Oral Abstract

Abstract Identifying Number: 1006274

UPDATE ON THE MCBIOS TIMEBER RATTLESNAKE GENOME PROJECT

Adam Thrash[1], William S. Sanders[1], Mark A. Arick II[1], Bindu Nanduri[2], Daniel G. Peterson[1], and Doug Rhoads[3]

[1]Institute for Genomics, Biocomputing, and Biotechnology, Mississippi State University, Starkville, MS, 39759
[2]Department of Basic Science, College of Veterinary Medicine, Mississippi State University, Starkville, MS, 39759
[3]Department of Biological Sciences, University of Arkansas, Fayetteville, AR, 72701

The MidSouth Computational Biology and Bioinformatics Society (MCBIOS) initiated a collaborative project in 2012 to sequence, assemble, and annotate the genome of the timber rattlesnake. The timber rattlesnake is a threatened species in Arkansas, and rattlesnakes are unique model organisms that facilitate research in different contexts, especially relating to low-energy lifestyles. Through the efforts of Dr. Doug Rhoads and many others at the University of Arkansas, preliminary whole genome shotgun sequencing was conducted. Since 2012, a number of groups have worked with that sequence data, but none of the sequence data was able to be assembled into a publication or even draft genome announcement quality sequence. In 2014, Mississippi State University, in conjunction with the University of Arkansas generated additional sequence data and was able to use it to improve the quality of the existing draft assembly. Two libraries were created – 350bp and 550pb PCR fragment free libraries – and sequenced on both the Illumina HiSeq (2x100bp) and the Illumina MiSeq (2x300bp). This data assembled into an assembly with N50 of 22490 using ABySS, and annotated using the MAKER pipeline. *NSF EPS-0903787 EPSCoR Research Infrastructure Improvement Program: Track-1 Modeling and Simulation of Complex Systems.*

Abstract Identifying Number: 1006300

DYNAMIC VOXELIZATION FOR VIRTUAL ROTATOR CUFF SURGERY

Alexander Yu[1], Doga Demirel[1, 2], Tansel Halic[1], and Sinan Kockara[1]

[1*]Department of Computer Science, University of Central Arkansas, Conway, Arkansas, 72035
[2]Department of Computer Science, University of Arkansas at Little Rock, Little Rock, Arkansas, 72204

In arthroscopic rotator cuff surgery, drilling a suture anchor into the humeral head is one of the critical tasks. We used voxel based method for realistic real-time virtual simulation of drilling. Voxelization allows the ability to create convex and concave surfaces and enables robust haptic interaction during the drilling. However, the voxelization cannot represent exact geometry of the humeral head. This worsens especially with large voxel sizes. On the contrary, the use of finer and fixed resolution voxel sizes could be computationally unattainable. Therefore, a dynamic voxelization is highly desirable for realistic and computationally efficient interaction. In this study, we introduce a novel voxelization method based on dynamic proximity hierarchy (DPH). DPH is a graph spanner based hierarchical representation of approximate shortest paths of the points in the geometry. The hierarchy construction time is *O(n logn),* where *n* is number of geometry nodes. In this dynamic method, the size of the voxels are changed based on the required resolution and level of detail. Based on the proximities of interacting drilling objects or vicinity of the arthroscope, finer resolution of voxels can be dynamically generated. The desired optima such as visual quality, realism of drilling simulation can be determined by traversing the dynamic proximity hierarchy and modify the voxels in real-time. In this study, we present the details of this dynamicity with DPH and the performance results.

Abstract Identifying Number: 1006264

THE ROLE OF POLYAMINES IN PNEUMOCOCCAL DEFENSE MECHANSIMS AGAINST OXIDATIVE STRESS GENERATED BY HYDROGEN PEROXIDE

[2]Anagha Gopakumar,Aswathy.N[1] Rai, Leslie A Shack[1]and Bindu Nanduri[1]

[1] Department of Basic Sciences, College of Veterinary Medicine, [2]Department of Biological Sciences, Mississippi State University, Mississippi State, MS, USA 39762.

Streptococcus pneumoniae is a causative agent of bacterial pneumonia, meningitis, otitis media and bacteremia. Limitations in available vaccinations and therapeutics had led to the emergence of increasing numbers of multi-drug resistant strains of this human pathogen. *Streptococcus pneumoniae* is routinely exposed to high levels of oxidative stress during colonization and invasive disease and has evolved very distinct mechanisms from its Gram positive counterparts in its stress response. Our previous findings demonstrated the requirement for functional polyamine transport operon *potABCD*, is crucial for pneumococcal survival *in vivo*. Polyamines are implicated in oxidative stress responses in many bacterial species including pneumococci. Polyamines are poly cations at physiological pH and can bind to the negatively charged DNA and regulate gene expression. The objective of this study is to profile the differences in gene expression between the *ΔpotABCD* and TIGR4 to identify the role of polyamine transport in pneumococcal adaptation to hydrogen peroxide stress. TIGR4 and *ΔpotABCD* were cultured in Todd Hewitt broth with 0.5mM Hydrogen Peroxide overnight .Mid-log phase cultures were used for RNA extraction. Gene expression was analyzed using real-time Quantitative Polymerase Chain Reaction (qRT-PCR). The genes we investigated include *tcs04, spxR, Rgg, psaR and ciaRH* which are all characterized as being involved in regulating oxidative stress. The results from qRT-PCR analysis will be discussed. *MS EPSCoR EPSCoR Grant #0903787*

Abstract Identifying Number: 1002150

BAYESIAN LEARNING IN AN UNDERDETERMINED SYSTEM OF GENE REGULATORY NETWORKS

Andrew Maxwell*[1], Ping Gong[2], Chaoyang Zhang1[1]

[1]School of Computing, University of Southern Mississippi, Hattiesburg, MS, 39406
[2]Environmental Laboratory, US Army Engineer Research and Development Center, Vicksburg, MS, 39180

Gene regulatory networks (GRNs) have been shown to have an underdetermined set of solutions, meaning that because of the nature of the expression data, there can be multiple unique solutions that can produce an inferred regulatory network. There are many inference methods to date, and each method contains its own complex solution to solve the problem of GRNs. The Bayesian Learning and Optimization Model (BLOM) is one such method that combines a state space model, expectation-maximization algorithm, and Kalman filter with time-series expression data to achieve a solution to GRNs. Past results have demonstrated that the BLOM algorithm achieved good performance on experimentally generated datasets. However, the performance with in silico-generated simulation datasets has been less than ideal. It has been known that different inference algorithms perform either better or worse depending on the type of data used, but there are steps that can be introduced to this Bayesian Learning model to obtain more accurate results in both experimental and simulation data. For an underdetermined GRN system, certain known network motifs - i.e., certain patterns of connected nodes and edges, within a network can cause an inference algorithm to produce results where a large amount of the network is falsely identified as potential regulatory paths. By eliminating these types of motifs, one can reduce the solution space and improve a network's inference accuracy. In an effort to improve the inference accuracy of BLOM, the incorporation of ensemble modeling techniques are utilized to remove uncertain edges from resulting networks and to identify different aspects of BLOM that can be improved. Several examples including DREAM4 datasets are used to investigate the efficiency and accuracy of BLOM with and without uncertain edges in GRNs.

Keywords: Gene Regulatory Networks (GRNs); Bayesian Learning and Optimization Model (BLOM); Network Inference.

Funding sources: US Army Environmental Quality and Installation Research Program and NSF EPSCoR.

Abstract Identifying Number: 100402

PROTEOMICS OF HOST AND PATHOGEN TO STUDY THE ROLE OF POLYAMINE TRANSPORT IN PNEUMOCOCCAL VIRULENCE IN A MOUSE MODEL

Aswathy N. Rai[1], Leslie A. Shack[1], Edwin Swiatlo[2] and Bindu Nanduri[1]

[1] Department of Basic Sciences, College of Veterinary Medicine, Mississippi State University, Mississippi State, MS, 39762.

[2] Division of Infectious Diseases, University of Mississippi Medical Center, Jackson, MS, 39216.

Streptococcus pneumoniae (pneumococcus) causes pneumococcal pneumonia. Design of effective therapeutic strategies for this pathogen is confounded by genome plasticity, and increasing antibiotic resistance. Polyamine transport system in the pneumococci is an attractive therapeutic target as deletion of polyamine transport operon (*ΔpotABCD*) in *S.pneumoniae* TIGR4 led to attenuation in a mouse model of pneumonia. Our previous findings show that TIGR4 persists for 24 hr post infection (p.i) in mouse lung while *ΔpotABCD* gets cleared. Here we report expression proteomics of pathogen response to deficient polyamine transport and host response to *ΔpotABCD.* Pathogen proteomics will identify pneumococcal proteins/virulence factors regulated by polyamines that lead to reduced virulence. Host proteomics is expected to identify differences in lung protein expression that could result in the observed difference in growth kinetics of TIGR4 and *ΔpotABCD* in mouse lung. Proteins were isolated from lung tissue from mice challenged with TIGR4 and *ΔpotABCD* (4 hr and 12 hr p.i) and bacteria cultured in vitro and subjected to1D LC ESI MS/MS. Our results show that reduced expression of pneumococcal capsular polysaccharide biosynthesis protein (Cps4F), and virulence factor pneumolysin (Ply), in *ΔpotABCD* could be responsible for the observed attenuation of *ΔpotABCD* strain. Protein networks identified in mouse lung tissue from *ΔpotABCD* infection at 4 hr clearly show activation of inflammatory molecules. Activation of inflammatory networks that leads to bacterial clearance is delayed in TIGR4, as we see this 12 hr p.i. Taken together, these findings illustrate the molecular mechanisms in the host and pathogen during pneumococcal infection in mice. *NIGMS grant # P20GM103646*.

Abstract Identifying Number: 100402

EVALUATION OF HOST RESPONSE DURING EXPERIMENTAL BOVINE RESPIRATORY DISEASE USING EXPRESSION PROTEOMICS

Aswathy N. Rai, Leslie A. Shack, Joseph S. Reddy, Wes Baumgartner, William B. Epperson, [1]Ty B. Schmidt and Bindu Nanduri.

College of Veterinary Medicine, Mississippi State University, Mississippi State, MS, 39762
[1]Department of Animal Sciences, University of Nebraska-Lincoln, Lincoln, NE 68583

Bovine Respiratory disease complex (BRD), is a multifactorial disease affecting cattle and is most frequently characterized by the clinical onset of bronchopneumonia. *Mannheimia .haemolytica*, a gram negative bacterial pathogen is most commonly associated with field cases. BRD diagnostics rely predominantly on clinical signs, which are subjective and error-prone. Here we describe mass-spectrometry based proteomics to study protein expression indicative of disease in calves challenged with *M. haemolytica*. Total proteins were isolated from post-mortem lung tissue and broncho alveolar lavage fluid (BALF), 7-days post challenge. Proteins were identified by 1D LC ESI MS/MS from regions of the lung that showed tissue damage (affected) and also an area that appeared normal. This sample collection strategy facilitated the comparison of differentially expressed proteins from affected and normal regions of lung. Our results showed that 40 proteins were common to affected and normal regions of the lung while 88 proteins were unique to affected region. This observation has implications for *in vivo* sample collection where the extent of tissue damage is not known *a priori*. Comparison of BALF protein expression from affected and normal regions showed that 116 proteins were common to both. Comparison of BALF and lung protein expression profiles clearly demonstrate that a number of acute phase proteins are common to both sample types. Taken together, these results show variation in protein expression consistent with the spectrum of tissue damage in BRD which needs to be accounted for when obtaining samples in vivo for identifying biomarkers for disease diagnostics and disease stratification. *MS EPSCoR Grant #0903787.*

Abstract Identifying Number: 1006297

INTEGRATION OF MICRORNA-MRNA INTERACTION NETWORKS WITH MICROARRAY DATA TO INCREASE EXPERIMENTAL POWER

Bernie J. Daigle, Jr.[1]

[1] Departments of Biological Sciences and Computer Science, The University of Memphis,
Memphis, TN, 38152

The detection of differentially expressed (DE) genes between two or more biological conditions is an essential step in the search for candidate disease genes, drug targets, and discriminative biomarkers. Although widely used for this task, DNA microarrays are notorious for generating noisy data. One strategy for reducing this noise is to assay many experimental replicates. However, as this approach can be costly and sometimes impossible, methods are needed which improve DE gene identification at no additional cost.

An important source of information for differential expression analysis comes from microRNA (miRNA) experiments. While the transcriptional roles of miRNAs are well documented, methods for incorporating miRNA datasets with traditional gene expression assays are lacking. Thus, I developed Noisy-Or Optimization for DifferentiaL Expression analysis (NOODLE), a novel Bayesian network-based approach for integrating miRNA-mRNA interaction networks with microarray data to improve DE gene identification. Given a microarray dataset, NOODLE acts by increasing belief that an mRNA is DE if one or more interacting miRNAs are themselves DE (and vice versa).

I first apply NOODLE to synthetic datasets, achieving more accurate DE gene identification than the popular *limma* method over a wide range of network topologies and configurations. Using two publicly available cancer datasets, I next demonstrate how NOODLE increases experimental power by up to a factor of four. Finally, I apply NOODLE to data interrogating expression differences between induced pluripotent and embryonic stem cells. My results uncover important biological differences between these cell types that would be missed using existing methods. *Supported in part by the Institute for Collaborative Biotechnologies through grant W911NF-10-2-0111 from the U.S. Army Research Office.*

Abstract Identifying Number: 100255

PROTEOMICS AND HOST-PATHOGEN INTERACTIONS

Bindu Nanduri

Department of Basic Sciences, College of Veterinary Medicine, Mississippi State University,
Mississippi State, MS, 39762

Despite investments and progress in genomics, bioinformatics, and computational biology, global burden of infectious diseases remains a major health care concern to date. Understanding the interplay between host and pathogen that underpins health/disease is critical for identifying potential targets for therapeutic, prophylactic and intervention strategies to eliminate/ or reduce the severity and economic impact of infectious diseases. Infection is the outcome of altered interactions of any of the following kind; between the host cells and invading microorganisms, interactions amongst host immune molecules, pathogen molecules, or host-microbiome interactions. A plethora of bioinformatics databases are available that are repositories protein-protein interactions. A brief overview of the existing computational resources for host-pathogen interactions (HPI) will be presented. Mass spectrometry based expression proteomics is an invaluable tool to measure genome expression readout at the protein level. Expression proteomics of the host and/or pathogen can identify protein response networks during infection and improve our understanding of disease progression. Furthermore, proteomics is a powerful to identify protein-protein interactions. A brief overview of mass-spectrometry based expression proteomics will be presented. In summary, this workshop focuses on studying inter-species interactions that are at the heart of infectious diseases with particular focus on proteomics, and host-pathogen protein-protein interactions. *NIGMS grant # P20GM103646 and MS EPSCoR Grant #0903787.*

Abstract Identifying Number: 100326

LANDSCAPE OF CIRCRNA CANDIDATES ACROSS 11 ORGANS AND 4 DEVELOPMENTAL STAGES IN FISCHER 344 RAT

Binsheng Gong[1], Joshua Xu[1], and Weida Tong*[1]

[1*]Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, US Food and Drug Administration, Jefferson, AR 72079

Circular RNA (circRNA) is a class of endogenous noncoding RNAs and has attracted great attention due to their potential biological function as regulators of microRNAs as recently reported in some studies. The next-generation sequencing technologies and novel bioinformatics approaches enable the detection of circRNAs in many species. Thousands of novel circRNA candidates have been revealed in mammalians as well as nematode. This study provides an overview of circRNA candidates detected through an RNA-seq dataset across 11 organs of Fischer 344 rats from 4 developmental stages. The induction of circRNA candidates displays clear organ-specific patterns and gender differences for some organs. Liver and muscle have the lowest numbers of circRNA candidates and brain has the most and the pattern was also observed for expressed genes and transcripts in our previous study. Among the 1,793 parental genes, only 58 were detected with backspliced junctions in all eleven organs and63 detected in ten organs but absent in one organ. 333 genes were detected with backspliced junctions only in one organ. The overlap of the induced circRNAs between male and female are less than 50% in each non-sexual organ, except for brain with an up to 67% concordance observed in aged rats. A trend of increase in circRNA candidates along the four developmental stages was observed in brain andliver for both sexes. In contrast, there is a drop in circRNA candidates in thymus for aged rats of both sexes. The number of circRNA candidates was stable in heart and lung. In the sex organs, the number of circRNA candidates remained stable across the aging points in Uterus, increased in the younger ages (Juvenile through Adult) in thymus and then significantly dropped for aged rats. Further knowledge of circRNA candidates in rat will undoubtedly advance the study of drug toxicity at the RNA regulation level.

GENE EXPRESSION ANALYSIS OF WILD TYPE AND AN IRON-DEPENDENT TRANSCRIPTIONAL REGULATOR DEFICIENT PNEUMOCOCCI USING RNA-SEQ

Caleb Benson[1,3], Aswathy N. Rai[1,2], Mark A. Arick II[1], Edwin Swaitlo[4], and Bindu Nanduri[2,3]

[1]Institute for Genomics, Biocomputing and Biotechnology, Mississippi State University, Mississippi State, MS, 39762
[2]College of Veterinary Medicine, Mississippi State University, Mississippi State, MS 39762
[3]Department of Biological Sciences, Mississippi State University, Mississippi State, MS 39762
[4]Department of Infectious Diseases, University of Mississippi Medical center Jackson, MS 39216

Steptococcus pneumoniae is a human pathogen that causes pneumonia, otitis media, and meningitis. During the course of infection, the bacterium has to adapt to various host niches where it encounters a wide range in concentration of available iron. Iron is known to be an important regulator of gene expression in virulent bacteria. *Steptococcus pneumoniae* TIGR4 encodes an iron-dependent transcriptional regulator (IDTR). Previous results from our lab showed that isogenic deletion of *idtr* leads to attenuation in mouse models of pneumonia and sepsis. The aim of this study is to identify genes that are regulated by *idtr* in *S. pneumoniae*. We isolated total RNA from mid-log phase cultures of TIGR4 and *Δidtr* grown in chemically defined medium. RNA-Seq was performed with three replicate cultures using Illumina MiSeq technology. Reads were mapped to *S. pneumoniae* TIGR4 reference genome using Bowtie2 and the number of reads aligned/gene was determined by using HTSeq. Using EdgeR differential expression analysis package, we identified significant changes in gene expression between the TIGR4 and *Δidtr*. A total of 545 genes were significantly altered in response to deletion of *idtr*. Of these, 476 were up regulated and 69 were down regulated. Expression of capsular biosynthesis genes such as *cps4A, cps4B cps4C,* and *cps4G* decreased in *Δidtr.* Comprehensive functional analysis of all differentially expressed genes will be conducted to identify *idtr* dependent gene expression in pneumococcus. *Funded by: MS EPSCoR Grant #0903787*

GENE EXPRESSION ANALYSIS OF WILD TYPE AND AN IRON-DEPENDENT TRANSCRIPTIONAL REGULATOR DEFICIENT PNEUMOCOCCI USING RNA-SEQ

Caleb Benson[1,3], Aswathy N. Rai[1,2], Mark A. Arick II[1], Edwin Swaitlo[4], and Bindu Nanduri[2,3]

[1]Institute for Genomics, Biocomputing and Biotechnology, Mississippi State University, Mississippi State, MS, 39762
[2]College of Veterinary Medicine, Mississippi State University, Mississippi State, MS 39762
[3]Department of Biological Sciences, Mississippi State University, Mississippi State, MS 39762
[4]Department of Infectious Diseases, University of Mississippi Medical center Jackson, MS 39216

Steptococcus pneumoniae is a human pathogen that causes pneumonia, otitis media, and meningitis. During the course of infection, the bacterium has to adapt to various host niches where it encounters a wide range in concentration of available iron. Iron is known to be an important regulator of gene expression in virulent bacteria. *Steptococcus pneumoniae* TIGR4 encodes an iron-dependent transcriptional regulator (IDTR). Previous results from our lab showed that isogenic deletion of *idtr* leads to attenuation in mouse models of pneumonia and sepsis. The aim of this study is to identify genes that are regulated by *idtr* in *S. pneumoniae*. We isolated total RNA from mid-log phase cultures of TIGR4 and *Δidtr* grown in chemically defined medium. RNA-Seq was performed with three replicate cultures using Illumina MiSeq technology. Reads were mapped to *S. pneumoniae* TIGR4 reference genome using Bowtie2 and the number of reads aligned/gene was determined by using HTSeq. Using EdgeR differential expression analysis package, we identified significant changes in gene expression between the TIGR4 and *Δidtr*. A total of 545 genes were significantly altered in response to deletion of *idtr*. Of these, 476 were up regulated and 69 were down regulated. Expression of capsular biosynthesis genes such as *cps4A, cps4B cps4C,* and *cps4G* decreased in *Δidtr.* Comprehensive functional analysis of all differentially expressed genes will be conducted to identify *idtr* dependent gene expression in pneumococcus. *Funded by: MS EPSCoR Grant #0903787*

Abstract

Abstract Identifying Number: 1006110

LCS BASED PROTEIN STRUCTURE PREDICTION

Cameron L. Walker, Venkata Kiran Melapu, Sravanthi Joginipelli, Karl Walker*[1]

[1*]Department of Bioinformatics, University of Arkansas at Pine Bluff, Pine Bluff, AR, 71602

There is an ever-increasing number of unsolved structures of proteins which are considered to be of low homology in the Protein Data Bank. Even the most accurate template-based protein structure prediction software show marginal performance against them. Accuracy of the structure predictions in this case, is always a function of the interdisciplinary knowledge shared by the research group. Sequence motifs are short, recurring patterns in DNA that are conjectured to have biological significance; These motifs often indicate specific binding sites for proteins such as nucleases and transcription factors. This tool gathers sequence segments shared between a target and its respective template(s) using LCS (Longest Common Substring). The longest common substring is the longest substring shared between two or more strings. It cross-references the afore mentioned sequence segments with against an interactive database of sequence motifs created as a part of this project from various public repositories. This allows users to access the templates with reasonable potential for future modeling. Although proteins are poly-amino acid sequences, for all practical purposes in this research, we considered them as character strings. This is an attempt to identify sequence motifs shared by target as well as the template(s) which is crucial for better construction of 3D models of proteins. *This project was made possible by the Arkansas INBRE program, supported by grant funding from the National Institutes of Health (NIH) National Institute of General Medical Sciences (NIGMS) (P20 GM103429) (formerly P20RR016460).*

Abstract Identifying Number: 1006252

HUMAN MICROBIOME: SEQUENCING-BASED HIGH-THROUGHPUT OMICS TECHNOLOGY AND BIOINFORMATICS USED IN THE ASSESSMENT OF THE SAFETY OF ANTIMICROBIAL DRUG RESIDUES IN FOOD

Carl E. Cerniglia

Director, Division of Microbiology, National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, AR72079

The human gastrointestinal tract ecosystem consists of complex and diverse microbial communities that have now been collectively termed the intestinal microbiome. Recent large-scale research endeavors using sequencing-based high-throughput molecular technologies have increased our understanding of the important role that the microbiome plays in human health and disease. Effective meta-omics technologies and bioinformatics tools have expanded our knowledge on microbial community composition, functional and metabolic activities on microbiome-host interactions. An area of public health interest has been the concern of the use of veterinary antimicrobial agents in food–producing animals may result in the presence of low–level of drug residues in edible foodstuffs. The potential risk to the consumer is that antimicrobial agents at residue–level concentrations could disrupt the colonization barrier and/or modify the antimicrobial resistance profile of the human intestinal microbiota and promote the emergence of antibiotic resistance bacteria. The presentation will highlight an integrated systems biology approach used for the safety evaluation and risk assessment of antimicrobial residues and their effect on the intestinal microbiome with the goal of insuring safety of the human food supply. Critical issues associated with the impacts of antimicrobial agent residues in foods on the intestinal microbial community and the potential increase in the population of drug-resistant bacteria will be discussed.

Development of the Tocoflexols, a Series of Novel Vitamin E Analogues with Improved Bioavailability

Awantika Singh[1,2], Philip J. Breen[2], Shraddha Thakkar[1,5], Mahmoud Kiaei[3,4], Martin Hauer-Jensen[2], and Cesar M. Compadre[2]

[1]UAMS/UALR Joint Bioinformatics Graduate Program, [2]Department of Pharmaceutical Sciences and [3]Department of Neurobiology and Developmental Sciences, [4]Center for Translational Neuroscience, [5]Department of Physiology and Biophysics, University of Arkansas for Medical Sciences, Little Rock AR 72205

Vitamin E is composed of two series of closely related compounds, tocopherols and tocotrienols, which show a widely varying degree of biological effectiveness. In recent years, the tocotrienols have gained considerable attention because of their beneficial effects in areas such as radiation protection, cholesterol reduction, neuroprotection, and cancer treatment and prevention. Unfortunately, the potential of the tocotrienols has been hampered due to short circulation half-life that severely limits their bioavailability. Using state-of-the-art computational techniques, we have developed the tocoflexols, which are designed to overcome the limitations of the tocotrienols. We developed analogues that more effectively use the natural mechanism of retention of vitamin E components in the body. Specifically this was accomplished by designing compounds that are better accepted by the alpha-tocopherol transfer protein, the liver protein that controls the plasma levels of vitamin E. By maintaining the bioactivity of the tocotrienols while achieving enhanced bioavailability, these compounds may have a strong potential as therapeutic agents. Structural modification of drugs to take advantage of endogenous transport systems is a novel and intriguing concept whose potential is just starting to emerge. Successful demonstration of its usefulness is likely to encourage development of similar strategies for the future drug design and development in other areas of biomedical sciences.

Abstract Identifying Number: 1006247

MOBIPROT: A FEASIBLE TOOL FOR PROTEIN SEQUENCE ALIGNMENT AND PHYSIOCHEMICAL PROPERTIES ON ANDROID PLATFORM

Chaitanya Mallikanti[1], Pooja[1], Preeti Rustagi[1], Swati Sinha[1], Latesh Kumar[2], Vivek Chandramohan[*1]

[1*]Biotechnology Finishing School (BTFS-V), Siddaganga Institute of Technology,
Tumkur, Karnataka, 572103

[2] Computer Science Department, Siddaganga Institute of Technology, Tumkur,
Karnataka, 572103

There is an accelerating growth of smart phones in the human world for entertainment and scientific applications which is making research works more effectively and productively. Android platform is most widely used operating system on smart phone and hand held tablet devices in important target for mobile application developers. According to International data Corporation (IDC) in 2015 second quarter android operating system (versions – 2.2 Froyo to 5.0 Lollipop) is dominating the smart phone market with 82.8%. There are various reported web based software tools, to find the amino acid sequence alignment for the best similarity and its protein properties which runs on desktop and laptops. However none of the app is currently available to provide mobile interface to one of the most widely used service in bioinformatics for the alignment and analysis to find best similarity between two amino acid sequence of different proteins, Partial Order Alignment (POA) graphic alignment and physiochemical properties of the protein sequences. A mobile app –MobiProt is developed and implemented for the android platform based mobile devices. In this app, global and local sequence alignment is based on the two input protein sequences using BLOSUM62 matrix. As there is a huge usage of smart phone and tablets devices, such bioinformatics app would raise productivity of researchers and facilitate the analysis of sequence data.

Abstract Identifying Number: 100262

NEW ADVANCES IN INFERENCE OF GENE REGULATORY NETWORKS FROM TIME SERIES DATA

Chaoyang Zhang[1], Bei Yang[1], Xi Wu[1], Andrew Maxwell[1], Nan Wang[1] and Ping Gong[2]

[1]School of Computing, University of Southern Mississippi, Hattiesburg, MS, 39406
[2]Environmental Laboratory, US Army Engineer Research and Development Center, Vicksburg, MS, 39180

Reconstruction of complex networks of genetic interactions and identifying unknown relations among genes are very important for understanding the underlying mechanism of biological processes. Inferring gene regulatory networks (GRNs) from time series microarray data or next generation sequencing data is a very challenging inverse problem due to its nonlinearity, high dimensionality, sparse and noisy data and significant computational cost. The existing GRN inference approaches such as probabilistic Boolean networks and dynamic Bayesian networks have various limitations and relatively low accuracy. In this work, we introduce a Bayesian learning and optimization model (BLOM) developed in recent years and its variant that uses Principal Component Analysis (PCA) to reduce the noise and dimension of the dataset. A total of 20 GRNs with different number of time points were used to evaluate the performance. The results show that the PCA-based method has compatible accuracy, is more computational efficient and stable with a different length of hidden variables, and can be applied to GRN inference from the dataset with a smaller number time points.

Abstract Identifying Number: 1006021

MISSING DATA INTERPRETATION FOR NON-REFERENCED OR SEMI-REFERENCED GENOMES

Charles Chen[1*], Shuzhen Sun[2], Enrique Jiménez Schwarzkopf[3] and Yousry A. El-Kassaby[4]

[1*] Department of Biochemistry and Molecular Biology, Oklahoma State University,
Stillwater, Oklahoma 74078, USA
[2]Department of Industrial Engineering, Oklahoma State University, Stillwater, Oklahoma
74078, USA
[3]School of Biological Sciences, Washington State University, Pullman, Washington
99164 USA
[4]Faculty of Forestry, the University of British Columbia, Vancouver, British Columbia
V6T 1Z4 Canada

The unprecedented throughput brought by Next-gen sequencing technologies has revolutionized genomic research for non-model organisms by generating hundreds of thousands single nucleotide polymorphism that could be associated with phylogenetically and agronomically important phenotypes. One drawback for a broad application of NGS, especially for study systems that are still lacking whole genome assemblies, is the amount of uncertainty in SNP determination. In this study, we evaluated the performance of Genotyping-by-sequencing (GBS) for species with massive genomic variation, compared both parametric and non-parametric algorithms capable of interpreting missing genotypes and discussed the applicability and usefulness in association mapping and prediction for highly polygenic phenotypes. *This work has been supported by Oklahoma Agricultural Experiment Station and Oklahoma Center for the Advancement of Science and Technology for C.C. and Johnson's Family Forest Biotechnology Endowment, FPInnovations' ForValueNet for Y.A.E.*

Abstract Identifying Number: 1006284

DIFFERENTIALLY EXPRESSED GENES IN LATITUDINAL POPULATIONS OF COMMON SUNFLOWER (*Helianthus annuus* L.) ARE ENRICHED WITH MICROSATELLITES

Chathurani Ranathunge[1], Gregory Wheeler[2], Andy Perkins[3] and Mark Welch[1]
(Skip single line)
[1]Department of Biological Sciences, Mississippi State University, Starkville, Mississippi, 39759
[2]Department of Evolution, Ecology and Organismal Biology, Ohio State University Columbus, Ohio 43210
[3]Department of Computer Science and Engineering, Mississippi State University, Starkville, Mississippi, 39759

Natural populations of widely distributed common sunflower (*Helianthus annuus* L.) are highly adapted to their local environments. In this study, a common garden experiment was conducted using six natural populations of common sunflower from two latitudes in Kansas and Oklahoma to test a potential gene regulatory role for transcribed microsatellites. The common garden experiment demonstrated that phenotypic variance among populations is largely explained by underlying genetic variation. RNA-Seq was conducted on 95 individuals and differential gene expression was inferred using the DESeq program. A gene ontology (GO) analysis was conducted on the significantly differentially expressed (DE) genes using the Blast2GO program. The DE genes were also mined for the presence of transcribed microsatellites using the SciRoKo program and a custom perl script identified the abundance of different motif types. RNA-Seq produced an average of 42.8 million reads per individual and DESeq identified 1521 significantly differentially expressed genes (1299 up-regulated and 222 down-regulated in Kansas). The GO analysis produced functional annotations for 349 of these DE genes based on the *Arabidopsis thaliana* protein database while SciRoKo identified 1253 microsatellites in 696 unique DE gene sequences. Trinucleotides were the most abundant (700) motif type represented in the identified microsatellites within the DE genes. The abundance of transcribed microsatellites within DE genes among populations grown in a common garden is consistent with a proposed causal role for these microsatellites in locally adapted populations of common sunflower. The study provided empirical evidence to test the role of transcribed microsatellites as cis-regulatory elements in gene expression regulation.

This work was funded by the NSF grant, MCB-1158521

Genome, Epigenome, Transcriptome, and Epitranscriptome Landscapes: from single cells, to entire cities, to space

Christopher E. Mason, Ph.D.
Associate Professor, Department of Physiology and Biophysics, Weill Cornell Medical College, New York, USA

The avalanche of easy-to-create genomics data has impacted almost all areas of medicine and science, and here we report the implementation of genomics technologies from the single-cell to an entire city. Recent methods and algorithms enable single-cell and clonal resolution of phenotypes as they evolve, both in normal and diseased tissues. We show that tumors evolve quickly at the genetic, epigenetic, transcriptional, and epitranscriptional level, enabling many means by which tumors can resist therapy. Notably, some of these changes can be resolved by single-cell analysis and enable prognostic relevance. We report new biochemical methods (eRRBS, MeRIP-seq) and algorithms (methylKit, eDMR, methclone) to examine these changes. Finally, we will discuss pilot data for creating enabling patients to become more involved in their 'omics data, including to an integrative genomics view of an entire city (based on our Pathomap project) that leverages longitudinal genomics and microbiome profiles of the NYC subway system. All of these pieces work together to guide the most comprehensive, longitudinal, mutli-omic view of human physiology with the NASA Twins Study.

Abstract Identifying Number: 1006260

EFFICIENT PAIRWISE STRUCTURAL ALIGNMENT OF RNA SEQUENCES BASED ON TOPOLOGICAL NETWORKS

Chun-Chi Chen[1], *Hyundoo Jeong*[1], *Xiaoning Qian*[1] *and Byung-Jun Yoon*[1,2,*]

[1]Department of Electrical and Computer Engineering,
Texas A&M University, College Station, TX 77843-3128, USA
[2]College of Science and Engineering,
Hamad bin Khalifa University (HBKU), Doha, Qatar
E-mail: byoon@qf.org.qa

Abstract

RNA secondary structure is well conserved across different species, where for many RNA families, it is known to be better conserved than the RNA sequence itself. For this reason, it is important to consider the underlying structure when aligning RNA sequences, especially for those with low sequence similarity. Simultaneous RNA alignment and folding algorithms aim to accurately align RNAs by predicting their consensus structure and aligning them at the same time. Unfortunately, the computational complexity of the optimal dynamic programming algorithm for simultaneous alignment and folding is $O(N^6)$ for aligning two RNA sequences of length N, which is too high to be used for large-scale analysis. In this work, we propose a novel method for pairwise structural alignment of RNAs by introducing the concept of topological networks that provide structural maps of the RNAs to be aligned. For each RNA, we construct a topological network based on the predicted RNA structure, which consists of sequential edges and structural edges weighted by the base-pairing probability. The networks can then be efficiently aligned by using network alignment techniques, yielding the structural alignment of the RNAs. The computational complexity of our proposed method is $O(N^2)$, which is significantly lower than the dynamic programming approach, while resulting in favorable alignment results. Furthermore, the proposed method is not restricted to nested structures, hence it can effectively handle RNAs with pseudoknots. We demonstrate the performance of the proposed method through extensive simulations based on known RNA families.

Abstract Identifying Number: 1006027

PHYLOGENETIC TREE CONSTRUCTION USING TRINUCLEOTIDE USAGE PROFILE (TUP)

Si Chen[1], Lih-Yuan Deng[1], Dale Bowman[1], Behrouz Madahian[1], Henry Horng-Shing Lu[2], and Tit-Yee Wong[3]

[1]Department of Mathematical Sciences, University of Memphis, Memphis, TN 38152
[2]Institute of Statistics, National Chiao Tung University, Hsinchu, Taiwan, 30010
[3]Department of Biological Sciences, Bioinformatics Program, University of Memphis, Memphis, TN 38152

A new method, called Trinucleotide Usage Profile (TUP), is proposed to build a genome-wide phylogenetic tree for a large group of species containing a large number of genes with long nucleotides sequences. We investigate the key issue of finding a macroscopic statistic to represent and characterize the whole-genome DNA information. The most popular method, called feature frequency profile (FFP), finds the frequency distribution for all words of certain length over the whole genome sequence using (over-lapping) windows of the same size. Unfortunately, in order to characterize the genome- wide information, the word length is often much larger than 3 (codon length). We argue, from an information theoretic viewpoint, that the frequency distribution of overlapping words can only dilute the gene information. We propose an essential modification on the popular FFP method while maintaining typical word length of 3. The main idea is to summarize the sequence in a matrix of three rows corresponding to three reading frames where each row is the distribution on the (non- overlapping) words of length 3 for the corresponding reading frame. Compared to FFP-3, our empirical study showed that the proposed TUP method can be useful to build phylogenetic trees with strong biological support.

An integrative method for comprehensively reconstructing transcripts

Dan Li, Mary Yang

The University of Arkansas at Little Rock

Reference-guided approach is often used to reconstruct human transcriptome. Without using a reference genome, de novo method enables novel transcripts discovery. Here we assessed the assemblers built from these two types of assembly methods, using simulated data and experimental RNAseq data, for long non-coding RNAs (lncRNAs) identification. Moreover, we developed an integrative approach, combining the results from different assemblers, to identify a more comprehensive lncRNA set. Compared to mRNAs, lncRNAs are typically shorter with fewer exons and less abundance. Using Polyester R package, we generated RNAseq reads based on known lncRNAs annotations. The reference-guided and *de novo* assembler identified 62.5% and 72.9% of the known lncRNAs, respectively. The relative low discovery rates may attribute to rigorous filters applied to single exons. To reduce false positive, we removed single exons that were not overlapped with known annotations and mapped to low mappablity and alignment regions. In our integrative approach, all transfrags from multiple samples assembled by these two assemblers were used as input for Cuffmerge utility. Then, the resulting assemblies were merged together, which resulted in a more comprehensive single collection for the following lncRNA identification procedure. Using the integrative approach, over 75% of known lncRNAs were identified, 88.1% of these identified lncRNAs overlapped >80% length of the known lncRNAs. Additionally, an experimental RNSseq data set, consisting of RNAseq reads of 57 human tissue samples, were analyzed. The integrative approach detected 1.3 and 1.1 fold more lncRNAs compared to reference-guided and de novo method, respectively. Thus our integrative approach outperformed the individual methods.

Acknowledgment: This work was supported by grants from NCRR (P20RR016460), NIGMS (P20 GM103429), 1R15GM114739 and Arkansas Science and Technology Authority (ASTA)

Abstract Identifying Number: 1006308

ADVANCING PINE GENOMICS: IDENTIFICATION AND CHARACTERIZATION OF WOOD FORMATION GENES FROM PINUS TAEDA BAC CLONES

Dinum Perera[1], Zenaida V. Magbanua[1], Mark Arick II[1], Philippe Chouvarine[1], C. Joseph Nairn[2], Jeffrey Dean[3], Jeremy Schmutz[4], Jane Grimwood[5], and Daniel G Peterson[1]

[1] Institute for Genomics, Biocomputing & Biotechnology, Mississippi State University, Mississippi State, MS 39762
[2] Warnell School of Forest Resources, University of Georgia, Athens, GA 30602
[3]Dept. of Biochemistry, Molecular Biology, Entomology, and Plant Pathology, Mississippi State University, Mississippi State, MS 39762
[4]US Department of Energy Joint Genome Institute, Walnut Creek, CA 94598
[5]HudsonAlpha Institute for Biotechnology, 601 Genome Way, Huntsville, AL 35801

The identification and characterization of essential gene structures underlying the chemical composition of major cell wall components such as cellulose, hemicellulose, and lignins should provide insight into wood properties in pines and other conifers that are a major source of timber, pulp, paper, and many other fiber-derived products. One hundred bacterial artificial chromosome (BAC) clones including fifty that were targeted for the above genes, twenty five from low-copy regions, and twenty five random clones were selected for sequencing and analysis from the loblolly pine (LP) BAC library that our group previously constructed. After masking for previously characterized repeats, combinations of similarity-based and ab initio gene prediction approaches were utilized to characterize coding regions in the BAC sequences. We discovered and characterized eight genes and a transcription factor related to wood formation: they are two GDP-mannose pyrophosphorylases (i.e. GMP2 and GMP1), caffeoyl-CoA O-methyltransferase (CCoAOMT), laccase (LAC7), korrigan endoglucanase (KOR1), two cellulose synthases (CesA1 and CesA2), phenylalanine ammonia lyase (PAL) and MYB transcription factor (MYB8). In addition, we have identified one hundred and five gene models including both putative protein-coding genes and likely pseudogenes. Information derived from this investigation, more specifically genes related to wood formation, may be utilized for genetic modification of LP wood properties through breeding or genetic engineering.

Funding source: National Science Foundation and United States Department of Agriculture

Abstract Identifying Number: 1006301

TIME AND VIDEO ANALYSIS OF VIRTUAL ARTHROSCOPIC TEAR DIAGNOSIS AND EVALUATION PLATFORM (VATDEP)

Doga Demirel[1,2], Alexander Yu[2], Tansel Halic[2], Sinan Kockara[2], Ahmadi Shayyar[3], and Larry J.Suva[3]

[1]Department of Computer Science, University of Arkansas at Little Rock, Little Rock, Arkansas, 72204
[2*]Department of Computer Science, University of Central Arkansas, Conway, Arkansas, 72035 72204
[3]University of Arkansas Medical Science, Little Rock, Arkansas, 72205

Shoulder arthroscopy is a minimally invasive surgery for diagnosis and treatment of the tissues/joints in the shoulder area. Surgery training in shoulder arthroscopy is challenging due to the constrained instrument motions attributed to the limited and narrow area within the shoulder, unconventional hand-eye coordination, and limited field of view arising from angle of the arthroscope. Conventional training methods such as mannequins and cadavers have limited use, low-fidelity in realism and are not cost efficient. However, Virtual Reality (VR) based surgical simulators offer a realistic, low cost, risk-free training and assessment platform where the trainees can repeatedly perform tasks and receive quantitative feedback on their performances. Training with VR simulators allow improvement in operating room performances and reduce the learning curve. In our ongoing study, we are developing a VR arthroscopic rotator cuff diagnosis and repair surgery simulator (VATDEP). As the first step of creating the VATDEP, we have established a Hierarchical Task Analysis (HTA). According to our HTA, we have analyzed the procedure and derived tasks/subtasks and goals associated with ideal and optional actions and performance scoring metrics. Scoring metric will be used for a quantitative evaluation of the VATDEP. In this work, we are presenting HTA, time and performance analysis of surgery videos from surgeons at different skill levels. The ultimate goal is to create a standard and valid scoring that identifies the skill level of a shoulder arthroscopy surgeon.

Abstract Identifying Number: 1006072

APPROACHING PRECISION ONCOLOGY WITH TISSUE PROTEOMICS

Donald J. Johann, Jr[1], **Josip Blonder**[2]

[1]Department of Biomedical Informatics and Hematology/Oncology, University of Arkansas for Medical Sciences, Little Rock, AR 72205, [2]Leidos Biomedical Research, Inc, Frederick National Laboratory for Cancer Research, National Cancer Institute, Frederick, Maryland 21702.

Background:

The vast majority of therapy plans for cancer patients are derived categorically. Currently, large and time lengthy clinical trials supply the oncology community with aggregate population and survival statistics. Although this approach has been successful there are known shortfalls. More specifically, since the inherent molecular heterogeneity of the solid tumor under study has not been examined, it is not surprising that results may vary widely across individual patients. Therefore, what is needed is a rationally-based molecular profiling of an individual patient's tumor, in a timely manner. The aim is to assign the best drug(s) to the individual patient's cancer based on objective molecular data, thus maximizing therapeutic efficacy and minimizing toxicity.

Methods:

Presented is a new form of a protein-based molecular profiling approach developed utilizing LC/MS-MS and novel computational methods utilizing breast tumor specimens. First, to better illustrate the molecular circuitry of the tumor, proteomic data was rendered as a virtual wiring diagram. This was done on a per tumor basis thus graphically illustrating the molecular traffic intensities. Second, computationally deriving a numeric similarity score is generated to allow for easier comparison of molecular circuitry.

Results:

Biologically relevant proteins (p53, ER, Her2, Ki67) including driver oncoproteins were found. Effective molecular profiling of breast cancer at the protein level was achieved and the established clinical tumor subtypes were discriminated by unsupervised clustering. A visualized simulation of an estrogen carcinogenesis network was also developed. Experimentally identified proteins and expression levels were mapped and objectified onto known cancer pathways contained in the network.

Conclusions:

Since the tissue quantity and time requirements are the same as IHC the approach is clinically practical. The approach is discovery based and complementary to IHC but is not dependent on antibodies. This method shows promise as a research tool in a clinical trial.

Abstract Identifying Number: 1006071

MAXIMIZING THE POWER OF CELL TYPE-SPECIFIC DIFFERENTIAL EXPRESSION DETECTION IN HETEROGENEOUS GENE EXPRESSION MEASURES

Edmund Glass[1], Mikhail Dozmorov[1]

[1*]Department of Biostatistics, Virginia Commonwealth University, Richmond, Virginia, 23219

Heterogeneous tissues, like blood, contain of a mixture of cells. Thus, measuring differences in gene expression level in heterogeneous tissues (e.g., healthy controls vs. disease cases) represents a challenge in deciding which cell type contributes to the observed gene expression differences. Gene expression level in each proportion of cells contributes to the heterogeneous gene expression level, which can be modeled using simple linear regression. However, several parameters affect the power of linear regression in detecting differentially expressed genes on a cell type-specific level. Specifically, sample size, collinearity (interdependence) among cell proportions, and "goodness of fit" of linear regression affect the power of cell type-specific differential expression detection, measured as an area under receiver operator characteristic curve (AUROC).

Increasing sample size expectedly increased the power of linear regression in detecting cell type-specific gene expression differences. Low collinearity among cell proportions also positively affected AUROC. These parameters maximize gene-specific "goodness of fit" of regressing heterogeneous gene expression measures on cell proportions by minimizing root mean squared error.

We also outline the conditions when "goodness of fit" increases with the use of orthogonal linear regression to detect cell type-specific gene expression differences.

Applied to high-throughput 'omics' data, our results allow maximizing true positive rate of detection of cell-type specific gene expression differences.

Abstract Identifying Number: 1006273

Quantitative Structure-Activity Relation Study of Quaternary Ammonium Compounds in Pathogen Control: Computational methods for the discovery of food antimicrobials

Ethan C. Rath[1] **and Yongsheng Bai**[1]

[1] Department of Biology, Indiana State University, Terre Haute, Indiana, 47809

Quaternary ammonium compounds (QACs) are surfactants that are made of at least one cationic nitrogen attached to a variety of different side groups, usually one or more hydrophobic chains. These compounds are generally used for surface decontamination, oral hygiene, and even in carcass preservation. Recently there have been a large number of studies that have implicated QACs in the development of resistance in bacteria as well as in harmful environmental effects. One compound in particular, cetylpyridinium chloride (CPC), has recently gained acceptance as a safe and practical method for use in consumable raw poultry product decontamination. This compound is highly lipophilic and leaves a residue that is typically removed by polyethylene glycol (PEG). The use of PEG, as well as the remaining residue, is potentially toxic to consumers if not properly removed. Using computational methods, we propose the use of quantitative structure-activity relation (QSAR) analysis to determine the antimicrobial effects of novel and untested QACs and QAC-like structures for further testing. By diversifying the available QACs we hope to develop better disinfectants, create more environmentally friendly compounds, and help to stall, or even halt, the development of resistance.

Abstract Identifying Number: 1006266

A FRAMEWORK FOR EVALUATING THE QUALITY OF THE PERSONAL GENOMES GENERATED BY DE NOVO ASSEMBLY TOOLS

Gokhan Yavas[1], Leihong Wu[1], Huixiao Hong[1], Weida Tong[1] and Wenming Xiao[1]

[1]National Center for Toxicological Research, US Food and Drug Administration, Jefferson, AR 72079

With the advent of the Next Generation Sequencing (NGS) technologies, it is now possible to generate millions of sequencing reads from a human genome, which can then be used towards many applications such as identifying the single nucleotide variations (SNVs) and structural variations. Another promising application is the *de novo* assembly of reads to build a personalized genome. For this purpose, many tools have been developed to build a novel genome or to evaluate the quality of assembly outcomes. Evaluation of assembly usually depends on the alignment of contigs to a reference genome, which demands greatly on computational resources such as runtime memory, CPU and storage space. Based on several comparative studies on existing assembly evaluation tools, it remains as a big challenge to architect a good framework with high runtime performance. Another limitation for these existing tools is that the alignment results for each contig may not be processed correctly and thus their derived quality statistics are inaccurate. Furthermore, some key quality parameters are lacking with existing assessment tools.

In this study, we present a framework that can maximize the usage of available computational environment by performing contig alignment and post processing in parallel. Our flexible design allows split jobs being run either on a high performance computing (HPC) cluster or a multi-core workstation. The input, given in the form of a set of contigs, can be partitioned into a user-defined number of chunks, each of which can then be aligned and processed in either the separate nodes of a HPC cluster or separate cores of a workstation. Based on carefully filtered alignment, it generates statistics such as the total genome coverage, gene and exon coverage, contig duplication and continuity as well as SNVs embedded in the assembly and SNV related statistics. Our framework also provides stand-alone quality statistics such as contig size distribution, N*x* statistics, etc. We compared multiple genomes assembled via various assembly algorithms such as SOAPdenovo, Falcon, and Celera assembler. The results demonstrated the capability of our tool for providing a complete package of quality metrics with high performance on different settings of computer environment.

A TEMPLATE-BASED PROTEIN STRUCTURE RECONSTRUCTION METHOD USING DEEP AUTOENCODER LEARNING

Haiou Li[1,2], Qiang Lyu[1], and Jianlin Cheng[2, *]

[1] Department of Computer Science and Technology, Soochow University, Suzhou, China,215006
[2*]Computer Science, Informatics Institute, University of Missouri, Columbia, USA, 65201

Protein structure prediction is an important research problem in computational biology, and is widely applied to various biomedical problems such as protein function study, protein design, and drug design. In this work, we developed a novel deep learning approach based on the deeply stacked denoising autoencoder for protein structure reconstruction. We applied our approach to template-based protein structure prediction using only the 3D structure coordinates of homologous template proteins as input without the need of additional constraints. The templates were identified for a target protein by PSI-BLAST search. 3DRobot was used to generate initial decoy models for the target from the templates. A stacked denoising autoencoder was trained on the decoys to obtain a deep learning model for the target protein. The trained deep model was then used to reconstruct the final structural model for the target protein. Benchmarked on the target proteins that have highly similar template proteins, the GDT-TS score of the predicted structures is greater than 0.95, suggesting that the deep autoencoder is a promising method for protein structure reconstruction. *(The work was partially supported by an NIH grant (R01GM093123) to JC).*

Abstract Identifying Number:1002126

COMPARATIVE GENOMICS ANALYSIS OF *AEROMONAS HYDROPHILA* STRAINS

Hasan C. Tekedar, Safak Kalindamar, Attila Karsi and Mark L. Lawrence

College of Veterinary Medicine, Department of Basic Sciences, Mississippi State University, Mississippi State, MS 39762

Aeromonas hydrophila is a Gram-negative facultative anaerobe ubiquitous in several environments, and it can cause infection in aquatic poikilothermic animals, mammals, and humans. Historically, *A. hydrophila* is an opportunistic pathogen of freshwater fish, but since 2009 the U.S. channel catfish industry has been affected by epidemics where *A. hydrophila* is a primary pathogen. We recently reported completed genome sequencing of an *A. hydrophila* strain isolated from one of these outbreaks in a commercial catfish pond, ML09-119. For comparison, we also completed the genome sequence of an *A. hydrophila* strain, AL06-06, which was isolated from a goldfish in 2006 from the Auburn University Southeastern cooperative fish disease Laboratory in Greensboro, Alabama. In the current study, we compared these two genomes to the genome sequence of *A. hydrophila* SSU, a human clinical isolate (Genbank # NZ_JH815591.1), *A. hydrophila* ATCC 7966 (Genbank # CP000462) and other *Aeromonas* strains. The catfish epidemic isolate ML09-119 has several unique features, including inositol catabolism pathway and specific toxin clusters. On the other hand, *A. hydrophila* strain SSU has unique type IV and type VI secretion systems, and it has mercury resistance genes. Strain AL06-06 carries unique type VII secretion elements and phage elements. Taken together, comparative genomics analysis results of *A. hydrophila* and *Aeromonas* strains from different hosts reveals genetic elements that may contribute to host adaptation and molecular mechanisms of virulence. *This work was supported by the Mississippi State University College of Veterinary Medicine, the USDA Agricultural Research Service CRIS project 6402-31000-009-00D, and the Alabama Agricultural Experiment Station (Hatch project number ALA021-1-09005).*

Abstract Identifying Number: 1006249

DE NOVO TRANSCRIPTOME ASSEMBLY OF *RAUWOLFIA SERPENTINE* REVEALS NOVEL TRANSCRIPT RELATED TO ALKALOID BIOSYNTHESIS AND GENE DISCOVERY

Hithesh Kumar[1], Rahul Yadav[1], Smrithy Simon[1], Shashi kumar[2], Manjunath Dammanlli[1] and Vivek Chandramohan[1*]

[1*]Biotechnology Finishing School (BTFS), Department of Biotechnology, Siddaganga Institute of Technology, Tumkuru, Karnataka, 572103
[2]GenEclat Technologies, Bengaluru, Karnataka, 560056

Rauwolfia serpentine is a medicinal plant which belongs to the Apocynaceae family, also known as'Sarpagandha'in Sanskrit. Recent studies shows the presence of various alkaloids in *R.serpentine* with therapeutic values. This current study focuses on the transcriptomic analysis of *R.serpentine,* as, there is no sufficient Transcriptomic and genomic data were available in public databases. The raw Transcriptomic reads of *R.serpentine* were downloaded from NCBI SRA (Sequence Read Archive) database with Accession number SRA045782 was used for analysis beginning with quality checking using FastQC tool. The raw data were trimmed using RNA-Rocket in Galaxy server. The total number of bases before and after trimming was 49,38,52,836 and 49,02,19,109 respectively. The trimmed reads were assembled using Seven Bridge Genomics online Trinity automated pipeline and offline Velvet oases Tool. The total trinity transcripts of 127035 and 97172 trinity genes with average length of 974.75 and N50 contig of 1822 were discovered. The GC content was calculated to be around 39.89%. CD-HIT (Cluster Database at High Identity with Tolerance) tool was used for clustering the Genes and Isoforms. The transcripts were extracted and used as queries in BlastX against the Ref-Seq protein database. Blast2GO tool was used for functional annotation of the obtained transcript sequences. SSR (Simple Sequence Repeat) identification and biosynthesis pathway were done using MISA (MIcroSAtellite identification) tool and KASS (KEGG Automatic Annotation Server) server to elucidate the genes involved in mono-terpenoid biosynthesis. The results obtained through various online as well as offline tools were validated using CLC Genomics Workbench tool. The transcripts generated here provides a resource for gene discovery and development of functional molecular markers.

Abstract Identifying Number: **100137 (poster presentation)**

FDALABEL DATABASE: A RICH RESOURCE FOR STUDY OF PHARMACOGENOMICS BIOMARKERS TO FACILITATE PRECISION MEDICINE AND DRUG SAFETY

Hong Fang, Joshua Xu, Zhichao Liu, Stephen Harris, Shraddha Thakkar, Guangxu Zhou, Daojun Liu, Paul Howard, Weida Tong

National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, AR 72079

Pharmacogenomics (PGx) is the study of individual genetic differences (acquired and inherited) in correlation to drug response. Understanding the association between PGx markers and phenotypes improves knowledge of underlying mechanisms of diseases and treatment responses for enhanced drug safety and precision medicine. Research on this topic has been a challenge, because of lack of easy access to PGx data. We have developed the FDALabel database (http://www.fda.gov/ScienceResearch/BioinformaticsTools/ucm289739.htm), which allows users to perform customizable, full-text searches in over 80,000 drug labeling documents for 1600 small molecule drugs. FDALabel provides PGx information contained in the FDA-approved drug labeling (package insert). The prescription drug insert information provides consensus and combined information about product indications, target populations, and adverse drug reactions (ADRs) from FDA regulators, drug manufacturers, and scientific experts. In this study, biomarker information was used as query on the relevant PGx sections in FDALabel. As a result, we identified more than 170 drugs with genetic biomarker information. Furthermore, 36 biomarkers were identified and divided into three categories (1) drug metabolism variability (*e.g*., CYP enzymes); (2) increased risk of adverse events (*e.g.,* G6PD, TPMT, HLA-B); (3) drug's mechanism of action (*e.g*., CD30). These biomarkers are likely to impact the specified patient sub-population response to the drug. Network analysis and visualization were used to illustrate the relationship amongst drugs, biomarkers, and associated adverse effects. In summary, the case of PGx biomarkers has demonstrated the potential of using FDALabel for the study of ADRs (*i.e.*, to identify new trend and frequency of genetic variability associated with increased risks to public health) in pursuit of improved pharmacovigilance and precision medicine.

The Development of a QSAR Model for Predicting Estrogen Receptor-α Binding Using Large Data Sets

Hui Wen Ng and Huixiao Hong*

Division of Bioinformatics and Biostatistics, NCTR/FDA, Jefferson, AR

*Correspondence: Huixiao.Hong@fda.hhs.gov; 870-543-7296

Abstract

Many chemicals have been found to interact with the endocrine system in the human body. The estrogen receptor α (ERα), which is amongst the most studied receptors, plays very important roles in the ER-mediated endocrine disrupting activities. Understanding and predicting estrogenic activity of chemicals facilitates the evaluation of their endocrine activity. We developed a classification model using Decision Forest (DF) to predict ERα binding potential of chemicals. The DF model was developed using a large training data set of 3308 chemicals obtained from the U.S. Food and Drug Administration's Estrogenic Activity Database. The model was tested through cross validations and also using external data sets of 1641 chemicals obtained from the U.S. Environmental Protection Agency's ToxCast project. The model showed good performance in both internal (87% balanced accuracy) and external validations (~70–89% relative balanced accuracies), where the latter involved the validations of the model across different ER pathway-related assays in ToxCast. The important chemical features that contribute to the prediction ability of the model were identified through informative descriptor analysis and were related to current knowledge of ER binding. Prediction confidence analysis revealed that the model had both high prediction confidence and accuracy for most predicted chemicals. In summary, the results demonstrated that the model accurately predicted ER binding of chemicals and could be useful for assessing the ER-medicated endocrine activity potential of environmental chemicals.

Abstract Identifying Number: 100335

PREDICTIVE TOXICOLOGY IN REGULATORY SCIENCE: ENDOCRINE DISRUPTORS KNOWLEDGE BASE

Huixiao Hong, Hui Wen Ng, Sugunadevi Sakkiah, Hong Fang, Roger Perkins, Weida Tong

Division of Bioinformatics and Biostatistics, NCTR/FDA, Jefferson, AR

Regulatory science has been around for many years and means different things to different people. In the FDA mind, "advancing regulatory science" is to develop new tools, standards and approaches to assess the safety, efficacy, quality and performance of FDA-regulated products. Predictive toxicology is the methodology that shifts adverse effects observation using experimental animals to alternative measurements of safety of chemicals such as in vitro assays and in silico models. Predictive toxicology offers the potential to replace, refine and reduce animal usage and has thus been widely explored for applications in regulatory science. Endocrine disruptors (EDs) can potentially have adverse effects on both humans and wildlife. They can interfere with the body's endocrine system through direct or indirect interactions with many endocrine system protein targets. The public and regulatory concerns over EDs led to regulatory actions and expanded research across Europe, Japan, and North America. In response to the emerging concerns on EDs, we initiated and developed the Endocrine Disruptors Knowledge Base (EDKB) project to demonstrate the applications of predictive toxicology in regulatory science. During the 20 years of this project, we have assayed a large number of chemicals for their binding activity to four proteins that play important roles in assessing endocrine disruption potential (e.g. estrogen receptor, androgen receptor, alpha-fetoprotein and sex hormone binding globin), curated comprehensive data (e.g. binding of different receptors, reporter gene assays, cell proliferation assays and in vivo) on endocrine disruption potential for more than 8,000 chemicals, designed sophisticated databases to efficiently utilizing the data, developed computational tools and algorithms (e.g. Decision Forest) to construct predictive models, built a variety of ED prediction models and applied them to numerous real regulatory applications. Our continuous effort on the development of EDKB is expected to facilitate more and better ED-related regulatory science both for FDA-regulated products and environment regulatory applications.

Abstract Identifying Number: 1006028

EFFECTIVE COMPARATIVE ANALYSIS OF PROTEIN-PROTEIN INTERACTION NETWORKS BY MEASURING THE STEADY STATE NETWORK FLOW USING A MARKOV MODEL

Hyundoo Jeong[1] and Byung-Jun Yoon*[1,2]

[1]Department of Electrical and Computer Engineering,
Texas A&M University, Bryan, Texas, 77801
[2]College of Science and Engineering,
Hamad bin Khalifa University (HBKU), Doha, Qatar

Comparative analysis of protein-protein interaction (PPI) networks provides an effective means of detecting conserved functional modules across species. Such modules typically consist of orthologous proteins with conserved interactions, which can be exploited to computationally predict the modules through network comparison. In this work, we propose a novel probabilistic framework for comparing PPI networks and effectively predicting the correspondence between nodes that belong to functional network modules that are conserved in the given PPI networks. The basic idea is to estimate the steady state network flow between nodes that belong to different networks based on a Markov random walk model. The random walker is designed such that it can make random moves to adjacent nodes within a PPI network as well as cross network moves between potential orthologs with high sequence similarity. Based on this Markov model, we estimate the steady state network flow – or random transitions – between nodes in different PPI networks, which can be used as a probabilistic score measuring their potential correspondence. Subsequently, the estimated scores can be used for detecting conserved network modules through network alignment. Through simulations using real PPI networks, we demonstrate that the proposed scheme leads to more accurate alignment results at reduced computational cost, outperforming the current state-of-the-art algorithms.

Acknowledgement:
This work was supported in part by National Science Foundation through the NSF Award CCF-1149544

Abstract Identifying Number: 1006262

VDJML – TOOLS FOR CAPTURING THE RESULTS OF INFERRING IMMUNE RECEPTOR REARRANGEMENTS

Inimary T. Toby[1], Felix Breden[2], Adam Buntzman[3], Brian Corrie[2], John Fonner[4], Namita Gupta[5], Uri Hershberg[6], Christopher Jordan[4], Min Kim[1], Steven H. Kleinstein[5], Nishanth Marthandan[2], Stephen A. Mock[4], Nancy Monson[1], William Rounds[1], Manuel Rojas[4], Aaron Rosenfeld[6], Florian Rubelt[7], Walter Scarborough[4], Richard Scheuermann[8], Jamie Scott[2], Mohamed Uduman[5], Jason Vander Heiden[5], Lindsay G. Cowell[1]

[1]University of Texas Southwestern Medical Center, Dallas TX, USA
[2]Simon Fraser University, Burnaby, BC, Canada
[3]University of Arizona, College of Medicine, Tucson, AZ, USA
[4]Texas Advanced Computing Center, University of Texas, Austin, TX, USA
[5]Yale University, New Haven, CT, USA
[6]Drexel University, Philadelphia, PA, USA
[7]Standford University School of Medicine, Stanford, CA, USA
[8]J. Craig Venter Institute, San Diego, California, USA

V(D)J recombination, also known as somatic recombination, is a mechanism of genetic recombination that occurs in vertebrates, which randomly selects and assembles segments of genes encoding specific proteins with important roles in the immune system. This site-specific recombination reaction generates a diverse repertoire of T cell receptor (TCR) and immunoglobulin (Ig) molecules that are necessary for the recognition of diverse antigens from bacterial, viral, and parasitic invaders, and from dysfunctional cells such as tumor cells. Despite the widespread use of immune repertoire profiling, there is currently no standardized format for output files from VDJ analysis. Researchers utilize software such as IgBlast and IMGT/High V-Quest to perform VDJ analysis and infer germline rearrangements. Each of these software tools produces results in a different file format, and can identify the same result using different labels. These differences make it challenging for users to perform additional analysis using the output file from one software to the next. We have addressed this problem by developing a standardized file format for representing results. The purpose of VDJML is to provide a common standardized format for different VDJ analysis applications and to facilitate downstream processing of the results in an application-agnostic manner. The VDJML file format is accompanied by a suite of analysis tools, which are accessible via command line and written in C++ and python. The VDJML suite will allow users to streamline their VDJ analysis and facilitate the sharing of scientific knowledge within the community. The VDJML suite and documentation are available from https://www.vdjserver.org/software. We welcome participation from others in developing the file format standard, as well as code contributions. *This project is supported by a grant from the National Institute of Allergy and Infectious Diseases (AI097403) and the Burroughs Wellcome Fund.*

Abstract Identifying Number: 1002151

AN INTEGRATED STATISTICAL PROBE OF ERROR CORRECTION METHODS FOR NEXT-GENERATION SEQUENCING DATA

Isaac Akogwu[1], Nan Wang[1], Ping Gong[2], Chaoyang Zhang[*1]

[1*] School of Computing, University of Southern Mississippi Hattiesburg, MS, 39406
[2] Environmental Laboratory, U.S. Army Engineer Research and Development Center,
Vicksburg, MS, 39180

Error correction of next-generation sequencing (NGS) data is invaluable for downstream data analysis such as genome assembly and identification of genetic polymorphism. Numerous methods have been developed making it imperative to comprehensively examine the effects of data properties and correction algorithms on method performance.

Twenty-seven datasets of paired-end sequencing reads with a 2% error rate and lengths of 50-, 150- and 250-bp were simulated using ART, an NGS read simulator. Reference genomes of Escherichia coli, Human Chromosome 21 and Drosophila melanogaster were used for the simulation with genome coverage depths of 20, 80 and 320 fold. Error correction was performed using six well-known error correction algorithms. Results were statistically analyzed to determine the effects of genome coverage, read length, genome size and the algorithms on correction performance measured by f-score, precision, and gain metrics

Factorial analysis of performance metrics and post hoc Tukey test indicated that read length and correction algorithm both had a significant impact on the result. Lengths of 150-bp and 250-bp produced the best results while Bfc, Bless and Bloocoo tools had better performance. Furthermore, we discovered that shorter read length, higher genome coverage and larger genome size generated negative gains as more errors were introduced than corrected. Lighter performed best with shorter reads. Bfc had the shortest correction time while Bless used the lowest amount of memory overall. Despite producing comparable results, Musket consumed the most time and memory among all the tools. Trowel required a specified k-mer range to produce accurate corrections making it most applicable to short reads.

This work demonstrates that all the tested tools possess certain strengths and weaknesses depending on the type of data being corrected. It also suggests that users should select error correction tools based on the computational resources available to them and their knowledge in tweaking the parameters specified by the chosen error correction method.

This work was supported by the US Army Corps of Engineers Environmental Quality Program under contract # W912HZ-08-2-0011 and the NSF EPSCoR Project (NSF #EPS-0903787) grant

Abstract Identifying Number: 1002151

BECOW: A WEB-BASED BIOINFORMATICS ERROR CORRECTION WORKFLOW TOOL FOR NEXT GENERATION SEQUENCE DATA CORRECTION

Isaac Akogwu[1], Nan Wang[1], Ping Gong[2], Chaoyang Zhang[*1]

[1*] School of Computing, University of Southern Mississippi Hattiesburg, MS, 39406
[2] Environmental Laboratory, U.S. Army Engineer Research and Development Center,
Vicksburg, MS, 39180

Complexity of command line based bioinformatics tools presents a monumental challenge to users, especially biologists without the necessary training to use such tools. This has led to increased availability of web-based frameworks over the years. However, there is no web-based tool for error correction of sequencing reads generated by Next Generation Sequencing (NGS) technologies.

Here, we introduce Bioinformatics Error Correction Workflow (Becow), a web-based sequence error correction workflow implemented in Python. It includes four well-recognized standalone error correction algorithms including Bfc, Bless, Bloocoo and Lighter, which were selected after a multivariate statistical analysis of the correction results was performed to choose the best performer(s) for any user-specified datasets. The workflow provides a form allowing a user to upload paired-end fastq and reference sequence files in addition to data specific parameters like genome-size and kmer-length. Upon submission, error correction is automatically performed and the best correction result and evaluation statistics are returned to the user. Becow has been evaluated using both simulated (generated using ART with 2% error rate) and experimental NGS (retrieved from Sequence Read Archive (SRA)) datasets. Statistical evaluation were performed using Error Correction Evaluation Toolkit (ECET). Further downstream analysis (genome assembly) was performed with the corrected data and the obtained results were comparable to those in the NCBI database.

Becow is designed to correct errors in NGS data, generate corrected data and produce statistical assessment of the correction. It may also be used to select the best correction method for a specific dataset. Current limitations of Becow include its applicability to only Illumina reads and the allowable maximum file size of 10Mb, both of which may be improved in the future.

This work was supported by the US Army Corps of Engineers Environmental Quality Program under contract # W912HZ-08-2-0011 and the NSF EPSCoR Project (NSF #EPS-0903787) grant

Abstract Identifying Number: 1006060

VISINT-X: VISUALIZING INTERACTIONS IN CROSS-LINKED PROTEINS

Islam Akef Ebeid[2], **Mihir Jaiswal**[1, 3], **Carolina Cruz-Neira**[2], **Boris Zybaylov**[3]

[1]Department of Bioinformatics, University of Arkansas at Little Rock, Little Rock, AR
[2]Department of Information Science, University of Arkansas at Little Rock, Little Rock, AR
[3]University of Arkansas for Medical Sciences, Little Rock, AR

Chemical cross-linking combined with mass spectrometry (CXMS) is an established method in protein chemistry to infer low-resolution 3D protein structures and topology of interactions in protein complexes. Interpretation of CXMS data is challenging due to two-fold identification problem. The first challenge is to identify interacting peptides in the background of non-interacting peptides and the second is to determine the site of interactions. Previously, we developed an algorithm called X-Link Peptide Mapping (XLPM) to analyze and interpret CXMS data. In this report we present VisInt-X, which processes XLPM output to visualize and to interpret cross-linking data. VisInt-X visualizes the results of the cross-linking analysis by including 2D representations of interacting partners, scores and sites of interactions, and by 3D structural modelling of the interactions onto protein structures. VisInt-X is the most comprehensive tool for detailed visualization of the chemical cross-linking data.

INBRE grants NCRR (P20RR016460) and NIGMS (P20GM103429) at NIH.

Abstract Identifying Number: 1006267

Digital Terrain Mapping as A Novel Disease Classification Tool: An Alzheimer's Disease Case Study

Itika Arora[1*], Zongliang Yue[1*], Thanh M. Nguyen[3], Wei Feng[4], Michael T. Neylon[1], and Jake Y. Chen[1,2,3,4]

[1] BioHealth Informatics Department, School of Informatics and Computing, Indiana University, Indianapolis, IN 46202
[2] Indiana Center for Systems Biology and Personalized Medicine, Indiana University, Indianapolis, IN 46202
[3] Department of Computer and Information Science, School of Science, Purdue University, Indianapolis, IN 46202
[4] Medeolinx, LLC, Indianapolis, IN 46202

Systems Biology has been increasingly used to characterize different disease states in clinical samples. Traditionally, researchers rely on DNA microarray analysis or RNA-seq analysis followed by statistical machine learning to perform sample classifications. These methods often face significant challenges including over-fitting, curse-of-dimensionality, and lack of robustness in prospective studies; moreover, they are difficult to use by regular biological researchers without informatics expertise.

In this study, we developed a new clinical sample classification technique, "Digital Terrain Mapping (DTM)", based on performing systems biology characterization and visual analytics of functional genomics study results. We hypothesize that DTM-based classification can achieve good performance with fewer training samples while overcoming the inherent weaknesses of conventional approaches. To establish our claim, we collected differential gene expression data from the Gene Expression Omnibus (GEO) for an Alzheimer's disease (AD) case study (dataset GSE28146). The original study used laser capture microdissection to selectively collect only CA1 hippocampal gray matter from formalin-fixed, paraffin-embedded hippocampal sections of the patients followed by DNA microarray analyses. We processed the entire gene expression data set containing 20,638 genes from 15 disease (7 severe cases and 8 moderate cases combined) and 7 control samples, using a standard microarray data analysis toolkit from R/Bioconductor. Then, we constructed an AD-specific molecular interaction network using the new GeneTerrain software to generate DTM images for each (case or control) sample or pooled (case and control) sample. To evaluate the effectiveness of using pooled DTM images as the disease or control classes, we divided all the samples into training (5 Case Samples+2Normal Samples) and testing sets (10 Case Samples+5 Normal Samples) for 50 iterations. DTM can achieve a classification performance (sensitivity*specificity=0.69), compared with using conventional microarray feature selection and gene-based classifications (sensitivity*specificity=0.34). Our finding implies the significance of developing DTM as a general framework towards disease biomarker discovery.

[*] These authors contributed equally.

De novo protein conformational sampling using a probabilistic graphical model

Jianlin Cheng, University of Missouri, Columbia, Missouri

Efficient exploration of protein conformational space remains challenging especially for large proteins when assembling discretized structural fragments extracted from a protein structure data database. We propose a fragment-free probabilistic graphical model, FUSION, for conformational sampling in continuous space and assess its accuracy using 'blind' protein targets with a length up to 250 residues from the CASP11 structure prediction exercise. The method reduces sampling bottlenecks, exhibits strong convergence, and demonstrates better performance than the popular fragment assembly method, ROSETTA, on relatively larger proteins with a length of more than 150 residues in our benchmark set. FUSION is freely available through a web server at http://protein.rnet.missouri.edu/FUSION/.

Abstract Identifying Number: 1006306

COMPREHENSIVE ASSESSMENT OF HEPATOTOXICITY INDUCED BY HERBAL AND DIETARY SUPPLEMENT

Ji-Eun Seo[1,2], Rodney Ballard[1], Oh-Seung Kwon[2], Weida Tong[1], and Minjun Chen*[1]

[1*]Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, US Food and Drug Administration, Jefferson, Arkansas 72079, USA
[2]Toxicology lab, Doping Control Center, Korea Institute of Science and Technology, Seoul 02792; Biological Chemistry, University of Science and Technology, Daejeon 34113, Republic of Korea

Herbal and dietary supplements (HDS) have been widely cited as the cause of drug-induced liver injury in clinic; however, in most countries HDS is generally not categorized as drug and its regulation is less strict. The rise in use of HDS in the U.S. market increases the possibilities to induce liver injury. Moreover, the use of HDS is more popular in women and minority population that consequently leads higher risk of hepatotoxicity in the groups. In this study, we conduct a comprehensive assessment of hepatotoxicity of HDS in literature to better understand its impact on public health, especially for women health and minority population. Hepatotoxic HDS information was gathered from various literature data and web sources, and approximately 70 hepatotoxic HDS were organized along with the application, proposed toxin/active ingredient, mechanism, gender cases of HDS, and the consumption of HDS by minorities. We found a number of HDS were associated with hepatocellular, cholestatic, or mixed type liver injuries, and the clinical patterns of hepatotoxic HDS were quite variable, even for the same HDS. Some hepatotoxic HDS such as aloe vera and kava kava showed herbal-drug interaction. Besides, cocaine, pennyroyal oil, kava kava, and skullcap were specifically linked to cytochrome P450 metabolism. Hepatotoxic HDS were predominant in Hispanic, Asian, and African female, but the incidence of liver toxicity from HDS is still difficult to estimate and uncertain to confirm causality. Accordingly, to make high quality database for hepatotoxic HDS, the appropriate causality study of the hepatotoxic HDS is needed and the research and the database development for the case report are ongoing.

Abstract Identifying Number:1006259

SPIROPLASMA RELATEDNESS THROUGH THE INVESTIGATION OF VIRAL INSERTS

Jordan Fansler[1], Astri Wayadande[2] and Ulrich Melcher*[1]

[1*]Department of Biochemistry and Molecular Biology, Oklahoma State University, Stillwater, Oklahoma,74078
[2]Department of Entomology and Plant Pathology, Oklahoma State University, Stillwater, Oklahoma, 74078

Spiroplasma bacteria have a high susceptibility to invasion from viral genomes. The multiple viral inserts are targets for rearrangement and deletion, calling into question traditional bifurcating taxonomic classifications. To determine whether the insertions in contemporary genomes preceded separation of *Spiroplasma* species or are recent events, we characterized viral inserts from available *Spiroplasma* genomes. To test this, genomic sequences for the available fifteen species were acquired from GenBank, viral inserts identified as to positions of insertion in both viral and host genomes. This allowed a histogram to be created to study the exact viral genome location of each insert. The surrounding host sequences were compared among species to identify any common coding regions in or around each viral insert. These results were compared to the results of phylogenetic analysis of selected spiroplasma and viral genes. These studies provide insight into the relationships among species of the *Spiroplasma* genus suggesting close relatedness between several of the species, especially *S. citri* and *S. kunkelii*. *Funding provided by Oklahoma State University.*

Abstract Identifying Number: 100333

DEVELOPING AN INTELLIGENT RECOGNITION SYSTEM FOR STORAGE PEST FRAGMENTS CONTAMINATING FOOD PRODUCTS

Joshua Xu[1,*], Hongjian Ding[2], Suinn Park[3], Daniel Marin[4], Howard G. Semey[2], Monica Pava-Ripoll[5], Amy E. Barnes[2], Darryl A. Langley[2], Zhichao Liu[1], Himansu J Vyas[2,*], and Weida Tong[1]

[1]Division of Bioinformatics and Biostatistics, NCTR, FDA, Jefferson, AR 72079
[2]Arkansas Regional Laboratories, ORA, FDA, Jefferson, AR 72079
[3]Samsung Austin Semiconductor, LLC, Austin, TX78754
[4]School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ 85281
[5]Office of Food Safety, CFSAN, FDA, College Park, MD 20740
*Contact: Zhihua.xu@fda.hhs.gov

Food contamination by insects is quite omnipresent. Some insects are vectors of foodborne pathogens or indicators of insanitation. During food processing, insects are usually broken into small fragments which can be recovered and recognized by trained FDA food analysts. Correct identification of such fragments is essential to the agency's decision making. However, the identification can be very time consuming and difficult as some morphological characteristics are destroyed during food manufacturing procedures. This project, initiated by the food analysts at FDA's Arkansas Regional Laboratories (ARL), addresses an urgent need for a fast, accurate method to detect and identify beetle species based on microscopic image of elytral fragments recovered from processed food products. The proposed system will increase the reliability, sensitivity, and throughput compared to methods currently used by the FDA for the identification of storage insect fragments found in FDA-regulated food products. This poster presents the pilot study design and promising results for the development of image analysis and machine learning algorithms. It also outlines future plan for study expansion and system development. This project has great practical significance in protecting public health through empowering the agency with more efficient tools for food safety. *This work is supported by FDA/ORA project IR01048 and FDA/NCTR protocol E0759101.*

Abstract Identifying Number: 1006281

LARGE-SCALE MICROARRAY DATA INTEGRATION FOR IMPROVED DIFFERENTIAL EXPRESSION ANALYSIS

Kevin Allan Townsend[1], **Bernie J. Daigle, Jr.**[1,2]

[1] Department of Computer Science, The University of Memphis, Memphis, TN
[2] Department of Biological Sciences, The University of Memphis, Memphis, TN

Public biomedical repositories, such as the National Center for Biotechnology Information's Gene Expression Omnibus (GEO), contain a growing abundance of genomic data. GEO alone contains 1,421,394 samples, of which 525,277 pertain to human gene expression microarrays.* Although there is a plethora of data available, a problem arises of efficiently extracting biological information from these vast collections of datasets. Currently, these repositories are not easily amenable to the large scale extraction of data to perform analyses.

We have developed an automated computational pipeline to extract mass quantities of microarray data from GEO and organize it in an efficient, structured, and local database. Our pipeline enables large scale differential expression analyses between all biological conditions represented in the database. Currently, we are conducting a preliminary analysis on a 107,606 sample subset of GEO.

The results of our preliminary analysis will allow us to estimate the prior probability of differential expression for each gene in the human genome. In the future, we will develop an enhancement of the popular limma differential expression analysis tool that incorporates this information. In general, the accuracy of limma results strongly depends on the experimental sample size. We expect our enhanced tool will facilitate more accurate differential expression analysis while requiring fewer samples.

* As of 09/24/2015

Abstract Identifying Number: 1006085

MODEL CONSTRUCTION AND VALIDATION OF CANNABINOID RECEPTORS FOR DRUG DISCOVERY

Khaled M. Elokely[1,2,3]**, and Michael L. Klein**[1,2]

[1]Institute for Computational Molecular Science and [2]Department of Chemistry, Temple University, Philadelphia, Pennsylvania 19122, United States.

[3]Department of Pharmaceutical Chemistry, Tanta University, Tanta 31527, Egypt

The human cannabinoid (CB) receptors belong to class A family of G-protein coupled receptors (GPCR). There are two recognized subtypes of CB receptors defined as CB1 and CB2. CB receptors are part of the endocannabinoid system and their modulation affects numerous physiological and pathophysiological events including appetite, pain, mood, and memory. Till now, there is no experimentally solved crystal structure of CB1 or CB2 receptors. Structurally, CB receptors are characterized by seven transmembrane (TM) helices, three intracellular (IC) and three extracellular (EC) loops. To date the therapeutically used CB drugs are derived from cannabis or their derivatives. These drugs suffer from psychoactivity adverse effects, which limited their use. We are aiming by constructing valid 3D models to aid in drug development of safer and potent CB modulators. We attempted to build reliable 3D models for CB1 and CB2 using the multiple-templates approach. We constructed the models using the complete sequence of the receptors, and then we built the dimer for the dynamic simulations (MD) calculations. In order to use the models in drug development, we validated their ability to distinguish between known active and inactive CB modulators. Benchmark docking of actives and decoys databases showed the suitability of using these models in virtual screening workflows. MD of the apo CB receptors in phospholipid membrane demonstrated the structural stability of our models. Protein-ligand interaction profiling revealed the amino acids commonly involved in such interactions.

Abstract Identifying Number: 1006280

A BIOINFORMATICS STATEGY TO ENHANCE DILI PREDICTION BY INTEGRATING DIVERSE PREDICTIVE MODELS

Kristin McEuen[1,2], Leihong Wu[2], Shradda Thakkar[2], Weida Tong[2], Minjun Chen[2]

[1] Department of Bioinformatics, University of Arkansas at Little Rock and University of Arkansas for Medical Sciences, Little Rock, Arkansas, 72204
[2]Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, Arkansas, 72079

Drug-induced liver injury (DILI), though rare, can result in severe clinical outcome such as acute liver failure and even death. Despite rigorous toxicity testing in the preclinical phase, DILI is still frequently encountered in clinical trials and has been a major cause of drug failure and market withdrawal. In silico modeling has demonstrated its utility in identifying the DILI risk associated with compounds undergoing development and helping to avoid the potential late-stage disasters. Current research suggests that DILI is a multifaceted disease and predictive models derived from a single data type may not adequately account for all aspects of DILI and therefore limit their utility. Using an integrated approach to combine models with diverse data types may improve prediction accuracy [Chen et al., Biomarkers in Medicine, 2014]. In this study, we applied a simple voting strategy to integrate results from three diverse DILI predictive models, including the "Rule-of-Two" [Chen et al., Hepatology, 2013], the DILI prediction system [Liu et al., PLoS Computational Biology, 2011], and a DILI predictive QSAR model [Chen et al., Toxicological Science, 2013]. The prediction accuracies of these three individual models ranged from 80.5% to 83.4%, as estimated from the dataset of 205 drugs annotated for DILI risk based on the FDA-approved drug labeling. After investigating all possible model combinations, the highest prediction accuracy achieved (88.8%) was derived from integrating the "Rule-of-Two," the DILI prediction system, and the QSAR model. Our study highlighted the importance of including diverse data and algorithms to improve DILI predictive models considering the complex DILI mechanisms involved.

This work was funded by the Oak Ridge Institute for Science and Education and supported by grants from NCRR (P20RR016460) and NIGMS (P20 GM103429) at NIH.

Abstract for oral presentation, session 2: NextGen Sequencing

Abstract Identifying Number: 1006010

BRIDGES – BIOMARKERS REUSE IN DIFFERENT GENE EXPRESSION SYSTEMS

Leihong Wu[1], Gokhan Yavas[1], Huixiao Hong[1], Weida Tong[1] and Wenming Xiao[1]

[1]National Center for Toxicological Research, US Food and Drug Administration, Jefferson, AR 72079

Numerous studies and clinical trials have been done using various microarray platforms for last two decades to profile gene expression in cohort sample sets in order to discover biomarkers for diagnosis and prognosis of cancer. Huge amount of microarray data with valuable clinical information have been accumulated. However, with recent rapid development of next-generation sequencing technologies (NGS), NGS-based gene expression profiling (RNA-seq) has supper performance and poises to replace microarray-based assays in near future. Therefore, for both economic and scientific reasons, there is a tremendous need to establish a mechanism to bridge these two technologies in order to re-utilize the legacy microarray data.

Our previous study indicated that with appropriate gene mapping and data transformation, gene signatures can be transformed between microarray and RNA-seq gene expression assays. Moreover, the predication models built on microarray data are applicable directly to RNA-seq data. However, application of RNA-seq data trained models to microarray data largely depends on the gene mapping and the training algorithms.

In this study, we developed Biomarkers Reuse In Different Gene Expression Systems (BRIDGES) to construct predictive models that can be applied across-platforms. BRIDGES uses Kolmogorov–Smirnov statistic test on Hub-Genes, the genes that show consistent expression order in microarray and RNA-seq data for the samples. We used the 498 neuroblastoma samples profiled with both microarray and RNA-seq to estimate the transferability of the prediction models trained with three different algorithms on four clinic endpoints. The predictive models using Hub-Genes had similar prediction performance between intra-platform and inter-platforms, regardless of gene mapping and modeling algorithms. In addition, we found that a limited number of samples are enough to identify reliable Hub-Genes. Furthermore we validated BRIDGES with an unrelated acute myeloid leukemia data set of 172 samples to predict Cyto-Risk of patient. Since selection of Hub-Genes independent to technology platforms, cross-platform gene mapping and model building algorithms, BRIDGES can be generally applied to any cross-platform predictions.

Abstract for Poster presentation

Abstract Identifying Number: 1006010

INVESTIGATING THE ROLE OF DE NOVO ASSEMBLY TO DISCOVER SINGLE NUCLEOTIDE VARIATIONS IN HUMAN GENOME

Leihong Wu[1], Gokhan Yavas[1], Huixiao Hong[1], Weida Tong[1] and Wenming Xiao[1]

[1]National Center for Toxicological Research, US Food and Drug Administration, Jefferson, AR 72079

Discovering genetic variants, including single nucleotide variants (SNVs) and structural variants (SVs), is one of the major applications for the Next Generation Sequencing (NGS) technologies. Currently, the dominated practice of variant discovery is through short sequence alignment against a common reference genome. However, due to the intrinsic limitations of reference genome, such as incompleteness of genome assembly, existing of structural variation among normal individuals, and interference of single nucleotide polymorphism (SNP) on reads mapping, high level of false positive and false negative of variant calling remains as a challenge for alignment-based approach. Recent studies with *de novo* assembled personal genomes have reported a large list of novel variants, indicating that assembly-based variant calling might be an alternative strategy to identify genetic variants. However, there is no ground truth or broad laboratory validation for novel variants discovered by assembled contigs. A systematic assessment is therefore critically needed to determine whether assembly-based approach is reliable.

In this study, we used simulated data to evaluate the validity of SNVs discovered with assembled contig using SOAPdenovo, one of the mostly used tools for short-read assembly. Combining ART and varSim, we simulated ~3 million variants in short reads at coverages from 2x to 50x. We then used both alignment-based and assembly-based approaches to identify SNVs and compared the rate of recall and precision at each coverage. Our preliminary results showed that: (1) at least 20x coverage is needed to generate assembled contigs with a good coverage of genome and genes; (2) the assembly-based approach detected ~10% of the SNVs that were missed by the alignment-based approach; (3) comparing to the alignment-based approach, the assembly-based has a significantly lower rate of recall and precision.

Although assembly-based approach can serve as a complimentary way for SNVs discovery, with SOAPdenovo as the assembly tool, it associates with a great risk of erroneous calling for novel variants. Variants called from assembled contigs are not reliable unless much improved assembly outcomes are warranted with low error rate of fragment joining, good completeness of genome and high fidelity of assembled sequences. Ideally, haplotype resolved assembly would be preferred.

Abstract Identifying Number: 1006327

DISCOVERED ALTERNATIVE SPLICING FROM TCGA SUGGESTS ACROSS-CANCER TYPES ALTERNATIVE SPLICING AND REGULATORY FACTOR ALTERATION IN THESE CANCER TYPES

Lizhong Ding[1], and Yongsheng Bai*[1]

[1*]Department of Biology, Indiana State University, Terre Haute, IN 47809, U.S.A

The TCGA datasets collected matched normal tissues and tumor tissues from 11,000 patients and characterized genomic alterations in 33 cancer types and subtypes. However, alternative splicing (AS) events are not studied to see if they are related to specific cancer or they occurs along with the expression change of the regulatory factor expressions or even the mutation of the regulatory factors, which might shed light on the biological insights into the AS alterations in the some cancer types.

TCGASpliceSeq collected alternative splicing events in cancer. It used the SpliceSeq to treat all the mRNA samples in the TCGA and dump the results into a database for users to download and analyze. We will use the TCGASpliceSeq results to search if there are the same AS event occurring in other cancer types and if there are common patterns of these AS events across cancer types; in these cancer types samples, we will calculate correlation between the mRNA expression levels and the existence of the given AS patterns. Information about the known genes that are implicated in the AS event will used to see if there are mutations on the genes, which cause the corresponding AS alterations and to see if these gene mutations are correlated with specific cancer type

Abstract Identifying Number: 1006286

HPIDB 2.0: CURRENT UPDATES IN DATABASE CONTENT, INTERFACE AND COMPUTATIONAL PREDICTION OF HOST PATHOGEN INTERACTIONS

Mais G. Ammari[1], Cathy R Gresham[2], Fiona M. McCarthy[1] and Bindu Nanduri[2,3]

[1]School of Animal and Comparative Biomedical Sciences, University of Arizona, Tucson, AZ 85721
[2]Institute for Genomics, Biocomputing and Biotechnology and [3]College of Veterinary Medicine, Mississippi State University, Mississippi State, MS 39762

Resources that annotate, predict and display host-pathogen interaction (HPI) information that underpin infectious diseases are critical for developing novel intervention strategies. HPIDB 2.0 (http://www.agbase.msstate.edu/hpi/main.html) is a publically accessible, comprehensive resource for HPI information, and contains 43,276 manually curated entries in the current release. Since the first description of the database in 2010, we have made several changes to HPIDB data and interface services to facilitate search, access and predict HPI. Notably, HPIDB 2.0 now provides manually curated annotations for targeted, experimentally verified HPI. As a member of the International Molecular Exchange (IMEx) consortium, HPIDB 2.0 manual annotations meet community standards to provide detailed contextual experimental information and facilitate data sharing between resources. In addition, revision of our method for loading external interaction data into HPIDB ensures inclusion of high quality HPIs from available resources. HPI data provided by HPIDB 2.0 include a broad range of hosts and their pathogens (currently we have data for 567 pathogen and 68 host species). In addition, where annotated data is scarce, HPIDB serves as a platform to infer additional HPI and support investigator-driven data analysis. To improve HPIDB 2.0 HPI sequence-based predictions, a newer version of BLAST for sequence and homology searches is implemented. We are evaluating existing computational methods for HPI prediction to help researchers evaluate the quality of transferred homologous HPI to their species of interest. HPIDB 2.0 updates include enhanced search capacity, addition of Gene Ontology functional information, and implementation of network visualization. HPIDB 2.0 data is freely available and is updated on a regular basis.

This project is supported by Agriculture and Food Research Initiative Competitive Grant no. 2015-67015-23271 from the USDA National Institute of Food and Agriculture.

Abstract Identifying Number: 1006295

MODE OF ACTION AND BIOMARKER DISCOVERY FOR ANTI-CANCER NATURAL PRODUCTS

Malia Potts*[1], Elizabeth McMillan[1], John MacMillan[2], and Michael White[1]

[1]Department of Cell Biology, UT Southwestern Medical Center, Dallas, TX, 75390
[2]Department of Biochemistry, UT Southwestern Medical Center, Dallas, TX, 75390
*Current affiliation: Department of Cell & Molecular Biology, St. Jude Children's Research Hospital, Memphis, TN 38105

Natural products have provided a rich source of anti-cancer therapeutic leads that act through diverse modes of action. Key challenges to developing natural products for therapeutic use include incomplete understanding of anti-cancer modes of action, heterogeneity of patient responses, and toxicity. We have integrated experimental and computational methods to address these challenges, first by mapping natural products to their cellular mechanisms of action on a library-wide scale, and second by identifying biomarkers capable of predicting best responders to specific anti-cancer natural products. Within the experimental setting of human HCT116 colon cancer cells, we measured quantitative expression-based cellular responses to marine-derived natural products and to siRNA-mediated genetic perturbations. We then used unbiased similarity matrices to match similar responses, thus enabling target prediction at the pathway level. In this way, we matched several novel and known compounds to their cellular mechanisms of action. One key example is discoipyrrole A, a novel compound that targets the DDR2 collagen-sensing pathway and selectively kills lung cancer cells harboring oncogenic driver mutations in the DDR2 kinase. Another key example is the previously-studied depsipeptide didemnin B, which we found induces rapid apoptosis in sensitive cancer cells from various tissues of origin through a dual mechanism of action, targeting both the translational elongation factor EEF1A1 and the lysosomal hydrolase PPT1. We then used the elastic net machine learning algorithm to identify a multi-feature expression biomarker capable of predicting sensitivity to didemnin B and to an analog that is currently in clinical trials for oncological diseases. Improved mechanistic understanding combined with enhanced ability to predict response has the potential to improve and expand therapeutic utility of anti-cancer natural products. *This research was supported by the Welch Foundation, the US National Cancer Institute, the Cancer Prevention and Research Institute of Texas, and Komen for the Cure.*

Abstract Identifying Number: 1006312

DESIGN OF NEW INHIBITORS OF CYCLIN-DEPENDANT KINASE 5 (CDK5) FOR ALZHEIMER'S DISEASE

Manal A. Nael[1] and Robert J. Doerksen[1, 2 *]

[1]Department of BioMolecular Sciences, Division of Medicinal Chemistry, School of Pharmacy, University of Mississippi, MS, USA, [2]Research Institute of Pharmaceutical Sciences, School of Pharmacy, University of Mississippi, MS, USA

Inhibition of protein kinases is a potential therapeutic strategy to improve memory and to slow disease progression in Alzheimer's disease (AD) patients. This strategy also might prove useful to delay the onset of AD at presymptomatic stages. Studies using animal models of AD largely confirmed that cyclin-dependent kinase 5 (CDK5) deregulation contributes to neuronal loss in the disease. Indeed, studies revealed that selective inhibition of CDK5 reduces levels of amyloid beta (Aβ)□ induced neuronal loss in cortical neurons.

In this research project, we analyzed the ligand-binding pocket of CDK5 in order to search for new inhibitors. The active water molecules were checked to determine if they were native or artifacts due to crystallization. We used the PyWater program to average the experimentally defined waters, and SZMAP software for calculating the thermodynamic properties of the active site. We found three conserved active site water molecules, confirmed by thermodynamic calculations. We maintained the three waters in the active site, prepared a drug-like database, and performed virtual screening using docking. We selected the top six scoring compounds and tested them for in vitro CDK5 inhibition. Two of the compounds had IC50 less than 3 μM.

Clinical Implementation of Pharmacogenetics in Precision Medicine

Mary V. Relling

St. Jude Children's Research Hospital

Although there is substantial hype that the widespread introduction of genomic information will revolutionize health care, the use of genetic tests to inform clinical decision making remains uncommon. Because of decades of research, there are multiple pharmacogenetic tests that could inform more intelligent prescribing now. At St. Jude, we implemented a protocol, PG4KDS (www.stjude.org/pg4kds) to perform preemptive array-based pharmacogenetic testing on all patients, and to create clinical decision support to facilitate evidence-based prescribing advice for individual patients. At the same time, we co-created the Clinical Pharmacogenetics Implementation Consortium (CPIC) (www.cpicpgx.org) as a shared project between PharmGKB (https://www.pharmgkb.org) and NIH's Pharmacogenomics Research Network (www.pgrn.org). CPIC's goal is to accelerate proper use of pharmacogenomics in the clinic by writing, curating, and updating and providing freely available, peer-reviewed, standardized, and detailed gene/drug pharmacogenetic clinical practice guidelines.[PMID: 21270786] An underlying assumption of CPIC guidelines is that genotypes will become preemptively available: guidelines focus on HOW available genetic test results should be used to optimize drug therapy, rather than WHETHER tests should be obtained. This preemptive system is what we are implementing at St. Jude for patient care now.

Abstract Identifying Number: 100267

PROTEINS THAT ABROGATE AGGREGATION BY BLOCKING UPS DEGRADATION IN NEURO-DEGENERATIVE DISEASES: A COMBINED COMPUTATIONAL AND MOLECULAR BIOLOGY APPROACH

Meenakshisundaram Balasubramaniam[1, 2], **Srinivas Ayyadevara**[2, 3], **Ramani Alla**[2], **and Robert J Shmookler Reis**[1, 2, 3]

[1]-UALR/UAMS join Bioinformatics Program
[2] Department of Geriatrics, University of Arkansas for Medical Sciences, Little Rock, AR 72205
[3] Central Arkansas Veterans Healthcare System, Little Rock, AR 72205

Age-dependent protein aggregation is a common feature of neurodegenerative diseases, and involves common proteins (e.g. TDP-43 and α-synuclein) along with disease-specific "seed" proteins such as $A\beta_{1-42}$, tau (Alzheimer disease), and huntingtin (Huntington disease). Roles of these shared proteins in disease progression remain largely unstudied. Using a computational and molecular biology approach in a *C. elegans* model of Huntington disease, we recently identified CRAM-1, a novel protein in insoluble aggregates initiated by polyglutamine fused to yellow fluorescent protein (Q40::YFP). Interestingly, CRAM-1 promoted aggregates in multiple neurodegenerative-disease models. It contains a tri-helical domain that is structurally analogous to the ubiquitin-binding (UBA) domain of RAD23A, which helps convey ubiquitinylated proteins to 26S proteasomes. CRAM-1 impairs proteasomal degradation, apparently by competing with RAD23A for polyubiquitin binding. CRAM-1 knockdown significantly reduced aggregation and associated traits in *C. elegans* strains that model Huntington's, Parkinson's and Alzheimer's diseases. SERF2, the closest human ortholog of CRAM-1, also promotes amyloid aggregation in a human cell-culture model of Alzheimer disease, SH-SY5Y-APP cells. SERF2 knockdown in these neuroblastoma cells reduced amyloid aggregation by ~52%. These findings provide new insights into the roles of protein aggregation and impaired proteasomal degradation in neurodegenerative diseases, and identify a novel therapeutic target for intervention.

Supported by grants from NCRR (P20RR016460) and NIGMS (P20 GM103429) at NIH & Merit and SRCSA grants (to RJSR) from the US Department of Veteran Affairs

Abstract Identifying Number: 1006254

IDENTIFICATION OF DISEASE BIOMARKERS VIA INTEGRATED ANALYSIS OF LONGITUDINAL CLINICAL AND GENOMIC DATA

Qingyang Luo[1], Fengqi Chang[2], Ilana Fortgang[2], and Michelle Lacey*[1]

[1*]Department of Mathematics, Tulane University, New Orleans, LA, 70118
[2] Department of Pediatrics, Tulane University School of Medicine, New Orleans, LA 70112

The challenge of identifying meaningful diagnostic or prognostic genotypic biomarkers for many diseases is complicated by the range of phenotypes that are observed in the patient population. Such phenotypic variation is often captured through clinical records, but these are not commonly employed in the analysis of genomic data. In previous work (Luo *et al* (2014), *Frontiers in Genetics*), we developed a Bayesian hierarchical B-spline approach to fit disease trajectory models for primates exposed to low doses of Mycobacterium tuberculosis (Mtb) based on their clinical profiles. Disease severity estimates derived from these fitted curves were employed to identify genes significantly associated with disease progression, increasing the value of information extracted from the expression profiles and contributing to the identification of predictive biomarkers for TB susceptibility. We now present a second application of our approach to the analysis of gene expression profiles associated with induced colitis in both wild type (WT) and genetically modified mice lacing the TNFR1 receptor. Disease trajectory models were estimated on the basis of body weight and hematochezia, and all animals were biopsied following euthanasia. Through an integrated analysis of clinical trajectories, pathology data, and gene expression profiles, we show significant associations between the severity and duration of symptomatic illness and tumor development. Our results demonstrate that the incorporation of individual disease trajectory estimates enhances existing approaches for biomarker identification and offers the potential to provide insights into personalized treatment strategies for complex diseases.

Abstract Identifying Number: 1006268

GENOMERUNNER WEB SERVER: REGULATORY SIMILARITY AND DIFFERENCES DEFINE FUNCTIONAL IMPACT OF SNP SETS

Mikhail G. Dozmorov[1,*], Lukas R. Cara[1] , Cory B. Giles[2], and Jonathan D. Wren[2,3]

[1]Department of Biostatistics, Virginia Commonwealth University, Richmond, VA, 23298
[2]Department of Arthritis and Clinical Immunology, Oklahoma Medical Research Foundation, Oklahoma City, OK, 73104
[3]Department of Biochemistry and Molecular Biology, University of Oklahoma Health Sciences Center, Oklahoma City, OK, 73104

The growing amount of regulatory data from the ENCODE, Roadmap Epigenomics and other consortia provides a wealth of opportunities to investigate the functional impact of single nucleotide polymorphisms (SNPs). Yet, given the large number of regulatory datasets, researchers are posed with a challenge of how to efficiently utilize them to interpret the functional impact of SNP sets.

We developed the GenomeRunner web server to automate systematic statistical analysis of SNP sets within regulatory context. Besides defining the functional impact of SNP sets, GenomeRunner implements novel regulatory similarity/differential analyses, and cell type-specific regulatory enrichment analysis. Validated against literature- and disease ontology-based approaches, analysis of 39 disease/trait-associated SNP sets demonstrated that the functional impact of SNP sets corresponds to known disease relationships. We identified a group of autoimmune diseases with SNPs distinctly enriched in the enhancers of T helper cell subpopulations, and demonstrated relevant cell type-specificity of the functional impact of other SNP sets. In summary, we show how systematic analysis of genomic data within regulatory context can help to interpret the functional impact of SNP sets.

Availability: GenomeRunner web server is freely available at http://www.integrativegenomics.org/.

Funding

This work was supported by the National Institute of Arthritis and Musculoskeletal and Skin Diseases (a subaward from grant # P30 AR053483 to MGD), an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences (a subaward from grant # P30 GM103510 to MGD), and the National Institute of Health (Grants # 1P20GM103636 and 8P20GM103456 to JDW) for partial funding of this work.

Abstract Identifying Number: 100312

A Simple Rule to Assess the Risk of Drug-induced Liver Injury Associated with the DRESS (Drug Reactions/Rash with Eosinophilia and Systemic Symptoms) Symptom

Minjun Chen[1], Marc Stone[2], Eileen E Navarro Almario[2], Shashi Amur[2], Victor Crentsi[2], Tina M Burgess[3], Ruyi He[2],Hong Fang[1], John Senior[2], Weida Tong[1]

[1]NCTR, [2]CDER; [3]CVM, US FDA

Drug-induced liver injury (DILI) is a severe adverse event found with Drug Reactions/Rash with Eosinophilia and Systemic Symptoms (DRESS) with the mortality rate of 10%, which has caused some drugs withdrawn from market (e.g. trovafloxacin). Unfortunately, this adverse event is often unpredictable based on the current regulatory methodologies, and therefore new predictive models are urgently demanded by the drug development and regulation. In this study, we tried to identify simple rules to assess the DILI risk associated with the DRESS symptom. For this purpose, three lists of drugs were identified to cause DILI and DRESS, one from the FDA Adverse Event Reporting System (FAERS) database, one from literature with causality assessment, and one collected from the FDA drug labels. Based on the first drug list from FEARS, we found that the drugs with a daily dosage (DD) ≥ 100mg alone or DD ≥ 100mg & reactive metabolites (RM) were significantly associated with DILI risk, but this observation is not applicable for another published "rule-of-two" (i.e. DD ≥ 100mg & logP≥ 3). Moreover, DD ≥ 100mg plus RM presents less sensitive but more specific than DD ≥ 100mg. This finding was further validated by the analysis from the left two datasets. Putting altogether, our study indicated that a simple rule of combining daily dose and reactive metabolites could be a potential predictor useful for assessing DILI risk associated with DRESS symptom.

Abstract submission guidelines:

Page Limit: One page, not more than 300-words

- **Margins:** 1"-All sides
- **Font**: Times New Roman, size 12
- **Title**: Times New Roman, size 12 in **bold**, all CAPS
- **Authors:** Include all the names and affiliations - **underline presenting author**

Abstract Identifying Number: 1006253

COMPARATIVE STUDY OF THE REPETITIVE ELEMENTS IN MODEL SPECIES GENOMES

Mohamed K. Aburweis[1], **Xijin Ge**[1]

[1]Department of Mathematics and Statistics, South Dakota State University, Brookings, SD, 57006

Repetitive elements (repetitive DNA) are abundant in a wide range of species, from bacteria to mammals. In human, for example, repetitive elements comprise about 50% of the human genome. Repetitive elements play a critical role in genome evolution, but most of their biological functions remain unknown. Identification and analysis of the abundance and the distribution of repetitive DNA is important to understand genome structure and function. Here we conduct a comparative study of the distribution of repetitive DNA and abundance as well its impact on genome structure and function in model species using bioinformatics strategies and computational tools. We also analyzed RNA-seq dataset for 10 normal human and 10 normal mouse tissues using TUXEDO pipeline with strand specific RNA-seq. Then we compare the gene expression for these tissues with the repetitive DNA data to find the relationship between them by using multiple regression analysis techniques, and we found some repeats have significant effect on gene expression.

Abstract Identifying Number: 1006325

A PROBABILISTIC FRAMEWORK FOR IDENTIFYING DRUG-DRUG INTERACTIONS AND PRIMARY SUSPECT DRUGS FOR MULTI-DRUG TREATMENT SETTINGS

Monisha Puttaraju[1] and Halil Bisgin*[1]

[1*]Department of Computer Science, University of Michigan-Flint, Flint, MI, 48502

Adverse drug reactions (ADRs) are known to be one of the major causes for hospitalizations some of which can also result in death. Therefore, this undesired outcome remains as an important issue in addition to other disease centered public health concerns. Especially, those patients who need to take multiple medications for their treatments, the risk level gets higher due to possible drug-drug interactions or lack of knowledge about the primary cause of ADRs in a multi-drug setting. These facts introduce the necessity to study ADR collections to avoid such severe consequences. In order to investigate the causalities between drugs and ADRs for aforementioned risks, we propose a framework that treats each reported case in FDA Adverse Event Reporting System (FAERS) as a text document. This analogy enables us to come up with a probabilistic view of the reported ADRs through author topic modeling (ATM) in which drugs are considered to be the authors expressing their side effects in words. Resulting associations will be used to account for strength of a tie between a drug and a side effect, which will eventually help us predict the primary suspect in a multi-drug treatment. Further, these probabilistic relations are expected to detect hidden drug-drug interaction.

Abstract Identifying Number: 1006261

CONVEX HULLS GENERATION AND CONVERGENCE OF DENSITY BASED SKIN LESION DETECTION

(Skip single line)

Mustafa Bayraktar*[1], Sinan Kockara*[3], and Kamran Iqbal*[2]

(Skip single line)

[1*]Department of Bioinformatics, University of Arkansas at Little Rock, Little Rock, AR, 72204

[2]Department of Systems Engineering, University of Arkansas at Little Rock, Little Rock ,AR,72204

[2]Department of Computer Science, University of Central Arkansas, Conway ,AR,72035

Dermoscopy is one of the effective and minimal invasive imaging technique in diagnosis of skin cancer, especially for pigmented lesions. Accurate skin lesion border detection key to extract important dermoscopic features of the lesion. In current clinical settings, border delineation is performed manually by dermatologists. This technique leads to intra- and inter-observer variations due to its subjective nature. Moreover, it is tedious in longitudinal studies. Because of aforementioned hurdles, the automation of lesion boundary assessment in dermoscopy images is a need. One of the efficient lesion border detection method is reported as density based lesion border detection. This method takes cue from DBSCAN and modifies it more efficient with convex hull operations. Even though this method computationally more efficient than DBSCAN, it still suffers from convex hull generation and set operations on generated convex hulls. In this study, we propose a new convex hull generation method, which is a combination of the two existing algorithms, i.e.,'Quickhull' and Chan's algorithm. To this end, we represent skin lesion as the clusters of convex hulls, and, based on the set operations e.g. union, intersection, complement etc., we determine the border of the lesion. We note that these are the most computationally demanding steps in fast density based lesion border delineation. The proposed combination of the two existing algorithms, 'Quickhull' and Chan's algorithm, can compute the convex hull in O(nlogh), where h is the number of point on the convex hull, for the worst case and in O(n) for the average case. . *Supported by grants from NCRR (P20RR016460) and NIGMS (P20 GM103429) at NIH.*

Abstract Identifying Number: 1006261

LEFT VENTRICLE BORDER TRACKING TO DETERMINE SAFE DELIVERY TRAJECTORY FOR TRANSAPICAL AORTIC VALVE REPLACEMENT

Mustafa BAYRAKTAR,

[1*]Department of Bioinformatics, University of Arkansas at Little Rock, Little Rock, AR, 72204

Image-guided pre-operative planning is of paramount significance in obtaining reproducible results in robotic cardiac surgeries. Planning helps the surgeon utilize the amended quantitative information of the target area, and assess and evaluate the suitability of offered medical therapy prior to surgery. In transapical access aortic valve implantation, determining the safest corridor for the voyage of robotic delivery module along the left ventricle (LV) is an important step to prevent potential adverse events from happening, i.e., harming heart wall and mitral valves, and malpositioning of the valve to be delivered. Motivating from that fact that, we processed short-axis (SAX) cardiac magnetic resonance (CMR) images which are with promising volumetric capability and no-radiation effect. We propose a system that incorporates robust left ventricle segmentation, a combination of an isotropic denoising and a hybrid active contour model, in addition to a dynamic safe path optimization for robotic delivery module based LV segments
Supported by grants from NCRR (P20RR016460) and NIGMS (P20 GM103429) at NIH.

Abstract Identifying Number: 1006261

A Framework Proposal for Longitudinal Tumor Response Monitoring

Mustafa BAYRAKTAR,

[1*]Department of Bioinformatics, University of Arkansas at Little Rock, Little Rock, AR, 72204

Nowadays, structured radiology reports are commonly used for exchanging the information that is created during the patient imaging and annotation and markup session on the image data. One of the main advantages of using formatted reports in image-ordering workflow is to have capability of disambiguating reported real world values (i.e. imaging modality, quantitative variables of tumor mass, spatiotemporal information of tumor on the image data, annotated findings and conclusions about tumor, image slices on that tumor is measurable, etc.) those help us create a unique tumor entity. However, especially in timely basis response monitoring, information extraction for a particular tumor from the successive reports is remarkably complicated due to change at patient`s position during imaging, different scanning technique, tumor`s own intricate growth regime such as split-up or merge, and not standardized radiologist interpretations. Many times, these cases drive radiologist complete markup and annotation session on a tilted region of interest, and different gray values (even for same structure), by his/her own instant perceptions. These circumstances result in low precision and recall rates when two tumor entities are compared for merging. For instance, a reader (here an oncologist) is observing a tumor in a follow-up report and wants to call baseline report to see change in tumor growth. The tumor features retrieved from the baseline report might contain such visual and semantic information which makes the oncologist conclude that this tumor must be different, whereas it might not. To get rid of this ambiguity, we determined the methods that will assist developing an automated communication between two lesions for resolution these lesions refer to same tumor or not, and assign a unique ID to every stage of a tumor for a practicable quantitative analysis in timely fashion.

Supported by grants from NCRR (P20RR016460) and NIGMS (P20 GM103429) at NIH.

Abstract Identifying Number:1006275

SUCCESSFUL CLASSIFICATION OF COCAINE DEPENDENCE USING BRAIN IMAGING: A MACHINE LEARNING APPROACH

Mutlu Mete[1], Unal Sakoglu[1], Jeffrey S. Spence[2], Michael D. Devous, Sr.[3], Thomas S. Harris[3], Bryon Adinoff[4, 5]

[1]Department of Computer Science, Texas A&M University-Commerce, Commerce, TX, 75428
[2]Center for BrainHealth, University of Texas at Dallas, Dallas, TX, 75235
[3]Department of Neurology, UT Southwestern Medical Center, Dallas, TX, 75390
[4]VA North Texas Health Care System, Dallas, TX, 75216
[5]Department of Psychiatry, UT Southwestern Medical Center, Dallas, TX, 75390

Neuroimaging studies have yielded significant advances in the understanding of neural processes relevant to the development and persistence of addiction. However, these advances have not improved diagnostic accuracy. The aim of this analysis was to develop a statistical approach, using a machine learning framework, to correctly classify brain images of cocaine-dependent patients and healthy controls. Single Photon Emission Computerized Tomography images obtained during rest or a saline infusion in three cohorts of 2-4 week abstinent cocaine-dependent participants without other The Diagnostic and Statistical Manual of Mental Disorders-V Axis-I disorders (n=93) and healthy controls (n=69) were used to develop a machine learning model. A cross-validation approach was used for accuracy assessment. A first step voxel-based analysis with Information Gain was conducted to select statistically significant and densely connected regions (minimal cluster size ≥20 voxels). Support vectors machines (SVM) were then exploited for both major feature selection and participant classification (control or cocaine-dependent). A framework suitable for educing potential brain regions that differed between the two groups was developed. The voxel-based analysis identified 1500 densely connected voxels in 30 distinct clusters after a grid search in SVMs parameters. Participants were successfully classified with 88% and 89% F-measure accuracies in 10-fold cross validation (10xCV) and leave-one-out (LOO) approaches, respectively. Sensitivity and specificity were 0.90 and 0.89 for LOO; 0.83 and 0.83 for 10xCV. The SVMs approach successfully classified cocaine-dependent and control participants. If confirmed, these findings support the future use of brain imaging and SVMs in the diagnosis of substance use disorders and furthering an understanding of their underlying pathology.

Abstract Identifying Number: 1006293

COMPARATIVE FUNCTIONAL GENOMICS OF MAMMALIAN SPERM TH2B, REGULATOR OF SPERM CHROMATIN DYNAMICS AND MALE FERTILITY

Naseer Kutchy[1], Marie Jacobsen[2,3], Ana Velho[1,4], Amanda Lawrence[5], Giselle Thibaudeau[5], Arlindo Moura[4], Abdullah Kaya[7], Andy Perkins[3], Erdogan Memili[1]

[1]Department of Animal and Dairy Sciences, Mississippi State University, Mississippi State, MS 39762; [2]Department of Biosciences, Rice University, Houston, TX 77005; [3]Department of Computer Science and Engineering, Mississippi State University, Missis sippi State, MS 39762; [4]Department of Animal Science, Federal University of Ceara, Fortaleza, BRAZIL 60040; [5]Institute for Imaging Analytical Techn ologies, Mississippi State University, Mississippi State, MS 39762; [6]Selcuk University, Konya, TURKEY 35920

Male fertility, ability to fertilize and activate the egg and support early embryo development, is vital for mammalian reproduction. Cattle genetics and physiology are similar to those of other mammals including humans. Reliable fertility data along with well-established *in vitro* systems are available for bull which was the model organism used in this study. Chromatin compaction is essential for sperm physiology and is mobilized by testis specific histone variant 2B (TH2B) to facilitate eviction of retained histones through interaction with other chromatin remodelers BRD4 and CHD5. Molecular and cellular mechanisms by which TH2B regulates histone to protamine replacement as well as the evolutionary diversification of TH2B are poorly defined. The objective of this study was to test the central hypothesis that abnormal cellular levels and locations of the evolutionarily conserved TH2B cause low fertility using immunocytochemistry, western blotting, and bioinformatic approaches. Immunocytochemistry experiments showed that TH2B is located in the bovine sperm head, and statistically insignificant, expression levels of TH2B (Mean ± SEM) were higher in sperm from the low fertility vs. high fertility bulls [220.56 (± 9.20) vs. 198.39 (± 10.0)]. TH2B was also detectable in the sperm using western blotting. Bioinformatic experiments revealed that TH2B and H2B are highly conserved across mammalian taxa. Multiple sequence alignment specified that bovine H2B is more than 80% similar to those of mouse and human, and more than 90% to those of other ruminant animals such as goat and sheep. Neighbor-joining phylogeny of H2B, CHD5 and BRD4 discovered that bovines have evolutionary closeness with human, mouse and small ruminants. Gene ontology analysis revealed that TH2B has the conserved histone H2B domain and this domain has a strong association with proteins involved in chromosome organization and histone ubiquitination. The findings help advance animal science and biotechnology.
Partial funding for this research was provided by: Mississippi Agricultural, Forestry, Experiment Station, MSU; NSF (DBI1262901); Brazilian Research Councils (CNPq, CAPES and FUNCAP). ICAR, India.

Abstract Identifying Number: 100119

MULTICLASS COMPUTATIONAL EVOLUTION: BENCHMARK EVALUTATION AND APPLICATION TO RNA-SEQ BIOMARKER DISCOVERY

Nathan Crabtree[1], Jason H. Moore[2], Nysia George[3], Douglas P Hill[4], John F Bowyer[5]

[1*]Bioinformatics, Department of Information Science, University of Arkansas at Little Rock,
Little Rock, Arkansas, 72204
[2]Institute for Biomedical Informatics, The Pearlman School of Medicine, University of
Pensylvania, Philadelphia, PA, 19104
[3] Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, US
Food and Drug Administration, AR, 72079
[4]Norris Cottton Cancer Center, Dartmouth College, NH, 03756
[5] Division of Neurotoxicology, National Center for Toxicological Research, US Food and Drug
Administration, AR, 72079

Classification and feature selection algorithms are useful for identifying small numbers of important biomarkers in large datasets. Computational evolution is a classification and feature selection technique that uses genetic programming and evolutionary algorithms. Millions of classifier equations are randomly generated by putting together basic building blocks like math functions, variables from the dataset and constants. The classifiers with the best classification accuracy and least complexity are chosen for mutation and survival into the next round of evolution, similar to biological evolution by natural selection. Mutation occurs by adding or dropping building blocks or by combining classifiers together into a single new classifier. After many rounds of evolution, a few classifiers will remain that outperform all others. Old computational evolution approaches only worked for binary class datasets, in this study, we developed a computational evolution system that can classify and select features from multiple classes simultaneously. This is useful for researchers who either have complex experiment designs or are comparing multiple disease states. The computational evolution system was compared to support vector machines (SVM), random k-nearest-neighbor (RKNN), and random forest (RF) in a 10 rep, 5 fold cross validation on three real datasets and two simulated datasets. Comparison criteria were testing-set accuracy, number of selected features, run time, and stability of the selected feature sets measured by Tanimoto distance. A final run of the CES algorithm yielded a set of selected biomarkers, which were validated in the literature, biology databases, and ontologies.

Abstract Identifying Number: 1006065

INCORPORATING TOPOLOGICAL INFORMATION FOR PREDICTING ROBUST CANCER SUBNETWORK MARKERS IN HUMAN PROTEIN-PROTEIN INTERACTION NETWORK

Navadon Khunlertgit[1], **and Byung-Jun Yoon**[1,2]

[1*]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, 77843-3128
[2]College of Science and Engineering, Hamad bin Khalifa University (HBKU), P.O. Box 5825, Doha, Qatar

Discovering robust markers for cancer prognosis based on gene expression data is an important yet challenging problem in translational bioinformatics. By integrating additional information in biological pathways or a protein-protein interaction (PPI) network, we can find better biomarkers that lead to more accurate and reproducible prognostic predictions. In fact, recent studies have shown that “modular markers” that integrate multiple genes with potential interactions can improve disease classification and also provide better understanding of the disease mechanisms. In this work, we propose a novel algorithm for finding robust and effective subnetwork markers that can accurately predict cancer prognosis. To simultaneously discover multiple synergistic subnetwork markers in a human PPI network, we build on our previous work that uses affinity propagation, an efficient clustering algorithm based on a message-passing scheme. Using affinity propagation, we identify potential subnetwork markers that consist of discriminative genes that display coherent expression patterns and whose protein products are closely located on the PPI network. Furthermore, we incorporate the topological information from the PPI network to evaluate the potential of a given set of proteins to be involved in a functional module. Primarily, we adopt widely made assumptions that densely connected subnetworks may likely be potential functional modules and that proteins that are not directly connected but interact with similar sets of other proteins may share similar functionalities. Incorporating topological attributes based on these assumptions can enhance the prediction of potential subnetwork markers. We evaluate the performance of the proposed subnetwork marker identification method by performing classification experiments using multiple independent breast cancer gene expression datasets. We show that our method leads to the discovery of robust subnetwork markers that can improve cancer classification.

Acknowledgment:
This work was supported in part by National Science Foundation through NSF Award CCF-1149544

Abstract Identifying Number: 1006065

A NOVEL METHOD FOR IDENTIFYING ROBUST SYNERGISTIC SUBNETWORK MARKERS FOR CANCER PROGNOSIS

Navadon Khunlertgit[1], and Byung-Jun Yoon[1,2]

[1*]Department of Electrical and Computer Engineering, Texas A&M University, College Station,
TX, 77843-3128
[2]College of Science, Engineering, and Technology, Hamad Bin Khalifa University (HBKU),
P.O. Box 5825, Doha, Qatar

Identification of markers for cancer prognosis based on microarray gene expression data is one of the challenging problems in computational biology. By integrating additional biological information such as known biological pathways or protein-protein interaction (PPI) network, the so-called "modular markers" have been shown to improve the classification performance and provide the better understanding of disease mechanisms at a modular level. In addition, in order to investigate cellular functional modules, the network topological analysis of PPI network has been shown to efficiently detect meaningful modules whose member are not directly connected to each other. In this work, we propose a novel algorithm to find robust and effective subnetwork markers that can predict cancer prognosis accurately. We utilize affinity propagation, a useful clustering algorithm based on the message-passing concept, to identify subnetworks of discriminative genes whose protein products are closely located in the PPI network. We also incorporate the utilization of topological information of this interactomic dataset. Given the PPI network, we adopt the assumptions about the topological attribute of proteins in functional subnetworks which are, 1) densely connected subnetworks may be potential functional modules, and 2) proteins that are not directly connected but interact with similar sets of other proteins may have a higher probability of sharing similar functionalities. These simple assumptions help us to cover the interaction of proteins that might not be captured by their corresponding gene expression levels. Our proposed method simultaneously identifies multiple non-overlapping subnetwork markers that can synergistically predict cancer prognosis. We evaluate the performance of identified subnetwork markers by performing experiments on multiple breast cancer gene expression dataset. The experimental results show that the proposed method can identify robust subnetwork markers that may lead to enhanced cancer classifiers.

Acknowledgment:
This work was supported in part by National Science Foundation through NSF Award CCF-1149544

Abstract Identifying Number: 1006326

FETAL HEART LOCALIZATION BY MAGNETIC DIPOLE FITTING

Recep Avci 1, James D Wilson 2,3, Neslihan Bisgin 4, Hari Eswaran 2

1 Department of Bioinformatics, University of Arkansas Little Rock, Little Rock, AR, 72204

2 SARA Research Center, University of Arkansas for Medical Sciences, Little Rock, AR, 72205

3 Graduate Institute of Technology, University of Arkansas Little Rock, Little Rock, AR, 72204

4 Donaghey College of Engineering and Information Technology, University of Arkansas at Little Rock, Little Rock, AR, 72204

Objective: Knowledge of fetal heart and brain orientation is important during the prenatal stage of life. fMEG data can be used to locate and track sources of fetal heart and brain signals.

Materials and Methods: All data was collected using the SARA system that was built for investigation of maternal-fetal physiology and installed in a magnetically shielded room at UAMS, Little Rock, AR. The study group consisted of 36 healthy pregnant women with gestational ages ranging from 27 to 38 weeks. In the post processing, maternal heart signal (mMCG) is removed by Orthogonal Signal Space Projection (OP) and extracted fetal heart signal (fMCG) is averaged on the R peaks for every beat. Then, Magnetic Dipole (MD) fitting method is used for estimating the location of the heart.

Results: Given the magnetic field arising from the fetal R wave as recorded by SARA, and using a spherical model for the heart, a MD representing the magnetic source of the heart is estimated. Based on this MD, the magnetic field of the fetal R wave is simulated and the least squares error is found to be less than 3% for all patients. The results suggest that fitting a MD for a spherical heart model is appropriate for localization of fMCG signal.

The same algorithm can be used for the fetal brain to aid in validation of fMEG. Being able to locate both the heart and the brain sources, we will be able to track fetal movement.

Abstract Identifying Number: 1006316

PROTEIN STRUCTURE-BASED VIRTUAL SCREENING FOR THE DISCOVERY OF NOVEL CB2 RECEPTOR AGONISTS

Ngoc Nguyen, Kuldeep K. Roy, Robert J. Doerksen*

Division of Medicinal Chemistry, Department of BioMolecular Sciences, School of Pharmacy,
The University of Mississippi, University, MS 38677

The endocannabinoid system is one of the most important physiological systems and is implicated in a variety of physiological processes, such as appetite, pain perception, immune response, epilepsy, drug addiction, mood and memory. The cannabinoid (CB) receptors labeled CB1 and CB2, part of the endocannabinoid system, are Class-A GPCRs expressed commonly in the central nervous system, peripheral tissue and immune cells. Although they were discovered many years ago, not until recently have neuroscience researchers given them much attention, realizing the limited role of CB2 in the central nervous system. This work describes a systematic validation of CB2 receptor models and the discovery of novel CB2 receptor ligands using protein structure-based virtual screening. The CB2 models were used for Induced Fit docking (with Schrödinger software) of three common CB2 agonists: CP55,940, (−)-*trans*-Δ^9-tetrahydrocannabinol and WIN55212-2, and then further utilized for docking of another set of eleven CB2 agonists selected from Manera and coworkers' work (*Eur. J. Med. Chem.* **2015**, *97*, 10-18) to identify the two best CB2 receptor models. Next, a proprietary database of 102,927 compounds, provided by LIMR Chemical Genomics Center, was prepared and filtered for drug-likeness, and then docked into the two validated CB2 models in three consecutive steps, HTVS, SP and XP docking, and the top-ranked 10% of the compounds (110 compounds) were studied further using Prime MM-GBSA binding free-energy calculations. Afterwards, the compounds were subjected to hierarchical clustering using the Tanimoto similarity metric and MACCS fingerprints to identify structurally diverse clusters of compounds. The final 38 compounds, selected from 20 clusters, were procured and evaluated using radioligand displacement assays at a fixed 10 μM concentration. The *in vitro* result showed seven compounds exhibiting >50% displacement of the radioligand from the CB2 receptor. In future work, the binding affinity and functional efficacy of these compounds will be determined.

Deleted: a

Acknowledgements

This study was supported by an Institutional Development Award (IDeA) Grant Number P20GM104932 from the National Institute of General Medical Sciences (NIGMS), a component of the National Institutes of Health (NIH). Its contents are solely the responsibility of the authors and do not necessarily represent the official view of NIGMS or NIH. It was conducted in part in a facility constructed with support from the Research Facilities Improvements Program (C06 RR-14503-01) from the National Institutes of Health (NIH) National Center for Research Resources.

Deleted: *and the Project 4: Rational Design and Testing of Novel Cannabinoid Ligands of the COBRE*

Deleted: ”

Abstract Identifying Number: 1006234

NETWORK BASED FUNCTIONAL PAN-GENOMICS: A NEW APPROACH TO BUILD CONNECTIONS BETWEEN GENOMIC DYNAMICS AND PHENOTYPIC EVOLUTION IN THE GENUS *MYCOBACTERIUM*

Ohgew Kweon, Seong-Jae Kim, and Carl E. Cerniglia

Division of Microbiology, National Center for Toxicological Research/U.S. FDA, Jefferson, AR 72079

Despite a flood of genome sequences and functional genomics data, considerable knowledge gaps exist between genome and phenome that hinder efforts toward the treatment of mycobacterial diseases and practical biotechnological applications. Here we introduce a method called Network Based Functional Pan-Genomics (NBFPG) which systematically integrates the three different types of concepts: network, pan-genomics, and functional genomics. This approach allows for phenotype-related functional pan-genomic comparison in the Mycobacterial Phenotype Network (MPN). We demonstrate NBFPG using the “Polycyclic Aromatic Hydrocarbon-degrading” phenotype in the MPN. NBFPG identifies a small fraction of possible evolutionary trajectories, showing strong connection between genome reduction and the “PAH-degrading” phenotype evolution. Conclusively, NBFPG allows for network-level, systematic understanding of the pleiotropic and epistatic effects of genomic dynamics on phenotype evolution, and such evolutionary constraints may be applicable to any phenotype of other organisms.

Abstract Identifying Number: 1006246

ITERATIVE RECONSTRUCTION OF THREE-DIMENSIONAL MODELS OF HUMAN CHROMOSOMES FROM CHROMOSOMAL CONTACT DATA

Oluwatosin Oluwadare[1], Jackson Nowotny[1], Sharif Ahmed[1], Lingfei Xu[1], Hannah Chen[1], Noelan Hensley[1], Tuan Trieu[1], Renzhi Cao[1] and Jianlin Cheng[1]

[1]Department of Compter Science, University of Missouri, Columbia, MO 65211-2060

Abstract

The entire collection of genetic information resides within the chromosomes, which themselves reside within almost every cell nucleus of eukaryotic organisms. From earlier studies, chromosomes have been observed to have their own territory and are not randomly positioned. In addition, each individual chromosome is found to have its own preferred three-dimensional (3D) structure independent of the other chromosomes. The structure of each chromosome plays vital roles in controlling certain genome operations, including gene interaction and gene regulation. As a result, knowing the structure of chromosomes assists in the understanding of how the genome functions.We developed a unique computational approach based on optimization procedures known as adaptation, simulated annealing, and genetic algorithm to construct 3D models of human chromosomes, using chromosomal contact data.

Our models were evaluated using a percentage-based scoring function. Analysis of the scores of the final 3D models demonstrated their effective construction from our computational approach. Specifically, the models resulting from our approach yielded an average score of 80.41 %, with a high of 91 %, across models for all chromosomes of a normal human B-cell. Comparisons made with other methods affirmed the effectiveness of our strategy. Particularly, juxtaposition with models generated through the publicly available method Markov chain Monte Carlo 5C (MCMC5C) illustrated the outperformance of our approach. Our methodology was further validated using two consistency checking techniques known as convergence testing and robustness checking, which both proved successful.

The implementation of our approach proved effective in constructing 3D chromosome models and proved consistent with, and more effective than, some other methods thereby achieving our goal of creating a tool to help advance certain research efforts. The source code, test data, test results, and documentation of our method, Gen3D, are available at our sourceforge site at: http://sourceforge.net/projects/gen3d/.

This work was supported by the National Science Foundation [grant DBI1149224 to JC]

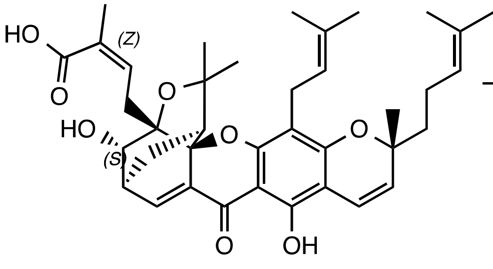
Abstract Identifying Number: 1006324

EXPLORING THE ALLOSTERIC INHIBITORY BINDING SITES FOR KNOWN NEGATIVE ALLOSTERIC MODULATORS WITHIN THE CANNABINOID CB2 RECEPTOR

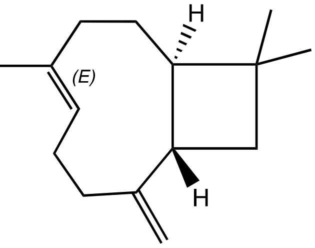
Pankaj Pandey, Kuldeep K. Roy, and Robert J. Doerksen*

Division of Medicinal Chemistry, Department of BioMolecular Sciences, School of Pharmacy, The University of Mississippi, University, MS 38677, USA

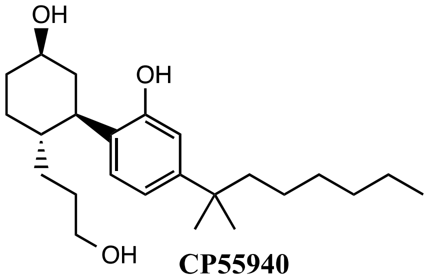
The CB2 receptor, a class-A membrane-bound G-protein coupled receptor, is currently an emerging therapeutic target in the cannabinoid research field to treat neuroinflammation and various other associated disorders. Target selectivity and off-target side effects are the two major limiting factors for orthosteric modulators, and, therefore, the search for allosteric modulators (AMs) is one of the widely used drug discovery approaches to avoid such limitations. To date, only a limited number of AMs of the CB2 receptor have been reported with micromolar activity, and knowledge of allosteric sites within the CB2 receptor is completely lacking. Therefore, we attempt to explore the allosteric inhibitory sites for known negative AMs, dihydro-gambogic acid (DHGA) and trans-β-caryophyllene (TBC), using computational approaches to predict and characterize sites within the CB2 receptor. We first mapped all potential allosteric sites within the validated CB2 receptor model while keeping a well-known CB2 agonist, CP55940, within the orthosteric site. Afterwards, we performed Glide docking using multiple conformers of DHGA and TBC into the five predicted allosteric sites within the CB2 receptor. Each of the two top-ranked protein-ligand(s) complexes, CB2–CP55940–DHGA and CB2–CP55940–TBC, was embedded into a hydrated lipid-bilayer, equilibrated and then simulated for 200 ns using the NAMD program. The molecular dynamics trajectories were further analyzed. Key findings and insights gained from the present study will be presented, which could be used in the identification of novel and high affinity and selective AMs of the CB2 receptor.



DHGA (S-isomer)



TBC



CP55940

Acknowledgements

This study was supported by an Institutional Development Award (IDeA) Grant Number P20GM104932 from the National Institute of General Medical Sciences (NIGMS), a component of the National Institutes of Health (NIH). Its contents are solely the responsibility of the authors and do not necessarily represent the official view of NIGMS or NIH." It was conducted in part in a facility constructed with support from the Research Facilities Improvements Program (C06 RR-14503) from NIH National Center for Research Resources.

Abstract Identifying Number: 1002178

PULSED INDUCTION, A METHOD TO IDENTIFY GENETIC REGULATORS OF SPECIFIC MATURATION EVENTS

Steven Pennington[1], Peter Hoyt[*1]

[1*]Department of Biochemistry and Molecular Biology, Oklahoma State University, Stillwater, OK, 74078

Determination is the process in which a stem cell commits to differentiation. The decision process regulating a cell to enter a determined state is complex and cell type-specific. Maintenance of undifferentiated cell populations are important for rapid tissue recovery which is ultimately dependent on this decision process. Our goal in this study is to use rapid-pulse induction of mouse erythroleukemia (MEL) cells using dimethylsulfoxide (DMSO), as a novel method to evaluate transcriptional changes using microarrays and identify candidate genes involved in maturation. Rapid temporal gene expression changes should be associated with regulation of the early stages of proerythroblast to erythrocyte differentiation or maturation. The pulsed induction method consists of a short induction period (e.g. 30 min) followed by removal of inducer and subsequent cell-culture growth for the duration of a standard differentiation time (i.e. 8 days). For reference, cells were split and RNA prepared immediately when the inducer is removed. Results show many genes are differentially expressed during short exposures to inducer. These include erythropoiesis specific genes such as GATA1, globin genes, and novel candidate genes suggesting that pulsed induction reveals early transcriptional events in the dynamic early signaling of erythropoiesis. Importantly our data identify transcriptional changes that are persistent, transient, or display a pendulum compensatory effect when cells are allowed to recover. The persistence of transcription following short induction suggests a mechanism for MEL-cell induction memory. Additional candidates for maintaining predifferentiated cells in a unique model cell system are supported by system modeling. *Funding provided by Oklahoma State University.*

Abstract Identifying Number: 1002178

PULSED INDUCTION, A METHOD TO IDENTIFY GENETIC REGULATORS OF SPECIFIC MATURATION EVENTS

Steven Pennington[1], Peter Hoyt[*1]

[1*]Department of Biochemistry and Molecular Biology, Oklahoma State University, Stillwater, OK, 74078

Determination is the process in which a stem cell commits to differentiation. The decision process regulating a cell to enter a determined state is complex and cell type-specific. Maintenance of undifferentiated cell populations are important for rapid tissue recovery which is ultimately dependent on this decision process. Our goal in this study is to use rapid-pulse induction of mouse erythroleukemia (MEL) cells using dimethylsulfoxide (DMSO), as a novel method to evaluate transcriptional changes using microarrays and identify candidate genes involved in maturation. Rapid temporal gene expression changes should be associated with regulation of the early stages of proerythroblast to erythrocyte differentiation or maturation. The pulsed induction method consists of a short induction period (e.g. 30 min) followed by removal of inducer and subsequent cell-culture growth for the duration of a standard differentiation time (i.e. 8 days). For reference, cells were split and RNA prepared immediately when the inducer is removed. Results show many genes are differentially expressed during short exposures to inducer. These include erythropoiesis specific genes such as GATA1, globin genes, and novel candidate genes suggesting that pulsed induction reveals early transcriptional events in the dynamic early signaling of erythropoiesis. Importantly our data identify transcriptional changes that are persistent, transient, or display a pendulum compensatory effect when cells are allowed to recover. The persistence of transcription following short induction suggests a mechanism for MEL-cell induction memory. Additional candidates for maintaining predifferentiated cells in a unique model cell system are supported by system modeling. *Funding provided by Oklahoma State University.*

Abstract Identifying Number: 1002144

THE CHALLENGES OF LONG NON-CODING RNA FUNCTION PREDICTION

Phil Williams, MidSouth Bioinformatics, University of Arkansas at Little Rock

Long non-coding RNAs (lncRNA) are involved in epigenetics and development. They are also implicated in cancer. In contrast to the microRNAs (miRNA), primarily involved in endogenous gene silencing, the lncRNAs have a variety of functions. Protein coding potential has been used to classify lncRNA from mRNA. The application of machine learning to function prediction using protein coding potential is under investigation. Machine learning is being applied to training predictors of a variety of types. These includes the binary classes of tissue-specific v.s. non-tissue-specific. Also within the set of tissue-specific, a predictor for nineteen specific tissues was trained. These tissues include brain, breast, kidney, liver, prostate and others. Prediction accuracy was assessed by calculating specificity based on Leave-one-out testing. Four tissues have specificity greater than 90%, these include brain 97.53%, liver 94.62%, and heart 93.33%. Tissues with specificity values between 80 and 90% are lung 87.04%, kidney 84.91% and Thyroid 84.62%. However, tissues prostate, placenta, breast and ovary have specificity values below 68%. Methods are under investigation to improve predictive accuracy for several classes of lncRNA. These include tissues specific classes of lncRNAs as well as GO ontologies. *This project was supported by the Arkansas INBRE program, with grants from the National Center for Research Resources - NCRR (P20RR016460) and the National Institute of General Medical Sciences - NIGMS (P20 GM103429) from the National Institutes of Health.*

Abstract Identifying Number: 100234

EARTHWORM TOXICOGENOMICS: A 21ST CENTURY APPROACH AND TOOLBOX FOR ENVIRONMENTAL PREDICTIVE TOXICOLOGY

Ping Gong*[1], Huixiao Hong[2], Nan Wang[3], Chaoyang Zhang[3] and Edward J. Perkins[1]

[1*]Environmental Laboratory, US Army Engineer Research and Development Center, Vicksburg, MS, 39180
[2]National Center for Toxicological Research, US FDA, Jefferson, AR, 72079
[3]School of Computing, University of Southern Mississippi, Hattiesburg, MS, 39406

As a sentinel species, earthworms have been considered among the best bioindicators or biomonitors for soil ecosystems owing to their close contact with the environment and essential roles in soil pedogenesis, structure, fertility and the terrestrial food chain. Earthworms have also been used extensively for assessing environmental risk and chemical toxicity in laboratory and field settings. In the past two decades, a comprehensive set of transcriptomic, proteomic, metabolomic and bioinformatic tools have been developed and applied to assess ecological impacts of contaminated soils on earthworms. In this presentation, we briefly review the development of earthworm toxicogenomics, then focus on what it can deliver to the mechanism-based next-generation risk assessment. Our emphasis is placed on the tools we developed in house for earthworm toxicotranscriptomics, and novel biomarkers discovered and mechanistic insights gained using these tools through toxicogenomics studies. Finally we will provide some remarks on the future perspectives of this interdisciplinary and promising field.

Funding sources: US Army Environmental Quality and Installation Research Program and NSF EPSCoR.

Abstract

Abstract Identifying Number: 1006288

COMPUTATIONAL MOLECULAR MODELING SIMULATIONS OF THREE DIMESIONAL PROTEIN STRUCTURES USING DATA MINING TECHNIQUES

Prasant Allaka[1], Jerry Darsey[2], Venkata Kiran Kumar Melapu[1], Sravanthi Joginipelli[1] and Karl Walker*[1]

[1*]Department of Mathematics and Computer Science, University of Arkansas at Pine Bluff, Pine Bluff, AR, 71601
[2]Department of Chemistry, University of Arkansas at Little Rock, Little Rock, AR, 72204

Proteins possess a complex chemical structure with unique repeated alpha amino acid units. The amino acids are directional, asymmetric about the central atom and possess a planar amide group. The spatial configurations of proteins are determined by the pairs of rotational angles about single bonds flanking the alpha carbon atom of each residue. This project is an attempt to predict the most probable three-dimensional conformational structure of the proteins using molecular dynamics simulations, which is based on the repeated random sampling technique. This method is based on calculating probabilities heuristically. Evolution-based Protein structure prediction methods seek to find conserved sequence patterns, while folding method simulate the physical process of folding. A folding pathway is a time series of protein folding events. Unlike most molecular simulation methods, our simulations create a pathway implicitly. While other methods enforce certain characteristics of the folding events during the simulation, including some genetic algorithms, our molecular dynamics simulations method addresses the problem of calculating thermal averages along the lines of statistical thermodynamics, by constructing a random walk through the configurational space of the considered protein model system. In this approach the apriori probabilities, which are proportional to the conditional probabilities, are replaced by the probabilities obtained from mining the data from Protein Data Bank. This generate a huge possible number of chains that can be possible secondary structures of the protein that is given as input. The algorithm incorporates those features, which restrict the energy states of confirmation based on restricting the probabilities given as input and can observe the response of the proteins in a different environment. *This project was made possible by the Arkansas INBRE program, supported by grant funding from the National Institutes of Health (NIH) National Institute of General Medical Sciences (NIGMS) (P20 GM103429) (formerly P20RR016460).*

Abstract Identifying Number: 1006269

HEPATOTOXICITY AND MITOCHONDRIAL TOXICITY OF FDA APPROVED TYROSINE KINASE INHIBITORS

(Qiang Shi, Jun Zhang, Xi Yang and William Mattes

Division of Systems Biology, National Center for Toxicological Research, US Food and Drug Administration, Jefferson, AR, 72079

Unlike traditional chemotherapy drugs that kill all types of cells, tyrosine kinase inhibitors (TKIs) are designed to target proteins that are aberrantly expressed in only cancerous but not normal cells. Therefore TKIs are expected to be highly effective against cancers and have few side effects in healthy organs. The US Food and Drug Administration (FDA) has approved 33 TKIs for cancer treatment. However, clinically significant hepatotoxicity has been associated with over 80% of TKIs. The mechanism is unknown. We recently found that the newly approved TKI regorafenib caused strong mitochondrial toxicity in rat primary hepatocytes at clinically relevant concentrations, and mitochondrial damage was the initial step leading to hepatocyte necrosis, indicating that some TKIs can have "off-target" effects that contribute to the organ toxicity. We then examined the mitochondrial toxicity of all FDA approved TKIs in isolated rat liver mitochondria. We focused on the oxygen consumption rate, inner membrane potential, inner membrane permeability transition, outer membrane integrity, and reactive oxygen species. It was found that about 40% of TKIs caused mitochondrial toxicity at clinically relevant concentrations. Mitochondrial toxicity appears to be an important mechanism and useful predictor of TKI hepatotoxicity. (*This project is supported by the U.S. FDA's Office of Women's Health and NCTR. The information in these materials is not a formal dissemination of information by FDA and does not represent agency position or policy.*)

Abstract Identifying Number: 1006296

FM-INDEX BASED LIGHT-WEIGHT ALIGNMENT FOR QUANTIFICATION OF RNA-SEQ DATA

Quang Tran, and Vinhthuy Phan

Department of Computer Science, University of Memphis, Memphis, TN, 38152

High-thoughput mRNA sequencing (RNA-seq) technologies has led to the availability of transcripts having been discovered recently. Accurate computational methods are intensive for transcriptome quantification from large set of RNA-seq reads, and also provide a critical component in RNA-seq differential expression analysis. Most existing RNA-seq quantification tools require either the correct alignments of reads to transcriptomes (alignment methods) or estimating k-mer counts within each transcript clusters (alignment-free methods). The first approach tends to be more accurate while the second approach tends to be more computationally efficient. We present a new light-weight method for quantifying transcript abundances from RNA-seq data. This method was designed to have the accuracy of alignment-based methods and efficiency of alignment-free methods in estimating transcript abundance. The method utilizes the FM-index data structure to map RNA-seq reads to a collection of transcripts with fast speed and high accuracy. Fast exact-substring search using an FM index supplies efficient computation matching alignment-free methods. At the same time, this method removes the limit of requiring fixed-length k-mers to index reference genomes. We demonstrated the efficiency and effectiveness of this method for quantifying transcript abundances on simulated data. The method is particularly faster and as accurate as current methods. Our software is written in Go, and is available as open-source software.

Abstract Identifying Number: 1006278

A MITOCHONDRIAL GENOME OF THE TIMBER RATTLESNAKE (*CROTALUS HORRIDUS*) AND A REPTILIAN MITOCHONDRIAL PHYLOGENY

Rachel Steele[1], Mark A. Arick II[2], Andy D. Perkins[3], Bindu Nanduri[4], Daniel G. Peterson[2], Douglas D. Rhoads[5], William S. Sanders[2, 6]

[1]Department of Biological and Environmental Sciences, Troy University, Troy, AL 36082
[2]Institute for Genomics, Biocomputing, and Biotechnology, Mississippi State University, Mississippi State, Mississippi, 39762
[3]Department of Computer Science and Engineering, Mississippi State University, Mississippi State, Mississippi, 39762
[4]College of Veterinary Medicine, Mississippi State University, Mississippi State, Mississippi, 39762
[5]Department of Biological Sciences, University of Arkansas, Fayetteville, Arkansas, 72701
[6]The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut, 06032

The timber rattlesnake (*Crotalus horridus*) is an important species of snake to study due to its unique physiology, which makes it a good model for studying hibernation in animals. A new mitochondrial genome was assembled and annotated for *C. horridus,* and a mitochondrial phylogeny depicting the relationships between *C. horridus* and other reptiles was constructed using the genes found within the mitochondrial genomes of these organisms. Mitochondrial DNA sequences were collected from a *C. horridus* specimen captured in Madison County, Arkansas and were mapped against the previously completed mitochondrial genome of another *C. horridus* specimen collected in Rutherford County, Tennessee, generating a new mitochondrial genome for *C. horridus*. Single nucleotide polymorphisms between the new genome and the reference genome were identified because the snake from Arkansas was thought to originate from population thought to have a higher inbreeding rate and a greater degree of homozygosity than the snake from Tennessee. The newly assembled mitochondrial genome was annotated, and the thirteen protein-encoding genes found within it were compared to the protein-encoding genes found in the mitochondrial genomes of sixty-six other species of snakes, two species of alligators, and five species of each of the following: lizards, birds, turtles, and crocodiles. A phylogenetic tree was constructed to visually represent the relatedness of *C. horridus* to these other snakes and reptiles. This phylogenetic analysis demonstrated the similarities and differences between the mitochondrial genomes of *Crotalus horridus* and snakes and other reptiles. *This work was supported by the National Science Foundation under grant DBI-1262901 REU Site: Undergraduate Research in Computational Biology at Mississippi State University.*

Abstract

Abstract Identifying Number: 100316

Single model quality assessment using protein structural and contact information with machine learning techniques

Renzhi Cao[1], Badri Adhikari[1], Debswapna Bhattacharya[1], Miao Sun[2], and Jianlin Cheng*[1,3]

[1*]Department of Computer Science, University of Missouri, Columbia, Missouri 65211, USA
[2]Department of Electrical and Computer Engineering, University of Missouri, Columbia, Missouri 65211, USA
[3]Informatics Institute, University of Missouri, Columbia, Missouri 65211, USA

Protein model quality assessment (QA) is playing a very important role in protein structure prediction for a long time. It is divided into two groups of methods: single model and consensus QA method. The consensus QA method usually has demonstrated better performance compared with single QA in the previous CASP competitions. In this paper, we propose a novel single model quality assessment method utilizing structural features, physicochemical properties, and residue contact predictions. For the first time, we apply contact information from two state-of-the-art protein contact prediction methods PSICOV and DNcon to generate a score as a feature for quality assessment. In total, we use 12 different features including the score generated from contact prediction, and a two layer neural network is used to train a model on CASP9 datasets. We blindly benchmarked our method on CASP11 dataset as the MULTICOM-CLUSTER server. Based on the evaluation of our method on CASP11, our method is ranked as one of the best single model QA method. Moreover, the good performance of features from contact prediction illustrates potential applications of using contact information in protein quality assessment. Finally, based on our evaluation on CASP11, our method has smaller loss compared with the standard consensus QA method DAVIS_consensus on top 150 models (stage2 on CASP11), showing the good model selection ability of our method.

The work was partially supported by an NIH grant (R01GM093123) to JC.

HIGH PERFORMANCE COMPUTING AND DATA MINING IN BIOINFORMATICS

Dr. Richard S. Segall[1]

[1]Department of Computer & Information Technology, Arkansas State University, College of Business, State University, AR 72467-0130

This presentation will first discusses an overview of supercomputing and how it differs from ordinary computing, and then discuss data mining software that can be used on supercomputers with large data intensity to discover biological patterns in genomics data at genetic level.

This talk illustrates some of the data mining that was performed by the author using SAS JMP® Genomics software on two collections of microarray databases that were available from the website of the National Center for Biotechnology Information (NCBI) for lung cancer and breast cancer. The software used in this research of SAS JMP® Genomics has been used to perform data mining using supercomputers such as those at the University of Minnesota Supercomputing Institute for Applied Computation Research.

The PowerPoint slides provides a selection of the data mining results that were obtained using SAS JMP® Genomics on the microarray databases retrieved from the Gene Expression Omnibus (GEO) repository on the website of National Center for Biotechnology Information (NCBI). The data mining performed in this research was performed to uncover meaningful patterns and results at a high level of density of data such as investigating the interlinked biological pathways represented by the DNA sequencing, and hence the use of supercomputers is a valuable tool in this research.

This presentation will also discuss the applications of supercomputing to the Human Brian Project, cancer genome analysis, sequence analysis and Genome Annotations, population genetics, and modeling of biological systems as presented in book edited by Segall et al. (2015). The talk will conclude with future directions of the research such as preliminary work on high performance data mining of plant imagery with available open-source plant image analysis software tools.

References:

Segall, RS, Cook. JS, and Zhang, Q. (2015), Research and Applications in Global Supercomputing, IGI Global, Hershey, PA, ISBN 978-1-4666-7461-5.

REGULATION OF STEROL TRANSPORT IN RESPONSE TO AGING IN SACCHAROMYCES CEREVISIAE

Thomas Hahn[1,2], Richard Segall [4], Helen Benes[2] and Fusheng Teng*[3]

[1]Dept. of Information Science, University of Arkansas at Little Rock (UALR), Little Rock, AR, 72204
[2]University of Arkansas for Medical Sciences (UAMS), Little Rock, AR, 72205
[3*]Department of Applied Science, University of Arkansas at Little Rock (UALR), Little Rock, AR, 72204
[4]Department of Computer and Information Technology, Arkansas State University (ASU), State University, AR, 72467-0130

Purpose: The objective is to understand how sterol transport is affected by caloric restriction (CR) and aging. Sterol synthesis generally increases as cells age. Yet, despite this increase, the sterol content of selected organelle membranes, such as lysosomal membranes, declines with age and adversely affects cellular functions. We are interested in a better understanding of changes in sterol transport to membranes of vacuoles (the yeast counterparts of mammalian lysosomes) in response to anti-aging and pro-aging manipulations.

Methods: We looked for changes in expression of genes that could be involved in lifespan extending mechanisms triggered by CR. We used microarray data to determine the responses to different combinations of mutations and to food availability and to examine differential expression of those genes. We constructed "heat maps" to identify the relevant genes and determined their roles in biological pathways, using this information to draw relevant gene interaction networks.

Results: We found that atg15 is required for lifespan extension in longevity mutants mimicking CR. Thus, we propose that autophagy relieves the ROS-stress induced by CR and plays a key role in lifespan extension.

Conclusions: Identifying changes in gene expression that maintain the sterol proportion and asymmetry across the vacuolar membrane in yeast could be the basis for novel drug development aimed at altering the expression of gene homologues in humans, such that adverse age-associated changes in the human lysosomal membrane could be postponed or even reversed. We have identified a number of gene candidates that can be easily studied in yeast for drug design.

This project was supported by the Arkansas INBRE program, with grants from the National Center for Research Resources - NCRR (P20RR016460) and the National Institute of General Medical Sciences - NIGMS (P20 GM103429) from the National Institutes of Health (NIH).

Abstract Identifying Number: 1006314

MULTI-LABEL SUPPORT VECTOR MACHINE CLASSIFICATION FOR INTELLIGENT HEALTH RISK PREDICTION

Runzhi Li*[1, 2], **Hongling Zhao**[1], **Chaoyang Zhang**[2], **and Yusong Lin**[1]

[1]Cooperative Innovation Center of Internet Healthcare, Zhengzhou University, Zhengzhou, Henan, 450000
[2*]School of Computing, University of Southern Mississippi, Hattiesburg, MS, 39402

Chronic diseases not only have a major impact on the patients' quality of life, but also comprise the bulk of healthcare costs and constitute the leading cause of mortality. In preventive medicine, it is important and feasible to identify and prevent the health and disease risks as early as possible through regular physical exams. How to analyze these physical records and predict the health risks remains a challenging problem in machine learning and data mining research. Many current researches are based on single label learning assuming that each medical record is associated with a single label (a type of disease) and all labels are disjointed. However, in many medical records, each one is related to multiple types of diseases such as hypertension and diabetes and other comorbidities so the single label learning method cannot be directly used for the data analysis and risk prediction. In this paper, a multi-label classification method was developed for more accurate health risk prediction, in which a new class label is created if a record is related to a combination of two or more types of diseases. With the new label set, the original multi-label classification problem is transformed to a multi-class classification problem that can be solved using existing classification methods, such as Support Vector Machine (SVM). To deal with the imbalance learning problem with varying class sizes, a penalty parameter was introduced and tuned for different classes during the training process of LIBSVM to improve the prediction accuracy. The 10-fold crossing validation was performed and the metrics of Precision, Recall and F-measure were used to evaluate the performance. As an example, the real physical exam records were employed which include 62 exam items and 11020 anonymous patients, and 10 diseases were predicted by the multi-label classification method. Experiment results showed that the multi-label classification method is more accurate than the method without using multi-label mapping.

Abstract Identifying Number: 1006294

LEVERAGING GRAPH TOPOLOGY AND SEMANTIC CONTEXT FOR PHARMACOVIGILANCE IN TWITTER STREAMS

Ryan Eshleman, Rahul Singh

Department of Computer Science, San Francisco State University, San Francisco, CA, 94132

Adverse drug events (ADEs) constitute one of the leading causes of post-therapeutic death and their identification constitutes an important challenge of modern precision medicine. Unfortunately, the onset and effects of ADEs are often underreported complicating timely intervention. At over 500 million posts per day, Twitter is a commonly used social media platform. The ubiquity of day-to-day personal information exchange on Twitter makes it a promising target for data mining for ADE identification and intervention. Three technical challenges are central to this problem: (1) identification of salient medical keywords in (noisy) tweets, (2) mapping drug-effect relationships, and (3) classification of such relationships as adverse or non-adverse.

We first use dictionary-based and algorithmic information extraction to identify occurrences of medically-relevant concepts in tweets. Next, a drug-effect graph (DEG) is constructed by mining users' tweet history. In the DEG, drugs and symptoms are connected with edges weighted by temporal distance and frequency. From this graph, edges are classified as either adverse or non-adverse with a classifier trained using both graph-theoretic and semantic features such as sentiment polarity of the source text. The proposed approach can identify adverse drug effects with high accuracy as measured by precision (.72), recall (.73), and F1 scores (.72). When compared with leading methods at the state-of-the-art, which employ graph-theoretic analysis alone, the proposed method leads to improvements ranging between 9% to 12% (in terms of the aforementioned measures).

Abstract Identifying Number: 1002186

COMPARATIVE REVERSE VACCINOLOGY ANALYSIS OF *AEROMONAS HYDROPHILA* ML09-119 GENOME

Hasan C. Tekedar, Safak Kalindamar, Attila Karsi and Mark L. Lawrence

College of Veterinary Medicine, Department of Basic Sciences, Mississippi State University,
Mississippi State, MS, 39762

Reverse vaccinology is a high-throughput computational method analyzing pathogenic bacteria genomes to identify potential vaccine candidate proteins. *Aeromonas hydrophila* is a Gram-negative opportunistic pathogen causing diseases in both animals and humans. The purpose of the current research was to determine conserved *Aeromonas* proteins that have the potential for the host immune response. First, we analyzed the genome of fish pathogen *A. hydrophila* strain ML09-119 by reverse vaccinology using the subcellular localization, transmembrane helices, and adhesion probability. The result was compared against the genomes of *A. salmonicida* A449, *A. hydrophila* ATCC 7966, and *A. veronii* B565. Among the total 4119 proteins in *A. hydrophila* ML09-119 genome, 59 vaccine candidate proteins were identified and compared to the other three *Aeromonas* genomes. Most of these proteins were predicted to be extracellular or outer membrane proteins. Fourteen of these proteins had orthologs identified in all three of the other *Aeromonas* genomes. The identified proteins had predicted functions as adhesion proteins, extracellular enzymes, toxins, receptors, lipoproteins, motility proteins, and membrane transport proteins. *This work was supported by the Mississippi State University College of Veterinary Medicine, the USDA Agricultural Research Service CRIS project 6402-31000-009-00D, and the Alabama Agricultural Experiment Station (Hatch project number ALA021-1-09005).*

Abstract Identifying Number: 1006287

COMPLETE GENOME SEQUENCE OF *FLAVOBACTERIUM COLUMNARE* STRAIN 94-081

Salih Kumru[1], Hasan C. Tekedar[1], Geoffrey C. Waldbieser[2], Mark L. Lawrence[1], and Attila Karsi[1]

[1]College of Veterinary Medicine, Mississippi State University, Mississippi State, Mississippi 39762, USA
[2]United States Department of Agriculture, Agricultural Research Service, Stoneville, Mississippi 39776, USA

Channel catfish is one of the most important commodities in Mississippi and the largest aquaculture industry in the United States. Columnaris disease, caused by *Flavobacterium columnare*, affects many commercially important freshwater fish species. *F. columnare* strain 94-081 was isolated from diseased channel catfish in 1994 from a commercial aquaculture production pond in Mississippi. *F. columnare* is divided into three genomovars, and strain 94-081 is representative of genomovar II, which tends to be highly pathogenic for catfish. To understand pathogenesis of strain 94-081, we sequenced its genome using 454, Illumina, and PacBio sequencing and obtained a circular genome of 3,321,600 bp. The genome contains 2,901 coding sequences and 82 RNAs. Forty potential elements responsible for defense mechanisms and virulence were detected. This is the first reported *F. columare* genomovar II genome sequence, and we expect it will help our understanding of *F. columnare* genomovar II pathogenesis in catfish. *This project was supported by the USDA (2006-35600-16571) and Mississippi State University College of Veterinary Medicine.*

Poster presentation

Abstract Identifying Number: 1006270

A Graph-theoretic Model of Nucleotide Binding Domain 2 of the Cystic Fibrosis Transmembrane Conductance Regulator

Samuel Kakraba[1], Debra Knisley[2]

[1]UALR/UAMS Joint PhD Program in Bioinformatics, University of Arkansas at Little Rock, Little Rock, AR 72204, USA .

[2]Department of Mathematics and Statistics, East Tennessee State University, Johnson City, TN 37614, USA.

One of the most prevalent inherited genetic diseases is Cystic fibrosis. This disease is caused by a mutation in a membrane protein, the cystic fibrosis transmembrane conductance regulator (CFTR). CFTR functions as a chloride channel regulating the viscosity of mucus lining the ducts of several organs. Most the common mutations of CFTR occur in the nucleotide binding domain 1 (NBD1). However, some mutations in nucleotide binding domain 2 (NBD2) also cause Cystic fibrosis.

In this work, we model NBD2 with a nested graph to study the effect of single point mutations on the NBD2. The vertices in the lowest layer each represents an atom in the structure of an amino acid residue, while the vertices in the mid layer each represents the residue. The vertices in the top layer each represent a subdomain of the NBD2. We use this model to quantify the effects of a single point mutation on the protein domain. Comparison of the wildtype structure with eight of the most common mutations is done. The graph-theoretic model provides insight into how a single point mutation can have such profound structural consequences. We are guided by the hypothesize that graph theory can be used to model and computationally quantify amino acids, thereby obtaining reliable molecular descriptors for studying effect of single point mutations on a protein domain and possibly suggest the line of action for drug design for this disease and other mutation-related diseases.

Supported by grants from NCRR (P20RR016460) and NIGMS (P20 GM103429) at NIH.

Abstract Identifying Number: 1006270

EFFECTS OF SMALL MOLECULES ON PROTEIN AGGREGATION AND PARALYSIS IN C. ELEGANS STRAIN EXPRESSING Aβ_{1-42} IN THE MUSCLE

Samuel Kakraba,[1] **Narsimha R. Penthala,[4] Peter A. Crooks,[4] Robert J. Shmookler Reis,[2,3] and Srinivas Ayyadevara[2,3]**

[1]UALR-UAMS Joint Program in Bioinformatics, Univ. of Arkansas, Little Rock, AR 72204
[2]Central Arkansas Veterans Healthcare System, Little Rock AR 72205
[3]University of Arkansas for Medical Sciences, Little Rock AR 72205
[4] Department of Pharmaceutical Sciences, College of Pharmacy, UAMS, Little Rock, AR 72205

Proteins require correct folding and maintenance in order to function effectively and efficiently. Most or all common neurological disorders, such as Alzheimer's and Parkinson's diseases, and possibly a wide range of other age-associated diseases, are attributable to protein aggregation that is cytotoxic, especially to nerve cells. Protein aggregation is a biological phenomenon in which misfolded proteins aggregate (i.e., adhere together in large conglomerates) either intra- or extracellularly. Our goal is to determine whether anti-inflammatory compounds (i.e. parthenolide, sclareol, Combretastatin, and thiadiazolidinones (TDZD) analogs) are effective at reducing protein aggregation as well as preventing paralysis in *C. elegans* strain CL4176, which expresses a human Aβ_{1-42} transgene in body-wall muscle. In addition to conducting studies on a library of small molecules as a first step in an iterative process of drug optimization, we have also assessed dose-response functions for active lead compounds in reducing protein aggregates.

Acknowledgement

This project was made possible by the Arkansas INBRE program, supported by grant funding from the National Institutes of Health (NIH) National Institute of General Medical Sciences (NIGMS) (P20 GM103429) (formerly P20RR016460).

MCBIOS 2016 Abstract

Abstract Identifying Number: 1006282

SILVER NANOPARTICLES: EFFECT ON INTESTINAL MICROBIOME AND DEVELOPMENT OF RESISTANCE

Sangeeta Khare[1], Kuppan Gokulan[1], Katherine Williams[1], Luis Valancia[1], Mary Boudreau[2] and Carl E Cerniglia[1]

[1]Division of Microbiology, National Center for Toxicology Research, Jefferson, AR, 72079
[2]Division of Biochemical Toxicology, National Center for Toxicology Research, Jefferson, AR, 72079

Silver has been utilized as an antimicrobial against pathogenic bacteria. Recent advances in the field of nanotechnology have led to its incorporation as silver nanoparticles (AgNP) into consumer-used products and health-supplements. The effects of intentional/unintentional exposure AgNP on the gastrointestinal homeostasis, as well as, the gut-microbial communities are unknown. Moreover, like many antimicrobial-agents, growing public health concerns is if the antimicrobial property of AgNP may also cause resistance.

The aim of this study was to see the effect of AgNP on the overall population of the microbial communities. Ileal mucosal samples were taken from the Sprague-Dawley rats (male and female) that were gavaged orally with discrete sizes of AgNP (10, 75 and 110 nm) and silver acetate for 13 weeks. DNA was extracted and subjected to 16S gene sequence-based analysis to examine bacterial communities in the ileum. The Weighted UniFrac-dissimilarity was used to calculate taxon abundance in different samples; whereas UniFrac method was used to assess the presence/absence of taxa. The Adonis test is utilized for finding significant whole microbiome differences. In general, bacterial community richness varied with the treatment. Based on presence and absence, a dominance of phylum Firmicutes was observed. When comparing the microbiome of dose effect, 38 Operation Taxonomic Units (OTUs), with significant in their abundance were detected. A significant difference based on these 38 OTUs was observed between three doses and controls. Principal component analysis showed a separation of microbiome between samples from 18 mg/ml and all other doses. A difference in presence/absence of OTUs at phylum-level was observed for Planctomycetes. A significant difference in OTU abundance at phylum-level was observed for Proteobacteria. Presence of silver resistance genes was detected in all experimental groups. Furthermore, we have confirmed the cross-talk between the bacterial species and host functional properties using the in vitro and ex vivo model.

Abstract Identifying Number: 1006239

MATRIX LINEAR MODELS FOR HIGH THROUGHPUT GENETIC SCREENS

Jane W Liang[1], Mark R. Segal[2], and Śaunak Sen[3]

[1*]Kaiser Permanente, Pleasanton, CA 94588
[2]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, CA 94143
[3]Department of Preventive Medicine, University of Tennessee Health Science Center, Memphis, TN 38163

We will outline statistical models for high throughput genetic screens. In such screens, one measures the fitness of a library of genetic mutants in a variety of growth conditions. Mutants might be grouped by gene or gene family; growth conditions might be grouped by antibiotic class or temperature. Our proposed models provide a simple framework for encoding such known relationships (eg. growth condition class and gene family of mutant) to enhance detection of associations that might otherwise be masked. We show that fast estimation algorithms can be developed by taking advantage of the structure of those models. Our method's performance in simulations and on an *E. coli* chemical genetic screen will be presented.

Abstract Identifying Number: 1006299

TEXTURE HOMOGENEITY ANALYSIS OF LESION BORDER IN DERMOSCOPY IMAGES FOR MALIGNANCY DETECTION

Sertan Kaya[1], Sinan Kockara[1*], Mutlu Mete[2], Tansel Halic[1], Halle E. Field[3] and Henry K. Wong[3]

[1*]Department of Computer Science, University of Central Arkansas, Conway, AR, 72035
[2]Department of Computer Science and Information Systems, Texas A&M University-Commerce, Commerce, TX, 75428
[3]Department of Dermatology, University of Arkansas for Medical Sciences, Little Rock, AR, 72205

Automated skin lesion border examination and analysis techniques have become an important field of research for distinguishing malignant pigmented lesions from benign lesions. An abrupt pigment pattern cutoff at the periphery of a skin lesion is one of the most important dermoscopic features for detection of neoplastic behavior. In current clinical setting, the lesion is divided into a virtual pie with eight sections. Each section is examined by a dermatologist for abrupt cutoff and scored accordingly, which can be tedious and subjective. This study introduces a novel approach to objectively quantify abruptness of pigment patterns along the lesion periphery. In the proposed approach, first, the skin lesion border is detected by the density based lesion border detection method. Second, the detected border is gradually scaled through vector operations. Then, along gradually scaled borders, pigment pattern homogeneities are calculated at different scales. Through this process statistical texture features are extracted. Moreover, different color spaces are examined for the efficacy of texture analysis. The proposed method has been tested and validated on 100 (30 melanoma, 70 benign) dermoscopy images. Analyzed results indicate that proposed method is efficient on malignancy detection. More specifically we obtained specificity of 0.96 and sensitivity of 0.86 for malignancy detection in a certain color space. The F-measure, harmonic mean of recall and precision, of the framework is reported as 0.87. *This research was partially supported by Arkansas Science and Technology Association Award# 15-B-25.*

Abstract Identifying Number: 1006099

Analysis of optimal alignment unfolds bias in existing variant profiles

Shanshan Gao, Quang Tran, and Vinhthuy Phan

Department of Computer Science, University of Memphis, Memphis, TN 38152, USA

Accurate detection of the insertion/deletion (INDEL) variants is a hard problem in the computational procedure to detect genetic variants. Our analysis revealed that public information of already detected INDELs from the 1000 Genome Project contained many INDELs that were constructed in a bias manner. This bias occurred at the level of aligning short reads to reference genomes to detect variants. The bias is caused by the existence of many theoretically optimal alignments between the reference genome and reads containing alternative alleles at those INDEL locations.

We examined several popular aligners and showed that these aligners could be divided into two groups: (1) those whose alignments yielded INDEL that agreed heavily with the information reported by the 1000 Genome Project, and (2) those whose alignments yielded INDEL that disagreed heavily with the information reported by the 1000 Genome Project. This finding suggests that the agreement or disagreement between the aligners' called INDEL and the reported INDEL is merely a result of arbitrary selection of one of the optimal alignments. Due to the importance of INDEL, our finding suggests that this phenomenon should be further addressed.

Abstract Identifying Number: 1006099

An alignment-based method for profiling microbial community using compressed FM-index

Shanshan Gao, Diem-Trang Pham, Vinhthuy Phan

Department of Computer Science, University of Memphis, Memphis, TN 38152, USA

Determining abundances of microbial genomes in metagenomic samples is an important problem in analyzing metagenomic data. This task can be a computationally expensive since microbial communities usually consist of hundreds to thousands of environmental microbial species. For instance, microorganisms in the human gut contain up to 100 trillion cells. Many methods that estimate the relative abundance of microbial genomes are based on indexing those genomes using fixed-length k-mers. These methods are very efficient and relatively accurate. A challenge of these methods is to identify the right value of k to index reference microbial genomes. Small values of k will result in inaccurate estimates, whereas large values of k would require infeasible computational resources. Our proposed method is essentially an indexing approach that is based on a variable-length value of k. The approach employs a compressed FM index of all reference microbial genomes. This index allows exact substring search between reads and reference genomes. The effect of this method is analogous to indexing microbial genomes using variable lengths k-mers.

In a preliminary test of our method, we indexed 244 oral bacteria genomes. Using 2x coverage of simulated reads, we were able to predict correctly the abundance of 209 out of 244 bacteria.

Liver Toxicity Knowledge Base (LTKB): A comprehensive database to understand multiple dimensions of Drug-Induced Liver Injury

Shraddha Thakkar, Minjun Chen, Hong Fang, Zhichao Liu, Gerry Zhou, Jie Zhang, Weida Tong
National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, AR 72079

Drug Induced Liver Injury (DILI) is one of the foremost reasons for acute liver failure along with cause of termination of many clinical trials and discontinuation of many approved drugs. Therefore, DILI is one of the major concerns during the drug development as well as the reviewing process. To address that issue, Liver Toxicity Knowledge Base (LTKB) was developed to improve our understanding of underlying mechanisms involving DILI and its prediction. This knowledgebase includes the DILI related information from ~3000 prescription drugs, such as drug physicochemical properties along with its dosage, side effect, and therapeutic uses. In addition, mechanistically relevant cellular end points from various *in vitro* assays (conventional, high-throughput and high-content assays), drug-elicited toxicogenomic responses from both primary hepatocytes and animals, and histopathology were also incorporated in the database. We also linked the LTKB data to the data from the ToxCast (from EPA) and Tox21 (from NIH, EPA, National toxicology program and FDA) projects. LTKB drugs were analyzed for its potential to cause DILI and classified based on its severity. DILI annotations were also identified from the FDA-approved drug labels that provide the safety information from clinical trials and post marketing surveillance. As a result, 749 drugs were identified with some level of DILI concern and 619 drugs further annotated for the DILI information from LiverTox database (from NIH). Database was developed on Accelrys Isentris 4.0 platform and provides the comprehensive DILI information and findings at various level of biological complexity at one location. In summary, this poster will provide the use cases for extracting the desired information from database for generating the better understanding of DILI. This knowledgebase can be a resource to improve DILI predictive model for drug discovery and drug safety.

Abstract Identifying Number: 1006323

DETERMINATION OF NEW BIOSYNTHETIC PATHWAYS OF ASCORBIC ACID USING BIOINFORMATICS MODELLING

Skylar Connor[1], Dr. Grant Wangila[2], and Dr. Karl Walker*[1]

[1*]Department of Computer Science, University of Arkansas at Pine Bluff, Pine Bluff, AR, 71601
[2]Department of Chemistry, University of Arkansas at Pine Bluff, Pine Bluff, AR, 71601

Literature results shows that plants that have been treated with ascorbic acid exhibit health growth and are resistance to insects resulting to better yield. Since ascorbic acid is produced by plants as they form fruits, it’s imperative for us to examine in details how this process occurs metabolically. In this project, computer models will be used to identify biosynthesis of ascorbic acid in plants, test for its presence in plant material and suggesting ways of how it can be enhanced in plants so that better plants can thrive. Detailed investigation of metabolic pathways of ascorbic acid, degradation in both plants and animals are vital tool s that will assist in establishing the threshold concentrations needed. The overall goal of this project is to characterize the biosynthetic and metabolic pathways of ascorbic acid that exist in nature. Furthermore, we plan to employ bioinformatics algorithms and tools to find or engineer new biosynthetic pathways of ascorbic acid and introduce these new pathways into plants that do not currently synthesize vitamin C.

Supported by grants from NCRR (P20RR016460) and NIGMS (P20 GM103429) at NIH.

Abstract Identifying Number: 1006277

PROMISE-ME: A Robust Method for Integrated Analysis of DNA Methylation, Gene Expression, and Multiple Biologically Related Clinical and Pharmacological Outcomes

Xueyuan Cao, Tong Lin, and Stan Pounds

Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN 38135

Projection onto the Most Interesting Statistical Evidence (PROMISE) is a robust method to perform integrated analysis of one form of genomic data with multiple biologically related pharmacologic and clinical outcome variables. PROMISE has been used successfully to identify genomic and transcriptomic features of pharmacogenetic relevance to the treatment of pediatric leukemia. Here, we extend that framework to develop PROMISE for methylation and expression (PROMISE-ME) as a robust method to perform integrated analysis of DNA methylation, gene expression, and multiple pharmacologic and clinical outcome variables. For each gene, PROMISE-ME evaluates the association of methylation with expression, the association of methylation with each outcome variable, and the association of expression with each outcome variable. Previous knowledge of the biological relationships among outcome variables is used to define a test statistic that is a linear combination of each of the pairwise association statistics described above. Permutation is used to determine the significance of this test statistic. The advantages of PROMISE-ME in terms of enhancing statistical power for meaningful biological discoveries are seen in simulation studies and an example from pediatric oncology research. *Funding provided by ALSAC and NIH.*

Abstract Identifying Number: 1006271

THE MICROBIOME AND ITS RELATIONSHIP WITH ANTIMICROBIAL RESISTANCE

Steven L. Foley

Division of Microbiology, National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, AR 72079

The human microbiome is made up of unique populations of microorganisms that reside in the different regions of the body and carry out a multitude of functions that impact human health. In recent years a great deal of attention has been paid to understanding the contributions of specific members of the microbiome to improving human health or leading to a disease state. One area of concern is whether the microbial populations within the gastrointestinal tract can serve as a reservoir for antimicrobial resistance genes, which can complicate antimicrobial therapy during an infection. Antimicrobial resistance genes are known to be carried by a wide range of bacterial taxa and are often located in mobile genetic elements, such as integrons, transposons and plasmids that can facilitate the spread of resistance among bacteria. The gastrointestinal microbiome can be exposed to bacteria containing antimicrobial resistance genes through the food and water supplies, person-to-person contact and from environmental sources. Researchers have utilized various different approaches to identify microorganisms present in the gastrointestinal tract, determine those taxa that are associated with antimicrobial resistance and assess the potential for horizontal spread of the resistance. These approaches include 16s rRNA high-throughput sequencing to identify the members of the microbiome, targeted amplification and sequencing of resistance genes and mobile genetic elements, and shotgun approaches to characterize the gastrointestinal metagenome. This presentation will examine laboratory and data analyses approaches that have been used to study the genetics of antimicrobial resistance and potential for resistance spread within the gastrointestinal environment.

Abstract Identifying Number: 1006263

IDENTIFICATION OF CRITICAL CHEMICAL FEATURES TO DIFFERNETIATE ANDROGEN RECEPTOR AGONSITS AND ANTAGONISTS: PHARMACOPHORE MODELING AND MODLECULAR DOCKING

Sugunadevi Sakkiah and Huixiao Hong*

[*]Division of Bioinformatics and Biostatistics, NCTR/FDA, Jefferson, AR, 72079

Androgen receptor (AR) is a ligand-dependent transcription factor and member of the nuclear receptor superfamily. It plays a vital role in male sexual development and regulates the gene expression in a variety of tissues. Binding of small molecules to AR initiates the conformational changes in AR that affect AR binding of co-regulator protein and DNA. AR agonists and antagonists are widely used in a variety of clinical applications (i.e. hypogonadism and prostate cancer therapy). In this study, we have applied an integrative approach that combines pharmacophore modeling and molecular docking techniques to identify the important chemical features to differentiate agonists from antagonists. We elucidated binding modes in AR for agonists from antagonists. Five crystal structures of AR bound with ligands (3 agonists and 2 antagonists) were used to generate the structure-based pharmacophore models that were validated by test and decoy data sets. The important chemical features which can distinguish antagonists from agonists were identified from the structure-based pharmacophore modeling. All the AR ligands present in the crystal structures (agonists and antagonists) and the test compounds were used in the molecular docking studies to find the binding modes of the agonists and antagonists in AR ligand-binding pocket. The molecular docking results revealed the relationship between the chemical properties of agonists and antagonists and their hydrogen bond interactions with the crucial residues in AR ligand-binding pocket. Our study demonstrated that this integrative approach has the ability to discriminate AR antagonists and agonists and is expected to facilitate safety evaluation and drug development targeting on AR.

Abstract #: 1006276

PRIORITIZATION, CLUSTERING AND ANNOTATION OF MicroRNAs USING LATENT SEMANTIC INDEXING OF PUBMED ABSTRACTS

Sujoy Roy[1], Brandon C. Curry[1], Behrouz Madahian[2], and Ramin Homayouni*[1, 3]

[1]Bioinformatics Program, University of Memphis, Memphis, TN 38152
[2]Quire Inc., Memphis, TN 38103
[3]Department of Biology, University of Memphis, Memphis, TN 38152

The amount of scientific information about microRNAs (miRNAs) is growing exponentially, making it difficult for researchers to interpret experimental results. In this study, we present a Latent Semantic Indexing (LSI) based automated text mining approach for prioritization, clustering and functional annotation of miRNAs. For 546 human miRNAs indexed in *miRBase*, corresponding text documents were created by concatenating titles and abstracts of publications referencing the miRNAs. The documents were parsed and a weighted keyword (term) by miRNA frequency matrix was created, which was subsequently factorized via singular value decomposition to extract pair-wise cosine associations between the term and miRNA vectors in reduced rank semantic space. LSI enables derivation of implicit associations between entities based on word usage patterns that have not been mentioned together explicitly. Using *miR2Disease* as a gold standard, we found that LSI identified keyword-to-miRNA relationships with high accuracy. In addition, we demonstrate that pair-wise associations between miRNAs can be used to group them into functional categories. These groups were automatically annotated by querying the reduced rank LSI space with the miRNA clusters and extracting relevant terms. Analysis of the clusters revealed known biological relationships as well as lesser known information. The accompanying web tool, *miRNA Literature Network* (*miRLiN*), provides an automated framework for interactively extracting and discovering functional information on human miRNAs in real time.

This work was supported by the University of Memphis Center for Translational Informatics.

Abstract

Abstract Identifying Number: 1006272

TEXT MINING METHODS FOR KNOWLEDGE EXTRACTION FROM FDA APPROVAL LETTERS

Suresh Subramani[1], Hailin Tang[1], Joe Meehan*[1], Salvatore Pepe[2], and Anne Pariser[2]

[1*]Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, Food and Drug Administration, Jefferson, AR, 72079
[2]Office of Translational Sciences, Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring, MD, 20993

When the U.S. Food and Drug Administration (FDA) approves New Drug Applications (NDAs) and Biologic License Applications (BLAs) it issues formal approval letters that identify the drug or biologic, its approved use, any required pediatric assessments, post marketing requirements, and other crucial information. However, it is difficult to extract this detailed information from FDA approval letters, due to the abundant information and the unstructured free text format. To overcome these issues, we applied automated text mining methods to extract the data (Drug information, Semantic metadata, Indications, Pediatric Research Equity Act (PREA) language categories, PREA and Post-market studies, etc.) from the various approval letters and used the data to build a user-friendly database. Natural language processing methods promise to provide regulatory science data to FDA staff in a useful format that encourages building of a learning environment within the FDA. *Funding: USFDA and ORISE*

Abstract

Abstract Identifying Number: 1006279

PAN-CANCER TRANSCRIPTOMIC NETWORK ANALYSIS

Tanzim Hassan[1], **Andy Perkins***[2]

[1]Starkville High School, Starkville, MS, 39759
[*2] Department of Computer Science and Engineering, Mississippi State University, Starkville, MS, 39759

Cancer's status as one of the top ten killers in the United States merits a continuation of research efforts to catalogue novel oncological driver genes and to develop improved diagnostic methods. To this effect, efforts have been made to organize genomic information about various cancers in massive online databases (e.g., The Cancer Genome Atlas). This wealth of genetic data stored in digital format allows for the application of powerful computational methods and tools in the overall fight against cancer.

Herein is presented a pan-cancer co-expression analysis of the transcriptomic landscape of breast, lung, and brain cancer using the data available on The Cancer Genome Atlas. Through a combination of in-house and existing tools for network analysis, natural-language processing based text mining, and gene enrichment analysis, several groups of highly co-expressed genes that have significant associations with oncological phenotypes are identified. Furthermore, a "guilt-by-association" framework to use the results generated in this study to identify novel cancer driver genes is outlined, and a pathway analysis approach to generating more results is proposed.

This submission correctly identifies several well-known cancer driver genes and outlines a framework to identify novel cancer driver genes. Overall, this work offers a comprehensive view of the co-expression transcriptomic landscape of several cancers while laying the foundation for continued work in the area.

Abstract Identifying Number: 100423

ARM-B: MINING BICLUSTERS WITH ASSOCIATION RULES IN GENE EXPRESSION DATA ANALYSIS

Tina Gui[1], Christopher Ma[1], Janet Nakarmi[2], Dawn E. Wilkins[1], Yixin Chen[1]

[1]Dept. of Computer and Information Science, University of Mississippi, Oxford, MS, 38655
[2]Department of Mathematics, University of Mississippi, Oxford, MS, 38655

Microarray technology has created a revolution in the field of biology and bioinformatics. Many data mining techniques, such as clustering, association rules, classification and neural networks, have been applied to microarray data analysis. The common constraints of these methodologies include the limited applicability for large microarray data sets and the restrictions prevent the full utilization of microarray data sets. In this paper, we propose a new biclustering method based on association rule mining to gain insight into gene expression data. The proposed approach can also discover overlapping clusters in high-dimensional data sets. To illustrate the usefulness of the method, we performed a variety of experiments testing the utility of association rule based biclusters for microarray data. The combination of these two techniques not only can be applied to various biological problems dealing with high-throughput microarray data, but also can help biologists and researchers to discover patterns in an efficient manner.

Abstract Identifying Number: 1006047

BAMS DATABASE- A DATABASE FOR BIOACTIVE MOLECULES

Ujwani Nukala,[1,2] Paola E. Ordoñez,[1,4,5] Shraddha Thakkar,[1,2] Darin E. Jones,[5], Monica L. Guzman,[3] and Cesar M. Compadre[1]

[1]Department of Pharmaceutical Sciences, University of Arkansas for Medical Sciences, Little Rock, AR 72205. [2]Joint Bioinformatics Graduate Program, University of Arkansas at Little Rock and University of Arkansas for Medical Sciences, Little Rock, AR 72205. [3]Division of Hematology/Oncology, Department of Medicine, Weill Cornell Medical College, New York, NY 10065. [4]Departamento de Quimica Aplicada, Universidad Tecnica Particular de Loja, Loja Ecuador. [5]Department of Chemistry, University of Arkansas at Little Rock, AR 72205.

Products from nature have been valuable resources of medicinal remedies since ancient times and they continue to be a major source for the development of new therapeutic agents. Many clinically important drugs, from the oldest to the recently approved drugs, are either natural products or their derivatives. Sesquiterpene lactones (SLs) are derived from plants that exhibit a wide range of biological activities including anti-inflammatory, anti-parasitic, anti-bacterial and anti-cancer effects. There are over 10,000 publications reporting on sesquiterpene lactones and their anticancer properties. The study of the structural factors responsible for their anti-cancer activity has been also the subject of numerous studies since early 70's. However, progress in the development of SLs is hampered by the current inability to fully understand their structure activity relationships.

The Bioactive Molecules Database (BAMS) is a web-based open access database of sesquiterpene lactones. The database is implemented using MySQL, the interface is implemented using PHP, and the layout of the website is created using HTML. The objective of this database is to support the research community interested in sesquiterpene lactones by maintaining the records of these molecules at one place. The database contains information such as chemical name, availability, 2D structure, biological activity and 3D molecular structures. Apart from browse function, it has a search function to filter records of interest by the chemical name. To maintain and update the database there is a deposition form available for the users to deposit their molecules. Currently the database includes sesquiterpene lactones belonging to different classes. This database with chemical structure and activity details of the sesquiterpene lactones provides an opportunity to develop QSAR models, docking experiments, virtual screening and other computational approaches searching for potential anticancer drugs. The Database is available at http://amanda.uams.edu/BAMS_Database

This project was supported by: University of Arkansas for Medical Sciences College of Pharmacy Research Fund, the Arkansas Science and Technology Authority, Project No. 15-B-01 and the Arkansas INBRE program, with grants from the National Center for Research Resources - NCRR (P20RR016460) and the National Institute of General Medical Sciences - NIGMS (P20 GM103429) from the National Institutes of Health. PEO is a recipient of a graduate student scholarship from Ecuador's Higher Education, Science, and Technology Ministry (SENESCYT). MLG is a recipient of the NIH Director's New Innovator Award Program, 1 DP2 00007399-0.

Abstract Identifying Number: 1006047

USE OF SURFACE SIGNATURE ANALYSIS TO STUDY THE STRUCTURE ANTI-LEUKEMIC ACTIVITY RELATIONSHIP OF SESQUITERPENE LACTONES

Ujwani Nukala [1,2], Paola E. Ordóñez[1,3,4], Shraddha Thakkar [1,2], David Mery[5], Darin E. Jones[4], Eloy Rodriguez[5], Monica L. Guzmán[6], and Cesar M. Compadre[1]

[1]Department of Pharmaceutical Sciences, University of Arkansas for Medical Sciences, Little Rock, AR 72205, [2]Joint Bioinformatics Graduate Program, University of Arkansas at Little Rock and University of Arkansas for Medical Sciences, Little Rock, AR 72205, [3]Departamento de Química Aplicada, Universidad Técnica Particular de Loja, Loja Ecuador, [4]Department of Chemistry, University of Arkansas at Little Rock, AR 72205, [5]Cornell University, Ithaca, NY, United States, [6]Division of Hematology/Oncology, Department of Medicine, Weill Cornell Medical College, New York, NY 10065.

Sesquiterpene lactones (SLs) are naturally occurring compounds that have shown potent anti-leukemic activity, and have the ability to target leukemic stem and progenitor cells. It has been postulated that, the SLs exert their effect by inhibiting the rapid-acting primary transcription factor NF-κB by alkylating cysteine-38 in the DNA binding loop and cysteine-120 in the nearby E' region which makes specific interactions with the DNA impossible. NF-κB is a heterodimer complex of p50 and p65 subunits that interact with the DNA, regulating the expression of several genes. For the drug-receptor interactions, apart from geometrical complementarity, matching interactions between the drug and the receptor are also required. In an effort to study this complementarity, we performed the surface signature analysis of the SLs and the region around cysteine-38 of the NF-κB chain-A, using MOLCAD program. Electrostatic surface signature analysis showed that the cysteine-38 residue provides an electron rich region for nucleophilic interaction of SLs with NF-κB and the lipophilic surface signature analysis of NF-κB active site showed that tyrosine-36 and cysteine-38 residues are providing a lipophilic region for hydrophobic interaction of SLs with NF-κB. This study shows that the characteristic surface features in the SLs have a matching lipophilic and electrostatic surfaces with NF-κB, i.e. a strong lipophilic surface, and a strong electron rich area. Thus, these characteristic surface patterns would explain why the various types of substantially different SLs interact with NF-κB. These findings could also be used to identify additional naturally occurring compounds with potential anti-leukemia activity and to guide synthetic approaches to develop more potent or specific compounds.

This project was supported by: University of Arkansas for Medical Sciences College of Pharmacy Research Fund, the Arkansas Science and Technology Authority, Project No. 15-B-01 and the Arkansas INBRE program, with grants from the National Center for Research Resources - NCRR (P20RR016460) and the National Institute of General Medical Sciences - NIGMS (P20 GM103429) from the National Institutes of Health. PEO is a recipient of a graduate student scholarship from Ecuador's Higher Education, Science, and Technology Ministry (SENESCYT). MLG is a recipient of the NIH Director's New Innovator Award Program, 1 DP2 00007399-0.

Abstract Identifying Number: 1006289

PROCEDURES FOR IDENTIFYING BIOMARKERS-DEFINED SUBGROUPS WITH DIFFERENTIAL TREATMENT EFFECT THROUGH RECURSIVE PARTITIONS IN PRECISION MEDICINE

Un Jung Lee, Yu-Chuan Chen, and James J. Chen*

Division of Bioinformatics & Biostatistics, National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, AR, 72079

Precision medicine is to customize a medical model for new tools and therapies to select best treatments, being tailored to the individual patient. Subgroup analysis plays an important role in precision medicine to assess the treatment effects in subgroups, which provides useful information to optimize the treatment assignment. In this study, we propose using tree-based recursive partitioning methods to identify patient subgroups with the enhanced treatment effect in clinical trials. Two subgroup identification strategies are presented. One is based on the Differential Effect Search (SIDES) algorithm where the subgroups are identified by maximizing the treatment effect between treatment group and control group. However, these subgroups are defined by different biomarkers with different sample sizes and adjusted p-values. We propose a majority voting method to identify a subgroup from the list identified. The second strategy first generates a classification tree using only the samples from the treatment group and then applies to the control samples. For each terminal node, the treatment effect between the control and treatment groups is compared. Those patients with a significant treatment effect in the terminal nodes constitute the subgroup of interest. We also conduct simulation experiments to evaluate these two methods and compare with others in term of sensitivity, specificity, and accuracy.

Title: Strategies for efficient partial alignment of reads for DNA/RNA quantification
Author: Vinhthuy Phan

The alignment of reads to a reference genome or transcriptome is an important step in many applications that require an estimate of gene expression based on RNA-seq data, or an estimate of abundance of microbial genomes in metagenomic analyses. In such settings, a complete alignment of reads is unnecessary. It is possible to infer relative abundance of transcripts or microbial genomes using only *partial* or *pseudo* alignments. While such partial alignments do not provide specific information to determine variants, for instance, they containing approximate mapped locations of reads and as such they can be useful for an approximate quantification of abundance of transcripts or microbial genomes.

In this talk, we will examine several latest strategies that attempt to partially align reads to a transcriptome. These strategies were implemented in three software packages Kalisto, Sailfish and RapMap. A common analysis of the strategies reveals that they rely on fixed-length k-mers to find partial alignments efficiently. The use of fixed-length k-mers might have led to techniques that could be unnecessarily complex. We shall show that finding exact-matches between a read and a reference sequence using an FM index is analogous to pseudo-alignment using variable-length k-mers and might lead to a much simpler and more efficient process.

Abstract Identifying Number: 1006035

IDENTIFICATION AND EVALUATION OF POTENTIAL LEAD COMPOUND FOR PARKINSON'S DISEASE BY INSILICO AND PROTEOMICS APPROACH

Vivek Chandramohan[1], B. S. Gowrishankar*[1], and H. Gurumurthy[2]

[1*]Research Scholar, Department of Biotechnology, Siddaganga Institute of Technology, Tumkur – 572103, Karnataka, India
[2]Department of Biotechnology, GMIT, Davangere - 577004, Karnataka, India

Parkinson disease (PD) is the second most increasing neurodegenerative disorder. Benchmark research proved that occurrence of Lewy body formation with dense *Alpha-Synuclein* miss-folding and its self-aggregation is responsible for its toxic effect on dopaminergic neuronal cells. Hence current study aimed at determining a drug candidate that can inhibit the self-aggregation of *Alpha-Synuclein* based on structure based drug design. We sought to evaluate Withaferin A, a from the *Withania somnifera* plant as an effective compound for inhibiting the miss-folding as well as its self-aggregation by using molecular docking and dynamic studies. The lead molecules were filtered against Lipinski's rule and ADMET properties for molecular docking studies. Our study shows that Withaferin A has best docking score of -14.7968 than that standard (STD) compound Levedopa -5.4554 with stable *Alpha-Synuclein* structure respectively. The best and standard molecules were obtained by docking in explicit solvent for 60ns nanoseconds at a temperature of 310 K using Molecular Dynamics cascade in Discovery Studio 3.5. Withaferin A was good to binding the target proteins, and maintains strong bonds causing very less to negligible perturbation in the protein backbone structures. Our result indicates that, after 15ns nanoseconds, there is no miss-folding and self-aggregation in *Alpha-Synuclein* proteins when Withaferin A binds to target protein. Therefore, from our study we hope to add one more option in the self-aggregation to tackle Parkinson's disease.

Of text and gene – Analysis of big genomics data with text mining methods

Weida Tong, National Center for Toxicological Research, FDA, Jefferson, AR, USA

Big data in genomics are diverse and complex. For example, in toxicogenomics, study design often profiles gene expression from assays involving multiple doses and time points, requiring analysis taking this characteristic into specific consideration in toxicity assessment. The genome is often referred to as a book of life: the genome has 30 billion letters (bases), ~25,000 words (genes) comprised by these letters, many sentences/paragraphs (biological processes) that can be constructed with these words to associate with diseases, and these sentences/paragraphs are repeated and spread across 23 chapters (chromosomes). Thus, one can conceptualize a relationship between genes and text; genes and text share many commonalities and characteristics. For example, the same word can appear in different sentences while the same gene can involve in different pathways. Such a commonality suggests that text mining tools could be useful alternatives to analyze genomic data. Topic modeling is a text mining approach, but, by analogy, could be effective in genomics data analysis due to the similar data structure between text and gene dysregulation. In this presentation, we will present the results by applying topic model to a very large toxicogenomics dataset that contains microarray gene expression data from >15,000 samples associated with 131 drugs tested in three different assay platforms (i.e., *in vitro* assay, *in vivo* repeated dose study and *in vivo* single dose experiment) with a design including multiple doses and time points. A set of "topics" (each consists of a set of genes) was determined, by which the varying sensitivity of three assay systems was observed. Specifically, the drug-dependent effect was more pronounced in the two *in vivo* systems than the *in vitro* system, while the time-dependent effect was reflected the strongest in the *in vitro* system followed by the single dose study and then the repeated dose experiment. The dose-dependent effect was similar across three assay systems. Although the results indicated a challenge to extrapolate the *in vitro* results to the *in vivo* situation, we did notice that, for some drugs but not for all the drugs, the similarity in gene expression patterns was observed across all three assay systems, indicating a possibility of using the *in vitro* systems with a careful design (such as the choice of dose and time point), to replace the *in vivo* testing strategy. The study demonstrated that text mining methodologies such as topic modeling provide an alternative way to other traditional computation means for data reduction in toxicogenomics, enhancing a researcher's capability to interpret the biological information in a reduced data features.

Abstract Identifying Number: 1002118

BEST PRACTICE IN MINING TOPICS FROM REGULATORY TEXTUAL DOCUMENTS

Weizhong Zhao, Yijun Ding, James J. Chen, Weida Tong, Roger Perkins, and Wen Zou

DBB, NCTR/FDA

Probabilistic topic modeling offers a viable approach to structure huge textual document collections into latent topic themes to aid text mining. FDA lore describes drug applications arriving in eighteen wheelers. Today the agency handles vast digital textual information from submissions representing some 25% of U.S. GDP, and untold terabytes of information from post market surveillance. Where experts are too few or slow, the means to extract information germane to regulatory questions is paramount. Here we describe extensive sensitivity studies to determine best practices for generating effective topic models. To test effectiveness and validity of topic models, we constructed a ground truth data set from PubMed that contained some 40 health related themes including negative controls, and mixed it with a data set of unstructured documents. The most useful models, tuned to desired sensitivity versus specificity, require an iterative process wherein preprocessing steps, the type of topic modeling algorithm, and the algorithm's model parameters are systematically varied. Models need to be compared with both qualitative, subjective assessments and quantitative, objective assessments, and care is required that Gibbs sampling in model estimation is sufficient to assure stable solutions. With a high quality model, documents can be rank-ordered in accordance with probability of being associated with complex regulatory query string, greatly lessoning text mining work. Importantly, topic models are agnostic about how words and documents are defined, and thus our findings are extensible to topic models where samples are defined as documents, and genes, proteins or their sequences are words.

Abstract Identifying Number: 1006304

TARGETED THERAPIES, SYSTEMS BIOLOGY AND THE FUTURE OF SAFETY ASSESSMENT

William B Mattees*[1]

[1*]Division of Systems Biology, National Center for Toxicological Research
FDA, Jefferson, AR, 72079

While the term "targeted therapy" has been used for over three decades, its use has exponentially grown in the field of cancer therapy with the advent of molecules that are designed to interfere with specific proteins in signaling pathways. A seminal example is that of imatinib (Gleevec), designed to block the BCR-Abl tyrosine kinase found in chronic myelogenous leukemia (CML); its effects on patient survival are dramatic. Dozens of such drugs have been approved or are in development, but the early enthusiasm is dampened by the growing spectrum of side effects seen with many of these new "targeted" therapies. Thus sunitinib, originally described as a VEGF receptor inhibitor, actually inhibits a number of protein kinases and is associated with cardiotoxicity among other side effects. Because these new therapies are not as specific as one would hope, and the complete signaling circuitry of even a few cells in a few tissues in one species remains to be described, their impact on any given individual animal or human cannot yet be predicted. Thus the challenge of 21st century safety assessment is to describe that circuitry as much as possible for critical organs in relevant species, to identify those species differences where non-clinical testing must be augmented, and identify targets that are key in adverse effects. Furthermore, relevant, translational biomarkers of adverse effects elicited by "targeted therapies" must be developed and qualified for both non-clinical and clinical use. Finally, the impact of differences in these circuits as found in human populations needs also be investigated and considered in the overall process of safety assessment. Despite the seemingly expansive nature of this vision, the tools are at hand, and the work can be done.

Abstract Identifying Number: 1006265

EVALUATION OF NON-INVASIVE MICRORNAS AS BIOMARKERS OF HEPATOTOXICITY; AN UPDATE ON TRANSLATIONAL ACETAMINOPHEN TOXICITY BIOMARKERS IN CHILDREN

Xi Yang[1], Qiang Shi[1], James Greenhaw[1], Prit S Gill[2], Sudeepa Bhattacharyya[2], Richard Beger[1], Laura P James[2]

1 Division of Systems Biology, National Center for Toxicological Research, Food and Drug Administration, 3900 NCTR Road, Jefferson, AR 72079, USA

2 Department of Pediatrics, University of Arkansas for Medical Sciences and Arkansas Children's Hospital Research Institute, Little Rock, AR 72202, USA

While generally considered to be safe at therapeutic doses, nonetheless acetaminophen (APAP) is responsible for 14% of acute liver failure cases in children. Recently, several studies have reported human circulating microRNAs (miRNAs) as novel biomarkers of drug-induced toxicity. This study tested the hypothesis that miRNA alterations in biofluids are associated with liver injury induced by APAP. Elevations of urinary miRNAs have been detected in rats administered toxic doses of hepatotoxicants. Here, we examined serum and urinary miRNAs as potential biomarkers of APAP toxicity in children. Real-time quantitative polymerase chain reaction (qRT-PCR) arrays and small RNA sequencing were used to detect global expression of miRNAs from three pediatric subgroups: i) healthy children (n=10, ALT median 18, range 10-37 IU/L), ii) hospitalized children receiving therapeutic doses of APAP (n=10, ALT median 26, range 6-177 IU/L) and iii) children hospitalized for APAP overdose (n=9, ALT median 2314, range 25-9909 IU/L). Out of 147 miRNAs detected in the APAP overdose group, eight showed increased levels in serum (miR-122, -375, -423-5p, -30d-5p, -125b-5p, -4732-5p, -204-5p, and -574-3p) compared to the other subgroups. Analysis of urine samples showed that four miRNAs (miR-375, -940, -9-3p and -302a) increased in the overdose group, compared to the other subgroups. In conclusion, this study suggests that miRNAs could represent a non-invasive clinical biomarker of APAP toxicity in children.

Abstract Identifying Number: 1006290

POST: A FRAMEWORK FOR SET BASED ASSOCIATION ANALYSIS IN HIGH DIMENSIONAL GENETIC DATA

Xueyuan Cao[1;4], E. Olusegun George[4], Mingjuan Wang[1;4], Dale B.Armstrong[4], Cheng Cheng[1], Jeffery Rubnitz[2], James Downing[3], and Stanley Pounds[1]

[1]Biostatistics, St. Jude Children's Research Hospital, Memphis, TN, 38105
[2]Oncology, St. Jude Children's Research Hospital, Memphis, TN, 38105
[3]Pathology, St. Jude Children's Research Hospital, Memphis, TN, 38105
[4]Department of Mathematics, University of Memphis, Memphis, TN, 38152

With high throughput technologies, investigators can address various biological questions in whole genome level which are not attainable before. In addition to gene level testing, set-based association study is getting attractive. Projection onto the Orthogonal Space Testing (POST) is proposed as a general and flexible procedure to perform association test for gene profiling data in a set level (gene sets, pathways etc.). The probe level signals of a predefined set are first projected to an orthogonal subspace which is spanned by first handful eigenvectors explaining a predefined fraction of variation of probes. The projected data are subjected to parametric association tests, adjusting for other presenting features if needed. A POST statistic is defined as sum squares of individual z-statistic weighted by corresponding eigenvalues (a quadratic form). The correlation structure of z-statistics is approximated by bootstrap resampling and the p-value is approximated by a generalized χ^2 distribution. The proposed method was further investigated and compared to SAFE and MRPP tests in a simulation study. POST was applied to a gene profiling data set of 105 pediatric AML patients from a combined cohort in St Jude Children's Research Hospital, which were measured by U133A array. Out of the 1057 biological processes, 22 were significantly associated with EFS (event-free survival) at q value 0.4 level. Five of the 22 were signaling pathways including 'Regulation of Wnt receptor signaling pathway' ranked number one showing that POST can identify meaningful gene sets with clinical phenotypes.

Abstract Identifying Number: 1006311

LARGE-SCALE SOYBEAN GENOMIC VARIATION ANALYSIS WORKFLOW IN SOYKB NGS BROWSER

Yang Liu[1,2,*],Saad M. Khan[1,2,*],Juexin Wang[2,3,*],Shiyuan Chen[2,3],Mats Rynge[4],YuanXun Zhang[3],Shuai Zeng[2,3],Joao V. Maldonado dos Santos[5],Babu Valliyodan[5,6],Nirav Merchant[7], Henry T. Nguyen[5,6], Dong Xu[1,2,3], Trupti Joshi[1,2,3,8,*]

[1] Informatics Institute, University of Missouri, Columbia, Missouri, 65201
[2] Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, Missouri 65201
[3] Department of Computer Science, University of Missouri, Columbia, Missouri, 65201
[4] Information Sciences Institute, University of Southern California, Los Angeles, California, 90089
[5] Division of Plant Sciences, University of Missouri, Columbia, Missouri, 65201
[6] National Center of Soybean Biotechnology, Columbia, Missouri, 65201
[7] iPlant Collaborative, University of Arizona, Tucson, Arizona, 85721
[8] Department of Molecular Microbiology and Immunology and Office of Research, School of Medicine, University of Missouri, Columbia, Missouri, 65201
* Equally contribution authors

Motivation: With the advances in next-generation sequencing (NGS) technology and significant reduction in sequencing costs, it is now possible to sequence large sets of crop germplasm for detecting genome-scale structural variations using resequencing, in order to better improve traits analysis. To facilitate large-scale NGS genomic variation analysis, we have developed an online workflow system named PGen using XSEDE high-performance computing (HPC) system, iPlant cloud storage resources and Pegasus workflow management system. The workflow allows users to call single nucleotide polymorphisms (SNPs) and insertion-deletion (Indels), perform SNP annotations and conduct copy number variation analyses on more than 50 germplasm lines at a time. In order to browse and access NGS data easily, we also developed an NGS browser of soybean knowledge base (SoyKB) connected to PGen workflow (http://soykb.org/NGS_Resequence/NGS_index.php). PGen workflows can be customized to other organism.

Results: We have developed both Linux and web-based implementation of PGen workflow integrated within Soybean Knowledge Base (SoyKB), available respectively at https://github.com/pegasus-isi/PGen-GenomicVariations-Workflow and http://dev.soykb.org/Pegasus/index.php. PGen workflow identified 10,218,140 SNPs and 1,398,982 Indels from one data set which contain 106 soybean lines sequences at 15x coverage. 297,245 non-synonymous SNPs as well as 220,936 synonymous SNPs were identified of 69,105 transcripts. 3330 copy number variation regions were also identified overlapped with 1905 transcripts and 438 transcripts have more than 3 copies changes. SNPs have been identified from multiple soybean resequencing projects for total 500+ soybean germplasm, and data is being utilized for trait improvement using genotype to phenotype prediction approaches.

Funding: *We acknowledge Missouri Soybean Merchandising Council (Grant # 368) and the United Soybean Board (Grant # 1320-532-5615) for the funding support.*

Abstract Identifying Number: 1006104

MMIRNA-VIEWER: A DATA VISUALIZATION TOOL BUILT TO PRESENT THE RELATIONSHIP BETWEEN MIRNAS AND MRNAS

Yongsheng Bai*[1,2] Steve Baker[3], Ethan Rath[1], Lizhong Ding[1], Andrew Carrillo[3], Lianfang Wang[3], Hui Jiang[4], Gary Stuart[1,2]

[1]Department of Biology, [2]The Center for Genomic Advocacy
[3]Department of Mathematics and Computer Science,
Indiana State University, Terre Haute, IN 47809, U.S.A
[4]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, U.S.A.

The Cancer Genome Atlas (TCGA) harbors an increasing amount of cancer genome data with both tumor and normal samples. In a previous study, we published a tool called *MMiRNA-Tar* (http:/bioinf1.indstate.edu/MMiRNA-Tar) for describing miRNA and mRNA co-relationships for tumor and normal samples in TCGA bladder urothelial carcinoma (BLCA) datasets. In order to fully deploy important relationships and attributes identified between miRNA and mRNA pairs for any given cancer types and study biological consequences across multiple cancers, we developed a web interface tool called *MMiRNA-Viewer* that can concurrently present the co-relationships of expression for miRNA−mRNA pairs of both tumor and normal samples into a single graph.

The input file of *MMiRNA-Viewer* should contain the expression information for mRNA and miRNA in normal and tumor samples, the correlation between mRNA and miRNA, and the predicted target relationship by a number of databases. The output graph shows two types of nodes summarizing expression information for miRNAs and mRNAs, and two types of links representing target relationships between miRNA and mRNA pairs in normal and tumor samples. Users can visualize detailed information about cancer-related gene expression changes, and also changes in the expression of transcription-regulating miRNAs for well-characterized cancer genomes.

With its user-friendly interface, dynamic visualization and advanced queries, we believe *MMiRNA-Viewer* offers an intuitive approach for visualizing and elucidating co-relationships between miRNAs and mRNAs. *MMiRNA-Viewer* is available at http:/bioinf1.indstate.edu/MMiRNA-Viewer.

Publication of this article was funded by the ISU to Dr. Yongsheng Bai

Abstract Identifying Number: 100336

EVALUATION OF TOXICOGENOMICS IN ADVANCED RESEARCG AND NOVEL APPLICATIONS

Yuping Wang[1*], Binsheng Gong[1], Zhichao Liu[1], Jian Yan[2], Tao Chen[2], Wen Zou[1], Huixiao Hong[1], Yuji Morikawa[3], Takeki, Uehara[4], Weida Tong[1]

[1]Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. Food and Drug Administration. Jefferson, AR 72079
[2]Division of Genotoxicity, National Center for Toxicological Research, U.S. Food and Drug Administration. Jefferson, AR 72079
[3]Informatics & Structure-based Drug Discovery, Discovery Research Laboratories for Innovative Frontier Medicines, Shionogi & Co. Ltd., Toyonaka, Osaka, Japan
[4]Global Project Management Department, Shionogi & Co., Ltd., Osaka, Kita-ku, Japan.

Advances in research and technologies have opened new possibilities in toxicogenomics methods and applications which empower the potential to revolutionize toxicology. With the emergence of high-throughput sequencing, it is now possible to discover many different biological components simultaneously. Here, we will summarize the evaluation of toxicogenomics in the context of new components and technology as well as novel promise in drug discovery and development. As a case study, we present the result from our analysis of miRNA and mRNA expression data generated from rat livers treated with amiodarone and benzbromarone, drugs sharing similar chemical structure but inducing different phenotypes of liver injuries. We did an integrated analysis of the differentially expression of miRNAs (DEMs) and mRNAs (DEGs) at multiple doses and time points by comparing with the control groups. Enriched toxicological functions analysis demonstrated compound-specific mode of action (MOA) which may provide insightful and add-on value to traditional toxicology currently in use. We emphasized the need of an envisioned strategy on how toxicogenomics can become a novel tool to define the best practices in the application of toxicogenomics in drug discovery and development.

Abstract

Abstract Identifying Number: 1006283

IDENTIFICATION OF A NON-CANONICAL MICRO-RNA IN AN *FGF2-SPECIFIC* SNP REGION AMONG BREAST CANCER PATIENTS

Yusuf Nawawi[1], Chelsea Drennan[1], Patrick Apopa[1], Rosalind Penney[2], Isaam Makhoul[3] and Mohammed Orloff[1,3]

[1]Department of Epidemiology, College of Public Health, University of Arkansas for Medical Sciences, Little Rock, AR, 72205
[2]Department of Environmental and Occupational Health, College of Public Health, University of Arkansas for Medical Sciences, Little Rock, AR, 72205
[3]Winthrop P. Rockefeller Cancer Institute, University of Arkansas for Medical Sciences, Little Rock, AR, 72205

Background: Breast cancer (BCa) is the most common cancer and second most common cause of cancer death among women in the United States. Although there has been a steady decrease of mortality rates of BCa, the gap for accurate diagnosis of BCa still persists warranting identification of new biomarkers. Deregulation of *FGF2* gene expression has been found to be associated with increased risk of the development and progression of BCa among women.

Methods: In our previous study, we identified SNP rs2922982 in *FGF2* is associated with BCa. In the present study, we seek to assess the rs2922982 genotypes in tumor and normal samples and correlate with both regulatory molecules that bind the rs2922982 site and the *FGF2* expression patterns.DNA and miRNA was extracted from 13 Formalin Fixed Paraffin Embedded breast tissue samples from the UAMS Tissue Biorepository after IRB approval.

Result: Through mirBase database we identified *hsa-miR-3680-5p* as the candidate miRNA that binds in the region that harbors rs2922982. rs2922982 is located in the 3' end, intronic region of the *FGF2* gene. ACA67, a potential regulatory molecule small nucleolar RNA, was also identified to be binding at the SNP location. The *hsa-mir-3680-5p* expression analysis revealed a ΔΔCt of 0.1331 ± 0.0557. DNA sequencing showed 1/13 samples have genotype GG, 2/13 with GT genotype and 10/13 with TT genotype at the rs2922982 location. There was no difference of miRNA expression level when the genotypes were compared (*p*-value 0.376).

Conclusion: Located in the 3' end, intronic region of the *FGF2* gene, rs2922982 may suggest a non-canonical miRNA binding site. This study, however, did not find a difference of miRNA expression level based on allelic variation on rs2922982. Future study with larger sample size and adjustment of other covariates will assess and correlate the *FGF2* expression with the miRNA levels and rs2922982-specific genotypes.

(Source of funding: Winthrop P. Rockefeller Cancer Institute)

Potential Reuse of Oncologic Drugs for the Treatment of Rare Diseases

Zhichao Liu*, Hong Fang, William Slikker, Weida Tong

National Center for Toxicological Research, US Food and Drug Administration, USA

Abstract

Cancer research has been a focus in the biomedical field resulting in many oncologic drugs in clinical use. In contrast, very few treatment options are available for rare diseases although they are progressive, disabling and life threatening. Therefore, we investigated the potential use of oncologic drugs for the treatment of rare diseases. A strong association between cancer and rare diseases was observed at the molecular level. Specifically, an overlap of approximately 60% was shown between 127 genes associated with cancer of many kinds and 2976 rare disease genes, and the same degree of overlap was also obtained when the analysis was conducted at the pathway level. By placing both gene lists mentioned above in a gene-gene network, over 95% gene pairs (one from each list) have two genes locating less than three genes apart in the network, indicating that cancer genes and rare disease genes likely involve similar biological processes. In addition, many drug targets for cancer were found to relate to rare diseases. The molecular level of association between cancer and rare diseases was further substantiated with existing clinical trial data and literature review. In summary, we ranked the rare disease classes by their potential to be treated with oncologic drugs. The study demonstrated that anticancer drugs are potential sources for the treatment of rare diseases, and the proposed framework offers an opportunity to identify potential therapeutics from cancer research for use in rare diseases.