NSF Sponsored Student Research Forum

Location: Salons 1-2

Co-Chairs

Professor May D. Wang

Georgia Tech and Emory University

Professor Zhaohui (Steve) Qin

Emory University

Professor Anna Ritz

Reed College

Student Co-Chair

Ying Sha

Georgia Tech

Schedule

4:00–4:05 pm	**NSF Travel Fund Awardee Forum Chairs and NSF Program Directors**	
4:05–4:10 pm	**Mansooreh Ahmadian**	Hybrid ODE/SSA Model of the Budding Yeast Cell Cycle Control Mechanism with Mutant Case Study
4:10–4:15 pm	**Sara A. Amin**	ProSol DB: A Protein Solubility Database
4:15–4:20 pm	**Huiyuan Chen**	A Flexible and Robust Multi-Source Learning Algorithm for Drug Repositioning
4:20–4:25 pm	**Arman Cohan**	Identifying Harm Events in Clinical Care through Medical Narratives
4:25–4:30 pm	**Chelsea Jui-Ting Ju**	Fleximer: Accurate Quantification of RNA-Seq via Variable-Length k-mers
4:30–4:35 pm	**Yi-Pin Lai**	A Compatibility Approach to Identify Recombination Breakpoints in Bacterial and Viral Genomes
4:35–4:40 pm	**Junfeng Liu**	Differential Compound Prioritization via Bi-Directional Selectivity Push with Power
4:40–4:45 pm	**Sijia Liu**	Dependency Embeddings and AMR Embeddings for Drug-Drug Interaction Extraction from Biomedical Texts
4:45–4:50 pm	**Laraib Malik**	Rich Chromatin Structure Prediction from Hi-C Data
4:50–4:55 pm	**Kasra Manavi**	Gaussian Mixture Models with Constrained Flexibility for Fitting Tomographic Tilt Series
4:55–5:00 pm	**Stephanie Mason**	Exploring Protein Cavities through Rigidity Analysis
5:00–5:05 pm	**Ahmed Metwally**	Detection of Differential Abundance Intervals in Longitudinal Metagenomic Data Using Negative Binomial Smoothing Spline ANOVA
5:05–5:10 pm	**Pourya Naderi Yeganeh**	Exploiting the structure of the interaction network in pathway enrichment analysis
5:10–5:15 pm	**Richard Platania**	Large-scale Deep Learning with Biomedical Data
5:15–5:20 pm	**Elham Rastegari**	A correlation Network Model Utilizing Gait Parameters for Evaluating Health Levels
5:20–5:25 pm	**Allison M. Rossetto**	Simple Voting with an Ensemble Convolution Neural Network for Lung Tumor Detection
5:25–5:30 pm	**Ying Sha**	Interpretable Predictions of Clinical Outcomes with An Attention-based Recurrent Neural Network
5:30–5:35 pm	**Zachary Stanfield**	Drug Response Prediction as a Link Prediction Problem
5:35–5:40 pm	**Hang Wu**	Learning Deep Representations for Causal Inference
5:40–5:45 pm	**Zheng Xu**	Seq2seq Fingerprint: An Unsupervised Deep Molecular Embedding for Drug Discovery

Hybrid ODE/SSA Model of the Budding Yeast Cell Cycle Control Mechanism with Mutant Case Study

Mansooreh Ahmadian

Virginia Tech

Computational Biology Lab,

2160L Torgersen, Department of Computer Science, Virginia Tech Blacksburg, VA 24060

540-4495-120

amadian@vt.edu

A. Research Summary

Complex systems emerging from many biochemical applications often exhibit multiscale features and present great challenges in modeling and simulation. The goal of this project is to face these challenges by developing rigorous mathematical theories and innovative numerical algorithms for hybrid modeling methods. We will develop mathematical foundations for error analysis of hybrid methods applied to chemical systems, and algorithms for partitioning a biochemical system into subsystems in different scales. We will apply appropriate algorithms to simulate each subsystem, and synchronize and combine results from all subsystems while maintaining high efficiency. Our algorithm development is motivated by realistic modeling and simulation of a complex biological control system: the cycle of growth and division in yeast cells. Our goal is to have a detailed cell cycle model that reflects dynamics at gene and mRNA levels, accounts accurately for known probabilistic features of cell proliferation in yeast cells, and accurately predicts the aberrant behaviors of mutant strains. Algorithms, theories, and software of hybrid methods developed in this project will be applied and tested in the modeling and simulation of this complex cell cycle model.

Intellectual Merit. The multiscale challenge we aim at in this project is motivated by a realistic model of a central aspect of cell physiology. The novelty of this project lies in the systematic study of hybrid methods in discrete stochastic simulation for biochemical systems. The proposed work will advance the frontier of computational methods for the simulation of complex biochemical systems and enable automatic regime switching among multiple scales in complex biochemical systems.

Broad Impact. Computational biology integrates biological discovery with mathematical modeling and simulation. Most biological systems inherently have multiscale features. The theories and algorithms proposed in this project enable efficient simulation for those biological systems and will benefit the model development in computational biology applications. Moreover, the techniques about hybrid methods are also applicable to multiscale simulation of complex systems in other areas. This research project will provide multi-disciplinary training in computer science, mathematics, and biology. As part

of this project, the models and simulation methods will be introduced in graduate and undergraduate courses on computational cell biology. Models, algorithms, and software developed in this project will be made publically available through open source software packages such as JigCell, StochKit, and CoPaSi, as well as project websites.

Paper Title

Hybrid ODE/SSA Model of the Budding Yeast Cell Cycle Control Mechanism with Mutant Case Study

B. Biosketch

MANSOREH AHMADIAN
PhD Student
Department of Computer Science, Virginia Tech, Blacksburg, Virginia, USA.
amadian@vt:edu
https://sites.google.com/a/vt.edu/mansooreh-ahmadian/academics

Professional Preparation

Imam Khomeini International University, Iran, Electrical Engineering, **B.S**., 2008
Science and Research Branch Azad University, Iran, Electrical Engineering, **M.S**., 2012
Virginia Tech, USA, Computer Science, **Ph.D**., Spring 2016-present

Professional Appointments

Research Assistant, Department of Computer Science Spring, 2017 – Present

Products

Pertinent Products:

- Mansooreh Ahmadian, Shuo Wang, John J. Tyson, Yang Cao, "Hybrid ODE/SSA Model of the Budding Yeast Cell Cycle Control Mechanism with Mutant Case Study," Accepted, ACM-BCB 2017, Boston, MA, August 20-23, 2017.
- Shuo Wang, Mansooreh Ahmadian, Minghan Chen John J. Tyson, Yang Cao, "A Hybrid Stochastic Model of the Budding Yeast Cell Cycle Control Mechanism," Accepted, ACM-BCB 2016, Seattle, WA, October 2-5, 2016.

Additional Significant Products:

- M. Ahmadian, M. Lima, M. Behnam, H. Khaloozadeh "Compressibility Prediction of Reduced Water Atomized Iron Powder Using Adaptive Neuro-Fuzzy Model," In Proceeding of European Powder Metallurgy Association (EPMA), Basel, Switzerland, September 2012.
- M. Ahmadian, S. Yousefli, H. Khaloozadeh "Hardness prediction of heat treated iron parts manufactured by powder metallurgy using neural network and fuzzy rule-based models," In Proceeding of European Powder Metallurgy Association (EPMA), Basel, Switzerland, September 2012.

Synergistic Activities

_ Student Member of ACM,

_ Member of SIAM

ProSol DB: A Protein Solubility Database

Sara A. Amin

Tufts University
161 College Avenue, Medford, MA
936-520-1717
Sara.amin@tufts.edu

A. Project summary

Engineering non-native synthesis pathways in microbial hosts has shown promise in producing commercially useful molecules. The selection of highly soluble protein sequences to implement catalyzing reactions along synthesis pathways can be facilitated by predicting the solubility of protein sequences in the host. Current solubility predictors apply machine-learning algorithms, such as Support Vector Machines (SVM) and Neural Networks (NN), to predict solubility using protein sequence features such as hydrophilicity, net charge and α-helix.

In this project we build a database, referred to as Protein Solubility Database (*ProSol* DB) that allows for quick lookup of predicted solubility values using Enzyme Commission (EC) numbers. We used *ccSOL omics*, a tool bases on machine learning models, to compute solubility prediction scores for various proteins from *UniProKBt* in *E. coli*. *ProSol DB* serves as a source of identifying protein sequences with high predicted solubility scores eliminating the need of recurring calls to protein solubility predictors. Combining the *ProSol* DB with synthesis pathway tools can assist in avoiding

experimental efforts spent on expressing low solubility enzymes when more soluble alternatives are identified. This work promises to expedite the design-build-test cycle of metabolic engineering efforts.

Intellectual Merit. This research addresses protein solubility from a computational point of view to assist biologist in the decision making process when selecting which protein to use to catalyze reactions. The *ProSol DB* will give suggestions of high soluble proteins that is traditionally verified using experiments. Biologists usually retain to protein sequences they know worked or ones that their functionality can be enhanced with some techniques. However those techniques can consume time, and if more than one protein is need to synthesize a pathway the same technique to enhance a protein's functionality might not be applicable to the other. Therefore, *ProSol DB* provided predicted solubility score from which biologists are able to pick proteins that are soluble instead of enhancing the solubility of proteins with low solubility.

Broader Impacts. The *ProSol DB* will provide information and suggestions for novel protein sequences that were never considered in the literate as candidates to validate experimentally. Those novel proteins can lead to finding new synthesis pathways to produce industrial molecules of interest in fast and less expensive settings. Finally, the *ProSol DB* can be combines with pathway synthesis tools to expedite the design-build-test cycle of metabolic engineering efforts.

Project/Poster Title: ProSol DB: A Protein Solubility Database

Bio-sketch

I'm Sara A. Amin, a fourth year PhD student at Tufts University. I'm in the computer science department and my research is an interdisciplinary between Computer Science and Synthetic Biology. I worked on several projects where the main theme is to develop an automated methodology of exploring different engineering options within large design spaces under the metabolic engineering field. My projects focus on designing and building tools that integrate metabolic synthesis pathways and identifying specific genes designs that controls the functionality occurring within a cellular host. I believe the work is important in the field of synthetic biology since it bridges the gap between the experimental and computational fields, and eventually it will give biologist the

ability to make decisions that will save time, effort and money. Presenting my work at ACM-BCB will help in getting the exposure I need to my work. It will also help me connect with peers in my field through the young professional research forum leading to fruitful future research collaboration. In addition, I'll be connecting with possible future employers during ACM-BCB as I'll be graduating within the coming academic year.

A Flexible and Robust Multi-Source Learning Algorithm for Drug Repositioning

Huiyuan Chen

Case Western Reserve University
2477 Overlook RD, Apt 402, Cleveland, OH, 44106
216-543-0744
hxc501@case.edu

A. Project summary

Overview: A constant challenge in drug discovery is drug repositioning that exploits new therapeutic indications for already approved drugs. In recent years, publications of multiple bioinformatics databases, such as Drugbank, Uniprot and OMIM, enable exploration of new biomedical insights of drug-target-disease relationships in drug repositioning. One of the objectives of this proposal is to design a unified framework that integrates information across different databases to better characterize both drugs and diseases. Due to the nature of data collection, it is very unlikely that all databases will be available for all drugs/diseases. The second objective of the proposal is to address the data incompleteness problem. The proposed method is to infer information for one data source could by borrowing information from other sources.

Intellectual Merit: The Intellectual Merit of the proposal includes a multi-kernel

method in combination with a novel multi-view learning algorithm to solve the drug repositioning
problem, which is formulated as a link prediction across a two layers network: drug-layer and disease-layer. The proposal has two objectives:

- Objective 1: A multi-kernel learning algorithm is designed to simultaneously incorporate multiple kernel matrices representing respective drug/disease layer from different databases.
- Objective 2: A novel multi-view learning algorithm is proposed to impute missing values in one source by borrowing information from others.

Broad Impact: The novel framework will have broad benefits both in pharmaceutical industry and academia research. In industry, it provides a promising strategy to reduce the total cost because existing drugs have already been approved and safety. In academia research, it proposes a feasible way to easily incorporate multiple data sources and also imputes the missing value across different databases, which can be also applied to some research fields such as social networks, multi-network clustering and recommender system since those fields often involve with multiple data sources.

Paper Title Accepted by ACMBCB2017:

A Flexible and Robust Multi-Source Learning Algorithm for Drug Repositioning

B. Bio-Sketch

Huiyuan Chen is a third-year Ph.D. candidate in Electrical Engineering and computer science at Case Western Reserve University. His research advisor is Dr. Jing Li. His research interests cover bioinformatics, graph mining, machine learning, especially in Multi-View Multi-Source Learning, Sparse Learning and Local Clustering Detection. Before joining CWRU in Aug 2015, he received his master degree and Bachelor Degree from University of Pittsburgh and Beijing Jiaotong University, and both major in mechanical engineering. He is now focusing on the analysis of biological networks, systems biology of complex diseases, and designs an efficient algorithm for drug combination in large-scale.

Identifying Harm Events in Clinical Care through Medical Narratives

Arman Cohan

Information Retrieval Lab
Department of Computer Science
Georgetown University

3700 O St NW,
STM 313.14
Washington DC, 20057
202-509-3830

arman@ir.cs.georgetown.edu

A. Project Summary

Preventable medical errors have been shown to be a major cause of injury and death in the United States. Medical errors are estimated to be the 3rd leading causes of death in the U.S which translates to an estimated incidence of 210,000 to 400,000 deaths annually. To address these major concerns, healthcare systems have adopted reporting systems in clinical care to help track and trend hazards and errors in patient care. The data from these systems are later used to identify the causes of harm and actions that should be taken to prevent similar situations. These reporting systems allow frontline clinicians to report events that are relevant to patient care including both near misses and serious safety events. Serious safety events are situations where a patient was harmed. For example, an event where a patient was harmed by adverse drug reaction to a medication administered by a nurse is a serious safety event.

Although reporting systems have been implemented with the goal of improving patient safety and patient care, hospital staff are faced with many challenges in analyzing and understanding these reports. These reports which are narratives in natural language are generated by frontline staff and vary widely in content, structure, language used, and style. While these texts provide valuable information about the safety event, it is challenging to perform large scale analysis of these narratives to identify important safety events. In this project, we propose and evaluate Natural Language Processing (NLP) methods to identify cases that caused harm to the patient based on medical narratives and triage these safety events into different severity levels.

We present a deep learning architecture for identifying the severity of harm from narratives regarding incidents in patient care. While there is a growing number of work in categorizing patient safety reports, none has looked at the modeling of general harm across all event types. Our methods are based on a general neural network model consisting of several layers including a convolutional layer, a recurrent layer, and an attention mechanism to improve the performance of the recurrent layer. Our methods are designed to capture local significant features as well as the interactions and dependencies between the features in long sequences. Traditional methods in general and domain specific NLP rely heavily on engineering a set of representative features for the task and utilizing external knowledge and resources. While these models have been shown to work reasonably well for different tasks, their success relies on the type of features that they utilize. Apart from the feature engineering efforts, these approaches usually model the problem with respect to certain selected features and ignore other indicators and signals that might improve prediction. We show how are methods are able to significantly improve over the existing methods on two large scale datasets of patient reports.

The impact of the methods that we investigate in this project is substantial to patient care. More accurate methods in the identification of harm can help the data analysis and reporting process, prevent harm to patients, better prioritize resources to address safety incidents, and subsequently improve general patient care.

This methods and outcomes of this project is expected to be of interest to a broad community within ACM-BCB including those working in biomedical NLP, text mining, text categorization, and health informatics in general.

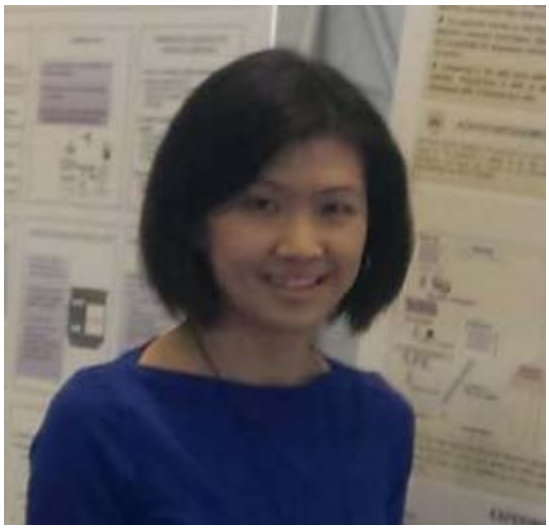
Paper Title

Identifying Harm Events in Clinical Care through Medical Narratives

Arman Cohan, Allan Fong, Raj Ratwani, and Nazli Goharian

B. Bio-sketch

Arman Cohan is a 4th-year Ph.D. student in computer science at Georgetown University, advised by Prof. Nazli Goharian. His broad research interests lie at the intersections of Natural Language Processing, Information Retrieval and Health Informatics. He has specifically worked on problems in domain specific document summarization, text mining and categorization, and NLP applications in the health domain.



Fleximer: Accurate Quantification of RNA-Seq via Variable-Length *k-mers*

Chelsea Jui-Ting Ju

UCLA

Department of Computer Science
3551 Boelter Hall Los Angeles, CA, 90095
508-310-3289
chelseaju@ucla.edu

A. Research Summary

Human genetics encompasses a wide range of studies; gene expression profiling is one of the studies that aims at interpreting the functional elements based on genetic information encoded in the genome. RNA-Sequencing (RNA-Seq) is the leading technology to quantify expression of thousands of genes simultaneously. A large number of computational challenges are introduced by this technology, specifically in handling a massive amount of read data and extracting biologically relevant knowledge from the data[1]. Inevitably, genomics has become one of the domains of Big Data science[2].

Intellectual Merit: The traditional data analysis of an experiment starts from aligning millions of RNA-Seq reads to the reference genome or transcriptome. This alignment step requires a substantial amount of computational resources and time[3]. Recent developments have moved to an alignment-free approach to alleviate the computational burden. Exisiting approaches utilize the notion of k-mers to infer

transcript abundances from read data, and employ only a fix size of k-mers. However, choosing the appropriate k can be challenging. We propose a method that can efficiently explore all possible kmers with variable lengths, and select a subset that can best describe the characteristics of different transcripts. Our method leverages the properties and structure of a suffix tree and the concept of splicing graph[4] for k-mers selection, and uses Aho-Corasick [5] and expectation-maximization (EM) algorithms for transcript abundance estimation.

Broad Impacts: Having an efficient and accurate approach is essential to facilitate a rapid sequencing analysis, which aids in the diagnosis and possible future treatments of diseases. The outcome of this project will advance the state-of-the-art of transcriptome quantification methods. Our approach incorporates valuable algorithmic techniques, which can be easily extended for tasks that involve processing vast sequencing data. These applications includes, but are not limited to bisulfite sequencing, ChIPseq, and miRNAseq. The participants of this project come from a diverse range of trainings, including machine learning, algorithm, data mining, and biology. I am with great pleasure to mentor both master and undergraduate students in this project and to motivate them with the research problems in bioinformatics. Since bioinformatics is an interdisciplinary research, it is always rewarding to serve as a mediator to bring students from different disciplines together. As a mentor of the B.I.G. summer program at UCLA, I plan on using the extended works from this project to inspire and encourage more undergraduates, who are underrepresented in STEM fields, to participate in bioinformatics research.

The paper or poster title/titles that is/are co-authored by the applicant and accepted by ACMBCB2017

One of my research projects, as listed below, has been accepted as a regular paper for oral presentation at ACM-BCB 2017. As the first author of this paper, I am very excited about the opportunity to present this work.

Authors: Chelsea J.-T. Ju, Ruirui Li, Zhengliang Wu, Jyun-Yu Jiang, Zhao Yang, and Wei Wang

Title: Fleximer: Accurate Quantification of RNA-Seq via Variable-Length k-mers

B. Biosketch

I am a PhD candidate in the Department of Computer Science at University of California, Los Angeles (UCLA), supervised by Professor Wei Wang. My primary research interests focus on developing and applying data mining and machine learning methods to address the computational challenges in genomic studies, specifically in gene expression quantification via next generation sequencing data. One of my research projects aims at improving the resolution of RNA-Seq read mapping between pseudogenes and their parent genes. The work has led to a conference presentation at ACM-BCB 2014, entitled "PseudoLasso: Leveraging Read Alignment in Homologous Regions to Correct Pseudogene Expression Estimates via RNAseq", and a journal publication at IEEE/ACM TCBB 2017, entitled "Efficient Approach to Correct Read Alignment for Pseudogene Abundance Estimates". Moving toward the alignment-free approach, I am dedicating my current research to explore and employ the notion of k-mers to develop more efficient and accurate approaches for analyzing sequencing data in different applications. Earlier, I received my Master of Science Degree in Computer Science from UCLA, and my Bachelor of Science Degree in Bioinformatics from University of Alberta, Canada. Prior to pursuing my postgraduate education, I worked on microarray data analysis to study the functional genomics and molecular physiology of forest tress in a botany lab, followed by industrial research at a start-up venture dedicated to producing clean nonfood based cellulosic biofuel.

A Compatibility Approach to Identify Recombination Breakpoints in Bacterial and Viral Genomes

Yi-Pin Lai

Department of Computer Science and Engineering, Texas A&M University
Interdisciplinary Life Sciences Building, College Station, 77844, TX, United States
979-985-7397
yplai.tw@tamu.edu

A. Research Summary

Project Summary: Recombination is an evolutionary force that results in mosaic genomes for microorganisms. The evolutionary history of microorganisms cannot be properly inferred if recombination has occurred among a set of taxa. That is, polymorphic sites of a multiple sequence alignment cannot be described by a single phylogenetic tree. Thus, detecting the presence of recombination is crucial before phylogeny inference. The phylogenetic-based methods are commonly utilized to explore recombination, however, the compatibility-based methods are more computationally efficient since the phylogeny construction is not required. We propose a novel approach focusing on the pairwise compatibility of polymorphic sites of given regions to characterize potential breakpoints in recombinant bacterial and viral genomes. The performance of average compatibility ratio (ACR) approach is evaluated on simulated alignments of different scenarios comparing with two programs, GARD and RDP4. Three empirical datasets of varying genome sizes with varying levels of homoplasy are also utilized for testing. The results demonstrate that our approach is able to detect the presence of recombination and identify the recombinant breakpoints

efficiently, which provides a better understanding of distinct phylogenies among mosaic sequences.

Intellectual Merit: The intellectual merit of the project is the average compatibility ratio approach, which extends the pairwise compatibility concept to a set of characters and indicates the tendency of compatibility within the regions efficiently. ACR utilizes recursive computation and does not require phylogeny construction. The higher ratio reflects that more characters are jointly compatible, suggesting more sites are congruent in a tree within the region. In contrast, the lower the ratio is, the less likely the recombination events happened in the region. A sliding window is used to average all of the compatibility ratios of regions within the given window size for each site. A site with a local minimum represents that its upstream region and downstream region within a given size are less jointly compatible comparing to other regions. So, sites with local minimums are the potential breakpoints, and consecutive sites between every two adjacent breakpoints represent a non-overlapping segment that is expected to be more congruent with a topology. As a result, a list of top potentially recombinant regions will be obtained. The ACR approach is tested on simulated datasets of two scenarios and three empirical datasets, and the performance is compared with GARD and RDP4. Comparing the results of simulation scenarios with ground truths, ACR approach performs better than other programs. The results of the analyses of bacterial and viral genomes demonstrate that our proposed approach is able to characterize biological sequence alignments and provide segments that reflect distinct phylogenetic histories.

Broader Impacts: The broader impacts of the project are the efficient detection of recombination breakpoints and elucidation of phylogenetic relationship among species for exploring further biological insights, including the relationship of genetic mutations and drug resistance, the connection between tree topology and homoplasy, and the application to detect introgression in plant or vertebrate genomes. Several bacteria and viruses have been reported that suggest recombination events occurred during evolution, so the global phylogenetic tree can be misleading if the recombinant sites are not considered. The accurate phylogeny enhances the understanding of relationship between genetic mutations and drug resistance, and hence provides better treatments for antimicrobial resistance. In addition, since the real biological datasets

are usually complex, characterizing the connections between homoplasy and size, length, and pattern of tree branch paves the way for probability model construction for unveiling more precise evolutionary history. Furthermore, the approach currently focuses on viruses and bacteria. For heterozygote individuals that are genotyped and tend to have more missing data due to coverage issues, the approach could be generalized to handle large amount of missing data for plant or vertebrate genomes.

Paper Title

Yi-Pin Lai and Thomas Ioerger. 2017. A Compatibility Approach to Identify Recombination Breakpoints in Bacterial and Viral Genomes. In Proceedings of ACM BCB conference, Boston, MA, USA, Aug 20-23 2017 (BCB'17), 10 pages. DOI: 10.1145/3107411.3107432

B. Biosketch

Yi-Pin Lai is a second-year Ph.D. candidate working with Dr. Thomas Ioerger in the Department of Computer Science and Engineering at Texas A&M University. Her research area is bioinformatics and computational biology, with interests in phylogeny, genetic mutations and antibiotic resistance, and metagenomics. She currently focuses on characterizing the recombination events in polymorphic sequence alignments of bacterial and viral genomes using an average compatibility ratio approach to better understand the distinct phylogenies among mosaic sequences. She also works on exploring the consistency of drug resistance polymorphisms among *Mycobacterium tuberculosis* isolates. She aims to apply computational and statistical methods in large-scale genomic data to reveal underlying biological mechanisms and address issues in biology.

Prior to her Ph.D. program, Yi-Pin developed an R/Bioconductor package (iGC) to identify differentially expressed genes driven by copy number alterations (CNAs). It is a tool that enables a concurrent analytics of both gene expression profiling and CNAs in the same individual to characterize CNA-driven genes. She also worked on a metagenomics project, comparing and identifying microbiome communities between Crohn's disease and ulcerative colitis tissue samples. Yi-Pin received her B.S. in electrical Engineering from National Tsing Hua University and her M.S. in Biomedical Electronics

and Bioinformatics from National Taiwan University advised by Dr. Eric Chuang in biological pathway analysis.

Publications:

[1] Lai, Y. P., Lu, T. P., Lee, C. Y., Lai, L. C., Tsai, M. H., and Chuang, E. Y. (2012). The MiRPathOgen: A systems biology tool for pathway analysis with gene and microRNA expressions, Human Genome Meeting (HGM), Sydney, Australia, Mar. 2012.

[2] Chiu, Y. C., Wu, C. T., Hsiao, T. H., Lai, Y. P., Hsiao, C., and Chen, Y. (2015). Comodulation analysis of gene regulation in breast cancer reveals complex interplay between ESR1 and ERBB2 genes. BMC Genomics, 16.

[3] Lai, Y. P., Wang, L. B., Wang, W. A., Lai, L. C., Tsai, M. H., Lu, T. P., and Chuang, E. Y. (2017). iGC--an integrated analysis package of gene expression and copy number alteration. BMC bioinformatics, 18(1), 35.

[4] Lai, Y. P. and Ioerger, T. R. (2017). A compatibility approach to identify recombination breakpoints in bacterial and viral genomes. In Proceedings of ACM BCB conference, Boston, MA, USA, Aug 20-23 2017 (BCB'17), 10 pages, accepted.

[5] Yadon, A.N., Maharaj, K., Adamson, J.H., Lai, Y. P., Sacchettini, J.C., Ioerger, T.R., Rubin, E.J., and Pym, A.S. (2017). Comprehensive characterization of pncA polymorphisms conferring resistance to pyrazinamide. Nature Communications, accepted.

Differential Compound Prioritization via Bi-Directional Selectivity Push with Power

Junfeng Liu

Indiana University - Purdue University Indianapolis
410 W 10th St, HITS 5000, Indianapolis, IN 46202
(317) 666-0420

liujunf@iupui.edu

A. Research Summary

Overview Computational compound prioritization methods are developed to provide ranking structures of compounds in order to facilitate the process of identifying drug candidates. Bioassay data is one of the primary sources that are used in early stages of drug discovery to reveal compound bioactivities and identify promising drug candidates. Besides compound bioactivities, bioassay data contains valuable information across bioassays. The hidden knowledge in the bioassay space has not been fully discovered by existing methods but may be extremely useful to improve drug development. In the proposed research, we develop a machine learning based computational framework for better compound prioritization that focuses on the local

ranking structure within a bioassay, and meanwhile, discover the global ranking structure by leveraging the ranking relations across multiple bioassays. Specifically, the compound prioritization problem is solved based on activity and selectivity simultaneously by leveraging ranking difference across bioassays. The proposed research will explore new computational methods to discover, analyze and leverage structures and relations among bioassay data for better compound prioritization or other usages in computational biology.

Intellectual Merit The proposed research is innovative, both in terms of employing original computational methods into concerned issues in drug discovery, and in terms of developing unique computational methodologies for core Computer Science research. For drug discovery, the research will provide novel perspectives and methodologies as to how we can utilize the large-scale data to solve important problems in drug discovery. These methodologies adopt highly data-driven approaches and maximally consider all available data in a systematic manner. Thus, they go beyond the conventional methods that are highly dependent on domain knowledge but largely ignore the signals carried by the data themselves. Therefore, the proposed project has a high potential to explore new knowledge and new methods that could be significant to drug discovery. For Computer Science research, the proposed research will contribute new solution framework and methods the areas of data mining and machine learning. Specifically, the proposed project will lead to novel methods for boosting ranking performance by incorporating relevant information in an optimization framework, deploying iterative procedures and greedy strategies for large-scale problems with multiple simultaneous tasks, etc. These methods are fundamental and generalizable to various Computer Science applications.

Broader Impacts The proposed research has a wide range of impacts, from scientific to societal. The proposed research is aimed at providing novel computational tools that can effectively discover the unrevealed information and novel knowledge among bioassays in order to improve compound prioritization. Thus, it will have scientific impacts on fully revealing the existing knowledge of Big Data, stimulating information extraction and developing novel analytical techniques. Additionally, the proposed research also tries to facilitate drug candidate identification and expedite drug

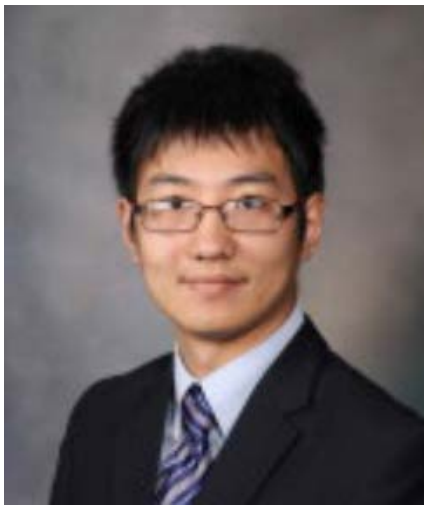
discovery with data-driven methods through better compound prioritization. It will also bring significant impacts on the economy and the society.

Paper Title

Differential Compound Prioritization via Bi-Directional Selectivity Push with Power

B. Biosketch

Junfeng Liu is a first-year graduate student from the Department of Computer and Information Science, Indiana University - Purdue University Indianapolis (IUPUI). He is also a research assistant in Prof. Xia Ning's lab, and a University Fellowship Award recipient of 2016-2017 academic year at IUPUI. Junfeng's research focus is on Machine Learning, Data Mining and Big Data Analytics, with applications in drug discovery and bioinformatics. Particularly, Junfeng is interested in discovering novel knowledge via leveraging information from multiple bioassays and improving drug development using Structure-Activity-Relationship (SAR) Modeling and Structure-Selectivity-Relationship (SSR) Modeling for compound prioritization. In his first publication, compound activity ranking models are developed by leveraging multiple bioassays. In these methods, assistance bioassays and assistance compounds are identified and incorporated intelligently to build models that can accurately prioritize active compounds in a bioassay. In his latest publication, Junfeng developed a compound activity ranking model which also favors the higher rankings of selective compounds. In the ranking model, compounds are ranked well based on their activities, and meanwhile, selective-related compounds are preferably pushed higher or lower in the ranking list with a bi-directional push strategy. The bi-directional push strategy is enhanced by the push-powers that considers the ranking structure across multiple bioassays. In these projects, data-driven techniques are extensively adopted to discover information from large-scale bioassay data.



Dependency Embeddings and AMR Embeddings for Drug-Drug Interaction Extraction from Biomedical Texts

Sijia Liu

Mayo Clinic / University at Buffalo, SUNY
200 First St. SW Rochester, MN 55905
liu.sijia@mayo.edu

A. Research Summary

The boost in the capacity and volume of biomedical texts has created a tremendous opportunity for computer science research to be applied to improve clinical practice efficiency. It is widely acknowledged that relation extraction of unstructured textual contents using Natural Language Processing (NLP) and text mining techniques are essential for using biomedical data for secondary purposes.

Abstract Meaning Representation (AMR) is a semantic formalism that expresses the logical meanings of English sentences in the form of a directed, acyclic graph. AMR aims to abstract away from syntactic idiosyncrasies and attempts to capture only the core meaning of a sentence. Thus, AMR-based embeddings could encode more information from the sentence, such as semantic roles, named entities, and coreference information, which is important for DDI detection.

In this study, we utilize dependency embeddings and novel AMR embeddings as features for DDI detection. We evaluate these embeddings on DDIExtraction 2013 challenge corpus with two different classification methods: SVM and Random Forest.

We compare the performance of using different embeddings in a baseline method. The experimental results show that both dependency and AMR embeddings are effective for detecting DDIs. The best performance was obtained by using word, dependency and AMR embeddings in the baseline method.

Intellectual Merit

The proposed research demonstrates that combining representation learning methods like word embedding and context embedding can effectively leverage semantic information to downstream applications like supervised machine learning methods. This technique provides new and effective method for sentence representation for semantic abstraction and summary, particularly in biomedical relation detection.

Broad Impact

The proposed research can be applied to medication clinical decision support systems. It has great potential to avoid drug-related medical errors by providing relevant DDI information to healthcare providers when prescribing medication. Besides, the proposed context embedding model can be applied to various relation extraction and NLP tasks.

Paper Title

"Dependency Embeddings and AMR Embeddings for Drug-Drug Interaction Extraction from Biomedical Texts"

B. Biosketch

Sijia Liu is a PhD candidate from Department of Computer Science and Engineering, University at Buffalo, SUNY working with Dr. Vipin Chaudhary. His proposed doctoral dissertation is titled "Unsupervised Relation Extraction from Biomedical Texts". He also works as an Informatics Specialist in Division of Biomedical Statistics and Informatics, Mayo Clinic under the supervision of Dr. Hongfang Liu, contributing both in medical informatics research and clinical Natural Language Processing (NLP) application development. He has broad interests in biomedical NLP and machine learning. Since 2016, he has authored/co-authored 11 publications in mainstream medical informatics journal and conferences such as Journal of the American Society for Information Science and Technology (JASIST), AMIA Annual Symposium, ACM-BCB, World Congress

on Health and Medical Informatics (MedInfo), AMIA Joint Summit of Clinical Research Informatics (AMIA CRI), and International Workshop on Semantic Evaluation (SemEval). He has won IEEE ICHI 2016 Student Travel Award, and was nominated as Student Paper Award of AMIA 2016 Summit on Clinical Research Informatics. He is the winner of the participated subtask (Task 10 Scenario B) in SemEval 2017.

Rich Chromatin Structure Prediction from Hi-C Data

Laraib Malik

Stony Brook University
New Computer Science Building, Engineering Drive, Stony Brook University, NY, 11794
631-480-3487
lmalik@cs.stonybrook.edu

A. Project Overview

Summary: Recently, advanced methods have been used to study the 3D structure of chromatin in the cellular nucleus, which is known to directly impact the transcriptional regulation of genes taking part in various biological processes within the cell. Studies of this structure revealed the presence of densely packed regions of the chromatin, called topologically associating domains. Several computational methods have since been developed to predict these domains using data output from chromatin conformation capture experiments [1, 2, 3]. However, these methods have an underlying assumption that the folding of chromatin is not nested. In our current work, we addressed this drawback and developed a method to efficiently predict a hierarchy of domains using a contact matrix from Hi-C, a high-throughput experimental assay that allows genome-wide conformation capture [4, 5].

Intellectual Merit: Developed an improved and efficient method for the prediction of a hierarchy of topologically associating domains using data from chromatin conformation capture experiments. Our tool has the ability to process higher resolution

data from recent studies and is mostly data-driven, requiring only a single parameter as input from the user, making it easy to use on larger datasets.

Broad Impact: With advancing technology, the quality of output from Hi-C experiments is improving and our tool has the capability to process these higher resolution datasets. As more studies are done using chromatin conformation, our methods can be used to gain insights on how the hierarchical structure of domains varies across cell types and species. The simplicity of the tool and required input makes it easy-to-use for a broad audience. The methods we developed and will continue to develop are released online as an open-source software, under a free software license.

Title Rich Chromatin Structure Prediction from Hi-C Data

B. **Bio-sketch**

a) Professional Preparation
Lahore University of Management Sciences Computer Science B.S. 2013
Stony Brook University Computer Science Ph.D. Ongoing
b) Appointments
Research Assistant Stony Brook University 2015 - present
Graduate Teaching Assistant Stony Brook University 2014
Undergraduate Teaching Assistant Lahore University 2012 - 2013
c) Other Publications
- Laraib Malik, Shravya Thatipally, Nikhil Junneti and Rob Patro. Graph regularized, semi-supervised learning improves annotation of de novo transcriptomes. *bioRxiv* , 2015 (Poster presentation at *BioData'16*)
- Alexis Santana, Darby Oldenburg, Varvara Kirillov, Laraib Malik, Qiwen Dong, Roman Sinayev, Kenneth Marcu, Douglas White, Laurie Krug. LPS/TLR4 signaling enhances RTA occupancy of the origin of lytic replication during murine gammaherpesvirus 68 reactivation from latency. *Pathogens*
- Avi Srivastava*, Hirak Sarkar*, Laraib Malik and Rob Patro. Accurate, Fast and Lightweight Clustering of *de novo* Transcriptomes using Fragment Equivalence Classes. *RECOMB-Seq* , 2016
d) Ph.D. Thesis Advisor: Rob Patro

e) Awards and Honors

Grace Hopper Celebration Student Scholarship, 2017

CRA-W Grad Cohort Scholarship, 2015

High Merit B.S. Graduate, 2013

Gaussian Mixture Models with Constrained Flexibility for Fitting Tomographic Tilt Series

Kasra Manavi

University of New Mexico

MSC01 1100
1 University of New Mexico
University of New Mexico
Albuquerque, NM 87131
(505) 277 8912
kazaz@cs.unm.edu

A. Research Summary

Overview: The objective of this PhD thesis is the model and associated algorithmic development for efficiently capturing and analyzing the complex conformational space of multi-molecular binding. The emphasis is on elucidating motions and structures involved in molecular assembly, which is characterized by sets of molecules binding to form a functional superstructure. This thesis focuses on the problem of allergen-antibody assembly, a precursor to immune response, e.g., allergic reactions.

Despite recent simulation and experimental advances for studying allergic reactions, a key component of allergic response is poorly understood: the structural role in the formation of assembled antibodies and allergens. Specifically, assembly occurs when multiple binding site (multivalent) allergens bind with bivalent antibodies. The binding

events create superstructures that initiate intracellular signaling. These advances unfortunately have not been able to provide a detailed enough understanding of the allergen structure's role in assembly.

Our proposed models and methods, inspired by robotic motion planning, will provide detailed information about the structure and assembly of aggregates. Specifically, we propose: (1) novel tunable-resolution models, (2) methods for generating and identifying assemblies, and (3) verification of our models with experimental data.

Intellectual Merit: Despite the steady development of tools and methods for modeling molecular assembly, few methods explore detailed geometry in cases involving hundreds of molecules. The novelty of the proposed research is in the incorporation and integration of (1) hybrid and multi-resolution models of molecular structures, (2) understanding of the impact of allergen structure, (3) application of the techniques to study a problem of medical interest and (4) general techniques for modeling/simulation applicable to other molecular interaction problems.

Broader Impacts: The proposed research will provide critical tools applicable to several molecular assembly problems. The focus of this work is on one such problem, antibody assembly. Geometric insights provided by our modeling will enable the rapid testing of disruptions to the allergenic cascade, thus leading to allergy treatments. Experimental collaborations will allow direct utilization and verification of the proposed work and may lead to the development of new experiments. PhD candidate Manavi has a highly successful track record of mentoring students backed by national recognition of his undergraduate mentee's achievements including several papers and 3rd place in the CRA Undergraduate Research Award.

Paper title

Gaussian Mixture Models with Constrained Flexibility for Fitting Tomographic Tilt Series

B. Bio-sketch

In his time in the PhD program, Kasra Manavi has been a mentor to several high school and undergraduate students working on research projects in his lab. Two (2) of these mentored students have being recognized on the national stage for their work with

him. His research area focuses on applying robotic motion planning techniques to model molecular interactions. During his graduate student career, he has authored and co-authored eight (8) conference and journal publications and was awarded Best PhD Presentation at the American Indian Science and Engineering Society national conference. He has also been the President of the Computer Science Graduate Student Association president and was an organizer of the Computer Science Student Conference. In his free time, he works with Navajo Language Renaissance, a non-profit organization working on revitalizing the Navajo language, where he helped implement an online Navajo language proficiency exam allowing students who don't have access to Navajo classes to earn credit for scholarships.

Exploring Protein Cavities through Rigidity Analysis

Stephanie Mason

Western Washington University
110 Forest Lane · Bellingham, WA · 98225
360 · 441 · 0770
stephanie.mason@wwu.edu

A. Research Overview

The objective of this project was to create a survey of data related to the biological and computational properties of protein cavities. To do this, we combined the outputs from an existing fast-graph rigidity analysis approach and related it to structural protein data in order to explore any existing relationships. Rigidity properties of protein cavities is an unexplored area of computational biology. The properties of protein cavities that have been defined computationally are mostly limited to surface area and their residue content. Rigidity analysis has been used to analyze the stability of proteins, but it has not been specifically applied to cavities—which are known to be the active sites of proteins in most cases—until this research.

Intellectual Merit

Using bioinformatics techniques, this work creates a database of cavity-rigidity information useful for developing new algorithmic approaches to making biological predictions related to mutagenesis, ligand binding, and protein-protein interactions. It

advances the knowledge of biology by providing new data useful for making decisions about where to focus wet lab work, which can save time and provide for better focused efforts. It also explores new biological questions through a computational approach. For example, what does this rigidity analysis say about protein function? How is function affected by changing rigidity, specifically at cavities, where ligand binding is known to occur? These are the types of questions that we have addressed and continue to address in our research.

Broad Impact

This work enhances the our scientific understanding of protein cavities, and disseminates this new understanding through an exploration of the data. In the long run these data will be available for public use, and can prove useful for both computational biologists inferring properties of protein structure as well as traditional biologists looking to focus on a subset of the structures in their specialized work.

Paper Title accepted by ACMBCB2017

Investigating Rigidity Properties of Protein Cavities (Accepted into ACMBCB2017 CSBW)

B. Biosketch

Stephanie Mason
Research Assistant & Web Developer
Western Washington University
516 High Street, Bellingham WA 98225
360 441 0770
stephanie.mason@wwu.edu
http://stephaniemasondesign.com

Professional Preparation

• California State University Long Beach, CA. Linguistics BA, 2009

• Western Washington University Bellingham, WA Biology/Math (major), Computer Science (Minor), Chemistry (Minor), BS Expected, 2017

Appointments

• 2016-Present: Research Assistant, Western Washington University

• 2015-Present: Web Developer, Western Washington University

Products

• Stephanie Mason, Tim Woods, Brian Chen, and Filip Jagodzinski. 2017. Investigating Rigidity Properties of Protein Cavities. In Proceedings of ACM-BCB '17, Boston, MA, USA, August 20-23, 2017.

• Web Development for Window, Western Washington University's Magazine: http://window.wwu.edu/

• Web Development for Western's Small Business Development Center: https://sbdcdev.wwu.edu/

Synergistic Activities

• Developed curriculum for middle school aged children to study computer science in an art context as part of an after-sschool program (2016)

• Developed curriculum for middle school and elementary aged children to study computer science, electronics, and art for summer workshops (2017)

• Volunteer for GEMS (Girls in Engineering, Math, and Science) event (2017)

• Member of Society of Women Engineers (2017-present)

• Member of campus chapter of Association of Women in Computing (2016 – present)

• Member of Association for Computing Machinery

Detection of Differential Abundance Intervals in Longitudinal Metagenomic Data Using Negative Binomial Smoothing Spline ANOVA

Ahmed Metwally

University of Illinois at

Chicago

ametwa2@uic.edu

A. Project summary

One of the objectives of the microbiome studies is to determine whether there is a particular microbial signature (e.g. taxa or genes) associated with a particular disease state and/or disease outcome. These biomarkers can play an important role in the development of preventative and therapeutic strategies. In addition, a major challenge in microbiome studies is the variability in microbial taxa among subjects, in addition to variability due to disease influences. A powerful strategy to address this challenge is the analysis of time series data in which the time intervals associated with temporal effects are identified. Modeling metagenomic data for disease-association studies is an active area of research. The standard parametric models may reduce variance if the data follows the corresponding parametric distribution, but the models may be substantially biased if the data does not support that distribution. On the other hand,

non-parametric models do not assume any prior distribution of the data and thus are not biased towards any distribution, but these models may suffer from a huge model variance.

In this proposal, we propose a new strategy to accurately identify the time intervals when the features are differentially abundant between two phenotypic groups. Correlating the features' differential time intervals with the time-specific clinical data may reveal information that can be used in improving intervention or treatment plans. The proposed method has the ability to handle the inconsistencies and common challenges associated with human studies, such as variable sample collection times and uneven number of time points along the subjects' longitudinal study. The method employs a negative binomial distribution in conjunction with a semi-parametric SS-ANOVA to model the count reads. The method performs the significance testing based on unit time intervals. Next, we plan to expand the current longitudinal differential abundant method to count for more covariates (Age, gender, race, disease severity, etc.). Moreover, the proposed method can be used to identify the significant time intervals in any longitudinal count data such as metagenomics, 16S rRNA, or RNAseq. Furthermore, we plan to package the R-code and deposit it on CRAN/Bioconductor repositories for public use.

The paper or poster title/titles that is/are co-authored by the applicant and accepted by ACMBCB 2017

BCB081: Detection of Differential Abundance Intervals in Longitudinal Metagenomic Data Using Negative Binomial Smoothing Spline ANOVA

BCBp217: Microbiome Dynamics as Predictors of Lung Transplant Rejection

B. Bio-sketch

Ahmed is a Bioinformatics Ph.D. candidate at the University of Illinois at Chicago. He got my B.Sc. in 2010 with the highest-class honors, and M.Sc. in 2014, both in Biomedical Engineering from Cairo University. In November 2015, his Ph.D. proposal received the UIC Chancellor's research award. Subsequently, in April 2016, he won the first-place award at the UIC Research Forum among participants from Engineering, Computer Science, Mathematics, and Statistics departments. In March 2017, he won the Scientific Excellence Award for the best poster at the UIC Department of Medicine

scholarly activities. Ahmed has also had the opportunity to work in the genomics software engineering industry through two internships at ThermoFisher Scientific. In these internships, he worked on developing software applications that advance personalized medicine for cancer patients. Additionally, in recognition of outstanding leadership, Ahmed has been elected (2017-2019) to be the global student representative for the IEEE Engineering in Medicine and Biology Society (EMBS), the world's largest international society of biomedical engineers where he establishes initiatives to promote the career development of bioengineering trainees and strengthening the fields of bioinformatics and bioengineering. Ahmed has a strong track record of research and interest in academic activities. During his undergraduate years, he received a DAAD fellowship in 2009 to conduct summer research at the Institute of Materials Research, Helmholtz-Zentrum Geesthacht, Germany. During 2010-2014, he worked as a demonstrator and then as an assistant lecturer in the Biomedical Engineering Department of Cairo University. During 2012-2013, he was appointed as a junior scientist at Nile University to work on developing bioinformatics algorithms in collaboration with Ulm University, Germany. Ahmed has received many awards from IEEE, ISCB, UIUC, APBioNET, ICTP, and LinkSCEEM for various educational and scholarly activities.

Exploiting the structure of the interaction network in pathway enrichment analysis

Pourya Naderi Yeganeh

UNC Charlotte
9201 University City Blvd.
Charlotte, NC, 28213
pnaderiy@uncc.edu

A. Project Summary

Overview. Understanding the functional abnormalities and the systematic dysregulations is an integral component of the study of complex diseases, including cancer. Enrichment analysis models are among the most common methodologies to carry out the functional inference of high-throughput experimental data. These models generally investigate the functional associations of diseases by evaluating the prevalence of manually curated annotations in the set of dysregulated genes of the experimental data. Although the biological functions are composed of interactions among biomolecular entities, traditional models for enrichment analysis disregard the interactions and focus on the entities, e.g. genes and proteins, as independent observations. The objective of this research project is to harness the organization of the biological entities and their interactions, as a network, to improve the functional inference of genomic data. In particular, I aim to contribute to this objective by exploring graph theory models that evaluate the structural position (i.e. centrality) of

genes in a pathway graph, based on their biological importance. To this end, I will utilize the curated pathways databases and interaction networks to identify the centrality models that best describe the difference between inter-pathway and intrapathway interactions. Moreover, I aim contribute to the objective by constructing an enrichment analysis model where the importance of the genes is assigned relative to a calculated centrality score. To identify evidence of pathway perturbations, I attempt to calculate an aggregate structural effect of the dysregulated genes based on the centrality scores. This perturbation evidence will be combined with additional dysregulation evidence to obtain a more precise and biologically relevant inference.

Intellectual merit. The proposed research leverages the biomolecular interactions towards a comprehensive enrichment analysis of genomic data, particularly, by analyzing the topological organization of the pathway graphs. Common enrichment analysis methods discard the interactions, which results in loss of information. Recent studies show that the utilization of interactions improves biological inference. One short coming of the current models is the use of limited factors, if any, to leverage the interaction network structure. The proposed research improves the current inference models by prioritizing the biological entities based on their structural position (centrality) in the pathways rather than ignoring the interactions. This research aims to infuse mathematical abstractions that describe global structures of pathways as a part of common inference models.

Broader impact. The proposed design will be available as a practical tool for biomedical researchers working on complex diseases genomic data, especially cancer researchers. Modern cancer studies emphasize on functional characterization of diseases in order to design effective therapeutic and diagnostic approaches. An interaction-sensitive pathway analysis model will provide a robust framework to detect the functional perturbations in the disease data and ultimately contribute to the design of effective drug target discovery and biomarker discovery.

ACCEPTED BCB SHORT PAPER: "Use of Structural Properties of Underlying Graphs in Pathway Enrichment Analysis of Genomic Data", Pourya Naderi Yeganeh and M. Taghi Mostafavi

B. **Biosketch**

Pourya grew up in Tehran, Iran, where he attended Sharif University of Technology. He graduated in 2013 with a B.Sc. in computer science and moved to the US to start his Ph.D. study in computing and informatics at the University of North Carolina at Charlotte. He has been working on computational models for early detection of ovarian cancer under supervision of Dr. Taghi Mostafavi. His current research focus is on graph models for functional inference of genomic data.

Large-scale Deep Learning with Biomedical Data

Richard Platania

Louisiana State University
Baton Rouge, LA 70808
225-939-8207
rplata1@lsu.edu

A. Project Summary

In recent years, the application of deep learning has been prolific in many domains. The data used by these domains comes in many forms, such as text, signals, or images. Deep learning applies neural networks with hidden layers to these data types. As a result, detailed features are learned, and in-depth analysis is possible. This application has been extremely successful with object detection and classification in images, voice recognition, recommendation systems, and many other popular goals.

Biomedical data can also benefit from the use of deep learning. However, it is of particular difficulty to apply due to the complexity and size of the data. For instance, features within the data can be much more complex and require more advanced deep learning techniques. Additionally, typical image-based deep learning tasks outside of biomedical data use images with few pixels. Medical images, such as mammograms, are of significant size, causing an increase in the time and complexity of training the neural networks.

The goal of this project is to effectively apply large-scale deep learning techniques towards biomedical data. In particular, we focus on medical image data, such as mammogram and fMRI, and DNA sequencing data. Object detection and classification tasks are to be performed on the data. An example of each, respectively, is detecting lesions in mammograms and diagnosing ADHD from an fMRI. These tasks will be made available to the public for use by medical professionals and researchers of relevant fields.

Intellectual Merit

The research involved in this project will provide detailed analysis of biomedical data. Medical professionals will be provided with tools to help them better diagnose and understand various medical conditions. Through this project, advanced deep learning techniques will be developed that may be applied to other applications within the field of Bioinformatics.

Broader Impacts

The deep learning applications developed in this project will be made available through a public website. This will allow physicians, bioinformaticians, or other relevant experts to make use of these tools for their research. Several undergraduate students will be involved in the project, which will prepare them for their future research. Some of the applications, such as fMRI and mammogram analysis, have public competitions that our project will be a part of. This will promote collaboration among other deep learning experts within the field of bioinformatics.

B. Biosketch

Richard is a 4th year Computer Science Ph.D. student at Louisiana State University, where he also received his Bachelor's degree in Computer Science with a minor in Mathematics. His current research focus is in large-scale deep learning with biomedical data. In particular, this includes automated detection and diagnosis of breast cancer, subtype diagnosis of ADHD, and cell type classification of DNA sequences. He has published papers in notable conferences such as ACM-BCB and ICDCS as well as a journal article in *Computation and Concurrency: Practice and Experience.* Other research interests include big data applications within bioinformatics, such as genome assembly and molecular dynamics simulations. Some of his practical expertise includes

Python and C++ programming, developing deep learning applications with TensorFlow, and writing applications for big data using tools in the Hadoop ecosystem.

A correlation Network Model Utilizing Gait Parameters for Evaluating Health Levels

Elham Rastegari

University of Nebraska at Omaha
1110 S 67th St, Peter Kiewit Institute, Room 354, Omaha, NE 68182
402-541-5648
erastegari@unomaha.edu

A. Project Summary

Intellectual Merit

Healthcare is moving rapidly from the long-standing reactive treatment approach to the early detection and preventative era. However, to fully embrace this trend, new approaches need to be developed. A step in this direction is to explore how to leverage data collected from wearables mobility monitoring devices to help in assessing health levels. This would pave the way for continuously monitoring individuals which, in turn, lead to helping physicians diagnose some health conditions in the early stages as well as assessment of treatment approaches during rehabilitation. However, a major missing piece in moving forward with this concept is the lack of a sophisticated data analytics model. This study contributes to the field by filling this gap, introducing a new correlation network model.

Broader Impact

Technology has advanced a lot but the effect of technology in healthcare domain is still minimum. This research emphasizes on utilizing new technologies and integrating various data (e.g., mobility, genetic background, diet) to advance healthcare. The broader impacts of this research are as follow:

1. The project represents one of the early attempts to use mobility data to assess health levels and potentially predict health problems. Its outcomes can potentially open doors for more data-driven healthcare projects.

2. The proposed method is a step towards personalized healthcare; because in the real-life setting, every individual can enter/change his/her own information such as age, gender, genetic background, height, and weight and this information together with continuously collected gait and mobility patterns over time will be used for assessing health levels of the individual. This model can be used in tracking changes associated with the mobility parameters of either healthy individual to assess the potential degradation in health for early diagnosis purposes or individuals with some health issues to assess the effectiveness of treatments. This would allow for the ability to assign/reassign each individual dynamically to a group or subpopulation with similar mobility patterns or health levels.

3. The proposed model is applicable to many health conditions, including aging, Parkinson's Disease (PD), Multiple Sclerosis, Huntington disease, stroke, and Amyotrophic Lateral Sclerosis so will help physicians in diagnosis and better assessment of treatment approaches.

The paper title accepted by ACMBCB2017 is "A Correlation Network Model Utilizing Gait Parameters for Evaluating Health Levels".

B. Biosketch

Elham Rastegari, born in Iran, is a Ph.D. student in the Biomedical Informatics program with concentration in Health Informatics at the University of Nebraska at Omaha. She has developed a passion for interdisciplinary approaches to solve issues in the healthcare domain. her focus is more on developing new techniques to improve health care. These techniques benefit from machine learning, big data analytics, signal processing, and social computing. Before starting her Ph.D., Elham was a lecturer at

Azad University of Tehran. She holds a bachelor and a master's degree in computer science. She enjoys swimming, hiking, Persian calligraphy, and planting in her free time.

Simple Voting with an Ensemble Convolution Neural Network for Lung Tumor Detection

Allison M. Rossetto

University of Massachusetts Lowell
1 University Avenue, Lowell MA
(248)971-5733
Allison_Rossetto@student.uml.edu

A. Research Summary

Lung cancer is the leading cause of cancer deaths in the United States. Approximately 225,000 people each year are diagnosed with lung cancer in the US. Currently, physical scans are read and tumors are identified and classified by human beings. Automated systems would help speed up and increase the accuracy of this process. The method we are developing is an ensemble of Convolution Neural Networks (CNN) that is built from two individual CNNs: one uses unprocessed images and the other images smoothed with a Gaussian filter. The ensemble method decreases the number of false positives in the automated labeling of the scans using a voting system. Our improved ensemble method has an average accuracy of 97.94% and false positive rate of 0.22%. This is a great improvement over the currently used methods and shows that our method has promise as an initial diagnostic technique.

Intellectual Merit

The proposed method demonstrates how the false positive rate of lung tumor identification can be decreased using multiple deep learning networks with predictions combined via a simple voting method. The method is a step forward in developing an accurate automated diagnosis system for lung cancer CT images.

Broad Impact

Lung cancer is the leading cause of cancer deaths in the United States and around the world. Early detection is a crucial part of increasing the chance of patient survival. However, physical screening methods suffer from very high false positive rates. These false positive diagnoses can lead to unnecessary treatments and procedures, which waste time, money, and resources and cause undue stress to patients and families. Our automated initial diagnosis techniques are, therefore, an essential part of increasing the accuracy of early detection.

ParBio Workshop Paper:

Allison M Rossetto and Wenjin Zhou. 2017. Ensemble Convolution Neural Network with a Simple Voting Method for Lung Tumor Detection. In Proceedings of ACM-BCB'17, August 20–23, 2017, Boston, MA, USA. DOI:http://dx.doi.org/10.1145/310 7411.3108174

B. **Bio-sketch**

I am currently a PhD student in Computer Science at the University of Massachusetts Lowell. I have a Bachelor's of Science in Engineering Biology and a Master's of Science in Computer Science from Oakland University. My research has focused on areas of computational drug design and automated tumor detection using deep learning. I have published several papers in both of these areas and have given invited talks on the role of computer science in healthcare at multiple universities.

I am originally from Michigan. I am involved in STEM outreach, mainly with underprivileged and underrepresented groups. I also volunteer with a feral cat organization, which does Trap-Neuter-Return (TNR) and I assist in the care of a feral cat colony.

Interpretable Predictions of Clinical Outcomes with An Attention-based Recurrent Neural Network

Ying Sha

Georgia Tech

1608 Alexandria CT SE,
Marietta, GA 30067
919-491-1631
ysha8@gatech.edu

A. Research Overview

The increasing accumulation of healthcare data provides researchers with ample opportunities to build machine learning approaches for clinical decision support and to improve the quality of health care. Several studies have developed conventional machine learning approaches that rely heavily on manual feature engineering and result in task-specific models for health care. In contrast, healthcare researchers have begun to use deep learning, which has emerged as a revolutionary machine learning technique that obviates manual feature engineering but still achieves impressive results in research fields such as image classification. However, few of them have addressed the lack of the interpretability of deep learning models although interpretability is essential for the successful adoption of machine learning approaches by healthcare communities. In addition, the unique characteristics of healthcare data such as high dimensionality and temporal dependencies pose challenges for building models on healthcare data. To address these challenges, we develop a gated recurrent unit-based recurrent neural network with hierarchical attention for mortality prediction, and then, using the diagnostic codes from the Medical Information Mart for Intensive Care, we

evaluate the model. We find that the prediction accuracy of the model outperforms baseline models and demonstrate the interpretability of the model in visualizations.

Intellectual Merit: This study bridges the gap between artificial intelligence and health care by developing a model not only appropriately characterizes healthcare data but also achieves a great balance between accuracy and interpretability.

Broader Impact: This study proposes a model that has great potential to advance personalized health care. Physicians can use the model to prioritize patient care based on predicted risks of individual patients. Ultimately, the model will aid in increasing healthcare efficiencies and lowering healthcare cost.

The title of the paper authored by the applicant is "Interpretable Predictions of Clinical Outcomes with An Attention-based Recurrent Neural Network".

B. Bio-Sketch

Ying Sha earned her B.S. degree in biology from Peking University and her M.S. degree in bioinformatics from Georgia Tech. She is currently a doctoral student in the school of biology, where she is conducting research pertaining to temporal data mining using intensive care unit (ICU) data. As a graduate research assistant, she actively collaborated with clinical institutes such as Children's Healthcare of Atlanta and developed tools for facilitating clinical decision support. She was also a summer intern at Dow Agrosciences, at which she worked closely with microbiology researchers for projects related to natural product discovery.

Drug Response Prediction as a Link Prediction Problem

Zachary Stanfield

Systems Biology and Bioinformatics Program, Case Western Reserve University

10900 Euclid Avenue, Olin Building Room 513, Cleveland, OH 44106
(828) 612-3066
zachary.stanfield@case.edu

A. Research Overview

This work is centered on the problem of drug response prediction, which involves using molecular features of a given sample (i.e. patient) to predict sensitivity to a drug. In the context of cancer, heterogeneity and complexity of the disease means that patients with a highly similar diagnosis may respond quite differently to the same treatment. Therefore, cancer treatment can directly benefit from advances in drug response prediction. Previous approaches to solve this problem employ traditional machine learning algorithms to learn a predictive model for each drug based on genomic data from cell lines and response data from drug screening assays. However, these approaches typically lack the ability to provide a functional or biologically relevant context to the features in the model. To overcome this, methods utilizing a biologically relevant network can be used to leverage existing biological knowledge to improve drug sensitivity prediction.

The development of an accurate drug response prediction algorithm can be greatly beneficial to medicine in that it may be employed to make real predictions for patients or investigated to determine what biological information is being used to make those predictions. Such knowledge may directly improve cancer treatment and, therefore, patient outcomes. Regarding reproductive research, discovery of the biological mechanisms regulating parturition will provide insights into how labor disorders such as preterm birth, a disease affecting over 15 million infants worldwide, occur. Research in this field may also use current knowledge of parturition to target specific areas of the cellular machinery (i.e. a pro-labor kinase), as discovered by these network approaches, to delay the onset of preterm labor, ultimately reducing the number of premature births.

Paper Title

Drug Response Prediction as a Link Prediction Problem

B. Biosketch

During my undergraduate career at North Carolina State University, pursuing a B.S. degree in Physics, I acquired extensive knowledge of mathematics, learned how to leverage computing to solve real-world problems, and developed a strong ability to approach problems in a systematic and creative manner. In the latter years of this training and education, I formed an interest in the application of physics/mathematics/computer science to biomedical problems and disease research. Therefore, my undergraduate career provided me with training and background knowledge well-suited for bioinformatics and initiated my interest in the biomedical sciences.

This interest led me to enroll as a PhD student in a systems biology and bioinformatics program where I could build on my training to become an interdisciplinary researcher able find creative solutions to a variety of biological problems. Throughout my coursework and PhD training, I have developed skills in areas such as data mining and machine learning while enhancing my knowledge of cellular and molecular biology and how it plays a role in disease progression. My research focuses on employing computational algorithms on complex networks to identify important biological mechanisms of disease using high-dimensional molecular data. My projects include a

network-based method to predict drug response for cancer cell lines (this work) and modeling cellular signaling in preterm birth.

Learning Deep Representations for Causal Inference

Hang Wu

Georgia Tech and Emory University

U.A. Whitaker Building, Atlanta, GA, 30332
404-368-7718
hangwu@gatech.edu

A. Research Summary

Abstract

This project aims to design and deploy an open source tool for Biomedical and Health Informatics, which explores deep learning techniques for causal inference in biomedical observations. The current causality methods use strong assumptions and lack flexibility. With the success of deep learning in representing relationship in other types of raw data, we will find variables carrying domain knowledge from observation data in biomedicine, and perform causal inference tasks (e.g. identifying the causal structures of individual observations).

In this project, we will integrate "knowledge graphs" from biomedical domain experts and large collections of health "observational data" to perform relevant causal inference tasks. Specifically, we will derive a representation for the medical conditions in the integration, and then apply them to infer the causes of death.

Broad Impact

Our project has two impacts, mainly in two different application scenarios:

1) Death reporting: medical examiners need to file death certificates for every death case reported, where s/he needs to identify the cause of death for the patient. Such

identification is challenging and vary physician to physician, so using an algorithm to automatic identify the causes can facilitate the process and introduce less errors.

2) Hospital care: the successful identification of critical conditions that might lead to death can help physicians and nurses prioritize their treatment and care for patients, thus improving the quality of care for patients.

Intellectual Merit

The intellectual merit of our project is mostly two-fold:

1) We combine data-driven approaches with biomedical domain knowledge for causal inference tasks. The application of deep learning-based representation learning is also novel compared to traditional statistical approaches.

2) The learned representation for medical conditions can also be used for other tasks, for example, building predictive models for patient outcomes needs a better understanding of causal relations between observed conditions.

Paper Title

Infer Cause of Death for Population Health Using Convolutional Neural Network

Hang Wu, May D. Wang

B. Bio-sketch

Hang Wu

U.A. Whitaker Building, Suites 4238
Georgia Tech, Atlanta, GA, 30332
hangwu@gatech.edu

Research interest: healthcare informatics, biomedical imaging, machine learning

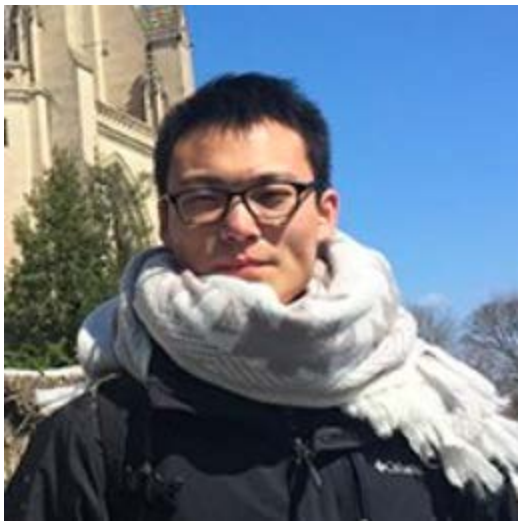
EDUCATION

09. 2015 - PRESENT **PH.D. STUDENT, BIOMEDICAL ENGINEERING**
Georgia Institute of Technology and Emory University

08. 2010 – 07. 2014 **B.E., SCHOOL OF INFORMATION SCIENCE**
Tsinghua University, China

SELECTED PUBLICATIONS

- Hang Wu, May D. Wang. "Multi-View Non-Negative Tensor Factorization as Relation Learning in Healthcare Data" Accepted by 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS).
- Sonal Kothari, Hang Wu, Li Tong, Kevin Woods, May D. Wang. "Automated Prediction for Esophageal Optical Endomicroscopic Images" In Proceedings of 2016 IEEE International Conference on Biomedical and Health Informatics
- Hang Wu, Chihwen Cheng, Xiaoning Han, Yong Huo, Wenhui Ding, and May D. Wang "Prediction in the Presence of Low Rank Missing Data with Applications to Post-Surgical Complication Prediction" In Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS). Doi: 10.1109/EMBC.2015.7319957
- Hang Wu, John H. Phan, Ajay K. Bhatia, Caitlin A. Cundiff, Bahig M. Shehata, and May D. Wang "Detection of Blur Artifacts in Histopathological Whole-Slide Images of Endomyocardial Biopsies" In Proceddings of 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS). Doi: 10.1109/EMBC.2015.7318465
- Hang Wu, Ji-jiang Yang, Jianqiang Li. "Low Redundancy Feature Selection with Grouped Variables and Its Application to Healthcare Data" In Big Data (Big Data), 2014 IEEE International Conference on (pp. 71-76). IEEE. Doi: 10.1109/BigData.2014.7004396



Seq2seq Fingerprint: An Unsupervised Deep Molecular Embedding for Drug Discovery

Zheng Xu

The University of Texas at Arlington

500 UTA Blvd. Arlington, TX 76010
817-879-5048
zheng.xu@mavs.uta.edu

A. Research Summary

Our project aims at finding a data-driven approach to represent each molecule for drug discovery tasks. Traditionally, the chemical screening hit test is extremely expensive and labor intensive. Therefore, using machine learning methods has recently raised great interests in the drug discovery community to automatically predict the drug properties for molecules. However, due to the structured nature of the molecules, it is hard to choose a fixed-length continuous molecular representation as an input for machine learning methods. Traditional representation methods require lossy compression or human guided feature extraction, which is to some extent no favorable in general drug discovery tasks due to lower performance or expensive labor cost. In this project, we develop a completely data-driven approach, utilizing the most recent deep recurrent neural network with an unsupervised training strategy. The proposed method is demonstrated a good performance on various tasks.

Intellectual Merit

Our method is developed based on a most recent state-of-the-art model from a seemingly unrelated area, natural language processing. This method is named sequence to sequence network, or seq2seq net, in short. Seq2seq net or its variants have topped many language translation benchmark recent days. But there is still no attempt to apply it to drug discovery area for molecular representation.

Broader Impact

The method developed in this project can potentially help back the natural language processing, e.g., the dropout wrapper added to Gated Recurrent Units. Also, we have demonstrated the success of seq2seq net in a seemingly unrelated area, drug discovery. This is a pioneer example for future extension to other areas.

ACM BCB Reference

Our accepted paper is:

Zheng Xu, Sheng Wang, Feiyun Zhu, Junzhou Huang, Seq2seq Fingerprint: An Unsupervised Deep Molecular Embedding for Drug Discovery, ACM BCB 2017.

B. Biosketch

Zheng Xu received the B.S. degree from Huazhong University of Science and Technology, Wuhan, China. He is a Ph.D. student in the Department of Computer Science and Engineering, The University of Texas at Arlington. His research interests include machine learning, computer vision, and biomedical informatics, with focus on the deep neural network, efficient algorithms for sparse models as well as their applications on bioinformatics.

May Dongmei Wang, Ph.D.

Professor, The Wallace H. Coulter Joint Department of Biomedical Engineering
Georgia Tech Petit Institute Faculty Fellow &
Director of Biomedical Big Data Initiative
Kavli Fellow, Georgia Cancer Coalition Distinguished Cancer Scholar, Fellow of AIMBE
Georgia Institute of Technology and Emory University
UA Whitaker Bldg. Suite 4106, 313 Ferst Dr., Atlanta, GA 30332-0535, USA
Tel: +1-404-385-2954
Email: maywang@bme.gatech.edu

Biography

Dr. May Dongmei Wang is a full professor in the Joint Biomedical Engineering Department of Georgia Tech and Emory University, a Kavli Fellow, a Georgia Cancer Coalition Scholar, Georgia Tech Petit Institute Faculty Fellow and Director of Biomedical Big Data Initiative, and a Fellow of the American Institute for Biological and Medical Engineering (**AIMBE**). She earned her BEng from Tsinghua University China, MS and PhD from Georgia Institute of Technology. Her research is in Biomedical Big Data Analytics with a focus on Biomedical and Health Informatics (**BHI**) for predictive, personalized, and precision health (**pHealth**). In FDA-organized MAQC international consortium, she led the comprehensive RNA-Seq data analysis pipeline study. Dr.

Wang has published over 200 peer-reviewed conference proceeding and journal papers in referred journals (e.g. Briefings in Bioinformatics, BMC Bioinformatics, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Journal of American Medical Informatics Association, Journal of Biomedical and Health Informatics, Journal of Pathology Informatics, PNAS, Annual Review of Medicine, Nature Protocols, Circulation Genetics, IEEE Trans. on Biomedical Engineering etc.) and conference proceedings, and she has delivered more than 200 invited and keynote lectures. She received Outstanding Faculty Mentor Award for Undergraduate Research at Georgia Tech, and a MilliPub Award (for a high-impact paper that has been cited over 1,000 times) from Emory University.

Dr. Wang has served as an Emerging Area Editor for **PNAS**, Senior Editor for Journal of Biomedical and Health Informatics, an Associate Editor for **TBME**, and a panelist in NIH, and NSF review panels. She has helped organize ACM Bioinformatics, Computational Biology and Health Informatics Conferences and IEEE International Conference on Biomedical and Health Informatics. She is elected as Vice Chair for 2018 Gordon Research Conference (GRC) on Advanced Health Informatics, and has served in IEEE Big Data Initiative (**BDI**) Steering Committee. Dr. Wang is Georgia Tech Biomedical Informatics Program Co-Director in Atlanta Clinical and Translational Science Institute (**ACTSI**), and Co-Director of Georgia-Tech Center of Bio-Imaging Mass Spectrometry. Her research has been supported by NIH, NSF, CDC, Georgia Research Alliance, Georgia Cancer Coalition, Emory-Georgia Tech Cancer Nanotechnology Center, Children's Health Care of Atlanta, Atlanta Clinical and Translational Science Institute (ACTSI), and industrial partners such as Microsoft Research and HP.

Zhaohui "Steve" Qin, Ph.D.

Associate Professor
Director of Graduate Studies
Department of Biostatistics and Bioinformatics
Rollins School of Public Health
Emory University
Atlanta GA 30322

Biography

Zhaohui (Steve) Qin is currently an Associate Professor in the Department of Biostatistics and Bioinformatics at Rollins School of Public Health, Emory University. He is also an affiliated faculty member at the Department of Biomedical Informatics, Emory University School of Medicine and Biostatistics and Bioinformatics Shared Resource, Winship Cancer Institute.

Dr. Qin received his B.S. degree in Probability and Statistics from Peking University in 1994 and Ph.D. degree in Statistics from University of Michigan in 2000. After postdoctoral training in Dr. Liu Jun's group at Harvard University, he joined the Department of Biostatistics at University of Michigan in 2003. In 2010, he moved to his current position in Emory University.

Dr. Qin has 15 years of experience in statistical modeling and statistical computing with applications in statistical genetics and genomics. Dr. Qin has published more than 100 peer-reviewed research papers covering statistics, bioinformatics, statistical genetics and computational biology. His publication has been cited more than 13,000 times according to Google Scholar. Recently, his research is focused on developing Bayesian model-based methods to analyze data generated from applications of next-generation sequencing technologies such as ChIP-seq, RNA-seq, Hi-C, WGBS, ATAC-seq, Repli-Seq and resequencing. Dr. Qin also actively collaborates with biomedical

scientists and clinicians on projects that utilize next-generation sequencing technologies to study cancer genomics.

Anna Ritz, Ph.D.

Assistant Professor
Biology Department
Reed College, Portland, OR

Biography

Anna Ritz is an Assistant Professor of Biology at Reed College in Portland, Oregon. Her current research develops network-based algorithms to identify signaling pathway dysregulation in diseases such as cancer. Before she arrived at Reed in 2015, she was a postdoctoral researcher in the Computer Science Department at Virginia Tech, where she worked on analyzing and modeling human signaling pathways. Anna obtained her Ph.D. and Sc.M. in Computer Science from Brown University, where her graduate work focused on identifying large structural rearrangements in human and cancer genomes from single-molecule ``third generation'' sequencing technologies.

Ying Sha

PhD. Student
Georgia Tech,
Atlanta, GA 30332

Biography

Ying Sha earned her B.S. degree in biology from Peking University and her M.S. degree in bioinformatics from Georgia Tech. She is currently a doctoral student in the school of biology, where she is conducting research pertaining to temporal data mining using intensive care unit (ICU) data, supervised by Dr. May. D. Wang. As a graduate research assistant, she actively collaborated with clinical institutes such as Children's Healthcare of Atlanta and developed tools for facilitating clinical decision support. She was also a summer intern at Dow Agrosciences, at which she worked closely with microbiology researchers for projects related to natural product discovery.