In order to efficiently visualize an entire genome structure, GMOL uses five separate scales. These scales are (using the human genome as an example): Genome Scale, Chromosome Scale, Loci Scale, Fiber Scale, and Nucleotide Scale. A special file format, “GSS”, was designed to store the five different scales. Additionally, new functions were developed and integrated into GMOL. These functions allow the user to browse the genome sequence and also enable PDB file transforming. With GMOL, a user can choose any point (denoted as a “unit” in this report) at any scale and scale it up or down to visualize its structure and get genome sequence from either Ensembl or a local database. Furthermore, a user can extract the PDB format file of any single unit for visualization with other tools.

Molecular Dynamic Simulation of β-amyloid peptide

Meenakshisundaram Balasubramaniam, University of Arkansas at Little Rock/Universiy of Arkansas for Medical Sciences,V.A Medical Center, Little Rock, Arkansas.
Shmookler Reis RJ, University of Arkansas for Medical Sciences, V.A Medical Center, Little Rock, Arkansas.

Alzheimer's disease (AD) is characterized by β-amyloid aggregation. Cleavage of Aβ within the Amyloid Precursor Protein (APP) at the C-terminus by the action of γ-secretase results in the release of $A\beta_{42}$ peptides. We used 1-ns atomistic molecular dynamics simulations to study the dynamic behavior of $A\beta_{42}$ peptide. Our result shows $A\beta_{42}$undergoes fluctuations and slow conformational change over the simulation time. This dynamic behavior and fluctuations in the structure may allow exposure of internal backbone amino acids to surface, which can then interact with the same peptide to form an $A\beta_{42}$ dimer and/or with, can interact with other proteins to form larger aggregates.

A Bioinformatics Study on Evolutionary Diversification of Multiple Inositol Polyphosphate Phosphatase1 as an Aid to Understanding its Functional Significance in Mammalian Systems

surya kilaparty, UALR
Awantika Singh, UALR/UAMS
Nawab Ali, UALR

Multiple Inositol Polyphosphate Phosphatase1 (Minpp1) in higher organisms dephosphorylates the most abundant inositol hexakis-phosphate (InP_6) as well as more phosphorylated ($InsP_7$ or $InsP_8$) or less phosphorylated ($InsP_5$ and $InsP_4$) than $InsP_6$. Changes in cellular levels of these InsPs have been implicated in the regulation of diverse cellular functions. Minpp1 is classified as a member of the histidine phosphatase super family of proteins and resembles phytases found in lower organisms in its function. Besides InsPs, Minpp1 also dephosphorylates other compounds and shows mutase activity. In mammalian systems, normally, it is localized inside endoplasmic reticulum as soluble luminal protein without any demonstrated access to InsPs, whereas in other organisms it may be membrane bound or secreted extracellularly. In order to understand physiological relevance of Minpp1 in higher organisms, we used a bioinformatics approach to explore the extent of evolutionary diversification in its structure and function. Human Minpp1 amino acid sequence was BLAST searched in NCBI and EMBL-EBI databases. Phylogenetic analysis revealed that MInpp1 is widely distributed and there exists a systematic pattern in sequence identity from lower to higher organisms. Multiple search tools (TASSEL, PROSITE, Pfam, SMART, PANTHER) were used to identify key functional motifs in Minpp1 which were further analyzed by ClustalX2 for comparison among species. Certain motifs that were predominant in higher organism were absent in lower organisms and vice-versa. This study demonstrates a diversification of the motifs and sheds light on evolutionary adaptability of Minpp1 function from lower to higher life forms.

A Constrained Importance Sampling Approach for Inference of Time-Varying Gene Regulatory Networks

Jehandad Khan, University of Arkansas at Little Rock
Nidhal Bouaynaya, University of Arkansas at Little Rock

We tackle the problem of inference of time varying genetic regulatory networks from a limited number of observations. Gene regulatory networks evolve over time in response to functional requirements in the cell and environmental conditions. Collected genetic profiles from dynamic biological processes, such as cell development, cancer progression and treatment recovery, underlie genetic interactions that rewire over the course of time. We formulate the problem of estimating time-varying networks in a non-linear state-space framework, where the non-linearity is due to the dynamics of gene expression. We subsequently use the Importance Sampling (IS) method to iteratively infer the gene interactions at each time instant. We further incorporate the prior knowledge of sparsity of the network by constraining the importance sampling approach to sparse states. The sparsity constraint reduces the dimensionality of the system and increases the accuracy of the estimation while bringing down the computational cost of the algorithm.

Challenges in Inferring Large-Scale Networks

Muhammad Baig Awan, University of Arkansas at Little Rock
Nidhal Bouaynaya, University of Arkansas at Little Rock

The problem of inference of gene regulatory networks suffers from poor sampling, i.e., a limited number of observations, and a large number of genes. These issues are even more severe in gene regulatory networks, which rewire over time due to cellular requirements or environmental conditions. In the time-varying scenario, we need to track hundreds of genetic interactions at each time point from only a limited number of noisy measurements. We formulate this tracking problem in a non-linear state-space framework, where the non-linearity is due to the gene expression dynamics. We employ the particle filter to estimate the hidden state parameters, i.e., the gene interactions at each time instant. However, due to the large dimension of the state vector, the particle filter suffers from the degeneracy or sample attrition problem. We suggest an alteration in weight calculation for high-dimensional states to avoid the degeneracy of the algorithm and reduce the computational cost of the particle filter.

NETS TOOL: A computational Approach to predict potential drug candidates

Sravanthi Joginipelli, Jerry A. Darsey, UALR/UAMS Bioinformatics program

Artificial neural networks (ANNs) are biologically inspired computer programs designed to simulate the way in which the human brain processes information. Predicting potential drug candidates with the help of artificial neural networks can reduce the time and effort of traditional methods. Unfortunately most of the traditional approaches rely to some extent on classical structure activity approaches that require experimental data, complex calculations and other parameters that limit the applicability of the methods for reasons of cost, time, and prior knowledge. Ideally, Computational methods could be employed, at least for screening, thus reducing the number of molecules requiring complete evaluation by biological assays and allow additional screens for molecules whose drug activity is currently unknown. This computational approach serves as interface between electronic structure modeling program (Gaussian) and neural network program (NETS). Generating training sets and tests sets for neural network program is a time consuming process, the effort of generating training and test sets can be reduced with this program written in PERL (Practical Extraction and Report Language). This project develops powerful predictive models in less time and predicts potential drug candidates by their physicochemical properties such as molecular orbitals HOMOS (Highest occupied molecular orbitals) and LUMOS (Lowest unoccupied molecular orbitals), Total energy, Dipole moment and inhibition concentration (IC50) and evaluates close connection to bioavailability as therapeutic agents.

Inference of Temporally rewiring genetic networks using time-varying differential equations

Annick Dongmo UALR

Collections of complex time-dependent genetic data from dynamic biological processes such as cancer progression, immune response, and developmental processes, are expanding given the improved data collection of new technologies. Clearly, the “static” view provided by the current time-invariant network topologies is obsolete for these

dynamical systems and we must develop methods and algorithms that reflect the dynamic or rewiring nature of genetic interactions. We model time-varying gene networks using time-varying coefficients differential equations with an additive noise term representing the errors in measurements and the model. Model networks resulting from differential equations are directed, allow for feedback loops, and categorize stimulations and inhibitions. The ability to recognize whether a gene interaction is stimulative or inhibitory is critical in understanding transcriptional regulatory interactions and designing appropriate drug targets. The inference problem is to estimate the time-varying network interactions. Given the limited number of observations at each time point, this problem is severely under-determined. We first transform the time-varying problem into a time-invariant one by decomposing the time-varying parameters in a suitable basis function. We then estimate the basis coefficients using a robust Singular Value Decomposition (SVD) approach that constraints the topology of the network to be sparse and hence reduces the number of required measurements for estimation. Our simulation results show the performance of the proposed approach for different basis functions.

In-silico evaluation of chemical interactions between active chemical constituents from ayurvedic medicinal plants and carrier protein ligands

Venkata kiran kumar melapu, Jerry A. Darsey

Modern medicine may be catering well to the needs of most of the people in developed and developing countries, but there are many tribal regions and remote places that are very far from the civilized world and hence far from access to modern medicine. People of those tribal parts have developed their own medicinal practices using their medicinal plant resources and succeeded to some extent in curing their common ailments. There is a great need to do a deep, consolidated and collaborative research on the active chemical constituents of the traditional medicinal plants. Evaluating the interactions of the chemical interactions of the chemical constituents of the medicinal plants and carrier proteins in humans may give interesting explanations to the anti-ailment activity of the medicinal plants. Furthermore, with the advent of various computational techniques to analyze the chemical compounds, we can predict the accuracy of the ligand protein relationships, their affinity and their orientation to each other very accurately. This information may further lead to novel drug discovery to the ailment from medicinal plants. This paper is an effort to bring together all the possible insilico methods of evaluation of all possible interactions between ligands and different human carrier proteins.

Uncovering Protein-Protein Interactions in Brassica napus Using Integrative Methods

Ning Zhang. Informatics Institute, Christopher S. Bond Life Science Center, University of Missouri, Columbia, MO, 65211, USA
Qiuming Yao. Department of Computer Science, Christopher S. Bond Life Science Center, University of Missouri, Columbia, MO, 65211, USA
Jay J. Thelen. Department of Biochemistry, Christopher S. Bond Life Science Center, University of Missouri, Columbia, MO, 65211, USA
Dong Xu. Department of Computer Science, Informatics Institute, Christopher S. Bond Life Science Center, University of Missouri, Columbia, MO, 65211, USA

Protein-protein interactions (PPIs) play essential roles in almost every biological process in a cell. Therefore, knowledge of PPI and protein interactome is a valuable source to understand biochemical and regulatory mechanisms of these processes. High-throughput technologies have generated huge amounts of PPI data. However, most studies focus on several well-studied model organisms, such as yeast and human. The challenge is how to utilize these PPI data for effective predictions of protein-protein interactions in a new organism. In this work, we applied integrative methods, combining sequence homology, gene co-expression, and phylogenetic profile to build a proteome-wide and high-confidence PPI network in rapeseed (*Brassica napus*) using public PPI data and other types of evidence. Due to the lack of genomic sequence and known PPIs in rapeseed, we also tested and evaluated our methods in *Arabidopsis thaliana* by comparing our prediction results with known interactions, gene ontology annotations and functional pathways.

Intergenic Region Analysis Pipeline for Bacteria (BIRAP): A tool for analyzing expression profile of intergenic regions generated by RNA-seq

Joseph S Reddy
Mark L Lawrence
Bindu Nanduri

Computational methods for gene expression analysis and variant calling using RNA-seq data are abundant. Recent re-annotation studies using the expression profile generated from RNA-seq has revealed shortcomings of automated methods and the necessity to validate existing annotation using experimental data. Expression profiling of intergenic regions helps in validating existing annotation as well as identifying novel functional elements. Hence there is a need for developing automated methods that would help achieve this goal. Also, tools for analysis of differential expression in unannotated regions are non-existent.

With an alignment map generated by BOWTIE and Samtools as input, we developed an automated pipeline in Perl to analyze intergenic regions. Our pipeline establishes a basal level of expression and identifies all intergenic regions expressed above a specified minimum length. It uses 'mpileup' output, which contains information on the number of fastq reads that align to each base in the genome (RPB), along with existing structural annotation of the corresponding genome, For each intergenic region identified, its location with respect to existing annotation is provided as GFF files. This also allows evaluation of the existing annotation using genome annotation tools such as Artemis.

For experiments involving multiple control/treatment samples, the pipeline normalizes expression across treatments and identifies differentially expressed intergenic regions. This not only helps evaluate existing annotation but also helps identify potential regulatory regions such as small non-coding RNAs known to play an active role in gene regulation. The pipeline will be tested on experimental data and results presented.

A “core genome” approach to decoding host-pathogen dual RNA-Seq data

Joseph S Reddy
Wei Wang
Adam Thrash
Andy D. Perkins
Mahalingam Ramkumar
Bindu Nanduri

Expression profiling, especially metatranscriptomics in host-pathogen scenario could help understand disease kinetics. A multi-factorial disease such as the bovine respiratory disease (BRD) in cattle involving a range of bacterial and viral pathogens along with environmental stress factors makes understanding the disease ecosystem a complex task. Using next-gen sequencing (NGS) approach for expression profiling of metatranscriptomic data of an infectious disease involving a host and multiple pathogens is a relatively novel concept and computational methods and pipelines for downstream analysis of such data is deficient. Such complex data from a dual RNA-seq experiment requires a read alignment approach that would provide a transcriptional landscape of the disease without having to look at expression profile of each individual pathogen.

Here, we propose a core genome approach to map reads to a composite genome comprising of common features across genomes of known bacterial pathogens associated with the disease. Orthologous regions across closely related bacterial genomes belonging to *Pasteurellaceae* associated with BRD are identified and artificial chromosomes built which would encompass features representing a core expression profile among these pathogens. Any unique regions are treated as subsets. Fastq reads generated from the dual RNA-seq experiment are then mapped to the host as well as these artificial chromosomes using traditional alignment algorithms such as BOWTIE. Expression profile generated from the composite genome is then traced back to their corresponding functional annotation using a trace back algorithm, providing an overall expression profile of the disease. The core genome approach is then evaluated against other alternatives.

Proteogenomic mapping of bovine respiratory disease pathogens

Joseph S. Reddy
Teresia Buza
Mahalingam Ramkumar
Mariola J Edelmann
James M Watt
Mark L Lawrence
Bindu Nanduri

Experimental structural annotation (identification and demarcation of functional elements within a genome) approaches validate gene models in newly sequenced genomes that are based on computational prediction. Transcriptome sequencing, tiling arrays and proteomics are expression based approaches for genome structural annotation. A major proportion of proteins described in the genome are often designated as 'hypothetical/predicted'. Confirming the expression of these proteins constitutes experimental validation of their existence and enhances the structural annotation. Here we describe proteogenomic annotation of three bacterial pathogens *Mannheimia haemolytica*, *Pasteurella multocida* and *Histophilus somni* that are causative agents of Bovine respiratory disease (BRD) in cattle. Total proteins were isolated from mid-log phase of growth in vitro from three replicate cultures. 2D LC ESI LCQ MS/MS analysis was conducted to identify proteins expressed in these three pathogens under experimental conditions. Mass spectra and tandem mass spectra were matched to the genome sequence translated in six reading frames, and peptides were identified using a randomized decoy database strategy. Peptides identified at 5% FDR were mapped on to the genome using the proteogenomic mapping tool at AgBase. The results were loaded onto Artemis genome annotation tool for further evaluation. We undertook a comparative genomics approach to validate gene models in these important veterinary pathogens.

PCA Based Analysis for Classification of Multiple Myeloma Microarray Data

RinkuSaha, Department of Information Science, UALR/UAMS Joint Graduate Bioinformatics Program, University of Arkansas, Little Rock, AR, 72204, USA
Songthip T Ounpraseuth, Department of Biostatistics, College of Public Health, University of Arkansas for Medical Sciences, Little Rock, AR, 72205, USA

DNA microarray technology considerably expedites the process of simultaneous discovery of the utility for thousands of genes. However the amount of data generated presents immense challenge for the biologists to carry out the analysis due to its high dimensionality. Dimension reduction methods are often employed at the beginning of data analysis and this aid in the selection of subset of genes with lower generalization error. Current work proposes a methodology based on PCA for gene subset selection which would help in efficient classification of normal healthy person samples from patient samples consisting of two genetic subtypes of multiple myeloma hyperdiploid(H-MM) and non-hyperdiploid(NH-MM). ANOVA was first performed on the Mayo Clinic multiple myeloma dataset obtained from NCBI Gene Expression Omnibus(GEO) followed by PCA. Highly correlated genes from each principal component explaining maximum variance were selected. Diagonal linear discriminant analysis and 3-KNN classification was then performed to obtain the overall misclassification error rate for the gene subset selected. An additional validation was done involving random gene selection to reduce dimensionality followed by classification. This helps in confirming that the variance explained by the reduced gene subset obtained from the proposed method is not random and that they carry more information that helps in efficient classification of the multiple myeloma dataset used in the study.

Supported by grants from NCRR (5P20RR016460-11) and NIGMS (8 P20 GM103429-11) at NIH

The Genome of Reniform Nematode, Rotylenchulus reniformis

Satish Ganji, Institute for Genomics, Biocomputing and Biotechnology, Mississippi State University
William S Sanders, Institute for Genomics, Biocomputing and Biotechnology, Mississippi State University
John Stokes, CVM Basic Science Department, Mississippi State University
Kurt Showmaker, Institute for Genomics, Biocomputing and Biotechnology, Mississippi State University
Ben Bartlett, Institute for Genomics, Biocomputing and Biotechnology, Mississippi State University
Hui Wang, Institute for Genomics, Biocomputing and Biotechnology, Mississippi State University
Martin Wubben, USDA-ARS, Mississippi State
Fiona McCarthy, Veterinary Science and Microbiology, University of Arizona
Zenaida Magbanua, Institute for Genomics, Biocomputing and Biotechnology, Mississippi State University
Daniel Peterson, Institute for Genomics, Biocomputing and Biotechnology, Mississippi State University

The reniform nematode (RN; Rotylenchulus reniformis) is a pest that causes considerable damage to cotton. For example, in 2011 yield losses of approximately 279,000 bales (total estimated value > $90 million) were attributed to RN damage. Ostensibly, sequencing the genome of R. reniformis represents a key step in identifying genes underlying RN's ability to infect host plants. Ultimately, knowledge of the RN genome may suggest means of minimizing R. reniformis damage through targeted disruption of RN-specific gene pathways. Towards this end, we determined the RN genome size and initiated whole genome sequencing of the nematode. Our flow cytometric analysis indicates that the genome size of R.reniformis to be almost twice the size of the genome of C. elegans (?100 Mb) and 3-4 times the size of the root-knot nematode (Meloidogyne incognita) genome (?50 Mb). Reniform nematode genome sequencing was performed using Illumina and Roche 454 technologies, and sequence reads were assembled using Newbler and ABySS. Here we present the current status of the R.reniformis genome sequencing project and discuss the present state of our draft assembly and its annotation.

Probabilistic Methods for Accurate Mapping of Metatranscriptomic Sequence Data

Wei Wang, Department of Computer Science and Engineering,
Mississippi State University
Adam Thrash, Department of Computer Science and Engineering,
Mississippi State University
Dilip Gautam, Department of Computer Science and Engineering,
Mississippi State University
Joseph Reddy, Department of Basic Science, College of Veterinary Medicine,
Mississippi State University
Arunkanth Ankala, Department of Human Genetics, Emory University
Tad Sonstegard, Beltsville Agricultural Research Center, US Department of Agriculture
Steven Schroeder, Beltsville Agricultural Research Center, US Department of Agriculture
Jeff Wilkinson
Mahalingam Ramkumar, Department of Computer Science and Engineering,
Mississippi State University
Bindu Nanduri, Department of Basic Science, College of Veterinary Medicine,
Mississippi State University
Andy D. Perkins, Department of Computer Science and Engineering,
Mississippi State University

High throughput sequencing technologies have allowed the generation of sequence data faster and at higher coverage than ever before. These methods generate millions of sequence reads, which must often be mapped to a reference genome sequence. When applied to metagenomic or metatranscriptomic studies, such as those involving a host-pathogen interaction, assignment of sequences to the proper genome becomes a challenge. The methods described here use a probabilistic approach to assign individual sequence reads to an appropriate genome. The focus of this project is to fully incorporate transcript length normalization into these probabilistic methods and handle the case in which a sequence read may map to multiple locations in one or more genomes. Both real and synthetic test cases have been identified or developed for validation of the probabilistic methods.

A Distributed CPU-GPU Framework for Pairwise Alignments on Large-Scale Sequence Datasets

Da Li - Dept. of Electrical and Computer Engineering - University of Missouri
Kittisak Sajjapongse - Dept. of Electrical and Computer Engineering – University of Missouri
Huan Truong - MU Informatics Institute - University of Missouri
Gavin Conant - MU Informatics Institute, Division of Animal Sciences – University of Missouri
Michela Becchi - Dept. of Electrical and Computer Engineering, MU Informatics Institute - University of Missouri

Several problems in computational biology require the all-against-all pairwise comparisons of tens of thousands of individual biological sequences. Each such comparison can be performed with the well-known Needleman-Wunsch alignment algorithm. However, with the rapid growth of biological databases, performing all possible comparisons with this algorithm in serial becomes extremely time-consuming. The massive computational power of graphics processing units (GPUs) makes them an appealing choice for accelerating these computations. As such, CPU-GPU clusters can enable all-against-all comparisons on large datasets. We present a hybrid MPI-CUDA framework for computing multiple pairwise sequence alignments on CPU-GPU clusters. Our design targets both homogeneous and heterogeneous clusters with nodes characterized by different hardware and computing capabilities. Our framework consists of the following components: a cluster-level dispatcher, a set of node-level GPU-dispatchers, and a set of CPU- and GPU-workers. The cluster-level dispatcher progressively distributes work to the compute nodes and aggregates the results. The node-level GPU dispatchers distribute alignment tasks to available GPUs and perform dual-buffering to hide data transfers between CPU and GPU. CPU- and GPU-workers perform pairwise sequence alignments using the Needleman-Wunsch algorithm. We propose and evaluate three designs for these GPU workers, all of them outperforming the existing open-source implementation from the Rodinia Benchmark Suite.

Performance of Hadoop Based REMD on Cloud Computing Platform

Jin Niu, Department of Computer Science, Southern University
Alvin Allen, Department of Computer Science, Southern University
Ebrahim Khosravi, Department of Computer Science, Southern University
Seung-Jong Park, Department of Computer Science, Louisiana State University
Shuju Bai, Department of Computer Science, Southern University

Recently, cloud computing has emerged as a prevalent computing paradigm with the advantage of virtualization, scalability, fault tolerance, and a usage based pricing model. It has been employed by scientists in various application areas. Replica Exchange Molecular Dynamics (REMD), a sampling method in molecular dynamics simulation, has been proved to be a powerful technique which can improve the sampling of the potential energy during the process of Molecular Dynamics. Due to the independency of each replica in REMD, it is suitable to perform REMD as a cloud computing application.

In our work, we implemented Hadoop based REMD on both cloud computing and parallel computing platforms. Time costs and data transfer rates were measured for comparing the performance of REMD on the two platforms. The scalability and fault tolerance of cloud computing were also verified in our experiment. We adopted LONI as the parallel computing platform and VCL in Southern University as the cloud computing platform. Software for molecular dynamics simulation is CHARMM (Vc35b3). To achieve the data parallelism, Hadoop framework was deployed to realize the REMD algorithm.

Our experiments showed that REMD algorithm can be implemented on cloud computing without performance degradation. Moreover, cloud computing platform demonstrated the characteristics of virtualization, scalability and fault tolerance, providing a dependable, adaptable, and robust environment for REMD. Our work will broad the user group of REMD due to the popularity of cloud computing.

Transcriptional and epigenetic variation: two integrative ways to explain the soybean innate immunity triggered by PAMPs

Saad Khan, Informatics Institute and C.S. Bond Life Sciences Center, University of Missouri-Columbia
Oswaldo Valdes-Lopez,C.S. Bond Life Sciences Center, University of Missouri-Columbia
Robert J Schmitz, The Salk Institute for Biological Studies, La Jolla, CA
Trupti Joshi, Informatics Institute and C.S. Bond Life Sciences Center, University of Missouri-Columbia
Joseph Ecker, The Salk Institute for Biological Studies, La Jolla, CA
Dong Xu, Informatics Institute and C.S. Bond Life Sciences Center, University of Missouri-Columbia
Gary Stacey, C.S. Bond Life Sciences Center, University of Missouri-Columbia

Pathogen-associated molecular patterns, or PAMPs, are molecules associated with groups of pathogens that are recognized by cells of the innate immune system. They play an important role in activation of immune response by identifying some conserved non-self molecules present on the cellular membranes of the pathogens (e.g. bacteria etc.). The immunity response resulting from the identification of these molecular patterns by plants is called PAMP/MAMP triggered immunity (MTI). It has been shown from earlier studies that MTI is heritable. Both genetic as well as epigenetic factors have been linked in the past with heritability of MTI in other plants. Based on these assumptions transcriptomics and epigenomics studies were conducted on two soybean genotypes with contrasting PAMP-triggered immunity as well as two of their F4 progenies. In this study we analyzed differences in gene expression and studied the patterns and heritability of methylation variants in a complex genetic population over multiple generations. This study has facilitated in differentiating between Mendelian and non-mendelian patterns of inheritance of methylation thus paving the way for understanding how methylation variants contribute to phenotypic variation.

Analysis of correlation between b and y ion in tandem mass spectra library

Mihir Jaiswal: University of Arkansas at Little Rock and University of Arkansas for Medical Sciences Bioinformatics Graduate Program, Little Rock, AR, USA
Boris Zybaylov (corresponding author: BLZybaylov@uams.edu): Department of Biochemistry and Molecular Biology, University of Arkansas for Medical Sciences, Little Rock, AR

Tandem mass spectrometry (MS/MS) is a key technology used in identification of peptides and proteins. Collision induced dissociation (CID) is a common peptide fragmentation technique used in MS/MS experiments. The CID fragmentation of a positively charged precursor ion occurs predominately at peptide bonds and produces b-ions(N-terminal fragments) and y-ions (C-terminal fragments). In this work we test the following hypothesis: “Detection of a particular b ion with the charge less than that of a precursor implies high probability of detection of the complementary y-ion with remaining charge”. Number of b-ions and complementary y-ions were calculated for each precursor spectrum from NIST Libraries of Peptide Tandem Mass Spectra and converted to a ratio (y-b ratio). The higher ratio would indicate that the underlying hypothesis is correct. The dependency of the ratio on charge, sequence length, number of peaks detected, and total number of b and y ions was studied. The ratio was found to be significant, and overall supportive of the hypothesis. We also found that the ratio was inversely correlated with the charge of the precursor. No significant correlation with other parameters was found. The complementary hypothesis was also investigated - “Detection of a particular y ion with the charge less than that of a precursor implies high probability of detection of the complementary b-ion with remaining charge”.

Protein interaction binding site prediction for comparative models by combining sequential and structural properties

Nan Zhao
Andi Dhroso
Zhiquan He
Chi-Ren Shyu
Dong Xu
Dmitry Korkin
University of Missouri, Columbia MO

During the last decade, both sequence-based and structure-based approaches to predict protein binding site have been developed. The sequence-based methods can be applied to virtually any protein sequence. However they are less accurate than structure-based methods. The current structure-based methods are designed exclusively for the experimentally obtained protein structures, thus having their limitations in coverage. The goal of this work is to bridge the coverage gap by developing a new hybrid approach for detecting binding sites in comparative models.

Here, we propose a sequence based protein binding site prediction approach that utilizes structure based methods' benefits. We utilize L1-regularized logistic regression to integrate sequence- and structure-based predictions for comparative models. The method relies on a series of features, including evaluation of comparative models, geometric features, solvent accessibility, hydrophobicity, secondary structure based on comparative models and name of residues. The non-redundant dataset of feature vectors for training and testing is automatically generated from the hetero-oligomer structures. The assessment of our binding site prediction strategy has demonstrated that it is able to use protein sequences as the only input and obtain comparable accuracies to the state-of-art structure based predictors across different quality levels of homology models. Our method could be useful in the large-scale functional annotation of proteins whose structures are represented only by the comparative models.

Transgenerational Effects of Chronic Low-Dose Radiation Exposure in Medaka Fish model

Oleksandr Moskalenko
Andrey Ptitsyn
Dmytro Grygoryev
John Zimbrick

Radiation is known to cause heritable mutations since the dawn of Genetics. However, mechanistic understanding of the low-dose radiation remains elusive. Medaka fish (*Oryzias latipes*) offers a convenient model for the whole-body transgenerational exposure testing. Using custom-designed microarrays we have studied gene expression patterns in three consecutive generations of Medaka fish continuously exposed to low-dose radiation. The experiment mimics natural conditions following radioactive contamination.

We have discovered consistent patterns of gene expression repeating through all three generations in the experiment. The signature is defined not by a large change of abundance of a finite number of biomarker genes, but rather by the set of molecular functions. In particular, the patterns of differential gene expression relative to first-exposed parental generation controls concern intracellular organelles, organelle and nuclear lumen, protein binding, ion binding, and general regulation of metabolic processes. The biological pathways behind the signature groups of genes are dominated by chemical perception, taste and smell sensation, and related signaling systems. Our observations are consistent with the hypothesis that low-dose ionizing radiation modulates chemosensory system. On the other hand, the chemosensory system is also under highest selective pressure under the conditions of a chronic radiation exposure. Analysis of transgenerational patterns of gene expression provides an important glimpse into the genes most likely affected in several stages of organism adaptation to environmental stresses.

A language-based approach to mining host-pathogen interactions from biomedical literature

Thanh Thieu
Samantha Warren
Joi Moore
Dmitry Korkin

In an infectious disease, the pathogen's strategy to enter the host organism and breach its immune defenses often involves interactions between the host and pathogen proteins. Currently, the experimental data on host-pathogen interactions (HPIs) are scattered across multiple databases, which are often specialized to target a specific disease. Collections of biomedical literature, such as PubMed, are primary sources for such data. Unfortunately, a common approach to mine the information on host-pathogen interactions by a manual query using keyword searches is time-consuming and unfeasible for a whole-PubMed mining of abstracts.

This work presents a language-based literature mining method that automatically detects if the title and abstract of an article contain host-pathogen interaction data as well as extracts the information about organisms and proteins involved in the interaction. The method first employs BANNER and SR4GN to extract named entities (e.g. proteins, organisms) and their relationships. After that, Link Grammar is used to analyze grammatical structure of each sentence. Finally, a set of semantic patterns inferred from training examples are utilized to link components of host-pathogen interactions.

In addition to identifying information on host-pathogen interaction, the method is also designed to spot uncertainty (*e.g.*, a hypothetical interaction). The method has been trained and tested on a high-quality manually curated dataset. The accuracy of 71% to 76% is achieved.

TOPOLOGICAL NETWORK ALIGNMENT BASED ON GRAPHLET DEGREE SIGNATURE

Si Li, School of Computing, The University of Southern Mississippi
Jonathan Z. Sun, School of Computing, The University of Southern Mississippi
Chaoyang Zhang, School of Computing, The University of Southern Mississippi

Thanks to the advancement of genome sequencing and bioinformatics techniques, tremendous data of large biological networks are now available for study, such as PPI networks and gene regulatory networks. Comparing these networks via network alignment is a key step to understand evolutionary changes, biological mechanisms and the complexity of diseases. Kuchaiev et al. recently developed a topological method of network alignment, the GRAAL algorithm, based on grapnhlet degree signatures. Their work was followed by multiple research groups resulting in a variety of adaptations. Yet aligning practical biological networks requests further improvements to enhance computational efficiency.

In this paper, we design three modifications to the basic GRAAL, the P, G and R modifications, in order to improve the algorithm's efficiency. We applied the modifications to datasets of real PPI networks and their synthetic counterparts which are made by adding random noises. Comparing to GraphCrunch2, the standard implementation of GRAAL, our experiments show that the combined use of these modifications can save up to 90% of computation time without losing too much accuracy. We consider this as a forward step to approach the alignment of practical-size biological networks, as well as an inspiration to the design of new global network alignment algorithms.

NeedleFinder, a Server for Binning and Statistical Analysis of Mass Spectra Data

Heng Luo, University of Arkansas at Little Rock/University of Arkansas for Medical Sciences joint Bioinformatics Program
Stephen Embry, University of Arkansas at Little Rock/University of Arkansas for Medical Sciences joint Bioinformatics Program
Stephen C. Grace, Department of Biology, University of Arkansas at Little Rock

Mass spectrometry is a technology to analyze what and how much are the components inside a biological sample. The mass spectra data produced by mass spectrometry can be utilized for metabolite identification, pathway analysis and even disease diagnosis. However, it may be a tough task to parse mass spectra data and compare across them to efficiently identify the difference. Here we introduce NeedleFinder, a web-based comprehensive analytical tool for metabolite mass spectra data, specifically built around LC-MS (Liquid chromatography–mass spectrometry) data, that allows the user to simply upload their NetCDF files and do a variety of analysis including generating total ion current chromatogram (TIC), base peak chromatogram (BPC), extracted ion chromatogram (EIC) and principal components analysis (PCA) with simply a few clicks. In additional, this server can generate a specific “barcode” by binning mass spectra data according to user input mass range into bins, which contain the total intensity for each range, so that the “barcode” is assigned to the corresponding sample with a certain condition or phenotype. By comparing the “barcodes”, users can easily locate the difference across samples and diagnose the inborn change on mechanism level. This program was originally written in Perl and is now online accessible by implementation of PHP, Perl, MySQL and R with a mission queuing and control system to maximally optimize the server performance. We are continuing adding functions to this server to make it a comprehensive analytic suite from taking raw mass spectra data to final decision-making results.

Coherence in Evolution: An Influenza Story

Gavin Conant, University of Missouri - Columbia
Dmitry Korkin, University of Missouri - Columbia

Historically, pandemic subtypes of Influenza A have posed serious health and economic stress worldwide. In 2009, the H1N1 subtype resulted in significant economic loss and 61,000,000 infections. The more recent interest, however, has been placed on subtype H3N2 and deadly subtype H5N1. Due to the lack of a universal vaccine or reliable treatment these common viruses still pose a serious risk. To fully understand these pandemic viruses, patterns of their evolutionary dynamics must first be understood. This is challenging task due to the remarkable ability of Influenza to reassort its genome. Here, we study the evolutionary patterns of different influenza subtypes by determining highly conserved and highly diverse regions using structural bioinformatics, unsupervised learning, and Metropolis Monte Carlo simulation.

The standard way of studying diversity is using Shannon entropy, however, we have chosen an expanded approach which takes into account entropy, substitution similarity, and percentage of gaps to find the diversity of each residue. To determine the regions, or patches, that are considered diverse, our method uses Metropolis Monte Carlo to sample the structural patches surrounding the most diverse residues on the surfaces of influenza proteins. Additionally, our method finds regions of 100% sequence conservation using a surface distance based clustering. The method is implemented as a computational pipeline and can be applied to find a coherent pattern of evolution across the protein surface of any virus. The obtained results and the computational pipeline may be useful in finding better treatments and a universal vaccine.

New method for efficiently sampling protein 3D conformations

Jingfen Zhang, Department of Computer Science and Christopher S. Bond Life Sciences Center, University of Missouri
Hongbo Li, Department of Computer Science and Christopher S. Bond Life Sciences Center, University of Missouri
Zhiquan He, Department of Computer Science and Christopher S. Bond Life Sciences Center, University of Missouri
Jing Maria Zhang, Department of Statistics, Yale University
Dong Xu, Department of Computer Science and Christopher S. Bond Life Sciences Center, University of Missouri

Protein tertiary structures are essential for studying functions of proteins at the molecular level. Computational prediction approach is indispensable since experimental methods are time consuming and expensive. One paradigm of computational approach is to sample conformations and then select the best one as predicted structure. However, the existing techniques such as Molecular Dynamics and Monte-Carlo simulation are often too expensive to sample broadly distributed conformations across the free energy landscape. In this paper, we propose a new method to efficiently explore possible structural conformations of a protein. Different from the free energy guided sampling, we obtain various distance constraints (for example, the distance constraints retrieved from target/pdb-template alignments or high-quality models) and sample the confirmations that are consistent with these constraints. At first, we decompose the target sequence to a small number of contiguous segments such that the constraints within segments are consistent (or can be clustered into limited number of groups) while the constraints between segments are diverse. Then, for each segment, we use the representatives of the clusters as the constraints for further sampling. And we also cluster the constraints between different segments and get the representatives. Thus, the conformation sampling is a process to combine the representatives of segments and representatives of inter-segments to build pair-wise distance constraints, and then to generate the structural models that satisfy the corresponding distance constraints by Multi-dimensional scaling techniques. Our contribution is to detect the conserved relationship among the constraints and sample the conformation in a much smaller, but more targeted conformation space.

Unexpected evolutionary recursive patterns in influenza A proteins

Xiaomin Wang, Department of Computer Science
Samantha Warran, Department of Computer Science
Dmitry Korkin, Informatics Institute, Department of Computer Science, Bond Life Sciences Center

Since their first documented pandemic outbreak in 1918, Influenza A outbreaks present one of the most critical health threats world-wide. Recently, an intriguing phenomenon was reported: survivors of the 1918 Spanish Flu were also immune to the 2009 Swine Flu.

This resulted in a structural biology study of the antigenic sites of an H1N1 surface protein, HA, showing a "recursive" pattern of the binding site mutations through the course of its evolution. Unfortunately, due to the significant time and efforts required, this experimentally-driven study only looked at a single functionally important region of only one out of ten H1N1 proteins from one of out of ten influenza A subtypes affecting human.

Here, we developed a comprehensive computational approach that integrates structural bioinformatics and dynamic programming to search for such recursive patterns existing on the surfaces of HA, NA, and NS1 proteins, across multiple subtypes of multiple species. Our results not only replicated the experimentally obtained pattern, but discovered novel previously unaccounted for patterns, suggesting that this new phenomenon expands beyond a single functional site of a single influenza subtype.

PHDcleav: A SVM-based method for predicting human Dicer cleavage sites using sequence and secondary structure of miRNA precursors

Firoz Ahmed[1], Rakesh Kaundal[2] and Gajendra PS Raghava[3*]

[1]Plant Biology Division, The Samuel Roberts Noble Foundation, Ardmore (OK), USA; [2]National Institute for Microbial Forensics & Food and Agricultural Biosecurity (NIMFFAB), Department of Biochemistry & Molecular Biology, Oklahoma State University, Stillwater (OK), USA; [3]Bioinformatics Centre, Institute of Microbial Technology, Sector 39-A, Chandigarh, India.

Abstract:

Dicer, an RNase III enzyme, plays a vital role in processing pre-miRNAs to generate miRNAs. Prediction of Dicer cleavage sites accurately is very important for identification of bona fide miRNAs in the human genome. In this study, a novel method has been developed to predict Dicer cleavage sites at 5p arm of pre-miRNAs using support vector machine. We used the dataset of experimentally validated human miRNA hairpins from miRBase (version 13) to extract fourteen nucleotides around Dicer cleavage sites. We developed number of models using various types of features and achieved maximum accuracy of 66% using binary profile of nucleotide sequence. The performance of prediction of Dicer cleavage site improved significantly from 66% to 86% when we integrated secondary structure information. This indicates that secondary structure plays an important role in form cleavage site. All models were trained and tested on 555 experimentally validated cleavage sites and evaluated using five-fold cross validation technique. In addition performance was evaluated on an independent dataset and achieved an accuracy of ~82%. Based on this study, we developed a webserver PHDcleav (http://www.imtech.res.in/raghava/phdcleav/). The webserver will help to understand the effect of genomic variations/SNPs in miRNA loci on Dicer cleavage sites and ultimately to mature mRNA formation. Moreover, it will also be useful to design artificial pre-miRNAs for potent gene silencing.

Keywords: pre-miRNA; Dicer cleavage site; Support Vector Machines; Nucleotide composition; Binary pattern; Accuracy.

A Comparative Study of Linear and Nonlinear Dimensionality Reduction Methods Using Gene Expression Data

Christopher Ma,University of Mississippi
Yixin Chen ,University of Mississippi
Dawn Wilkins, University of Mississippi

The “curse of dimensionality” has often arisen when analyzing biological data. Principal Component Analysis (PCA) has been widely used to decouple a set of correlated variables via a linear transformation of the variables. The transformation is defined by a set of principal directions, each of which is computed through maximizing the variance of the samples along that direction. The coordinates along each principal direction are called principal components. When the data cloud has an intrinsically linear structure, a small number of principal components are usually sufficient to account for most of the variance in the data. However, PCA is less effective in handling nonlinear data. Recent advance in restricted Boltzmann Machine provides a nonlinear dimensionality reduction framework. It models the structure of high dimensional data through a network that links stochastic binary input with hidden variables using symmetrically weighted connections. When the training process converges, the hidden variables generate a new representation of the data. Dimensionality reduction is achieved by controlling the number of hidden variables. In this work, we perform a systematic comparative study on dimensionality reduction using PCA and restricted Boltzmann Machine on gene expression data. Support vector machines and random forests are used to compare the discriminative power of the reduced feature space.

Computational analysis of 2D protein gel images for identification of differentially expressed proteins associated with resistance to aflatoxin accumulation in maize

Alka Tiwari Department of Biochemistry, Molecular Biology, Entomology and Plant Pathology, Mississippi State University, Mississippi State, MS 39762.
W. Paul Williams USDA-ARS, Corn Host Plant Resistance Research Unit, Mississippi State, MS 39762.
J. Erik Mylroie USDA-ARS, Corn Host Plant Resistance Research Unit, Mississippi State, MS 39762.
Gary L. Windham USDA-ARS, Corn Host Plant Resistance Research Unit, Mississippi State, MS 39762.
Ashli Brown Department of Biochemistry, Molecular Biology, Entomology and Plant Pathology, Mississippi State University, Mississippi State, MS 39762.
Xueyan Shan Department of Biochemistry, Molecular Biology, Entomology and Plant Pathology, Mississippi State University, Mississippi State, MS 39762.

Two-dimensional protein gel electrophoresis is a powerful tool for revealing differences in proteomic profiles expressed in tissues of interest under different treatments. Differentially expressed proteins across multiple samples can be identified and quantitatively evaluated by computational analysis of sets of 2D protein gel images. The objectives of this study were: 1) evaluation of different computational image processing methods (Matlab Image Processing Toolbox, Pinnacle, and RegStatGel) for the efficiency in quantification of protein expression levels, and 2) identification of differentially expressed maize proteins associated with resistance to *Aspergillus flavus*. infection and aflatoxin accumulation through the quantitative data analysis of 2D gel images. Two resistant (Mp715 and Mp719) and two susceptible (Va35 and Mp04:87) maize inbred lines were selected for this study. Developing kernels were collected from the primary ears of corn plants at 14 days after inoculation with *A. flavus*. Proteins were extracted using TCA/acetone precipitation in combination of a phenol extraction step. The 2-D protein gel electrophoresis was performed using PROTEAN IEF Cell (Bio-Rad) and PROTEAN II XL cell (Bio-Rad). Proteins were visualized with florescent dye Oriole (Bio-Rad). Gel images were obtained with an Alpha imager. We established a working protocol for quantitative proteomic studies of maize proteins by computational and statistical analysis of 2D protein gel images. It is a novel analysis procedure for profiling and comparative investigations of differentially expressed proteins which can be used to facilitate the development of DNA markers for maize resistance breeding.

Classification and Feature Selection Using Hybrid Top Pairs on Microarray Data

Tian Gui, University of Mississippi
Xiaofei Nan, University of Mississippi
Dawn E. Wilkins, University of Mississippi
Yixin Chen, University of Mississippi

Gene expression classification and feature selection are the most significant ways to diagnose diseases using microarray technology. Machine learning has transformed microarray technology into a clinically convenient tool. Continuing our previous work on classification using top scoring feature pairs, we developed a new approach for improving classification accuracy of microarray data using Hybrid Top Pairs (HTP) method, which combines top scoring pairs with original features. Top Scoring Pairs (TSPs) are sets of paired genes that can be employed to accurately classify instances into one of the given classes. In gene expression studies, the results of using a k-TSP classifier (a classifier built using the k best pairs) are promising. The two major advantages of applying the k-TSP algorithm are simplicity and interpretability – each TSP is calculated on one pair of genes and the classifier is based on the relative expression of the gene pair. The HTP methodology aims to extend the abilities of the original TSP technique by augmenting the original features with TSP attributes. To illustrate the effectiveness of HTP, we applied this method to three different sets of microarray data involving human cancer. Results from these three experimental datasets have shown using the HTP method is competitive with the original TSP technique. The HTP methodology used in this study provides a simple, yet powerful, technique for gene selection. The methodology provides a viable alternative, especially when applied to small sample sizes and when interpretability is important.

Top-Down Transversing the Gene Ontology to Extract Biological Knowledge for Biomedical Models

Teresia Buza, College of Veterinary Medicine; Institute for Genomics, Biocomputing and Biotechnology, Mississippi State University
Tony Arick, Institute for Genomics, Biocomputing and Biotechnology, Mississippi State University
Cathy Gresham, Institute for Genomics, Biocomputing and Biotechnology, Mississippi State University
Fiona McCarthy, Department of Veterinary Science and Microbiology, University of Arizona
Bindu Nanduri, College of Veterinary Medicine; Institute for Genomics, Biocomputing and Biotechnology, Mississippi State University

Modern molecular biology research is moving from high-input, low-output to low-input, high-throughput 'omics data generation. Simultaneously biologists face the challenge of converting this data into high-level biological knowledge. Due to limited resources, few model genomes are comprehensively annotated to serve as references, especially for studying human diseases. Limited resources hinder development of new animal models for human diseases. Developing a highly curated, cross-referenced resource that links key features of human disease causing genes to non-model animal genes and provision of computational methods for mining this resource provides opportunities to understand human-non-model animal disease relationships. Gene Ontology (GO) terms are neutral key features that provide functional information about relationships of disease genes. GO terms are organized in Directed Acyclic Graphs (DAG), in which a term can be a child of one or more parents in a hierarchical relationship. The terms provide more general to most specific biological knowledge when transversing hierarchically from root to leaf. A gene product may be annotated to different levels in the DAG, making it difficult for users to pinpoint the most specific function of a gene. Here, we develop a tool that integrates the most specific experimentally verified GO terms for human disease causing genes, their curated pathways, phenotypes and genetic disorders with orthologous genes in cow, a non-model animal. This initial work provides a foundation for developing a highly curated, cross-referenced resource for studying human diseases in non-model animals

Identification of protein coding genes in the plant-parasitic nematode Rotylenchulus reniformis through comparative proteogenomic mapping with Caenorhabditis elegans

William S. Sanders - Institute for Genomics, Biocomputing, and Biotechnology, Mississippi State University, MS, USA
Satish Ganji - Institute for Genomics, Biocomputing, and Biotechnology, Mississippi State University, MS, USA
Jennifer P. Arnold - Institute for Genomics, Biocomputing, and Biotechnology, Mississippi State University, MS, USA
Mark A. Arick II - Institute for Genomics, Biocomputing, and Biotechnology, Mississippi State University, MS, USA
Kurt C. Showmaker - Institute for Genomics, Biocomputing, and Biotechnology, Mississippi State University, MS, USA
Martin J. Wubben - United States Department of Agriculture – Agricultural Research Service, Crop Science Research Laboratory, Mississippi State, MS, USA
Daniel G. Peterson - Institute for Genomics, Biocomputing, and Biotechnology, Mississippi State University, MS, USA

Peptides identified through high-throughput mass spectrometry can be utilized to complement traditional structural genome annotation methods through proteogenomic mapping and provide experimental evidence that a given gene is being transcribed and translated into a protein. In model organisms, proteogenomic mapping can aid researchers by identifying novel exons and aid in the identification of intron-exon boundary correction and discovery. In non-model organisms, the peptides identified through mass spectrometry can be mapped back to the genome sequence of a model organism, providing insight into genes conserved in the non-model species. *Rotylenchulus reniformis* is a plant parasitic nematode affecting US cotton production, and causes an estimated annual $130,000,000 USD crop loss. Proteins were isolated from *R. reniformis* eggs and subjected to proteolytic digestion and analysis through mass spectrometry. Genomic *R. reniformis* DNA was isolated and sequences were generated using two platforms, combining both Illumina Sequencing-by-Synthesis and Roche 454 Pyrosequencing technologies. Through the proteogenomic mapping of these peptides to the proteome and genome of the model nematode species, *Caenorhabditis elegans*, and the use of the R. reniformis genome sequence data, we have identified a set of gene sequences to serve as a validation set for our ongoing efforts to sequence and assembly the genome of *Rotylenchulus reniformis*.

Circles within circles: Crosstalk between protein Ser/Thr/Tyr-phosphorylation and methionine oxidation

R. Shyama Prasad Rao, University of Missouri
Dong Xu, University of Missouri
Jay J. Thelen, University of Missouri
Ján A. Miernyk, University of Missouri

Reversible posttranslational protein modifications such as Ser/Thr/Tyr phosphorylation and Met oxidation are critical for both metabolic regulation and cellular signaling. Typically these modifications are studied separately. Herein we describe the potential for cross-talk between these modifications. The proximity of Met to Ser/Thr/Tyr within the proteome had not previously been addressed. In order to consider the possibility of a generalized interaction, we performed a trans-kingdom sequence analysis of known phosphorylation sites in proteins from bacteria, fungi, plants, and animals. The proportion of phosphorylation sites that include a Met within a 13-residue window centered upon Ser/Thr/Tyr is significantly less than the occurrence of Met in proximity to all Ser/Thr/Tyr residues. Met residues are present at all positions (-6 to +6, inclusive) of the 13-residue window that we have considered. Detailed analysis of sequences from eight disparate plant taxa revealed that only a small proportion of phosphorylation sites include a Met residue conserved at any specific position. The GO enrichment analysis indicated that the phosphorylation and Met oxidation crosstalk might be prevalent in kinases and proteins involved in signaling. The large proportion of known phosphorylation sites with Met in the proximity fulfills the necessary (but may not be sufficient) condition for cross-talk. These proteins/sites might be candidates for cross-talk between oxidative signaling and reversible phosphorylation.

MOLECULAR MODELING OF THE INHIBITORY EFFECTS HUMAN CYTOCHROME CYP3A4 BY DILLAPIOL DERIVATIVES

E. Nathalie Pineda,Chemistry Department, University of Arkansas at Little Rock and Dep. Pharm. Sci. University of Arkansas for Medical Sciences
A. Francis Carballo, Chemistry Department, Ottawa University, Ontario, Canada
S. Lui, Biology Department, Ottawa University, Ontario, Canada
R. Lilin, Biology Department, Ottawa University, Ontario, Canada
S. Gonzalez, Titulacion de Ingenieria Quimica, Universidad Tecnica Particular de Loja
J. Thor Arnason, Biology Department, Ottawa University, Ontario, Canada
Tony Durst, Chemistry Department, Ottawa University, Ontario, Canada
Darin Jones, Chemistry Department, University Of Arkansas at Little Rock
Cesar M. Compadre, Pharmaceutical Sciences, University of Arkansas for Medical Sciences

CYP3A4 is involved in the oxidation of many drugs and toxins in the intestine and liver, and its inhibition may cause pharmacokinetic interactions by altering the metabolism of drugs with serious clinical implications. In this research, enzymatic assays were used to determine the inhibitory effect of some previously isolated and synthetized dillapiol derivatives against human cytochrome CYP3A4. Comparative Molecular Field Analysis (CoMFA) as implemented in the program SYBYL was used to investigate the structural factors involved in the inhibition of CYP3A4. For the analysis the structures were optimized and the electrostatic charges were calculated using DFT with the B3LYP approximation, and the 6-311G basis set with the program GAUSSIAN 03. This model was used to understand the structural factors that determine the degree of inhibition of CYP 450 3A4 that this type of compounds may produce.

Combining Next-generation Sequencing and Comparative Genomics to Identify Novel microRNA.

Darren E. Hagen, University of Missouri, Division of Animal Sciences
Christine G. Elsik, University of Missouri, Division of Animal Sciences, Division of Plant Sciences, MU Informatics Institute

Given the importance of microRNAs (miRNAs) in post-transcriptional gene regulation, it is likely that they play key roles in behavior adaptations, such as those exhibited in social insects, including ant and bee species, which are members of the order Hymenoptera. Small RNA molecules such as miRNAs pose a challenging computational problem. Bioinformatic algorithms struggle to identify these elements due to the reduced information of short sequences. Next generation sequencing technologies have identified novel miRNAs in many organisms, but short sequences often result in spurious mapping and inaccurate predictions. Furthermore, the analysis pipelines traditionally used fail to identify rarely-expressed miRNAs, whose signal is often drowned out by miRNAs with higher expression. Predicting which protein-coding genes are targeted by a particular miRNA is even more difficult, especially in animals. Mismatches, gaps, and incomplete matching found in known miRNA-messenger RNA (mRNA) interactions result in a tremendous amount of false positives.

The availability of nine sequenced hymenopteran insect genomes has provided an unprecedented opportunity to identify hymenopteran miRNAs. Here we present our approaches based on next generation sequencing technologies combined with comparative genomics to identify conserved miRNAs, including the more rare transcripts, and to identify putative miRNA targets using orthology of both miRNA and protein-coding genes (the potential targets).. In our analysis of seven ant species, honeybee, and parasitoid jewel wasp, we discovered 27 previously unidentified miRNA. Analysis of 3'UTR sequences resulted in improved target predictions for both conserved and novel hymenopteran miRNAs.

Quantitative RT-PCR data analysis of RNA transport pathway genes associated with resistance to aflatoxin accumulation in maize

Matthew Asters, Department of Biochemistry, Molecular Biology, Entomology and Plant Pathology, Mississippi State University, Mississippi State, MS 39762;
W. Paul Williams, USDA-ARS, Corn Host Plant Resistance Research Unit, Mississippi State, MS 39762;
Andy Perkins, Department of Computer Science and Engineering, Mississippi State University, Mississippi State, MS 39762.
J. Erik Mylroie, USDA-ARS, Corn Host Plant Resistance Research Unit, Mississippi State, MS 39762;
Gary L. Windham, USDA-ARS, Corn Host Plant Resistance Research Unit, Mississippi State, MS 39762;
Xueyan Shan,Department of Biochemistry, Molecular Biology, Entomology and Plant Pathology, Mississippi State University, Mississippi State, MS 39762;

Aspergillus flavus is a pathogenic fungus to seed crops that produces aflatoxins, which are carcinogenic food contaminants that result in significant economic losses in corn production. This study focused on identifying maize genes associated with resistance factors that inhibit *A. flavus* infection and aflatoxin contamination. This research focuses on the qRT-PCR analysis of RNA transport pathway genes, which are crucial to a variety of plant defense responses. Resistant (Mp718, Mp719, and Mp04:104) and susceptible (Va35, Mp04:85, and Mp04:89) maize inbred lines were used in this study. Experimental conditions consisted of two treatments (inoculated and un-inoculated with *A. flavus*), six maize inbred lines with three replicates for each, and two sample collection time points (2 and 7 days after inoculation). Expression data for 50 genes were obtained by qRT-PCR technique using the Roche LightCycler 480 instrument. Specific R codes were developed for batch-reading qRT-PCR data in R, along with the efficiency correction of cp values, and the normalization with reference genes. These codes have not been available in R for pre-processing Roche form of qRT-PCR data. Linear models were used to fit qRT-PCR data for ANOVA analysis. Network analysis was conducted to investigate the co-regulated genes in the RNA transport regulatory networks. Results show that a NUP85-like nucleoporin gene was highly expressed in the resistant line Mp719. This supports that the Nuclear Pore Complex (NPC) related regulatory activities are involved in the maize host resistance to *A. flavus*. Elucidation of genes responsible in the regulation of host-fungus interactions will lead to new strategies to prevent *A. flavus* infection maize.

De novo Transcriptome Assembly of the Plant-Parasitic Nematode Rotylenchulus reniformis

Kurt C. Showmaker; Institute for Genomics, Biocomputing, and Biotechnology; Department of Biochemistry, Molecular Biology, Entomology and Plant Pathology, Mississippi State University, USA;
Satish Ganji; Institute for Genomics, Biocomputing, and Biotechnology, Mississippi State University, USA;
Mark A. Arick; Institute for Genomics, Biocomputing, and Biotechnology, Mississippi State University, USA;
William S. Sanders; Institute for Genomics, Biocomputing, and Biotechnology, Mississippi State University, USA;
Zenaida Magbunga; Institute for Genomics, Biocomputing, and Biotechnology, Mississippi State University, USA;
Martin J. Wubben; USDA-ARS, Mississippi State, USA
Daniel G. Peterson; Institute for Genomics, Biocomputing, and Biotechnology, Mississippi State University, USA;

Rotylenchulus reniformis, commonly known as the reniform nematode, is a pathogen of cotton, soybean, and sweet potatoes in the Southeastern United States. An estimate of cotton production loss due to *R. reniformis* parasitism in the United States in 2011 was 279,000 bales. Here, we present a *de novo* transcriptome assembly, annotation, and analysis for a population of diploid *R. reniformis* nematodes. The RNA used for the assembly was isolated from the egg, second-stage juvenile (J2), J3, vermiform adult and sedentary female (SF) developmental stages of the nematode. Homologies were found between *R. reniformis* and plant parasitic (e.g. *Meloidogyne* and *Heterodera* spp.), animal parasitic (*Brugia malayi*) and free living (*Caenorhabditis elegans*) nematodes. Of importance we report homologs between *R. reniformis* and previously described genes from the soybean cyst nematode, *Heterodera glycines,* involved in the establishment and maintenance of its specialized feeding structure, the syncytium, in the host root. Phylogenetically the *R. reniformis* transcriptome currently represents the most complete gene set for any nematode species within the family Hoplolaimidae.

Identification of oncogenic pathways of breast cancer in a genome wide association study

Vinay Raj, University of Arkansas for Medical Sciences
Susan Kadlubar, University of Arkansas for Medical Sciences

Breast cancer is the most common form of cancer in women and the second leading cause of cancer-related deaths in the U.S. Genome-wide association studies (GWAS) have identified a number of single nucleotide polymorphisms (SNPs) that are associated with risk of developing breast cancer. However, their clinical utility to predict risk, response to treatment, or treatment toxicity, remains undefined. There is a need to better understand the biology of breast cancer and to develop potential new targets for genetic testing with regard to cancer risk and prognostic value. Pathway analysis on GWAS data can elucidate relevant biological processes, and identify new candidate target genes. We sought to identify multiple cancer-related pathways in breast cancer from genetic variant signatures. We employed a pathway-based approach for a total of 3258 statistically significant SNPs ($P < 0.01$) from a breast cancer genome wide association study. Genes were represented by their most strongly associated SNP and the over representation of gene based associations in each pathway was calculated using Fisher's exact test. A total of 347 pathways containing 1645 genes were included in our study. Of these, five pathways ('Protein kinase A Signaling', 'GNRH Signaling', 'α-Adrenergic Signaling' and 'G-Protein Coupled Receptor') were highly enriched with associations. These results suggest that genetic alterations associated with these pathways may contribute to breast cancer susceptibility.

Analysis of Military Unique Chemical-Induced Neurotoxicity through Gene Regulatory Network Reconstruction

Haoni Li
Ping Gong
Chaoyang Zhang

Military-related activities produce many different chemicals, a portion of which inevitably contaminate soil. Neurotoxicity has been associated with such energetic compounds as TNT, RDX, and their degradation products. Monitoring, assessing and predicting the risks these chemicals pose when released require fundamental knowledge on how neurotoxicity occurs. We are interested to identify and discover how components involved in neurotransmission within the soil invertebrate *Eisenia fetida* interact and are affected by neurotoxicants. Understanding this network of interactions is essential for development of predictive risk models in the future. The objective of the current study was to gain some insights into the mechanism of chemical-induced reversible neurotoxicity through analysis of reconstructed gene regulatory networks in earthworms received different treatments (control, Carbaryl and RDX). Gene interactions networks were reconstructed using a Dynamic Bayesian Network (DBN) approach and were analyzed to identify differential edges (interactions) between the control and the exposed animals. Three synaptic pathways enriched with differentially expressed genes were chosen for reconstruction using time-course gene expression datasets. They are dopaminergic synapse, GABAergic synapse and synaptic vesicle cycle pathways. Our analytical results suggest that the regulation of protein kinase C (PKC) by voltage-sensitive calcium channel through Ca^{2+} broke down during exposure to RDX. Instead, calcium channels influenced protein phosphatase 2 catalytic subunit (PP2B) and protein kinase A (PKA) directly by Ca^{2+}, not through PKC regulation. Such computational inferences constitute a basis for formulating new hypothesis that can guide further biological validation experiments.

An evolution-inspired computational framework for computer-aided molecular design

Dmitry Korkin
University of Missouri - Columbia

It takes between 8 to 10 years on average to bring a drug from conception to market, with an average price tag of US $1.2 billion to $1.4 billion and above per drug. For the new drug leads to be clinically tested, scientists have to develop assays for the target disease and screen millions of compounds before the testing can even begin in animals. Thus, a cheaper and faster computational alternative that could significantly enhance the initial stages of drug design is needed. Recent computational approaches address the problem of computer-aided molecular design by calculating a set of physico-chemical and geometrical features and employing a machine learning framework. Such frameworks are often based on a classical vector-space formalism, whereas the objects in chemistry and molecular biology have a structural nature.

In this work, we present a novel symbolic machine learning approach for computer-aided molecular design. Called Chemical Evolving Transformation Systems (ChemETS) the framework is based on the structural representation of chemical compounds and, most importantly, their functional classes (*e.g.*, therapeutic categories). Moreover, ChemETS includes a structure-based supervised learning method to infer a compact representation of the whole functional class of compounds, based on a limited set of the class and non-class representatives. Finally, our formalism employs the class representation to classify the compounds into therapeutic categories. The class representation is defined as a progenitor, a structure common to all members in the class, and a set of transformations, a set of commonly used chemical modifications used to construct each individual member of the positive training set. The learning is achieved through a structure-driven analog of the Metropolis Monte-Carlo simulation. The inductively learned class representation for a training set of compounds of a given therapeutic category allows us to generate new compounds from the same therapeutic category. The key distinctive feature about the ChemETS framework is that the representation of the class members is evolutionary so that each of the class members is constructed from the final set of transformations. We have demonstrated the utility of our approach by constructing the class presentations of several therapeutic categories.

Differential Reconstructed Pathways for Deriving Toxicity Thresholds in Chemical Risk Assessment

Yi Yang, School of Computing, University of Southern Mississippi, Hattiesburg, MS 39406
Ping Gong, Environmental Services, SpecPro Inc., San Antonio, TX 78216
Xiaowei Zhang,State Key Laboratory of Pollution Control and Resource Reuse, Nanjing University
Nan Wang, School of Computing, University of Southern Mississippi, Hattiesburg, MS 39406
Chaoyang Zhang, School of Computing, University of Southern Mississippi, Hattiesburg, MS 39406

Pathway alterations reflected as changes in gene expression regulation and interaction are resulted from cellular exposure to toxicants. Such information is often used to elucidate toxicological modes of action. From risk assessment perspectives, alterations in biological pathways are an unexploited resource for setting toxicant threshold, which may be more sensitive and mechanism-informed than traditional toxicity endpoints. Here, we report a proof of concept study where a microbial genome-wide live cell reporter array was used to collect dynamic time-series gene expression data in Escherichia coli cells. The E. coli cells received four treatments, i.e., control, 10, 100, and 1000 mg naphthenic acids/L. Expression of 1800 GFP (green fluorescence protein)-labeled promoter genes were measured every 10 min for three hours (i.e., 18 time points). Differentially expressed (DE) genes were identified using Gaussian Process regression and then mapped to biological pathways in the EcoCyc database. DE genes-enriched pathways were selected for pathway reconstruction using a Bayesian Learning and Optimization model, which was developed in our Computational Biology and Bioinformatics lab. The reconstructed pathways of 4 different treatments were compared to infer the lowest concentration, at which significant pathway alterations were observed. Finally, the pathway alteration-derived threshold was compared with those derived from other toxicological endpoints such as cell growth. Findings from this study demonstrate that our approach has a great potential in providing a novel tool for threshold setting in chemical risk assessment.

Utilizing RNASeq in Eukaryotic Genome Annotation and Genome Databases

Christopher P. Childers; Division of Animal Sciences, University of Missouri, Columbia, MO
Darren E. Hagen; Division of Animal Sciences, University of Missouri, Columbia, MO
Justin T. Reese; Division of Animal Sciences, University of Missouri, Columbia, MO
Anna K. Bennett; Division of Animal Sciences, University of Missouri, Columbia, MO; Department of Biology, Georgetown University, Washington, DC
Christine G. Elsik; Division of Animal Sciences, University of Missouri, Columbia, MO; Department of Biology, Georgetown University, Washington, DC; Division of Plant Sciences, University of Missouri, Columbia, MO; MU Informatics Institute, University of Missouri, Columbia, MO

Important applications of RNASeq (transcriptome sequencing using Next Generation Sequencing technologies) include identification of transcribed regions within a genome, delineation of intron/exon structures of genes, detecting alternatively spliced transcript variants, and determining levels and breadth of gene expression across tissues and developmental stages. Here we will present approaches we have used to apply RNASeq in several eukaryotic genome annotation projects and databases. Some of our work includes testing open-source RNASeq assembly and alignment software, optimizing parameters, and developing pipelines to automate processing. In our opossum (*Monodelphis domestica*) transcriptome project, we used normalized libraries from a diverse set of tissues and developmental stages to generate an expression atlas, which we present in genome browsers of the Monodelphis Genomic Resources Database (OpossumBase.org). We further developed our RNASeq-based approaches to support computational gene prediction and manual gene annotation for genomes hosted in our Hymenoptera Genome Database (HymenopteraGenome.org), which provides genome informatics and annotation resources for insects of the order hymenoptera, including parasitoid jewel wasp (*Nasonia vitripennis*), three honey bee species (*Apis mellifera*, *A. florea*, *A. dorsata*), two bumble bee species (*Bombus impatiens*, *B. terrestris*), and nine ant species species (*Acromyrmex echinatior*, *Atta cephalotes*, *Camponotus floridanus*, *Cardiocondyla obscurior*, *Harpegnathos saltator*, *Linepithema humile*, *Pogonomyrmex barbatus*, *Solenopsis invicta*, and *Wasmannia auropunctata*). As part of our Bovine Genome Database project (BovineGenome.org), we are generating transcriptome assemblies based on a large number of tissues, to improve gene annotation and to create resources describing relative expression levels and the breadth of expression across tissues and relative expression levels.

Role of Global and Targeted hypomethylation in cancer

Huidong Shi, GHSU Cancer Center, Department of Biochemistry and Molecular Biology, Georgia Health Sciences University, Augusta, GA 30912
Dong Xu, Informatics Institute, Computer Science Department and Christopher S. Bond Life Sciences Center, University of Missouri-Columbia, MO-65201

Global-hypomethylation and Focal-hypermethylation are both frequent in human cancers. Previous studies revealed that there is an overall decrease in methylcytosine content of upto 70% compared with the level in normal somatic cells. On comparing DNA-methylation data of 31 chronic-lymphocyte leukemia (CLL) against normal B-cell, we observed that differentially methylated regions were more randomly present over the genome among different samples for hypomethylation in contrast to more conserved hypermethylation of individual genes. However, among the large number of random hypomethylated regions, few regions were conserved in most of the samples. These targeted/site-specific hypomethyled regions were found to be significantly enriched for B-cell receptor (BCR) signaling pathways with few other interconnected pathways. Rnaseq expression data for each samples show significantly increased expression for all genes in these pathways. This is important as we know that BCR signaling is a central pathogenic mechanism in B-cell malignancies, including CLL, that promotes leukemia cell survival and proliferation and modulates CLL cell migration and homing. Conserved hypomethylated regions were also highly enriched for DNA and protein binding motifs and histone-methyltransferases indicating the expression regulation through inter-connected histone deacetylation and methylation. Along with this, there was very high enrichment for alternative -splicing to regulate expression by splicing due to hypomethylation. Hence, this study shows that apart from global hypomethylation of repeat sequences, there also exists site-specific hypomethylation of certain genes that might contribute to some significant deleterious effect that can potentially result in initiation and prognosis of cancer and other diseases and thus have higher biological significance.

PLpred: a bioinformatics system for the identification and classification of plastid type proteins

Kaundal, R.[1] and Verma, R.[2]
[1]National Institute for Microbial Forensics & Food and Agricultural Biosecurity (NIMFFAB),
[1,2] Department of Biochemistry & Molecular Biology, Oklahoma State University, Stillwater (OK), USA.

ABSTRACT:

Plastids are important component of plant cells being the site of manufacture and storage of chemical compounds used by the cell, and contains pigments such as those used in photosynthesis, starch synthesis/storage, cell color *etc*. They are essential organelle of the plant cell, also present in algae. Recent advances in genomic technology and sequencing efforts have generated a pile of sequence data that needs to be annotated at a faster pace. In view of this, it is important to develop a prediction system that can distinguish between plastid and non-plastid type proteins accurately. Comparing the amino acid composition of plastid and non-plastid proteins shows significant differences, which is used as a basis to develop various prediction models using Artificial Intelligence. A range of protein features such as composition-based amino acid-, dipeptide-, split-, pseudo-composition, similarity-based PSI-BLAST, evolutionary information-based Position Specific Scoring Matrix (PSSM), and hybrid combinations are used in a Support Vector Machine (SVM) framework. To identify plastid *vs*. non-plastid protein types, we achieve the best prediction accuracy of 85.28% with a Matthews Correlation Coefficient (MCC) of 0.71. Further, we develop classification models to characterize the identified plastid proteins into various functional types (chloroplast, chromoplast, etioplast, amyloplast). A hybrid vector combining composition and homology-based information achieves the best prediction accuracy of 90.13% with a MCC of 0.76. The best performing models have been implemented as a webserver for use by the research community, accessible at http://bioinfo.okstate.edu/PLpred/. We believe this tool will be very useful in the functional annotation of various genomes.

Keywords: Plastid, Chloroplast, Chromoplast, Etioplast, Amyloplast, Prediction, Amino acid composition, Bioinformatics, Support Vector Machine.

Constructing Three-Dimensional Structures of Human Chromosomes from Chromosomal Contact Data

Tuan Trieu, Department of Computer Science, University of Missouri, Columbia
Jianlin Cheng, Department of Computer Science, Informatics Institute and C. Bond Life Science Center, University of Missouri, Columbia

Chromosomes are not positioned randomly within the nucleus, but they tend to adopt preferred conformations. To study the topologies and how chromosomes fold, here we describe a method to generate three-dimensional structure of individual chromosomes from contact maps. An objective function was derived for structures and this objective function was then optimized to generate structures that satisfy as much as possible the corresponding contact maps. We used contact maps from a Hi-C dataset to generate three-dimensional structures for chromosomes at resolution of one megabase. Parameters used in our method were derived from a FISH data to construct structures as close to the true structure as possible. Depending on chromosome length, our generated structures satisfy from eighty to one hundred percentages of contacts and non-contacts from the corresponding contact maps. We also verified our structures with the two compartments feature of chromosome and found that our structures exhibit this feature.

Statistical Methods for Ambiguous Sequence Mappings

Tamer Aldwairi, Mississippi State University
Dilip Gautam, Mississippi State University
Michael Johnson, Tennessee Tech University
Bindu Nanduri, Mississippi State University
Mahalingam Ramkumar, Mississippi State University
Andy D. Perkins, Mississippi State University

Mapping RNA sequences to a reference genome often results in high percentages of short reads assigned to multiple locations within the genome. These reads are known as "ambiguous mappings" and are often discarded by sequence mapping tools and pipelines. The amount of ambiguous mappings within these sequences can sometimes be significantly large, occupying in certain cases as much as half of the original sequence reads. We are developing task specific computer programs as an alternative approach. This statistical approach is based upon identifying significantly expressed genomic locations. We handle ambiguous data through a multi-step process starting with a standard short read alignment tool to identify all the possible mappings within the genome for each sequence read. Custom programs are then used to identify expressed genomic locations and then using statistical methods we compare gene expression in the regions of interest and compare with a number of randomly-selected genomic locations. Using these comparisons will help us in establishing a value at which a gene is significantly expressed and determine the locations that are most likely to be the best mapping for each ambiguous sequence.

Evaluation and application of microarray data: normalization to gene network inference.

Tanwir Habib, BTS LLC, Vicksburg MS 39180 USA
Natàlia Garcia-Reyero, Institute for Genomics, Biocomputing and Biotechnology, Mississippi State University, Starkville, MS 39759, USA
Gerald Ankley, U.S. Environmental Protection Agency, National Health and Environmental Effects Research Laboratory, Duluth, MN, USA
Daniel Villeneuve, U.S. Environmental Protection Agency, National Health and Environmental Effects Research Laboratory, Duluth, MN, USA
Edward J Perkins, Environmental Laboratory, US Army Engineer Research and Development Centre, Vicksburg, MS 39180, USA

Reconstructing gene regulatory networks from large-scale expression data holds great promise. One type of important information underlying the expression data is the 'genetic network', that is, the regulatory networks among genes. Current evaluation studies lack comprehensive metrics to assess the role of normalization procedures for gene network inference.

Here, three normalization techniques have been performed on a raw single channel microarray dataset. The dataset was generated by exposing fathead minnow (*Pimephales promelas*) to an aromatase inhibitor Fadrozole for 8 days and leaving them in clean water for 8 more days, to measure recovery. RNA was extracted from ovary tissue and microarray analysis using a 15k custom array (Agilent platform) was performed. Data was log2 transformed and normalized using three different methods: quantile, fastlo, and scale. In a fourth analysis, data were only log2 transformed and no normalization technique was applied. In order to estimate the impact of normalization procedures on the correlation structure, we first identified statistically significantly expressed probe sets using 1-way ANOVA and then computed correlation and mutual information between all pairs in the set using the network inference algorithms, CLR, ARACNE, MRNET, MIC, and TINGe. Each filtered network was evaluated for functional relationships between genes and we examined whether they shared common GO biological processes. We found that the two normalizations techniques quantile and fastlo had very similar results. Quantile normalized data with CLR algorithm worked best in identifying gene-gene interactions based on the literature evidences.

We conclude that the normalization procedure strongly affects the data correlation structure. Thus, choosing the right normalization procedure is a key step towards the inference of accurate cellular networks.

A Graphical Processing Unit Supported Neuroimaging Software in JAVA

Mutlu Mete, TAMU-Commerce
Harish Ankam, TAMU-Commerce
Unal Sakoglu, TAMU-Commerce

Despite its popularity, the versatile clustering, feature selection, machine learning, and data analysis tools found in WEKA have not yet been used by the neuroimaging analysis community. The developed software package with a GUI closes the gap between the neuroimaging research community and this versatile software by converting the NIFTI data to .arff format, which can then be analyzed by WEKA algorithms. In addition, we aim to extend the ability of WEKA to include an independent component analysis (ICA) package. The data conversion and ICA packages is integrated into WEKA as graphical user interface (GUI) to help neuroimaging researchers easily explore and apply the versatile algorithms of WEKA to their data. The GUI also lets users open and visualize NIFTI data using graphical processing units. Feature selection and correlation analysis in functional magnetic resonance images showed that a 14x speedup was gained within GPU experiments. In terms of GPU utilization within a JAVA platform, this software is one of first studies in neuroimaging literature.

Fractal approach for automatic malignancy determination in dermoscopy images

Sinan Kockara
Muhyeddin Ercan
Sait Suer
Melissa Perkins
Ashley Lawrence

Accurate diagnosis of melanocytic lesions is amongst the most difficult problems for dermatologists. Misdiagnosis of these lesions results in one of the causes of medical malpractice for this group of physicians. While there are myriad publications defining the dermatologic criteria that reproducibly distinguish ("presumably") benign melanocytic nevi from malignant melanomas, these criteria are not universally accepted nor easily recognized in all cases. Border irregularity is one of these criteria; however, currently there is no objective and quantitative measurement exists for lesion border irregularity. Therefore, focus of this study is to investigate shape irregularity features of suspected skin lesions from dermoscopy images. Border irregularity characteristics of the lesions will be investigated and analyzed in hopes of developing objective and quantifiable criteria that evaluate diagnostically challenging lesions and effectively distinguish benign from malignant lesions. More specifically, this study will algorithmically delineate lesion borders and then quantitatively measure irregularity of the extracted border.

Transcriptome Analysis of an Oleaginous Filamentous Fungus for Novel Biomass Consolidated Bioprocessing Model

Shangxian Xie, Texas A&M and Huazhong Univ. of Sci. Tech.
Su Sun, Texas A&M and Huazhong Univ. of Sci. Tech
Hu Chen, Texas A&M
Xing Qing, Texas A&M
Scott Sattler, USDA ARS
Xiaoyu Zhang, Huazhong University of Sci. Tech.

Lignocellulose represents the most abundant biomass on earth and the major feedstock for advanced biofuel. Consolidated bioprocessing (CBP) of lignocellulose into biofuel also is thought to be the future biomass conversion model. From our previous study, we found an oleaginous filamentous fungus *Cunninghamella echinulata* FR3, which could accumulate around 30% lipid from the raw sorghum straw and degrade 30% of the total lignin, as the potential candidate of biomass CBP conversion. To further study the genetic property of *C. echinulata* FR3, we carried a comprehensive transcriptome analysis of this fungus growing in the sorghum straw. It showed there were 45 glycoside hydrolase (GH) families related to the carbohydrate degradation, and there is one endoglucanase belonged to GH9 family expressing relative high level during the sorghum straw conversion. We also found around 80 different genes related to the lignin degradation and metabolism, including peroxidase VPL1and beta-esterase may be responsible for the lignin depolymerization. What's more, we also found that the lipid synthesis genes, as well as several polyunsaturated fatty acid synthesis genes also were expression in relative high level. These results further approved that *C. echinulata* FR3 could converse the lignocellulose into lipid. However, the cellulose and lignin degradation enzymes were not expression so high comparing to the wood rot fungi. And there also is no main lignin degradation enzymes like laccase, manganese peroxidase and lignin peroxidase found in the transcriptome data. These analysis results provide us a solid genetic guide for reverse design of this fungus to improve the lignocellulose conversion efficiency to lipid. It also provides a novel CBP model for the complete biomass conversion with oleaginous filamentous fungi.

Armband tracker and reminder for patients with Dementia and Alzheimer's disease

Victor Nutt
Shubhalaxmi Kher

Dementia and Alzheimer's is a slow progressive brain disease characterized by impairment of memory and eventually results into disturbances in reasoning, planning, language, and perception. Gradually for such patients all day to day activities becomes restricted. Slowly the patient is unable to manage his/ her life independently and needs constant monitoring and assistance. Simple tasks become hard to remember and create confusion. For example,it becomes hard to remember if the patient had just got into the house or is going out once the shoes are put on. In other words they need confirmation of the any activity done.

A compact wearable tracker and reminder system for patients with Dementia and Alzheimer's disease is proposed. The tracker and reminder system has two components; 1) the sensor and network system in the home and 2) reminder system. Wireless sensor network is deployed in the house to track the physical activity of the patient. The sensor network keeps track of the movements of the subject, whereas the reminder system keeps track of the timetable for various activities and generates a voice signal to announce and later to confirm the activity. The reminder system assists by generating fixed audio alarms like; breakfast Time, breakfast done, shower time, time to sleep, etc. The system uses multiple sensors and a microcontroller. User friendly armband system using a microcontroller collects activity information and assists the patient with alarms and reminders. Such a system will help assist the patients and reduce their dependency on others.