6th Annual Louisiana Conference on Computational Biology and Bioinformatics

April 6-7, 2018

Friday, April 6, 2018
1:00 pm - 8:00 pm

Saturday, April 7, 2018
9:00 am - 5:00 pm

Louisiana State University
Digital Media Center
Baton Rouge, LA

Table of Contents

Friday April 6, 2018

1:00 – 1:10 pm ....Welcome and Opening Remarks

Session I: Evolutionary Genomics, Infectious Diseases and Cancer Informatics

1:10 – 2:10 pm   Jessica Kissinger - University of Georgia
Unleashing Your Inner Data Scientist: Challenges and Opportunities from the Field of Infectious Diseases

2:10 – 2:40 pm Chindo Hicks - Louisiana State University Health Sciences Center, New Orleans
Bioinformatics Approaches for Integrating Germline and Somatic Mutation Information in Cancer

2:40 – 3:10 pm   Christian Clement - Southern University at New Orleans
Infection Characteristic of HSV-1 Neuronal Viral DNA and Analysis of ICP4, ICP0 In Silico

---

3:10 – 3:30 pm...Break

Session II: The Human Microbiome and Metagenomics

3:30 – 4:30 pm   Morgan Langille - Dalhousie University
Disease Signatures of the Human Microbiome

4:30 – 5:00 pm   Filipa Godoy-Vitorino - Inter American University of Puerto Rico
Characterization of the Cervicovaginal Bacterial and Fungal Profiles
Associated with HPV Infections in a Hispanic Population

5:00 – 5:30 pm   Chris Taylor - Louisiana State University Health Sciences Center, New Orleans
Visualization of Daily Changes in the Vaginal Microbiome of Women Leading up to Bacterial Vaginosis

---

5:30 – 6:30 pm...Dinner

6:30 – 8:00 pm...Poster Session

Saturday, April 7, 2018

8:00 – 9:00 am...*Breakfast Available*

9:00 – 9:10 am ....Welcome and Opening Remarks

Session III: Big Data, Bioinformatics Education, and Genomics Software

9:10 – 10:10 am   Ben Busby - National Center for Biotechnology Information/NIH
The Future is Now! Biomedical Data Science in the 21st Century

10:10 – 10:40 am   Rebecca Christofferson - Louisiana State University
Elucidating the Effect of Time When Investigating Within-Mosquito Zika Kinetics
Altered by the Integration of Rearing and Extrinsic Incubation Temperature

10:40 – 11:10 am   Elia Brodsky - Pine Biotech
A Collaborative Model for Bioinformatics Education: Combining
Biologically Inspired Bioinformatics with Project-Based Learning

11:10 – 12:10 pm   Paul Fullerton - Partek
Partek Next Generation Sequencing Data Analysis Solution Overview

---

12:10 – 1:10 pm...*Lunch*

Session IV: Translational Bioinformatics, Imaging and Data Visualization

1:10 – 2:10 pm   Jake Chen - University of Alabama Birmingham
Unravelling Complex Patterns in Omics Data with Integrative GNPA Tools

2:10 – 2:40 pm   Phillip Kilgore - Louisiana State University, Shreveport
Computer-assisted Inference of Association and Causation in Time-Series Data

2:40 – 3:10 pm   Fabrizio Donnarumma - Louisiana State University
Multi-modal Mass Spectrometry for Translational Proteomics

3:10 – 3:30 pm...*Break*

Saturday, April 7, 2018

Session V: Protein Structure and Machine Learning

3:30 – 4:00 pm   Yong-Hwan Lee - Louisiana State University
      MD Simulation of Protein Control by Covalent Modifications: A Preliminary Study with PFKFB3

4:00 – 4:30 pm   Manisha Panta - University of New Orleans
      Prediction of Hierarchical Classification of Transposable Elements using Machine Learning Approach

4:30 – 5:00 pm   Reecha Khanal - University of New Orleans
      Prediction of RNA Binding Protein using Machine Learning Techniques

---

5:00 *pm*.. *Concluding Remarks*

Oral Presentation Abstracts

Session I: Evolutionary Genomics, Infectious Diseases and Cancer Informatics

6TH ANNUAL LA CONFERENCE ON COMPUTATIONAL BIOLOGY & BIOINFORMATICS

keynote

Jessica Kissinger, PhD
- **University of Georgia**
- **Director of the Institute of Bioinformatics**
- **Professor, Department of Genetics**

I am interested in parasite genomics and the biology of genome evolution. How do genomes evolve? Can we trace gene order, identify and determine the fate of horizontally transferred genes (as well as any existing endogenous orthologs) and identify genes that are phylogenetically restricted? The answers to these questions are important because of the insight they provide into parasite biology and the identification of potential drug targets. My research focuses on, but is not limited to, Apicomplexan parasites.

OP-01   **Unleashing Your Inner Data Scientist: Challenges and Opportunities from the Field of Infectious Diseases**
Jessica Kissinger
University of Georgia

You can plan a trip to almost anywhere in the world with just a few keystrokes and share information between airlines, hotels and rental cars, yet if you want to know the gene expression of an important pathogen, the host's response to the pathogen or the host's microbiome, the endeavor could take months. At a time when nearly all funded researchers are required to share their data, actually using it is nearly impossible. This talk will explore some the challenges, opportunities and a few solutions towards making data related to some infectious diseases and model organisms accessible to the average bench scientist.

OP-02   **Bioinformatics Approaches for Integrating Germline and Somatic Mutation Information in Cancer**
Chindo Hicks
LSUHSC New Orleans

Cancer is a complex genetic disease involving both germline and somatic mutations. Recent advances in next generation sequencing have enabled deciphering of cancer genomes. Large multicenter projects such as The Cancer Genome Atlas (TCGA) and the International Cancer

Oral Presentation Abstracts

Genome Consortium (ICGC) have performed detailed analyses of the somatic mutations in cancer genomes. These discoveries are increasing our understanding of the molecular basis of tumors and enabling discovery of clinically actionable biomarkers and targets for the development of novel therapeutics. However, to date, information on somatic mutations has not been leveraged to elucidate the possible oncogenic interactions between somatic driver mutations and germline mutations to understand their joint role in tumorigenesis. Here we integrated germline mutation information derived from genotype data with somatic mutations from TCGA and our own experiments on whole exome sequencing using transcriptome data as the organizing principle. The objective was to investigate the possible oncogenic interactions and mechanisms of cooperation between germline and somatic mutations during tumor development and progression. Our working hypothesis is that germline and somatic driver mutations interact and cooperate during tumorigenesis, and that these complex array of interacting genetic factors affect molecular networks and biological pathways driving tumor development and progression. The analyses revealed functionally-related genes containing germline and somatic driver mutations involved in tumors. Additionally, the analysis revealed molecular networks and biological pathways enriched for germline and somatic mutations. We conclude that integrative bioinformatics analysis is a powerful tool for decoding the possible oncogenic interactions and mechanisms of cooperation between germline and somatic mutations during tumor development and progression.

OP-03   **Infection Characteristic of HSV-1 Neuronal Viral DNA and Analysis of ICP4, ICP0 In Silico**
Christian Clement
Southern University at New Orleans

Background: NZW-rabbit-model harboring neuroinvasive H129, epinephrine iontophoresis induced to reactivate, produced neovascularization in specifically the left brain. A novel HSV-1 replication and viral DNA accumulation in neuronal/corneal damage and triggering of electrical seizures was hypothesized. Exploring this HSV-1 infection phenotype using a transgenic mouse model carrying the human APOE4 gene (ApoE e4/e4) infection screen with two virus strains 17Syn+ high-phenotypic-reactivator (HPR) and 17ΔPst(LAT-) low-phenotypic-reactivator (LPR), stipulated ICP4, ICP0 roles (Louisiana Biomedical Research conference, 15th Annual Meeting, January 20-22, 2017, Baton Rouge LA; Poster Session Abstracts, #4: 'ICP4 and ICP0 HSV-1 Viruses, Virus-Induced Inflammation and DNA in the Brain'; 5th annual LA Conference on Computational Biology & Bioinformatics, April 7-8, 2017, Xavier University New Orleans, LA; Poster Session Abstracts, #PA-04 'Evaluating ICP4 and ICP0 HSV-1 Viruses, Virus-Induced Inflammation and DNA in the Brain'). Methods: HSV-1 DNA copy numbers were determined by calculating the number of HSV-1 polymerase genes per sample. Forward/reverse primer pairs were 5'- AGA GGG ACA TCC AGG ACT TTG T -3'/5'- CAG GCG CTT GTT GGT GTA C -3' and signal levels was quantified with the fluorescent probe sequence 5'-6-FAM/ACC GCC GAA CTG AGC A/3' BHQ-1'. Results: Hippocampus is associated with long-term memory and spatial navigation and one of the first regions to suffer damage in Alzheimer's disease. HSV-1 DNA detection of the viral genome of >10 copy numbers for both 17Syn+ (HPR) and 17ΔPst(LAT-) (LPR) in the left hippocampus and for only17Syn+ (HPR) in the left cortex were not significantly affected by heat stress treatment, concentrations of Acyclovir dosing and mode of drug administration, unlike the control C57Bl/6N mice which were viral DNA free. Conclusion: This is a first report of novel HSV-1 infection characteristic in the presence of the human APOE4 gene.

Oral Presentation Abstracts

Session II: The Human Microbiome and Metagenomics

OP-04	**Disease signatures of the human microbiome**
	Morgan Langille
	Dalhousie University

Abstract The human microbiome, the community of microbes that reside in and on the human body, protects against infections, interacts with the immune systems, and has been linked to various diseases including obesity, inflammatory bowel disease, and cancer. Microbiome studies generate large datasets of sequencing data that require efficient, novel, and complex bioinformatic methods. An overview of microbiomics along with examples of bioinformatic methods that incorporate phylogenetics, comparative genomics, and machine-learning applied to predicting treatment outcome in Crohn's disease will be presented.   Bio Dr. Morgan Langille is a Canada Research Chair in Human Microbiomics and an Assistant Professor in the Department of Pharmacology at Dalhousie University, Halifax, NS, Canada. His expertise is in microbial genomics and bioinformatics, with his work cited over 3300 times, including a first authored bioinformatic paper which has been extensively used in the human microbiome field (cited >1400 in 4 years). Dr. Langille also founded and is Director of the Integrated Microbiome Resource (IMR), which provides sequencing and bioinformatics for microbiome projects, processing over 34,000 samples in 3 years from 421 different research projects. Dr. Langille leads a dynamic and interdisciplinary research team with currently eight members

spanning expertise in computer science, statistics, sequencing, and microbiology. He leads several human microbiome projects including a recently published study that demonstrated the prediction of treatment outcome in pediatric Crohn's disease and identifying risk factors for infection complications in pediatric acute lymphoblastic leukemia.

OP-05   **Characterization of the cervicovaginal bacterial and fungal profiles associated with HPV infections in a Hispanic population**
Filipa Godoy-Vitorino
Inter American University of Puerto Rico

The cervicovaginal microbiota of the female genital tract has an important role in the induction or prevention of disease, likely playing a role in susceptibility to Human papillomavirus (HPV) infections. Among Hispanics, Puerto Ricans have the highest rates of cervical cancer and an HPV prevalence higher than in the continental USA. We hypothesized that the structure of the cervicovaginal bacterial and fungal biota could be related to HPV infections, atypical cell changes and neoplasia. We sequenced both the 16S rRNA V4 region and the ITS-2 fungal regions of cervical and vaginal of 62 Puerto Rican women and genotyped cervical HPVs.   We identified a high prevalence of high-risk HPV and co-infections with as many as 9 HPV types and a significant increase in the bacteria diversity in association with CIN3 pre-cancerous lesions. We found a significantly higher fungal diversity in high-risk HPV infections and Atypical Squamous Cells of Undetermined Significance (ASCUS).   Our data suggests that cervicovaginal bacteria and fungi respond to the host epithelial microenvironment and viral infections, and could play a role in cervical dysplasia.

OP-06   **Visualization of Daily Changes in the Vaginal Microbiome of Women Leading up to Bacterial Vaginosis**
Christopher Taylor
LSUHSC New Orleans

Bacterial vaginosis (BV) is the most common cause of vaginal discharge in reproductive aged women and has been associated with preterm delivery, low birth weight, and an increased risk of acquisition and transmission of sexually transmitted infections. BV is characterized by distinct changes in the vaginal microbiome which can fluctuate substantially on a daily basis. We describe a longitudinal study involving daily sampling of sexually active women who have sex with women (WSW) from the Birmingham metro area and present some visualization techniques for highlighting the changes that occur throughout the course of subject sampling illustrating the importance of longitudinal analysis of the vaginal microbiome.

Oral Presentation Abstracts

Session III: Big Data, Bioinformatics Education, and Genomics Software

OP-07   **The Future is Now! Biomedical Data Science in the 21st Century**
Ben Busby
NCBI, NLM, NIH

This talk will begin with a discussion of new computational and visualization tools from NCBI and others in the bioinformatics community. Next generation methods for extracting data from databases, such as 'edge computing,' will be discussed, using magicBLAST as an example. We'll also discuss other high-throughput ways to access data and metadata, such as the EDirect command line API wrapper, taxonomic organization of SRA datasets, and online .bam visualization, as well as text analysis leveraging the MetaMap and the PubMed and PubAg corpora. We'll also discuss how these data extraction mechanisms can be paired with community resources, and general frameworks such as Galaxy, CyVerse, Bioconda, and Bioconductor.   Leveraging this Infrastructure through Project Based Data Science Education Over the past three years, NCBI has run or been involved in 18 data science hackathons. In these hackathons, participants assemble into teams of five or six to work collaboratively for three days on pre-scoped projects of interest to the bioinformatics community. About 80% of teams produce an alpha or beta prototype, and approximately ten percent publish a manuscript. NCBI hackathons have generated over 100 products, about 50% of them stable or still in development. Some of these can be found at http://ncbi-hackathons.github.io.

Oral Presentation Abstracts

Hackathons provide an immersive learning environment and promote networking opportunities. We will discuss the educational aspects of these hackathons, options for setting up hackathons at your own institution, and tricks to make them successful. NCBI and other parts of NLM and NIH are also involved in other programs pertaining to project-based data science education. These include the NIH data science mentorship program, the visiting bioinformatician program, the code preservation and discovery project, and the microbial metagenomics discovery challenge. These will be discussed with an eye toward scaling up biomedical data science training worldwide.

OP-08   **Elucidating the effect of time when investigating the within-mosquito Zika kinetics altered by the interaction of rearing and extrinsic incubation temperatures**
Rebecca Christofferson
Louisiana State University-Baton Rouge

Because of the increasing threat that Zika virus (ZIKV) poses to extra-tropical regions, there is a need for better understanding of the effect of sub-tropical and temperate temperatures on the establishment potential of ZIKV within more temperate and/or seasonal Ae. aegypti populations. Additionally, a holistic view of vector competence, especially the role of time, should be considered when assessing effects of experimental treatments on within-mosquito viral fitness. To that end, mosquitoes were hatched and kept at a rearing temperature (RT) of 24°C or 28°C during the aquatic juvenile stages and as young adults before exposure to ZIKV at 3-5 days post emergence. After, mosquitoes were incubated at either the RT or the opposite temperature during the extrinsic incubation period, which defined the extrinsic incubation temperature (EIT). Thus, there were four combinations (RT24-EIT24, RT24-EIT28, RT28-EIT24, RT28-EIT28). Infection and dissemination rates were statistically analyzed via logistic regression. To further describe the within host kinetics, the continuous process of infection and dissemination through a population of mosquitoes was modeled to obtain predicted probabilities over time, and the infection time 50 and dissemination time 50 were estimated. Results: The odds of becoming infected were significantly different between RT24°C and RT28°C when EIT was constant at 24°C (OR = 0.3). Among the four groups, statistical significance for infection was seen between RT24-EIT28 compared to RT28-EIT24 (OR = 0.36). The fastest infection group was RT28-EIT24 with an IT50 of 6.13 days), while the fastest group to reach 50% dissemination was RT24-EIT28 at 12.21 days. Conclusions: There was a general and expected trend of higher EIT resulting in higher odds of a mosquito developing a disseminated infection compared to the lower EIT. However, there was also evidence that an interaction of RT and EIT affected infection rates. The groups predicted to have the fastest infection and dissemination through the population were not necessarily correlated with the highest odds of infection/dissemination. This indicates that the processes of infection and dissemination are complex, and when assessing arbovirus efficiency within the mosquito, several factors should be considered, including the temporal dynamics of the within-mosquito viral kinetics.

Oral Presentation Abstracts

OP-09   **A collaborative model for bioinformatics education: combining biologically inspired bioinformatics with project-based learning**
Elia Brodsky
Pine Biotech

Despite the growing impact of bioinformatics in the biological science community, integration of a full bioinformatics curriculum is cost prohibitive for many universities due to cost of infrastructure and computational resources. Furthermore, many programs prioritize the technical aspects of bioinformatics over the analysis logic, thus limiting the emphasis on critical thinking, problem solving, and in-depth inquiry. To address the gap in bioinformatics education and train students to approach complex problems, we present a new model for curriculum development that combines our online learning environment with traditional pedagogical approaches delivered through academic partnerships. The T–BioInfo platform allows users to combine computational analysis modules into pipelines to develop solutions for 'omics data and machine learning problems. State-of-the-art tools for analysis, integration, and visualization of data are offered through a user-friendly interface. In parallel, online educational modules provide a theoretical framework for the analysis methods and experimental techniques. This model for bioinformatics training was implemented at Loyola University New Orleans, a liberal arts institution, for the first time in January 2018. Twelve undergraduate students and five faculty members participated in a new one-semester bioinformatics course. After completing a set of online modules, students conducted research projects on topics such as patient derived xenograft models, immune responses in cancer, and precision medicine. Gains in critical thinking and problem-solving skills were observed and participants were enthusiastic to continue research. Our collaborative model for bioinformatics education combines best-practices in online and in-class learning with a powerful computational platform. This model could be implemented in undergraduate and graduate curricula to enhance research, build partnerships with industry, and strengthen the scientific workforce.

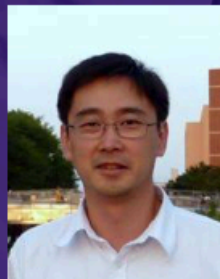**OP-10**   **Partek Next Generation Sequencing Data Analysis Solution Overview**
Paul Fullerton
Partek

Over 6,000 scientific articles have cited Partek software in their research. Why? Because it empowers scientists to perform sophisticated statistical analyses with intuitive point-and-click actions without command-line knowledge. Join us for a seminar where we will show you how the intuitive graphical user interface and interactive tools of Partek Flow software can simplify your NGS data analysis and overcome common analysis challenges, especially on the Single-cell RNA Sequencing (scRNA-Seq) data. We will demonstrate how to analyze a scRNA-Seq data set with multiple biological replicates and detect genes that are differentially expressed between cell populations across sample groups.

Oral Presentation Abstracts

Session IV: Translational Bioinformatics, Imaging and Data Visualization

OP-11 **Unravelling Complex Patterns in Omics Data with Integrative GNPA Tools**
Jake Chen
University of Alabama at Birmingham

Dr. Jake Y. Chen is a Professor of Genetics and Computer Science, Chief Bioinformatics Officer of the newly established Informatics Institute, and Head of the Informatics Section of the Genetics Department at the University of Alabama at Birmingham. He holds a BS degree in Biochemistry and Molecular Biology from Peking University, and both MS and PhD degrees in Computer Science and Engineering from the University of Minnesota. He has more than 20 years of bioinformatics R&D experience, including biological data mining, computational systems biology, and translational bioinformatics, with more than 150 peer-reviewed publications. His research focuses on building quantitative biomolecular systems models from genomic and clinical big data, thus helping understand, simulate, and predict complex disease biology outcomes. Prior to join UAB, he holds tenured faculty positions at Indiana University School of Informatics and Computing and at Purdue University Computer and Information Science Department. He is also an entrepreneur who created several startup companies to make emerging biomedical data easy to interpret and use by growing Medicine 2.0 stakeholders.

Oral Presentation Abstracts

OP-12   **Computer-assisted Inference of Association and Causation in Time-Series Data**

Phillip Kilgore
Louisiana State University-Shreveport

Many experimental designs include input and subsequent association of data from different time-points. These occasionally infer causation, a claim that one event arises out of another. As the move towards data-driven analysis progresses, the number of potential relationships which may arise out of it may increase. This leads to the question: are there computational tools available which can predict these relationships?   We begin by briefly describing causation and scientific definitions applicable to the sciences. We additionally differentiate this from association. We discuss a sample application of the causation problem with the gateway drug hypothesis (also gateway drug theory or stepping-stone theory), for which there has been considerable debate. We then present a synopsis of available computational methods to infer causation and present our approach 'GatewayNet.'   We then present the application of GatewayNet to an empirically gathered longitudinal data set from urine drug screening results of 70,000 individual trauma center patients at LSU Health Sciences Center in Shreveport. We were not able to infer any casual relationships between the data; however, we were able find to two notable associations which are known to have been initiated by doctors. Finally, we discuss how the environment the data was gathered in may have affected the results, and what can be done to improve data collection for the purpose of making associations or inferring causation.

OP-13   **Multi-modal Mass Spectrometry for Translational Proteomics**

Fabrizio Donnarumma
Louisiana State University-Baton Rouge

Mass spectrometry (MS) is an invaluable tool for translational medicine and biology. It can be used to identify novel biomarkers, link pathways to diseases and monitor energy and metabolites fluxes at the cellular level. In particular, MS-based proteomics has taken full advantages of the sensitivity and high throughput capability of MS at the identification as well as the quantification level. Masses can be recorded in several ways and specimens can be analyzed in imaging or bulk mode. MS imaging (MSI) can record mass spectra displaying all the masses present in a given pixel, producing hundreds of images in a single experiment. At the same time, identification relies solely on the mass value recorded. Liquid chromatography (LC) tandem MS (MS/MS) analysis of bulk samples allows for sensitive detection and quantification, but any information about localization is lost. We have developed a workflow to bridge MSI and LC-MS/MS using laser ablation and capture. MSI experiments are conducted using a matrix-assisted laser desorption/ionization (MALDI) MS instrument, which can analyze tissue sections and measure hundreds of proteins in each pixel. The MALDI imaging output is used to guide region of interest (ROI) selection, which is extracted using mid-infrared laser ablation. Captured samples are digested using magnetic beads sample preparation and the tryptic peptides are analyzed with LC-MS/MS. A custom software application (Proteomics and Imaging Tools, PIT) has been created to query protein databases. Detected proteins are mapped back to the MALDI image, increasing the overall protein identification confidence. In addition, the

software can include posttranslational modified sequences, which allows for a more accurate correlation between imaging and LC-MS/MS datasets. Novel modules of the software are being developed to generate MALDI images using LC-MS/MS to guide the mass selection, increasing the visualization capability of the workflow.

Session V: Protein Structure and Machine Learning

OP-14 **MD Simulation of Protein Control by Covalent Modifications: A Preliminary Study with PFKFB3**
Yong Lee
Louisiana State University-Baton Rouge

Dynamic covalent modification is one of the most widely adopted mechanisms for controls of various protein functions and, yet, the involved molecular mechanics remains an area of under development, because of the experimental difficulties of structural studies. To stimulate the area by lessening the experimental burdens, a test whether molecular dynamic (MD) can simulate those regulatory covalent modifications was made on 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase 3 (PFKFB3), an emerging target of fights against cancer. PFKFB3 is expressed only in tumor-like proliferating cells as an essential component of oncogenic cell transformation. PFKFB3 controls the level of fructose-2,6-bisphosphate (F-2,6-P2), the allosteric activator molecule for the three hallmarks of cancer: vigorous glycolysis, rapid cell cycle, and active blood vessel formations. The activity for F-2,6-P2 production is increased by asymmetrical di-methylation at Arg131 and Arg134 (N-CH3) and decreased by S-glutathionylation (S-Gsh) at Cys206. To understand the molecular mechanics involved in the N-CH3-dependent activation, a small scale MD was carried out. The resulting MD model was evaluated and supported by functional characterizations and comparisons to the model of the S-Gsh-dependent down-regulation. Our study suggests a possibility that MD can be an alternative approach to structural biology experiments for the study of protein control by covalent modifications. For better evaluation, we will expand our MD tests to other proteins with the known structures of their modified forms.

OP-15 **Prediction of Hierarchical Classification of Transposable Elements using Machine Learning Approach**
Manisha Panta
University of New Orleans

Transposable Elements (TEs) or jumping genes are the DNA sequences that have intrinsic capability to move within a host genome from one genomic location to another - genomic location can be either same or different chromosome. The study shows that TEs have role in genome function and evolution as their presence can modify the functionality of genes and increase size of genome. Thus, proper classification of the identified jumping genes in a genome is important to understand their particular role in germline and somatic evolution. The classification of TEs is majorly based on the mode of transposition, number and type of genes they contain and similarities in sequence. In this work, we studied multiple novel sequence-

derived features such as k-mers frequency, nucleotide composition and conservation profile to generate a hierarchical datasets. We proposed a machine learning based method to predict hierarchical classification of transposable elements using support vector machines (SVMs). We compared the proposed method with the existing methods based on the measures specific to hierarchical problems which includes whether the classification can stop at the internal node of the hierarchy or must continue until a leaf node is reached. The comparative results indicate that the proposed method significantly outperforms the state-of-the-art methods.

OP-16   **Prediction of RNA Binding Protein using Machine Learning Technique**
Reecha Khanal
University of New Orleans

RNA-binding proteins play important roles in many biological processes like gene regulation, protein synthesis and sequence encoding during both transcription and post-transcription processes. Identifying RNA-binding proteins from only sequence information is an incredibly challenging problem in computational biology. Although, existing literature show significant progress in the field, the problem is still distant from being fathomed. In this work, we present a machine learning technique to predict RNA-binding proteins, based on comprehensive set of features encoded from protein sequence. To develop a robust classifier, we encode the protein sequence with important features such as physiochemical properties, evolutionary information, torsion angle flexibility, disorder probability, monogram, bigram and more. The comparative results of the proposed method with the sate-of-the-art methods based on 10-fold cross-validation, independent test and case studies show that the predictor is able to correctly predict more number of RNA-binding and non RNA-binding proteins. Therefore, the proposed predictor can be applied for prediction of RNA-binding proteins only from sequence.

Poster Abstracts

PA-01   *Stacking based Effective Prediction of Disorder Proteins from Sequence*
Bhupendra Acharya
University of New Orleans

Intrinsically disordered or unstructured proteins do not adopt well-defined, stable three-dimensional (3D) structure. However, they perform important biological functions through their flexible conformations during binding process. Disorder proteins are of great interest in biomedical, pathology and drug formulation to combat critical diseases. Because they play significant role in regulation, signaling and drug design, the accurate identification of disordered regions becomes significantly important. In this study, we introduce a disorder predictor that employs ensemble machine learning based approach, called stacking and novel features to accurately identify disordered regions in protein. In stacking, the information from multiple predictive models are combined by a new predictor to generate an ensemble predictor. Furthermore, we explore novel and important features such as evolutionary information collected from position specific scoring matrix, half sphere exposure, torsion angle flexibility, charge and polarity of side chains, residue wise contact energy, secondary structure probabilities etc. to characterize the residues of a protein. The merits of the proposed approach has been rigorously examined by the validation, independent test and case studies which indicates that the proposed method provides significant improvement over the state-of-the-art methods.

PA-02   *TAFPred: Prediction of Torsion Angular Fluctuation to measure the Flexibility of Protein using Machine Learning*
Md Kauser Ahmmed
University of New Orleans

Proteins are dynamic molecules having a varying degree of structural flexibility that allows conformational changes. Structural flexibility enables multi-functional within proteins, which are essential for interactions between proteins and peptide, DNA, RNA or carbohydrate. Protein flexibility is mostly dictated by the fluctuations in the Cartesian coordinates. However, as the backbone of a protein can be expressed mostly by torsion angles ϕ and ψ, the fluctuation of these torsion angles can be used to describe the flexibility of the backbone of a protein. In this study, we propose a machine learning based method to predict the fluctuation of torsion angles by learning the features generated from protein sequences. For improved training of the proposed machine learning method, we utilized various useful features such as disorder probability, conservation profile based on the position-specific scoring matrix, secondary structure probabilities, monogram, bigram, position-specific estimated energy, half sphere exposures and so on. We compare the performance of gradient boosting regression (GBR) with the artificial neural network (ANN) based state-of-the-art. The GBR based predictor achieved 10-fold cross-validated correlation coefficient of 0.648 and 0.65, and mean absolute error of 22.12° and 23.77° for the angle fluctuation of ϕ and ψ, respectively. We have found that the GBR method proposed in this work outperforms the ANN-based previous best method. The torsion angle fluctuations have been successfully employed as features in various applications such as secondary structure prediction, disulfide bond prediction, protein disorder prediction and more. Thus, it is expected that our improved predictor will be useful to improve the accuracies of various bioinformatics

applications including protein structure and function prediction, while the predicted fluctuations of torsion angles are used as restraints.

PA-03   *Effective Brain Hemorrhage Diagnosis Using Machine Learning Approach*

Duaa Alawad
University of New Orleans

Due to their importance in medical diagnosis, computer-aided detection and diagnosis (CAD) system has drawn the attention of many researchers over the past few decades. One of the admired CAD systems is the one in which the medical images are processed and analyzed fast and comprehensively for accurate diagnosis. Such systems can be developed by effectively incorporating the elements of artificial intelligence, machine learning and digital image processing. The developed framework can then be easily customized for deployment in various types of medical imaging systems such as X-ray, Magnetic Resonance Imaging (MRI), Computed Tomography (CT) scan and ultrasound dealing with different types of diseases, due to their underlying common paradigm. In this study, we propose a machine learning based approach to a CAD system to detect the existence and type of hemorrhages in human brains from the produced CT scan images. The proposed machine learning predictor is designed and developed by extracting various important features from region of Interest (ROI) such as area, filled area, perimeter, orientation, convex area, bounding-box-area and so on through digital image processing. Although there exist some articles discussing the problem of detecting hemorrhages from a CT scan of the brain, very few of them address the problem of identifying the type of hemorrhages. While performing the validation and test, it is found that our proposed system attains higher accuracy in detecting the brain hemorrhage including their variations. Based to test cases, the hemorrhage detection rate is found to be 100% accurate and the hemorrhage type classification is found to be 96% accurate. Such a robust system can be utilized to extend the effectiveness of medical diagnosis.

PA-04   *Ab initio Protein Structure Prediction using Binary Genetic Algorithm for Energy Function Minimization*

Joel Andrepont
University of New Orleans

Crucial protein-function in essentially all biological process is closely related to the three-dimensional (3D) shape of the protein. Ab initio modeling seeks to predict the structure of a protein without relying on homologous sequences or native fragments of some protein. The development of ab initio method hinges on effective conformational space sampling and an accurate energy function that guides the search process. Existing literature clearly demonstrate the possibility to sample and predict good quality protein structure without using fragments from known protein structures. Towards this goal, we developed an ab initio method that employs a memory assisted binary Genetic Algorithm (MabGA) for conformational space sampling, searching for the global minimum or the native fold of the protein. Our approach samples the conformational space of the protein using novel mutation and crossover operations based on angular rotation and translation capabilities that allow for diversification and concentrated

searching on possible minima. This extra layer of abstraction causes the operations to behave differently than direct operations onto real valued structures such as those found in a Real Genetic Algorithm (RGA). Both the mutation and the crossover methods operate according to an ordering provided by the known best parents from previous generation. In addition, we developed and employed a novel knowledge-based energy function, called 3DIGARS, which is an optimized combination of crucial properties such as hydrophobic versus hydrophilic properties, sequence-specific predicted accessibility and ubiquitous phi-psi angular characterization. The samples obtained from the conformational space are scored using 3DIGARS and passed through the BGA operators for natural selection and creation of new generations. Rigorous examination and case studies demonstrates the effectiveness of the proposed method for predicting the corresponding 3D structure of a given sequence.

PA-05   ***Tulane Cancer Center Cancer Crusaders Next Generation Sequence Analysis Core***
Melody Baddoo
Tulane University

Next Generation Sequencing (NGS) has revolutionized transcriptomic and genomic studies with the promise of unprecedented insights into biology and human disease. Despite its potential, the complex nature of data analysis is a significant barrier for most biologists, slowing the adoption of this technology by many laboratories. We established the Tulane Cancer Center Cancer Crusaders Next Generation Sequence Analysis Core in an effort to eliminate the analysis bottleneck for COBRE and Louisiana Cancer Research Consortium (LCRC) investigators. The core was built upon a principally DIY (do it yourself) model whereby the core serves as an enabling mechanism to allow investigators to perform a broad spectrum of methods ranging from routine to highly specialized data analysis pipelines. This model is achieved through a facilitative approach: 1) providing appropriate computational resources, 2) maintaining and updating existing open source software and testing of newly published software, 3) study design consultation, 4) project specific pipeline development, 5) one-on-one training during analysis, 6) continuous accessibility of expertise, 7) troubleshooting, 8) custom script development, and 9) an educational seminar series (Next Generation Sequence Analysis Learning Series). The Core is equipped with 12 Mac Pro computers, each with sufficient memory to process all aspects of upstream and downstream NGS data analysis. These resources are complemented with access to the existing Tulane Sphynx cluster and the upcoming Tulane HPC, 'Cypress' for the processing of very large datasets. The Core receives funding from both the National Institutes of Health Institute of General Medical Sciences (NIHGMS) COBRE Program (Mentoring a Program in Cancer Genetics, Program Director: Prescott Deininger, Ph.D.) and the Cancer Crusaders.

PA-06   ***Molecular Docking and Simulation of Zika Virus NS3 Helicase***
Syed Badshah
Louisiana State University-Baton Rouge

The Zika Virus (ZIKV) has gained attention for the last few years due to the congenital microcephaly and Guillain Barre Syndrome that resulted in humans. The non-structural protein-3 (NS3) helicase of ZIKV play an important role in viral RNA replication. Here, we performed

Poster Abstracts

hundred nanosecond molecular dynamics simulation and molecular docking of the NS3 helicase of ZIKV with 1,4-benzothiazine derivatives. The root mean square deviation (RMSD) analyses showed the stability of the NS3 helicase. The simulations showed that the flexible and rigid domains of the protein play a crucial role during the RNA replication process. All such domains with ligand binding pockets can be targeted for drug design. The molecular docking show that the strong hydrogen bonding and arene-cation interactions are responsible for the binding between NS3 and 1,4-benzothiazine derivatives, which provides a new dimension for potent drug design for ZIKV.

PA-07   *Elucidating the mechanism of HSV1 induced RNA Polymerase II relocation*
Claire Birkenheuer
Louisiana State University-Baton Rouge

Human herpes simplex virus 1 (HSV1) is a common human pathogen, which causes serious disease in some patients. It uses host RNA Polymerase II (Pol II) for expression of its genome. How the virus relocates Pol II to its own genome and alters host gene expression for productive infection is poorly understood. Understanding these mechanisms will contribute to the development of herpes virus therapeutics. One way to study this is with precision nuclear run (PRO-seq). This technique uses biotinylated nucleotide incorporation in nuclear run-on reactions from infected cells. The incorporation stalls Pol II on the base being transcribed. Bioinformatics analysis creates maps defining the exact location of Pol II with base pair resolution. We mapped Pol II location in HSV1, and mock infected cell genomes at 3 hours post infection, with and without cycloheximide (CHX) treatment. A pipeline originating from the Danko laboratory incorporates cutadapt and bowtie to process and align the sequences respectively. Sequences were aligned to the human and HSV1 genomes, and IGV was used to view Pol II location. SeqMonk was used to analyze changes in Pol II occupancy using DeSeq2 statistical analysis. After infection, Pol II is removed from some host genes but is un-changed or recruited to others. Pol II termination is extended on many host genes but appears normal on others. CHX treatment, which inhibits immediate early viral protein expression, does not preclude the loss of Pol II on some genes, but abrogates it on others. The majority of HSV1 genes are occupied by Pol II at 3 hours post infection, but CHX treatment causes a reduction of Pol II on early and late genes, and increases occupancy on viral immediate early genes, UL39, and on the L/ST region. These data suggest virus binding and entry, incoming viral proteins, and immediate early genes all work collectively to change the location of Pol II shortly after HSV1 infection.

PA-08   *Evaluating the Role of Fluid Shear Stress on DNA Mutation*
Melyssa Bratton
Xavier University of Louisiana

Of the 200,000 new cases of breast cancer diagnosed each year in the U.S, approximately 80% of the cancers are estrogen receptor alpha (ER) positive. Therefore, most first line treatments are endocrine therapies that target ER. Selective estrogen receptor modulators (SERMs) and selective estrogen receptor down-regulators (SERDs) are used in premenopausal women, and aromatase inhibitors (AIs) are often used in postmenopausal women. Unfortunately, many of

18

these cancers revert to an endocrine resistant, metastatic phenotype after one or more successful first line treatment(s). Patients who present with endocrine refractory disease have no good option for therapy other than chemotherapy regimens, which ultimately fail. Interestingly, around 25% of patients with endocrine resistant, metastatic disease harbor a mutated ER within the cancer cells, a mutation that was not present in the primary tumor. One of the most common ER mutations converts tyrosine 537 to serine (Y537S), rendering the receptor constitutively active. During metastasis, cells from the primary tumor are shed into the bloodstream and undergo hemodynamic shear stress on route to the secondary site of metastasis. Some of these cells, known as circulating tumor cells (CTCs), will seed in secondary organs and produce metastatic lesions. Several studies have recently linked fluid shear stress (FSS) with DNA instability and tumor metastasis. We propose modeling FSS in breast tumor cells using a unique pulsatile microfluidic device, with the ultimate aim of understanding how FSS changes genomic dynamics that may affect not only ER but also other genes responsible for breast cancer metastasis.

PA-09   *T-BioInfo: A bioinformatics platform that enables scalable and efficient project-based learning for biomedical data analysis and problem solving skills.*

Elia Brodsky

Other

Bioinformatics skills are in high demand in life science research, but even at the level of basic data stewardship, these skills are taught only in about 25% of educational programs (M.D. Brazas et al, Nature, 2017). Limiting factors for the integration of bioinformatics practice-oriented courses into curriculum include the lack of qualified trainers experienced in the subject, cost of computational resources and infrastructure maintenance, and standard pre-requirement of coding competencies. T-BioInfo is a research platform that can also be used for education (https://t-bio.info) in combination with online educational modules allow overcoming these limitations providing theoretical background and hands-on experience in a project-oriented educational environment. The T-BioInfo platform is a combination of interconnected algorithms that can be assembled into pipelines for analysis of multiple types of 'omics' data and their integration. The platform enables application of advanced bioinformatics methods without the need of coding skills. In parallel, online educational modules provide a theoretical framework and practical guides needed for different types of analysis. The course participants are invited to analyze publically available datasets presented as exemplar projects throughout the course. This model for bioinformatics training was implemented online and received a high ranking from surveyed users (N=102), also the courses were piloted at Loyola University New Orleans, in January 2018. Twelve undergraduate students and five faculty members participated in the pilot program. The course was accompanied by assessment of participant's skills prior to the course, tracking the online activity of each participants and an additional assessment in the end of the program. We found that such educational courses can be effective online and efficiently implemented in university bioinformatics programs.

Poster Abstracts

PA-10   *Infection Characteristic of HSV-1 Neuronal Viral DNA and Analysis of ICP4, ICP0 In Silico*

Christian Clement
Southern University at New Orleans

Background: NZW rabbit model harboring neuroinvasive H129, epinephrine iontophoresis induced to reactivate, produced neovascularization in specifically the left brain. A novel HSV-1 replication and viral DNA accumulation in neuronal/corneal damage and triggering of electrical seizures was hypothesized. Exploring this HSV-1 infection phenotype using a transgenic mouse model carrying the human APOE4 gene (ApoE e4/e4) infection screen with two virus strains 17Syn+ high-phenotypic-reactivator (HPR) and 17ΔPst(LAT-) low-phenotypic-reactivator (LPR), stipulated ICP4, ICP0 roles (Louisiana Biomedical Research conference, 15th Annual Meeting, January 20-22, 2017, Baton Rouge LA; Poster Session Abstracts, #4: 'ICP4 and ICP0 HSV-1 Viruses, Virus-Induced Inflammation and DNA in the Brain'; 5th annual LA Conference on Computational Biology & Bioinformatics, April 7-8, 2017, Xavier University New Orleans, LA; Poster Session Abstracts, #PA-04 'Evaluating ICP4 and ICP0 HSV-1 Viruses, Virus-Induced Inflammation and DNA in the Brain').   Methods: HSV-1 DNA copy numbers were determined by calculating the number of HSV-1 polymerase genes per sample. Forward/reverse primer pairs were 5'- AGA GGG ACA TCC AGG ACT TTG T -3'/5'- CAG GCG CTT GTT GGT GTA C -3' and signal levels was quantified with the fluorescent probe sequence 5'-6-FAM/ACC GCC GAA CTG AGC A/3' BHQ-1'.   Results: Hippocampus is associated with long-term memory and spatial navigation and one of the first regions to suffer damage in Alzheimer's disease. HSV-1 DNA detection of the viral genome of >10 copy numbers for both 17Syn+ (HPR) and 17ΔPst(LAT-) (LPR) in the left hippocampus and for only17Syn+ (HPR) in the left cortex were not significantly affected by heat stress treatment, concentrations of Acyclovir dosing and mode of drug administration, unlike the control C57Bl/6N mice which were viral DNA free.   Conclusion: This is a first report of novel HSV-1 infection characteristic in the presence of the human APOE4 gene.

PA-11   *Sequence-Based Prediction of Protein Carbohydrate Binding Sites Using Machine Learning Techniques*

Lijing Cui
University of New Orleans

Protein-Carbohydrate binding plays an important role in many biological processes such as cellular adhesion and host-pathogen recognition, inferences to drug targeting and gene expression. There exist various experimental as well as computational techniques such as metal nanoparticle probing, molecular docking and site mapping, however, these techniques either require known complex structures which are costly as well as labor intensive to obtain. Thus, in this work, we present a sequence-based method to accurately predict protein-carbohydrate binding sites using a machine-learning approach. To accurately predict protein-carbohydrate binding sites, we explored several important sequence and predicted structural features such as: evolutionary feature called sequence specific scoring matrix (PSSM), protein disorder region, charge and polarity of side chain, torsion angle flexibility, half sphere exposure, position specific estimated energy and so on. In addition to feature engineering, feature selection technique is utilized to identify the features that significantly influence the accuracy of the proposed predictor. The predictive capability of the predictor is tested through case studies, 10-fold cross

validation and independent test dataset. The results obtained from validation and test, demonstrate robustness of proposed predictor in protein-carbohydrate binding site predictions.

PA-12   ***EGFR targeted near infrared (NIR) BODIPY-peptide conjugates for colon cancer imaging.***
Achyut Dahal
University of Louisiana at Monroe

Epidermal Growth Factor Receptor (EGFR) overexpression is observed in several cancers such as colorectal cancer, non- small cell lung cancer, breast cancer and ovarian cancers. The overexpression of EGFR can be effectively exploited for diagnostic and imaging purposes to detect different types of cancer. Two peptides, EGFR-L1 and EGFR-L2, were designed, which can selectively bind to the extracellular domain of EGFR. These peptides were conjugated to near infrared (NIR) 4,4-Difluoro-4-bora-3a,4a-diaza-s-indacenes (BODIPY) using linkers. BODIPY dyes maintain a highly tunable core structure, and are intrinsically stable under various physiological conditions such as chemical, physical, biological, and thermal making them promising candidates for biomarker labeling. However, the solubility of BODIPYs limits their use for in vivo studies. Here we have designed BODIPY-peptide conjugates with a glucose biomolecule attached to the BODIPY framework to improve the solubility and cell permeability of the conjugates. The ability of the designed BODIPY-peptide conjugates to bind to EGFR was evaluated using computational docking studies. AUTODCK was used to perform the docking studies on QB2 via LONI. The binding of the BODIPY-peptide conjugates to EGFR was confirmed by surface plasmon resonance (SPR) studies. Results from our studies indicated that the incorporation of the glucose biomolecule improved the solubility of BODIPYs and these conjugates bind specifically to the EGFR extracellular domain. Hence, these conjugates can be used for in vivo imaging studies of colorectal cancer, which overexpresses EGFR. The work was supported by funding from the National Institutes of Health (NIH) grant number R01 CA179902 (Vicente).

PA-13   ***Improved Mathematical Model Enhances Understanding of Endoreplication in Arabidopsis Trichomes with 4D Visualization***
Renee Dale
Louisiana State University-Baton Rouge

Arabidopsis trichomes switch from cell division (mitotic cycles) to DNA replication (endocycles) during development. Although the cell cycle has been modeled extensively, no previously published model has been able to recover all mutant phenotypes. In this project we set out to determine the system components required to recover 19 mutant phenotypes involving the cyclin-dependent kinase inhibitors SIM and KRP, the Anaphase-Promoting Complex/Cyclosome (APC/C) activator CCS52, and the mitotic cyclin CYCB starting with a published model. Four-dimensional visualization was used to characterize wild-type and in silico mutant cell cycle behavior. This analysis suggested that the APC targets an unidentified negative regulator of KRP for destruction during the G1 phase of the cell cycle as part of a feedback loop involving the S-phase CDKA/CYCD complex and its inhibitor KRP. The F-box protein FBL17 is one candidate for such a factor. Our analysis also suggests the relative importance of cellular control of model

components, in particular the APC/CCS52. Our model will be useful to plant biologists in the study of endoreplication and cell cycle controls.

PA-14   ***Multi-modal Mass Spectrometry for Translational Proteomics***
Fabrizio Donnarumma
Louisiana State University-Baton Rouge

Mass spectrometry (MS) is an invaluable tool for translational medicine and biology. It can be used to identify novel biomarkers, link pathways to diseases and monitor energy and metabolites fluxes at the cellular level. In particular, MS-based proteomics has taken full advantages of the sensitivity and high throughput capability of MS at the identification as well as the quantification level. Masses can be recorded in several ways and specimens can be analyzed in imaging or bulk mode. MS imaging (MSI) can record mass spectra displaying all the masses present in a given pixel, producing hundreds of images in a single experiment. At the same time, identification relies solely on the mass value recorded. Liquid chromatography (LC) tandem MS (MS/MS) analysis of bulk samples allows for sensitive detection and quantification, but any information about localization is lost. We have developed a workflow to bridge MSI and LC-MS/MS using laser ablation and capture. MSI experiments are conducted using a matrix-assisted laser desorption/ionization (MALDI) MS instrument, which can analyze tissue sections and measure hundreds of proteins in each pixel. The MALDI imaging output is used to guide region of interest (ROI) selection, which is extracted using mid-infrared laser ablation. Captured samples are digested using magnetic beads sample preparation and the tryptic peptides are analyzed with LC-MS/MS. A custom software application (Proteomics and Imaging Tools, PIT) has been created to query protein databases. Detected proteins are mapped back to the MALDI image, increasing the overall protein identification confidence. In addition, the software can include posttranslational modified sequences, which allows for a more accurate correlation between imaging and LC-MS/MS datasets. Novel modules of the software are being developed to generate MALDI images using LC-MS/MS to guide the mass selection, increasing the visualization capability of the workflow.

PA-15   ***Prediction of Protein Secondary Structure from Sequence using Machine Learning Approach***
Michael Flot
University of New Orleans

Secondary structure (SS) refers to the three dimensional form of local segments of proteins. Accurate prediction of SS from sequence is still an unsolved problem and is widely researched in bioinformatics. In fine-grained state, SS can be classified into 8 different types: 310-helix, alpha-helix, pi-helix, beta-strand, beta-bridge, beta-turn, high curvature loop and irregular. Whereas, in broad state SS are classified into three categories: alpha-helix, beta-strand and coil. In this study, we propose a machine learning based predictor for three and eight classes of secondary structure prediction, based on comprehensive feature encoding. To encode protein resides, we utilize important features such as solvent accessibility, conservation profile, half sphere exposure, torsion angle fluctuation, disorder probabilities and so on. The usefulness of the proposed approach is assessed by following widely used technique called 10-fold cross-validation

and independent testing. The results obtained from the rigorous examination validates the robustness of the proposed method.

PA-16   *Sequence and Structure based Protein Peptide Binding Residue Prediction*
Suraj Gattani
University of New Orleans

Protein-peptide interactions are one of the most important type of biological interactions, which play a significant role in many cellular signaling pathways (e.g., receptor tyrosine kinase (RTK) signaling) and other cellular processes (e.g., endocytosis). Such interactions have some special properties such as promiscuity, which means many different peptide sequences are able to bind to the same protein domain due to the conformational flexibility of the peptides. These special properties make the study of protein-peptide interactions challenging as well as interesting. Most of the available methods for protein-peptide interaction modeling need complex structure, which are unknown in some cases. Protein-peptide complex structures can be predicted by using docking techniques. However, these methods are accurate only if binding sites are known reasonably well. To decrease the cost of prediction, we propose computational study and prediction of the binding sites using effective machine learning based method to make accurate prediction of protein-peptide binding residues through the inclusion of both the sequence and the structure based features. Some of the useful features explored in this work are: evolutionary information extracted from position specific scoring matrix (PSSM) profile, secondary structure probabilities (SS), torsion angle fluctuation (dphi & dpsi), half sphere exposure (HSE), position specific estimated energy (PSEE) and so on. Although there exist many predictors, which are either based on sequence-based prediction or structure-based prediction, but none of them are of the combination of the both. While we provide the combined approach, we also keep the structural input optional with the sequence input requirement as the structural input are scarce. Regarding prediction resolution, we predict binding at the residual level. The robustness of the proposed method has been confirmed through proper validation and testing.

PA-17   *Integration of Pocket-Matching and Virtual Screening For Drug Repositioning*
Rajiv Govindara
Louisiana State University-Baton Rouge

A disease or disorder is defined as rare in the USA while it affects less than 200,000 Americans at any given time. 50% of rare diseases affect children. Rare diseases are often not pursued by the pharmaceutical industry due to extremely limited individual markets. Computational-based drug repositioning aims at discovering previously approved drugs for new indications, has emerged as a cost-effective and quicker alternative to traditional drug discovery. Specifically, matching of drug-binding pockets approach for repositioning is appealing because they putatively nominate the most promising candidate drugs for a given indication. Here, we present an integrated methods of pocket-matching and structure-based virtual screening for large-scale repositioning of existing drugs to treat rare diseases. The resulting dataset comprises 31,142 putative drug-target complexes linked to 980 orphan diseases. To illustrate how potential therapeutics for rare diseases can be identified, we discuss a possibility to repurpose a steroidal aromatase inhibitor

to treat Niemann-Pick disease type C. Taken together, the exhaustive exploration of the drug repositioning space exposes new opportunities to combat rare diseases with existing drugs. DrugBank/Orphanet repositioning data are freely available to research community at https://osf.io/qdjup/.

PA-18 ***Developing an Exploratory Structural Equation Modeling (ESEM) framework for investigating biological networks of complex traits***

Jacqueline Harris
Grambling State University

The field of Systems Biology explains higher order interactions in terms of design principles. By design the genetic architecture that causes disease precedes risk components in time and can be described in terms of a structural latent model. Exploratory Structural Equation Models (ESEMs) have the potential to combine relevant clinical, genetic, and environmental variables in a joint analysis that can be used to explore higher order biological systems. ESEMs are composed of two levels 1) a measurement model where observed variables (i.e. SNPs/phenotypes) are used as measures to indicate latent factors (i.e. unobserved underlying biology) and 2) a structural model where covariates that influence latent factors are analyzed. This statistical framework is complementary to multilevel data, and can be used to address a major obstacle of joining the analysis of genetic, biological, environmental, clinical, and socioeconomic variables to understand disease etiology.

PA-19 ***Bioinformatics Approaches for Integrating Germline and Somatic Mutation Information in Cancer***

Chindo Hicks
LSUHSC New Orleans

Cancer is a complex genetic disease involving both germline and somatic mutations. Recent advances in next generation sequencing have enabled deciphering of cancer genomes. Large multicenter projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) have performed detailed analyses of the somatic mutations in cancer genomes. These discoveries are increasing our understanding of the molecular basis of tumors and enabling discovery of clinically actionable biomarkers and targets for the development of novel therapeutics. However, to date, information on somatic mutations has not been leveraged to elucidate the possible oncogenic interactions between somatic driver mutations and germline mutations to understand their joint role in tumorigenesis. Here we integrated germline mutation information derived from genotype data with somatic mutations from TCGA and our own experiments on whole exome sequencing using transcriptome data as the organizing principle. The objective was to investigate the possible oncogenic interactions and mechanisms of cooperation between germline and somatic mutations during tumor development and progression. Our working hypothesis is that germline and somatic driver mutations interact and cooperate during tumorigenesis, and that these complex array of interacting genetic factors affect molecular networks and biological pathways driving tumor development and progression. The analyses revealed functionally-related genes containing germline and somatic driver

mutations involved in tumors. Additionally, the analysis revealed molecular networks and biological pathways enriched for germline and somatic mutations. We conclude that integrative bioinformatics analysis is a powerful tool for decoding the possible oncogenic interactions and mechanisms of cooperation between germline and somatic mutations during tumor development and progression.

PA-20 *Elucidating Protein Druggability with eFindsite*
Omar Kana
Louisiana State University-Baton Rouge

Identifying the viability of protein targets is one of the preliminary steps of drug discovery. Determining the ability for a drug to bind to a protein, termed druggability, requires a non-trivial amount of time and resources. Inability to identify druggability has accounted for a significant portion of failures in drug discovery. This problem is only further exacerbated by the large sample space of proteins involved in human diseases. With these barriers, the druggablity space within the human proteome is largely unexplored and has made it difficult to develop drugs for rare orphan diseases. Hence, we present a new feature developed in efindsite that uses supervised machine learning to predict the druggability of a given protein with high accuracy. With eFindsite, we elucidated the human druggability space to be around 6000-7000 proteins. These findings illuminate a myriad of new possible drug targets within the human proteome, including those related to incredibly rare and underfunded orphan diseases.

PA-21 *Computational Investigation of novel Casein Kinase 1o/E Inhibitors for treatment of Alzheimer's Disease*

Shrutivandana Kauloorkar
Xavier University of Louisiana

Alzheimer's disease (AD) is a progressive neurodegenerative disorder known to have notable symptoms like short term memory loss otherwise referred to as dementia. Abnormal hyperphosphorylation (P-tau) of the tau protein leads to the aggregation of amyloid plaques which is the hallmark of AD and several other neurodegenerative disorders. Casein kinase 1δ (CK1d) and casein kinase 1ε (CK1E) are closely related Ser-thr protein kinases belonging to Casein kinase 1 family of eight isozymes. The CK1o and CK1E isozymes are expressed in the brain. Overexpression of constitutively active CK1e leads to an increase of Aβ peptide production. CK1E inhibitors disrupt the amyloid precursor protein cleavage while an active form is known to augment APP production. This effect is specific for one of the brain CK1 isoforms (CK1ε). The upregulation of this particular isoform makes it an attractive target for the treatment of Alzheimer's disease.CK1o is thought to play role in neurofibirllary tangle formation, dopamine signaling, neuro transmitter release and cancer. Our research group has identified quinones as inhibitors of CK1δ and CK1E kinases. Some of these compounds have shown to be more selective for CK1E than CK1o. We have performed docking studies on these compounds with

the CK1o and CK1E X-ray crystal structures using the MOE modeling software to study the role of residues in selectivity. The results of the docking studies are presented here.

PA-22   *Informatics Tools For Mapping Molecular HLA Typing Data to UNOS Antigen Equivalencies*
Navchetan Kaur
Tulane University

United Network for Organ Sharing (UNOS) operates Organ Procurement and Transplantation Network (OPTN) for equitable organ allocation from deceased donors. For histocompatibility evaluation, molecular human leucocyte antigen (HLA) typing of the donors must be converted to antigen equivalencies for entry into UNet. Maintaining an up-to-date mapping table for this purpose has so far been deemed infeasible by UNOS, though OPTN provides general guidelines. To automate this conversion process for histocompatibility laboratories, we developed a mapping table and supporting informatics tools that would handle both unambiguous and ambiguous molecular typings.   Data sources referred included an OPTN policy document, an IMGT/HLA antigen mapping table maintained by WHO and the World Marrow Donor Association (WMDA) and the 2008 HLA Dictionary. Mapping of population-specific ambiguous HLA typings was done using allele frequencies from 26 US populations obtained from National Marrow Donor Program (NMDP). Validation of antigen mapping was done by mapping NMDP high resolution frequencies to antigens then comparing with UNOS calculated panel reactive antigen (CPRA) antigen frequencies using a normalized similarity index, 'If'.   We present a UNOS antigen mapping table for 16415 HLA alleles of the A, B, C, DRB1, DRB3/4/5, DQA1, and DQB1 loci from the v.3.30.0 release of the IMGT/HLA database. A web application and microservices to map HLA typing results were also developed (http://www.transplanttoolbox.org). Similarity score between the NMDP and the UNOS CPRA reference panels for 4 US broad population categories (Caucasians, Hispanics, African Americans and Asian/Pacific Islanders) ranged from 0.85 to 0.97. The table provides a standardized reference for antigen mapping that can be automatically updated to incorporate new alleles added to IMGT/HLA. Informatics tools for managing molecular HLA typing data may improve accuracy and reduce the manual effort needed to enter data into UNet.

PA-23   *Prediction of RNA Binding Protein using Machine Learning Technique*
Reecha Khanal
University of New Orleans

RNA-binding proteins play important roles in many biological processes like gene regulation, protein synthesis and sequence encoding during both transcription and post-transcription processes. Identifying RNA-binding proteins from only sequence information is an incredibly challenging problem in computational biology. Although, existing literature show significant progress in the field, the problem is still distant from being fathomed. In this work, we present a machine learning technique to predict RNA-binding proteins, based on comprehensive set of features encoded from protein sequence. To develop a robust classifier, we encode the protein sequence with important features such as physiochemical properties, evolutionary information,

torsion angle flexibility, disorder probability, monogram, bigram and more. The comparative results of the proposed method with the sate-of-the-art methods based on 10-fold cross-validation, independent test and case studies show that the predictor is able to correctly predict more number of RNA-binding and non RNA-binding proteins. Therefore, the proposed predictor can be applied for prediction of RNA-binding proteins only from sequence.

PA-24   ***Detecting Associations and Causation In Urine Drug Screening Data***
Phillip Kilgore
Louisiana State University-Shreveport

The gateway drug hypothesis (also gateway drug theory or stepping-stone hypothesis) is the idea that using certain drugs may increase the probability that other drugs will be used in the future. It has been hotly contested since 1975, when Kandel published a sequence predicting a bidirectional sequence of drug use. Since that time, the question of whether or not substance abuse follows such a sequence (and whether or not these observations can be truly regarded as causal) has remained controversial within the study of substance abuse.   To assess this problem, we developed GatewayNet, a sequential rule mining program which attempts to not only determine association, but also highlight instances where rules may represent causation. This is performed using structure learning to infer a weighted directed graph representing associations and their support. Although this form of structure learning is common for this application, we add three additional tools to support testing the gateway drug hypothesis: the concepts of initiation and gateway rules, a windowed method of sampling each history within the transaction database, the and a new 'certainty' measure which depicts how strongly one may reasonably conclude that one event may cause another.   We first apply GatewayNet to a synthetically derived data set with known patterns to validate its behavior. We then apply GatewayNet to an empirically-gathered data consisting of urine drug screening results from over 70,000 patients gathered in a trauma center setting between 1998 and 2011. We were not able to discover any gateway rules; however, we were able to discover initiations of two medicines that fell into obsolescence into their replacement, confirming that these sorts of transitions can be discovered.

PA-25   ***Comparative Study on Using Force Fields of Three Varying Course-graining Levels in the Computational Study of the Behavior of DPPC Lipid Bilayer in the Presence of DMSO***
Hye-Young Kim
Southeastern Louisiana University

One of the main factors of a successful simulation is employment of a good set of force fields (ff) and the current understanding of the nature of coarse-graining in multiscale simulations is still limited.   Our goal is to systematically investigate if and how differently (or similarly) simulations adopting different coarse-graining levels lead to a certain behavior of a lipid bilayer in the presence of DMSO, such as the area per lipid molecule, the membrane thickness, and ordering. Previous studies on this system were done in slightly different study system sizes, compositions, temperatures, etc. Therefore, to eliminate any factor other than the coarse-graining level, we performed MD simulations on the identical study system using same set of parameters other than the adaptation of three different molecular modeling methods: (1) fully united-atom

simulation using gromacs ff for all molecules, (2) fully coarse-grained simulation using Martini ff for all molecules, and (3) hybrid MD simulation using OPLSaa ff for DMSO and Martini ff for water and lipids molecules. The summary of our findings will be presented.

PA-26   ***GDash: A Genomics Dashboard that Integrates Modeling and Informatics***
Zilong Li
Louisiana Tech University

Genomics is a sequence based informatics science and a structure based molecular science. Historically these two scientific approaches were distinct. There are few tools that unite them in a way that is readily accessible and yet scientifically robust. Here we introduce the concept of a genomics dashboard as a tool specifically designed to integrate informatics and physical modeling approaches for the study of chromatin. GDash is a prototype genome dashboard that integrates our Interactive Chromatin Modeling(ICM) tools, the Dalliance genome browser and JSmol for the display of 3D structures of chromatin, sequence data and informatics tracks. The exchange of data between informatics and physical models is bi-directional such that structure data can be displayed as tracks in Dalliance (e.g. Roll, Slide, or Twist) and informatics data can be mapped onto molecular structures (e.g. color by genetic function). GDash allows users to rapidly fold any sequence of DNA into atomic or coarse grained models of chromatin based on the material properties of DNA and nucleosomes and experimental or theoretical nucleosome positioning data, interactively manipulate nucleosomes (add, delete and move) and assign different conformational states to each nucleosome (e.g. tetrasome, octasome, chromatosome). The coarse grained model of chromatin implemented in GDash utilizes LAMMPS to optimize and sample 3D structures in real time. GDash is a novel tool for the cross-validation of physical modeling and informatics datasets, for building knowledge based potentials for chromatin folding, and for investigating structure-function relationships for regions of the genome ranging from kilobases to megabases.

PA-27   ***Decomposing Small Molecules for Fragment-Based Drug Design with eMolFrag***
Tairan Liu
Louisiana State University-Baton Rouge

Hit identification, lead generation, and lead optimization are the key steps at the outset of a drug discovery process. Virtual screening that can rapidly evaluate millions of compounds has become an integral part of lead identification protocols. De novo methods such as fragment- and atom-based techniques have been developed to generate novel chemical compounds for virtual screening. Fragments with empirical connectivity patterns can help improve the performance of fragment-based chemical synthesis tools such as eSynth. Hereby, we offer eMolFrag, a new open-source molecular fragmentation software tool to generate molecular fragments to build targeted screening libraries.

Poster Abstracts

PA-28   ***A Specific Service Resource of Microbial Genomics Research in Louisiana-MGRG***

Meng Luo

LSUHSC New Orleans

The Microbial Genomics Resource Group (MGRG) at the medical school of LSUHSC was established in 2012. Our mission is to offer comprehensive services and support for microbial genomics research using Next Generation Sequencing (NGS) and bioinformatics analysis. Currently, we have two Illumina NGS platforms available of Miseq and Nextseq. In the past more than five years, we have established a mature technical system of molecular biology and bioinformatics by developing cost effective protocols and QA/QC pipelines; integrating data analysis and visualization software. Our technical serviced include microbial DNA/RNA isolation; measurement of microbial copy number; sequencing library construction of amplicon (16S rDNA, ITS, specific regions for virus), DNA genome (metagenome shotgun, CHIP sequencing), RNA (RNA-seq, metatranscriptome, microRNA). Based on clients' request, we can provide molecular service for specific needs such as primer design, protocol development and optimization and consultation of technical strategy. Our bioinformatics team offers consultation of experimental design, data analysis training, services for standard metagenomics analysis, data management, and access to biocomputing resources. By the end of March 2018, we have finished 34 projects and more than 3200 samples, which don't include those we had finished using Roche 454 GS-FLX NGS system by the end of 2013. As co-authors, we have published 15 peer reviewed papers. As current PI, co-PI and co-I, we are involving several projects awarded by NIH R01, UH2 etc. We have involved in a broad range of microbial studies in obesity, alcohol user, diet, immunology, aging, vaginal disease, dental disease, infant and environment, fish production, ecology in river etc. Our collaborators and clients are faculties, scientists and graduate students interior and exterior LSU system. If you are interested in MGRG's services and concern about charges, please visit http://metagenomics.lsuhsc.edu/mgrg/.

PA-29   ***DADA2-to-PICRUSt: An Experimental Pipeline to Perform De Novo PICRUSt on Denoised Amplicon Sequence Variants***

Vincent Maffei

LSUHSC New Orleans

The evolutionary information encoded by the bacterial 16S rRNA gene is a major target of deep-sequencing studies investigating microbial community structure and function. Recently developed methods that correct or 'denoise' nucleotide base-calling errors commonly observed in Illumina-sequenced amplicon libraries are now popular data processing and quality control steps in high-throughput 16S-sequencing analyses. The software suite Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt), which infers the functional potential of sampled microbial communities from 16S amplicon sequences, is widely used by the microbial sequencing community to generate hypotheses on microbiota community function. However, in its current form, PICRUSt is unable to infer sampled metagenomes directly from denoised 16S amplicon sequence variants (ASV) without subjecting ASVs to potential information

loss through additional clustering and filtering into pre-defined Operational Taxonomic Units (OTU) with pre-calculated genome annotations. A potential solution available within the PICRUSt suite is to calculate de novo genome annotations for the ASVs themselves on a per-study basis. This novel pipeline may ultimately leverage the greater accuracy of ASVs to improve PICRUSt metagenome prediction results. We hypothesized that this pipeline would improve PICRUSt metagenome prediction accuracy over the original, published PICRUSt workflow. Three public Illumina-sequenced projects with paired microbial 16S and shotgun metagenomic (SMG) sequence data were accessed: 1) the Human Microbiome Project (n = 50, human oral, stool, skin, and airway), 2) Jovel et. al. 2016 (n = 19, mock community and human stool), and 3) Gilbert et. al. 2015 (n = 12, plant root). Kyoto Encyclopedia of Genes and Genomes (KEGG) gene tables for SMG samples were generated via HUMAnN2. Matching 16S amplicon library sequences were denoised and clustered into ASVs using DADA2. ASV genome annotations were calculated using the PICRUSt genome prediction and ancestral state reconstruction (ASR) tutorials. For comparison, ASVs were best-kmer-matched (BKM) to reference sequences with the original PICRUSt pre-calculated genome annotations using the DADA2 implementation of the RDP classifier. The original PICRUSt pipeline was also performed on OTU-clustered 16S sequences. Metagenome prediction accuracy was assessed by Spearman rho correlation between the resulting paired 16S and SMG KEGG gene frequency tables. We found that the original PICRUSt metagenome predictions were more closely correlated with the SMG-measured gene frequencies than the ASR and BKM predictions (p < 0.05 in all datasets) although the effect size was very low (0.01-0.02 mean Spearman rho paired difference). A high (>0.7) mean rho correlation was observed individually for all three methods in all three datasets. We conclude that the ASR and BKM pipelines failed to specifically improve metagenome predictions over the original PICRUSt pipeline. However, both ASR and BKM are promising and potentially valid approaches to de novo sequence PICRUSt analysis given their comparable accuracy to the original PICRUSt pipeline.

PA-30   *Mapping the germline and somatic mutation landscape in indolent and aggressive Prostate Cancer*
Tarun karthik kumar Mamidi
LSUHSC New Orleans

Prostate cancer (PCa) is the second most common cancer in men and the fourth most common tumor type worldwide. Many PCas are indolent and do not result in cancer mortality, even without treatment. However, a significant proportion of patients with PCa have aggressive tumors that progress rapidly to metastatic disease and are often lethal. Despite remarkable progress in diagnosis and patient care in recent years, the identification of genetic markers that distinguish indolent from aggressive PCa remains a major challenge. Recent advances in high throughput sequencing technologies are rapidly expanding our understanding of PCa biology by providing comprehensive descriptions of somatic mutations in tumors. However, large volumes of somatic mutation information has not been leveraged and integrated with germline mutations for genetic markers predictive of aggressive tumors. The objective of this study is three-fold: (i) To identify and functionally characterize genes containing germline and somatic mutations distinguishing indolent and aggressive PCa, (ii) To discover and characterize the molecular networks and biological pathways influenced by germline and somatic mutations in each disease

subtype, and (iii) To model pathway crosstalk between germline mutation controlled pathways and the biological pathways regulated by somatic mutations. We combine germline mutations from GWAS and somatic mutation from TCGA using gene expression data from TCGA as the organizing principle, and leverage this approach with network and pathway analysis. Preliminary results of these analyses revealed gene signatures of functional genes containing germline and somatic mutations distinguishing indolent from aggressive tumors. Network and pathway analysis revealed molecular networks and biological pathways enriched for both germline and somatic mutations. The study reveals that joint analysis show how germline and somatic mutations are likely to cooperate in driving aggressive PCa.

PA-31   ***Preliminary Transcriptome Analysis of Flavopiridol Treatment in an Mouse Ocular Alkali Burn Model of Neovascularization***

Harris McFerrin
Xavier University of Louisiana

Herpes simplex virus type 1 (HSV-1) infection and replication induce inflammation and neovascularization in the cornea leading to corneal blindness. HSV infection is the leading cause of viral blindness in the U.S., with nearly 20,000 new cases reported annually. We have demonstrated that 100nM flavopiridol (FP) reduces the steps in vascularization: endothelial cell (EC) migration, invasion and tubule formation in vitro. FP also reduced HSV-1 replication in vitro and the clinical pathology and corneal neovascularization in a mouse ocular model of HSV-1 infection. To determine whether FP's actions in vivo is also due to its effects on cellular targets, we induced neovascularization in mouse eyes with 0.1N NaOH in the presence and absence of 0.01% FP. Mouse eyes were scored for neovascularization in a masked fashion, and neovascularization was significantly inhibited by FP. To determine pathways that were inhibited, RNA was extracted from the cornea and subjected to Next Generation RNA Sequencing Analysis. The combined inhibition of both cellular and viral mechanisms of HSV-1-induced neovascularization and stromal keratitis is a potential addition to the limited arsenal available for treating the new cases of HSV-1-associated corneal neovascularization diagnosed each year in the U.S. alone.

PA-32   ***A Machine Learning Approach for Disulfide Bond Prediction***

Avdesh Mishra
University of New Orleans

Disulfide bonds are covalent bonds formed during post-translational modification by the oxidation of a pair of cysteines. These bonds between cysteines are one of the major forces responsible for stabilizing protein conformations, and therefore plays an important role in ab initio protein structure prediction (aiPSP) and protein folding. Improved prediction of disulfide bonds can help improve the accuracy of aiPSP, since they impose geometrical constraints on the protein backbone which greatly reduces the search space. In this study, we developed a machine learning based method, for disulfide bond prediction using support vector machines. For an effective training, various useful features are extracted which includes: conservation profile, solvent accessibility, torsion angle flexibility, disorder probability, sequential distance between

cysteines and more. The process of disulfide bonds prediction is carried out in two stages: first, individual cysteines are predicted as either bonding or non-bonding; second, the cysteine-pairs are predicted as either bonding or non-bonding by including the results from individual cysteine bonding prediction as a feature. During both the stages, a widely used technique, called feature windowing, is applied to include the neighboring residue features. The comparison based on only the features employed shows that our method achieves jackknife validation accuracy of 83.88%, which is 5.68% better than the existing predictor, which is based on nearest neighbor algorithm. Moreover, for individual cysteine bonding and cysteine-pair bonding prediction, our predictor achieves a 10-fold cross-validation accuracy of 83.01% and 90.07%, respectively. Altogether, our method achieves an overall 13.48% improvement in comparison to the state-of-the-art method. Thus, our method can be utilized to annotate the sequences whose structure are unknown, which can further aid in experimental studies of the disulfide bond and structure determination.

PA-33   *eModel-BDB: A Database of Comparative Structure Models of Drug-Target Interactions*
Misagh Naderi
Louisiana State University-Baton Rouge

Despite significant advances in protein structure determination, it is highly unlikely that the experimental structures of all known protein sequences (>108 in NCBI RefSeq at present) will become available in the near future. Similarly, for 1.3x106 drug-binding data in the Binding Database, only 2,291 ligand-protein crystal structures with affinity measurements are currently available in the Protein Data Bank. On that account, there is a clear need for high-throughput computational methods to complement existing repositories by constructing the atomic-level models of pharmacologically relevant drug-protein complexes. Here, we describe eModel-BDB, a database of 200,008 high-quality comparative models of drug-bound proteins built on interaction data obtained from the Binding Database. These complex models were generated with state-of-the-art modeling techniques under a rigorous quality control. We also conducted an independent, retrospective validation against recently released experimental structures, demonstrating that eModel-BDB indeed contains high-quality structural data. eModel-BDB has a significant re-use potential in structure-based drug discovery and repositioning, drug target identification, and protein structure determination.

PA-34   *Prediction of Hierarchical Classification of Transposable Elements using Machine Learning Approach*
Manisha Panta
University of New Orleans

Transposable Elements (TEs) or jumping genes are the DNA sequences that have intrinsic capability to move within a host genome from one genomic location to another - genomic location can be either same or different chromosome. The study shows that TEs have role in genome function and evolution as their presence can modify the functionality of genes and increase size of genome. Thus, proper classification of the identified jumping genes in a genome is important to understand their particular role in germline and somatic evolution. The classification of TEs is majorly based on the mode of transposition, number and type of genes

they contain and similarities in sequence. In this work, we studied multiple novel sequence-derived features such as k-mers frequency, nucleotide composition and conservation profile to generate a hierarchical datasets. We proposed a machine learning based method to predict hierarchical classification of transposable elements using support vector machines (SVMs). We compared the proposed method with the existing methods based on the measures specific to hierarchical problems which includes whether the classification can stop at the internal node of the hierarchy or must continue until a leaf node is reached. The comparative results indicate that the proposed method significantly outperforms the state-of-the-art methods.

PA-35   *StackDPPred: Stacking based Prediction of DNA-binding Proteins from Sequences*
Pujan Pokhrel
University of New Orleans

Identification of DNA-binding proteins from only sequence information, is one of the most challenging problems in the field of genome annotation. DNA-binding proteins play an important role in various biological processes such as DNA replication, repair, transcription and splicing. Existing experimental techniques for identification of DNA-binding proteins are time-consuming and expensive. Thus, prediction of DNA-binding proteins from sequences alone using computational methods can be useful to quickly annotate and guide the experimental process. Most of the methods developed for predicting DNA-binding proteins use the information from the evolutionary profile, called the position-specific scoring matrix (PSSM) profile, alone and the accuracy of such methods has been limited. Here, we propose a method, called StackDPPred, which utilizes features extracted from PSSM and residue specific contact-energy to help teach a stacking based machine learning method for the prediction of DNA-binding proteins. In a standard benchmark test of 1063 (518 DNA-binding and 545 non DNA-binding) proteins using jackknife validation, StackDPPred achieved an ACC of 89.96%, MCC of 79.90% and AUC of 94.50%, which outperforms the state-of-the-art approaches. Furthermore, when tested on a recently designed independent dataset PDB186, StackDPPred outperformed existing approaches by achieving an ACC of 86.56%, MCC of 0.7364 and AUC of 88.78%. Therefore, the proposed StackDPPred can be used for effective prediction of DNA-binding proteins from sequence alone.

PA-36   *eToxPred: Fragment-Based Machine Learning Prediction of Chemical Toxicity*
Limeng Pu
Louisiana State University-Baton Rouge

The efficiency of drug development defined as a successful approval of new pharmaceuticals in the rate of financial investments has significantly declined. Nonetheless, recent advances in high-throughput experimental techniques and computational modeling promise reductions in the costs and development times required to bring new drugs to market. One of the increasingly important component of modern drug discovery is the prediction of toxicity of drug candidates. In this work, we describe eToxPred, a new approach to reliably estimate the toxicity and synthetic accessibility of small organic compounds. eToxPred employs machine learning algorithms trained on molecular fingerprints to evaluate drug candidates. The performance is assessed against multiple datasets containing known drugs, potentially hazardous chemicals, natural products,

and synthetic bioactive compounds. Encouragingly, eToxPred predicts the synthetic accessibility with the mean square error of only 4% and the toxicity with the accuracy of as high as 72%. eToxPred is a valuable tool that can be employed at the outset of drug discovery. It can be incorporated into protocols to construct custom libraries for virtual screening in order to filter out those drug candidates that are potentially toxic or would be difficult to synthesize. eToxPred is freely available as a stand-alone software at https://github.com/pulimeng /etoxpred.

PA-37   *Precision Medicine Guided by Next Generation Sequencing is Now Provided On-Site at FWCC*
Phoebe Rollyson
LSUHSC Shreveport

Feist-Weiller Cancer Center now offers precision therapy guided by next generation sequencing. We are performing on-site solid tumor sequencing using CancerPlex, a 435-gene panel developed and analyzed by our Cambridge, MA based partner, KEW. Our service consists of DNA extraction from tumor samples, library preparation, and DNA sequencing. Sequencing data is fed directly into KEW's bioinformatics pipeline for: sequence alignment; variant calling and filtering; sample QC; and variant classification, prioritization, and interpretation. Our report identifies clinically actionable variants unique to each patient's tumor; providing information about current FDA-approved therapies for the patient's disease and for other cancers, along with any associated resistance to therapy, and therapies available in clinical trial for each actionable variant. CancerPlex also screens for microsatellite instability, tumor burden, and viral integration.
      In a test comparison, we ran our panel on patient samples that had previously been sequenced by another well established sequencing laboratory. DNA was extracted from six FFPE tissue samples, by coring and by microdissection of each tissue block. Sequencing libraries were prepared from all six cored DNA extractions and two of the corresponding microdissected DNA extractions. Corresponding samples sequenced with ~99% concordance. Five of the six samples reported significant findings that were not previously reported and we were able to successfully sequence one sample that was reported 'QNS' (Quality Not Sufficient for testing) by the other lab. In addition, VUSs (Variants of Uncertain Significance) which are highlighted in the previous report are listed for informational purposes only at the end of our report in order to streamline the report, allowing more focused attention on actionable variants.

PA-38   *Reduced Epstein-Barr virus replication in human papillomavirus immortalized keratinocyte organotypic raft culture Guidry JT, Songock WK, Ma X, Nathan CO, Bodily JM, and Scott RS*
Rona Scott
LSUHSC Shreveport

An epidemic rise in oropharyngeal squamous cell carcinoma (OSCC) infected with human papillomavirus has been observed over the past decade. The majority of HPV+OSCC arise from the tonsils and base of tongue (BOT), which are lymphoid-rich regions that are known to harbor Epstein-Barr virus (EBV). We previously reported HPV/EBV co-infection in a subset of tonsillar and BOT tumors. As EBV and HPV replicate in differentiated epithelia, viral encounters may interfere with the outcome of infection. To examine the effect of HPV-immortalization on EBV replication, we used organotypic raft culture to produce stratified epithelia in vitro. HPV-immortalized

tonsillar keratinocytes were EBV-infected by co-culture with EBV-positive Akata Burkitt's lymphoma cells four days after differentiation induction. Raft tissues were harvested 6 days post-EBV infection for analysis of viral life cycles. In HPV-negative tonsillar keratinocytes, robust EBV replication was observed that was blocked by treatment with acyclovir, a nucleoside analog that inhibits EBV lytic replication. In contrast, a dramatic decrease in EBV DNA levels was observed in HPV-positive keratinocytes, which was not further reduced by acyclovir. EBV infectivity was similar between HPV-negative and HPV-positive rafts. Human telomerase-immortalized normal oral keratinocytes supported robust EBV replication in organotypic raft culture suggesting a specific effect of HPV rather than immortalization on EBV replication. Expression of E7 alone was sufficient to decrease EBV DNA levels. A reduction in early lytic gene expression was consistent with reduced EBV replication in HPV-positive rafts. An increase in EBER transcripts, associated with EBV latency, was noted. Thus, HPV-immortalization blocks EBV replication shifting EBV towards a persistent latent infection. Latent EBV would allow expression of its viral oncogenes, which may synergize with HPV towards the rapid development and progression observed in HPV+OSCC.

PA-39   *EGFR heterodimerization, inhibition, dynamics and cell signaling*
Sitanshu Singh
University of Louisiana at Monroe

Proteins interact with one another in an obligatory fashion maintaining a stable interaction for a long period, or interactions that are transient. These interactions control many biochemical pathways. Detailed knowledge about the structure of the interaction surface of proteins and its energetics is necessary to understand the regulatory mechanism of the biochemical pathway with the ultimate goal of modulating or blocking the biochemical pathways for therapeutic purposes. As a model of PPI, homo- and heterodimers of EGFR family of receptors will be discussed. Dimerization and signaling by EGFR, HER2, and HER3 have important implications for different types of cancer. We have designed peptidomimetics and cyclic peptides from plant protein template to inhibit EGFR dimerization. These peptidomimetics have shown nanomolar range antiproliferative activity in HER2 positive cancer cells. To model the heterodimerization of EGFR, we have used crystal structure of extracellular domain (ECD) of HER2, EGFR and HER3 and transmembrane TM structure of the same proteins. Using the model of ECD and TM we have built the heterodimers of EGFR: HER2 and HER2:HER3. Molecular dynamics simulations were carried out on these heterodimers using NAMD software on QB2 via LONI. Dynamics of the heterodimers revealed that the domain IV of ECD of EGFR, HER2, and HER3 undergo breathing motion during the dynamics. This type of breathing motion may be responsible for sending the signal from outside of cell to inside the cell. This research work was supported by funds from NCI/NIH under grant number 1R15CA188225-01A1 and also by the National Institute of General Medical Sciences of NIH under grant number 8P20GM103424.

PA-40   *In-silico Hemispherical logP Analysis to Identify Potential Anti-Cancer Drug Hits Within a Small Library of Flavonoid Compounds*
Christopher Stratton
Louisiana State University-Shreveport

In recent years, there has been a shift in the drug discovery paradigm from a strict in-vitro process toward an in-silico based process. The in-silico process allows the researcher to perform in depth testing of a compound, or a library of compound analogs, such that potential optimal lead drug(s), from a library to test compounds, can be quickly identified. This provides a means to determine which compounds to apply more funding and time towards developing. Many lead drug failures are pharmacokinetic in nature or fall into the realm of one or more of the following ADMET areas: absorption, distribution, metabolism, excretion, or toxicity. We utilized ADMET Predictor, a software package produced by Simulations Plus, Inc., to perform this analysis. The proliferation studies were carried out using the PANC1 cell line. This study reviews a library of flavonoid compounds, determining the ones reflecting the best bioavailability, lowest ADMET risk, and greatest impact on cancer cell proliferation. Flavonoid compounds include a family of molecules that are polycyclic in structure and have a skeleton consisting of a chromanone substituted with a benzene ring. These molecules have been studied extensively and found to contain therapeutic properties such as anti-cancer and anti-oxidant effects.

PA-41   ***ALL ATOM MOLECULAR DYNAMICS SIMULATIONS OF NUCLEOSOME POSITIONING***
Ran Sun
Louisiana Tech University

Nucleosomes are the building blocks of eukaryotic genomes. Given that a nucleosome contains 147 base pairs of DNA, there are over 4^{147} possible sequences. It is impossible to carry out an exhaustive study. Today's supercomputing resources support simulations of ensembles representing 10's to 100's of nucleosomes. Here we demonstrate a workflow that enables us to characterize 21 different positions of the super strong nucleosome positioning sequence, labeled 601. The positions include 10 upstream positions, 10 downstream positions, and the ideal position. Tools are presented for generating all atom models of the nucleosome by docking an arbitrary sequence of DNA onto the histone core of 1KX5. The workflow manages the simulation, analysis and publish of results. In all 21 simulations, the superhelix geometry evolves in less than 100ns of simulation to a conformation that is significantly different from the original geometry of 1KX5. Statistical analysis, principal component analysis, and Fourier Filtering are employed to further investigate changes in conformation and dynamics of DNA associated with positioning and mispositioning of 601. All simulation and analysis data is published via our TMB-iBIOMES server: http://dna.engr.latech.edu/ibiomes.html

PA-42   ***Cervical and anal bacterial communities in patients with high-risk and low-risk HPV infections***
Frances Vazquez
Other

The human cervix and anus are physical interfaces between the host and the environment. These micro niches share a susceptible transformation zone, characterized by a metaplastic epithelial site, underlying its proneness to HPV carcinogenesis. We hypothesized that bacterial communities differ in the anus and cervix, and may play a role in the susceptibility to HPV. To relate the microbiota to HPV infections, we characterized resident cervical and anal bacteria by

using Illumina sequencing of the 16S ribosomal RNA V4 region in population-based sample of 400 women form the San Juan Metropolitan area. Bacterial profiles were generated using a high-resolution bioinformatics methodology (Resphera Insight v2.2) for species-level taxonomic assignment.   Preliminary analyses showed that anal bacterial richness is significantly higher than that of the cervix, regardless of HPV risk. In fact, we found no differences in richness or diversity related to HPV in any of the body sites. Anal communities were enriched with Finegoldia magna and Bacteroides vulgatus while cervical bacterial communities were dominated by Lactobacillus iners, Lactobacillus spp. and Gardnerella vaginalis. L. iners decreased from a 34% relative abundance in HPV negative patients to as low as 4% in patients with HPV infections. Lactobacillus_crispatus was identified as significantly associated with High-risk HPV infection in the cervix. In the anus, Gardnerella vaginalis was enriched in Low-Risk HPV infections. Obese and severely obese women had an enrichment of Howardella ureilytica in the cervix and Prevotella disiens in the anus. Our data suggests that the bacterial biota does not change significantly in structure nor in composition with HPV infections likely due to the functional redundancy and resilience of the bacterial communities. Extension to this knowledge, including fungal community diversity in the two body sites, will likely help clarify host-microbiome relationships during HPV infections.

PA-43   ***Integrating Germline and Somatic Mutation Information for the Discovery of Biomarkers in Triple Negative Breast Cancer***

Jiande Wu
LSUHSC New Orleans

Breast cancer is the most diagnosed cancer among women in the US and the second leading cause of cancer-related in women. One of the more significant challenges is understanding the etiology of triple-negative breast cancer (TNBC) the most aggressive form of breast cancer. We have integrated genome-wide association studies (GWAS) information with gene expression data to infer the causal association between gene expression and the disease and to establish putative functional bridges between GWAS discoveries and the biological pathways. However, somatic mutation information from The Cancer Genome Atlas (TCGA) has not been maximally leveraged and integrated with germline mutation from GWAS discoveries to understand the broader biological context in which germline and somatic mutations operate in TNBC. The objective of this study is: (1) To identify genes containing germline and somatic mutations involved in TNBC and determine whether they are functionally related, (2) To discovery and characterize the molecular networks and biological pathways that are influenced by germline and somatic mutations, and (3) To investigate whether germline mutations controlled pathways intersect with biological pathways regulated by somatic mutations. Our working hypothesis is that genes containing germline and somatic mutations involved in TNBC are functionally related and interact with one another in molecular networks and biological pathways. To address this hypothesis, we use germline mutations from GWAS, somatic mutation from TCGA and gene expression data on TNBC from TCGA as the organizing principle, and leverage this approach with network and pathway analysis. This joint analysis of both the germline and TNBC tumor genomes should help determine whether and the extent to which pathways involved in TNBC risk and those driving tumor initiation, progression, and response to therapy or prognosis intersect in TNBC.

Poster Abstracts

PA-44   *Driver Gene Mutations Based Clustering of Tumors*
Kun Zhang
Xavier University of Louisiana

Motivation: Somatic mutations in proto-oncogenes and tumor suppressor genes constitute a major category of causal genetic abnormalities in tumor cells. The mutation spectra of thousands of tumors have been generated by The Cancer Genome Atlas (TCGA) and other whole genome (exome) sequencing projects. A promising approach to utilizing these resources for precision medicine is to identify genetic similarity-based subtypes within a cancer type and relate the pinpointed subtypes to the clinical outcomes and pathologic characteristics of patients.   Results: We propose two novel methods, ccpwModel and xGeneModel, for mutation-based clustering of tumors. In the former, binary variables indicating the status of cancer driver genes in tumors and the genes' involvement in the core cancer pathways are treated as the features in the clustering process. In the latter, the functional similarities of assumed cancer driver genes and their confidence scores as the 'true' driver genes are integrated with the mutation spectra to calculate the genetic distances between tumors. We apply both methods to TCGA data of 16 cancer types. Promising results are obtained when these methods are compared to state-of-the-art approaches as to the associations between the determined tumor clusters and patient race (or survival time). We further extend the analysis to detect mutation-characterized transcriptomic prognostic signatures, which are directly relevant to the etiology of carcinogenesis.

Oral Presentation Index by Author

Oral Presentation Index by Author

Poster Abstracts Index by Author

Poster Abstracts Index by Author

Poster Abstracts Index by Author

Poster Abstracts Index by Author

Organizing Committee

Nayong Kim
Louisiana State University-Baton Rouge

Gus Kousoulas
Louisiana State University-Baton Rouge

Karen Jones
Louisiana State University-Baton Rouge

John Quebedeaux
Louisiana State University-Baton Rouge

Ram Ramanujam
Louisiana State University-Baton Rouge

Ramesh Subramanian
Louisiana State University-Baton Rouge

Christopher Taylor
LSUHSC New Orleans

Alexis White
Louisiana State University-Baton Rouge

Sponsors

Louisiana Biomedical Research Network
LSU – Center for Computation and Technology
LSU – Office Research and Economic Development
Center for Experimental Infectious Disease Research