NSF Sponsored Student Research Forum

Location: Room 225

Co-Chairs

Professor Zhaohui (Steve) Qin
Emory University

Professor May D. Wang
Georgia Institute of Technology and Emory University

Student Co-Chair

Po-Yen L. Wu
Georgia Institute of Technology

Schedule

4:00–4:05 pm	**NSF Travel Fund Awardee Forum Chairs and NSF Program Directors**	
4:05–4:09 pm	**Sherif Abdelhamid**	Web Based Distributed Systems for Modeling Contagions in Networked Populations
4:09–4:13 pm	**Maezieh Ayati**	Assessing The Collective Disease Association of Multiple Genomic Loci
4:13–4:17 pm	**Elisabeth Brooks**	An Extensible Software Infrastructure for Testing the Evolutionary Consequences of Developmental Interactions
4:17–4:21 pm	**Shaolong Cao**	Unified tests for fine scale mapping and identifying sparse high-dimensional sequence associations
4:21–4:25 pm	**Minghan Chen**	Two-dimensional Model of Bipolar PopZ Polymerization in Caulobacter crescentus
4:25–4:29 pm	**Zihe Chen**	Mining k-Median Chromosome Association Graphs from a Population of Heterogeneous Cells
4:29–4:33 pm	**Evangelos Georganas**	Scalable parallel algorithms for de novo assembly of complex genomes
4:33–4:37 pm	**Alexej Gossman**	Identification of significant genetic variants via SLOPE, and its extension to Group SLOPE
4:37–4:41 pm	**Yarong Gu**	A New Method for DNA Sequencing Error Verification and Correction via an On-Disk Index Tree
4:41–4:45 pm	**Chao Ji**	A maximum-likelihood approach to absolute protein quantification in mass spectrometry
4:45–4:49 pm	**Trey Kind**	Fast and Efficient Compression of High-Throughput Sequencing Reads
4:49–4:53 pm	**Kanak Mahadik**	Scaling Genomic Sequence Search with Fine Grained Parallelization
4:53–4:57 pm	**Paola Pesantez Cabrera**	Scalable Algorithms for Building and Clustering Heterogeneous Biological Networks
4:57–5:01 pm	**Kristen Petersen**	Super Deduper, fast PCR duplicate removal in fastq files
5:01–5:05 pm	**Shi Qiao**	Integrated Querying of Disparate Association and Interaction Data in Biomedical Applications
5:05–5:09 pm	**Subrata Saha**	Genome Compression
5:09–5:13 pm	**Adib Shafi**	A systems biology approach for the identification of significantly perturbed genes
5:13–5:17 pm	**Veronika Strnadova-Neeley**	Scalable Clustering in Genetic Mapping
5:17–5:21 pm	**Jesmin Jahan Tithi**	Engineering High-performance Parallel Algorithms with Applications to Bioinformatics
5:21–5:25 pm	**Saima Sultana Tithi**	Incorporating Known SNPs in Short Read Aligners
5:25–5:29 pm	**Kun Wang**	Computational Inference of Expression and Alternative Splicing using High Throughput Data
5:29–5:33 pm	**Hsin-Yi Yeh**	Using Motion Planning to Rank Ligand Binding Affinity
5:33–5:37 pm	**Joseph Crawford**	Simultaneous Optimization of Both Node and Edge Conservation in Network Alignment via WAVE
5:37–5:41 pm	**Hao Wang**	Machine Learning for Data Analysis in Social Health Networks
5:41–5:45 pm	**Yanshan Wang**	Medical information retrieval: Challenges and Solutions
5:45–5:49 pm	**Aston Zhang**	Discovering De Facto Diagnosis Specialties

Web Based Distributed Systems for Modeling Contagions in Networked Populations

Sherif Elmeligy Abdelhamid

Virginia Tech

1880 Pratt Drive, Blacksburg, VA 24061-0477
540-684-0495
sherief@vbi.vt.edu

ABSTRACT

This talk provides an overview of my research work, which lies at the intersection of computation, biology, and education. In particular, I am interested in designing and building software systems to enable domain experts within the fields of biology and public health to easily access and effectively use high performance computing (HPC) to perform simulations and large scale data analyses. GDSC, EDISON and MARS are examples of such systems. Other aspects of my research focus on how to use these systems as learning tools for students and teachers.

BIOGRAPHY

Sherif is a PhD candidate in the Department of Computer Science at Virginia Tech, where he also received his second Master of Science degree in Computer Science. Currently, he is working as a Graduate Research Assistant in Network Dynamics and Simulation Science Laboratory (NDSSL) at VBI, Virginia Tech. Before joining NDSSL, he worked at the Digital Library and Research Laboratory (DLRL) and the Center of Geospatial Information Technology (CGIT), both at Virginia Tech. Sherif is interested in building software tools and systems that help informatics and health professionals understand their data (professionals who may not be computing experts). His paper, entitled “EDISON: A Web-Based Application for Computational Health Informatics At Scale” has been accepted to the ACM-BCB 2015 Conference. Prior to his PhD, he worked as Graduate Teaching Assistant at Arab Academy for Science and Technology in Egypt, where he received his first Master of Science degree in Computer Science. During his MS study at AAST, he developed research interests in image analysis and machine learning for bioinformatics. After he obtained his MS degree, he worked as a Lecturer in the College of Computing and Information Technology, AAST.

Assessing The Collective Disease Association of Multiple Genomic Loci

Marzieh Ayati

Case Western Reserve University

10900 Euclid Ave, Cleveland, OH 44106
216-854-1092
marzieh.ayati@case.edu

ABSTRACT

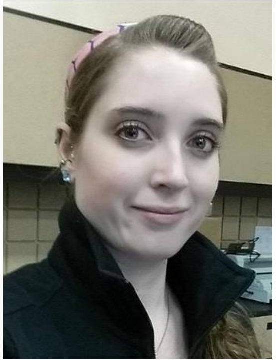
Genome-wide association studies (GWAS) facilitate largescale identification of genomic variants that are associated with complex traits. However, susceptibility loci identified by GWAS so far generally account for a limited fraction of the genotypic variation in patient populations. Predictive models based on identified loci also have modest success in risk assessment and therefore are of limited practical use. In this paper, we propose a new method to identify sets of loci that are collectively associated with a trait of interest. We call such sets of loci “population covering locus sets” (PoCos). The main contribution of the proposed approach is three-fold: 1) We consider all possible genotype models for each locus, thereby enabling identification of combinatorial relationships between multiple loci. 2) We use a network model to incorporate the functional relationships among genomic loci to drive the search for PoCos. 3) We develop a novel method to integrate the genotypes of multiple loci in a PoCo into a representative genotype to be used in risk assessment. We test the proposed framework in the context of risk assessment for two complex diseases, Psoriasis (PS) and Type 2Diabetes (T2D). Our results show that the proposed method significantly outperforms individual variant based risk assessment models.

BIOGRAPHY

I am a fourth year PhD student of Computer Science at Case Western Reserve University and I am advised by Dr. Mehmet Koyuturk. I hold a B.Sc. and M.Sc. in Computer Science from Sharif University of Technology in Iran.

My passion and interest toward bioinformatics turned out when I took “Introduction to bioinformatics” course and got familiar with fundamental algorithmic methods in computational biology. I was encouraged to learn more about this modern and fast growing science after I attended in some bioinformatics seminars and workshops. An enormous number of challenges, thrown out by rapid development in this field made me choose my master thesis in computational biology titled “Improvement of multiple sequence alignment and prediction of protein-protein interactions”. Taking courses like “Bioinformatics for Systems Biology” gave me deeper information about omic data and I have become interested to pursue my research and phd in the analysis and integrating the variety of omic data and its application in health and medical.

In order to keep my information updated and communicate with researchers in another area of this interdisciplinary field, I have volunteered to take part in the executive board of different student organizations such as Computational Biology Graduate Student Organization at CWRU and Regional Student Group (RSG) which acts in conjunction with the Great Lakes Bioinformatics Consortium (GLBioC). This year, I became a student activity chair of ACM BCB’15.



An Extensible Software Infrastructure for Testing the Evolutionary Consequences of Developmental Interactions

Elizabeth Brooks

Central Washington University

1000 East Harvest Loop Unit 2303, Ellensburg, WA, 98926
360-420-5688
brookse@cwu.edu

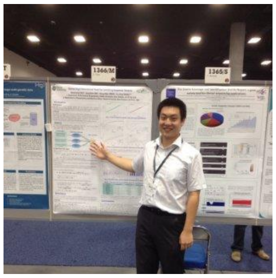
ABSTRACT

Quantitative genetic models commonly use the additive genetic variance-covariance matrix (G-matrix) of a set of traits to predict evolution in response to selection. However, non-linear interactions between developmental factors underlying the production of traits can produce dramatic changes to the G-matrix. To our knowledge there are no freely available tools for predicting the effect of non-linear interactions on evolutionary dynamics. Therefore, we have developed a code base and built two models for testing hypotheses about the effect of specific non-linear developmental interactions on trait (co)variances and ultimate evolutionary trajectories.

BIOGRAPHY

Elizabeth Brooks is currently studying Computer Science at Central Washington University (CWU), with minors in Biology, Music, and Mathematics. Her research interests include mathematical modeling, quantitative genetics, and evolutionary developmental (evo-devo) biology. She is planning to earn a PhD in Computational Biology or Bioinformatics directly upon completion of her bachelor's degree.

Elizabeth's current research is focused on modeling the evolutionary development of phenotypic traits. She has presented this research at several conferences, including Evolutionary Biology in the Pacific Northwest (EVO-WIBO, 2014), CWU's Symposium of University Research and Creative Expression (SOURCE, 2014 & 2015), and The Consortium for Computing Sciences in Colleges – Northwest Region (CCSC-NW, 2014).



Unified tests for fine scale mapping and identifying sparse high-dimensional sequence associations

Shaolong Cao

Tulane University

1440 Canal St, suit 2027, New Orleans, LA, 70112
504-453-5259
scao2@tulane.edu

ABSTRACT

In deep sequencing data, genomic marker sets of high-dimensional genotypic data and sparse functional variants are quite common. Existing sequencing association tests are incompetent to identify such marker sets and individual causal loci, although they appeared powerful to identify small marker sets with dense functional variants. In deep sequencing studies of admixed individuals, cryptic relatedness and population structure notoriously confound the association analyses of high-dimensional marker sets.

We propose a unified test (uFineMap) for accurately localizing causal loci and a unified test (uHDSet) for identifying high-dimensional sparse associations in deep sequencing genomic data of multi-ethnic individuals. These novel tests are based on scaled sparse linear mixed regressions with Lp (0<p<1) norm regularization. They jointly adjusts for cryptic relatedness, population structure and other confounders to prevent false discoveries and improve statistical power for identifying promising individual markers and marker sets that harbor functional genetic variants of a complex trait.

Under a wide range of simulated scenarios, the proposed tests appropriately controlled Type I error rate and appeared more powerful than several existing prominent methods. We illustrated their practical utilities by the applications to DNA sequence data of Framingham Heart Study for osteoporosis. The proposed tests identified 11 novel significant genes that were missed by the prominent famSKAT and GEMMA. Four out of six most significant pathways identified by the uHDSet have been reported to be related to BMD or osteoporosis in the literature.

BIOGRAPHY

My research interest has been focused on bioinformatics and genomics since September 2011. I have been committed to develop new sparse representation method theoretically and implement it to identify the association between the phenotype and genetic variants data using next generation sequencing data. Given the high dimensionality of genetic variants data, sparse representation has a unique advantage to handle this kind of data. In recent years, sparse representation models act successfully in many research fields.

Attending ACM-BCB 2015 conference is a great honor and opportunity for me to show our results and communicate with peer scientists, especially in bioinformatics and computational biology area. I expect to benefit from some valuable comments and meet more future co-researchers. In addition, I am very anxious to learn new hot topics and knowledge related to rare variants and population genetics from scholars during this prestigious conference. Over last two years, I gave three oral presentations at ACM-BCB 2013 & 2014 and ASHG 2013. The past experiences of ACM-BCB conference definitely impress me and attract me to attend this conference again. The informative feedback was helpful for us to develop journal papers.

Two-dimensional Model of Bipolar PopZ Polymerization in *Caulobacter crescentus*

Minghan Chen

Virginia Tech

302 Broce Dr. Blacksburg, VA
540-581-4852
cmhshirl@vt.edu

ABSTRACT

The asymmetric location of proteins is crucial to the *Caulobacter* cell cycle. At cell division, the landmark protein PopZ is located at the old ends of the newborn cells, and later in the cell cycle PopZ adopts a bipolar pattern in the predivisional cell. The polar localization of PopZ plays a determining role in the intracellular location of certain key cell cycle regulators and in tethering the replicated chromosome to the two ends of the cell. To study this mechanism, we propose a model of spatiotemporal organization in two spatial dimensions, based on a Turing mechanism of pattern formation in coordination with chromosome replication and segregation. Both deterministic and stochastic simulations capture the observed variations in the location and timing of PopZ polymerization.

BIOGRAPHY

My name is Minghan Chen, and I’m a second year graduate student majoring in Computer Science at Virginia Tech. My Ph.D. research is in areas of computational biology and mathematical modeling on biochemical systems. I have been working in the computational biology lab, focusing on the development of multiscale, multiphysics modeling and simulation methods and tools that help biologists to build, simulate and analyze complex biological systems, to simulate system dynamics and to analyze system functions. Particularly, I’m interested in molecule-based modeling and compartment-based modeling on reaction-diffusion systems.

In collaboration with my coworkers, I first built the *Caulobacter* crescentus cell cycle model in two-dimensional space. Both deterministic and stochastic systems were used to model the control mechanism for *Caulobacter* cell cycle, to analyze their interactions and to provide simulation results that can be compared with experimental observations in quantitative details. System parameters, such as reaction rates and initial conditions, play an important role in biology models. I also worked on parameter optimization on a budding yeast cell cycle model. To show the research result raw data, I made a data-based animation on webpage to visualize modeling results. The dynamic visualization tool, which can be applied to various cell type, shows species’ activities, distribution and population over the whole cell cycle.

I believe that the ability to communicate and collaborate with others is important to the success of researchers. Attendance in seminars, forums and other research activities really improves my communication skills and greatly help on my research potential. Besides the research work, traveling is my favorite that gives me adventure, new experience, perspectives.

Mining k-Median Chromosome Association Graphs from a Population of Heterogeneous Cells

Zihe Chen

State University of New York at Buffalo

203 Davis Hall, Buffalo, NY, 14260
zihechen@buffalo.edu

ABSTRACT

Finding the structural pattern from a set of objects is a commonly encountered prototype learning problem in machine learning and pattern recognition. In graph domain, such a structure is called median graph. Existing research has demonstrated that computing an accurate median graph could be rather challenging. In our paper, we present a new technique for mining k-median graphs from a population of heterogeneous cells. Each median graph is a representative structure of chromosome associations of a subset of the cells in the population. Comparing to existing techniques, our technique has several unique advantages. Firstly, it reveals, for the first time, the level of associations (or degree of associations) among the chromosomes. Secondly, it generates multiple median graphs simultaneously, and therefore can be used to handle heterogeneous data. Our technique is based on a number of interesting ideas, such as adaptive sampling, semi-definite programming model, embedding, and local search on uncertain data.

BIOGRAPHY

I am a fourth-year Ph.D. student at the State University of New York at Buffalo. My research interests mainly lie in the areas of Algorithms, Computational Geometry and their applications in Computational Biology and Biomedical Imaging. In the past couple of years, I have been focusing on developing algorithmic tools for determining the internal topological structure of chromosomes and the associations of chromosome territories.

Scalable parallel algorithms for de novo genome assembly

Evangelos Georganas

University of California, Berkeley, USA

2405 Sacramento St. APT D, Berkeley, CA 94702
510-316-4830
egeor@eecs.berkeley.edu

ABSTRACT

A critical problem for computational genomics is the problem of de novo genome assembly: the development of robust scalable methods for transforming short randomly sampled “shotgun” sequences into the contiguous and accurate reconstruction of complex genomes. De novo assembly has been unable to keep pace with the flood of data, due to vast computational requirements and the algorithmic complexity of assembling large scale genomes and metagenomes. We address this challenge by developing HipMer, an end-to-end high performance de novo assembler designed to scale to massive concurrencies. Our work is based on the Meraculous assembler, a state-of-the-art de novo assembler for short reads developed at the Joint Genome Institute. In this talk I will describe the algorithms and the optimization techniques we employed in order to parallelize the various modules of the Meraculous pipeline. Experimental large-scale results on the NERSC Edison Cray XC30 system using human and wheat genomes demonstrate efficient performance and scalability on thousands of cores. HipMer computes the assembly of the human genome in only 8.4 minutes using 15,360 cores of Edison while the original Meraculous code requires 24 hours.

BIOGRAPHY

Evangelos Georganas is a fifth year PhD student in the Computer Science Department at the University of California, Berkeley. He is co-advised by Prof. Katherine Yelick and Prof. James Demmel. He is affiliated with the Berkeley Benchmarking and Optimization group at UC Berkeley and the CLaSS Group at Lawrence Berkeley National Laboratory. His research interests include High Performance Computing and scientific applications. He is currently working on scalable parallel algorithms for genomics and he has also worked on communication-avoiding algorithms. He received his B.S. degree in Electrical and Computer Engineering from National Technical University of Athens, Greece in 2011.

Identification of significant genetic variants via SLOPE, and its extension to Group SLOPE

Alexej Gossmann

Tulane University

6823 St.Charles Ave., 70118, New Orleans
504-615-0758
agossman@tulane.edu

ABSTRACT

The method of Sorted L-One Penalized Estimation, or SLOPE, is a novel sparse regression method for model selection introduced in a sequence of recent papers (Bogdan et al. 2013 and 2014, and Candes et al. 2015). It estimates the coefficients of a linear model that possibly has more unknown parameters than observations. In many settings the SLOPE method is shown to successfully control the false discovery rate (the proportion of the irrelevant among all selected predictors) at a user specified level. In this paper we evaluate its performance on genetic data, and show its superiority over LASSO which is a related and popular method. Often in genetic data sets, group structures among the predictor variables are given as prior knowledge, such as SNPs in a gene or genes in a pathway. Following this motivation we extend SLOPE in the spirit of Group LASSO to Group SLOPE, a method that can handle group structures between the predictor variables, which are ubiquitous in real genetic data. Our simulation results show that the proposed Group SLOPE method is capable of controlling the false discovery rate at a specified level, and its superior variable detection capabilities over Group LASSO.

BIOGRAPHY

Alexej Gossmann is a fourth-year PhD student in the Department of Mathematics at Tulane University in New Orleans. He earned a BS in mathematics at Technische Universitaet Darmstadt in Germany in 2012, and In Spring 2014 he completed a Master’s degree in statistics at Tulane University. Alexej’s research interests lie in statistics and machine learning, and their application in imaging and genomics; his focus includes regularized regression methods for feature selection and prediction, and network or pathway based methods.

A New Method for DNA Sequencing Error Verification and Correction via an On-Disk Index Tree*

Yarong Gu

The University of Michigan, Dearborn, USA, 48128
313-695-6557
yarongg@umich.edu

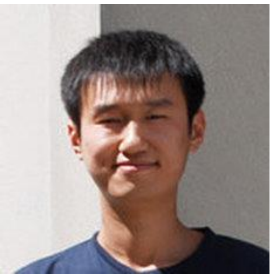
ABSTRACT

Existing sequencing error correction techniques demand large expensive memory space. We introduce a new disk-based sequencing error correction method to solve the problem. The key idea is to utilize a special on-disk index structure, called the BoND-tree, to store and access a large set of k-mers and their associated metadata on disk. With the BoND-tree, a set of special box queries to retrieve the relevant k-mers and their counts are efficiently processed. A comprehensive voting mechanism is adopted to determine and correct an erroneous base in a genome sequence. Experiments demonstrate that the proposed method is quite promising in verifying and correcting sequencing errors in terms of accuracy and scalability.

* This work was sponsored by NSF and done in conjunction with X. Liu, Q. Zhu, Y. Dong, C. T. Brown, and S. Pramanik.

BIOGRAPHY

Yarong Gu is a graduate student pursuing an M.S. degree in the Department of Computer and Information Science at the University of Michigan-Dearborn. She received her bachelor's degree in Computer Science from Southeast University in 2014, and was an exchanged student at UM-Dearborn in her senior year. She continued her graduate study at UM-Dearborn since then, and is currently a research assistant in the Database Research Lab. Her research interests include big data and bioinformatics with an emphysis on exploring genome sequence analysis applications, aiming to combine her background in data management with her research interest in bioinformatics, and to bring new ideas to genome sequence analysis from the point of data science. In addition, she just finished an summer internship in Alibaba Inc., focusing on mining data through the communication between customer and customer services.



A maximum-likelihood approach to absolute protein quantification in mass spectrometry

Chao Ji

Indiana University

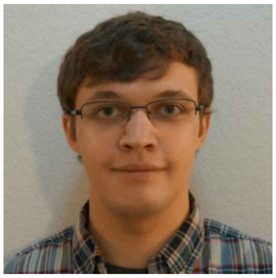
150 S. Woodlawn Avenue, Bloomington, IN 47405
jic@indiana.edu

ABSTRACT

Label-free absolute protein quantification refers to a process of estimating protein abundances in complex biological samples based on the data acquired from a liquid chromatography mass spectrometry (LC-MS) analysis. Most approaches to label-free quantification rely on measuring peak areas from an extracted-ion chromatogram. However, because of the differences in physicochemical properties associated with different peptide ions, observed peak areas in a single experiment are determined not only by peptide abundances, but also the intrinsic biases of analytical platforms. Therefore, accurate modeling of these biases provides an opportunity to developing new computational methods for precise absolute protein quantification. In this work, we developed a new algorithm for absolute quantification of proteins. The approach is based on the concept of peptide response rate, which characterizes the peptide-specific signal detection bias in an LC-MS experiment. We argue that peptide response rate is an intrinsic and reproducible property of peptide ions that can be reliably predicted using non-linear regression and features extracted from the sequence of the parent protein. Protein abundances are estimated using a maximum likelihood model in which the observed peak areas of peptide ions are adjusted using predicted peptide response rates. We evaluate our approach on a large LC-MS dataset as well as simulated data and provide evidence that the ac- curacy of absolute protein quantification is improved when peptide-specific response rates are taken into account.

BIOGRAPHY

Chao Ji is a currently PhD student at Indiana University, School of Informatics and Computing, advised by Prof. Haixu Tang and Predrag Rajvojac. He is interested in applying machine learning and data-driven approaches for solving a wide range of practical problems. In particular, his doctoral work focuses mainly on Mass Spectrometry based computational proteomics, and he has developed a number of methods for addressing peptide identification and protein quantification problems.



Fast and Efficient Compression of High-Throughput Sequencing Reads

Trey Kind

Colorado State University

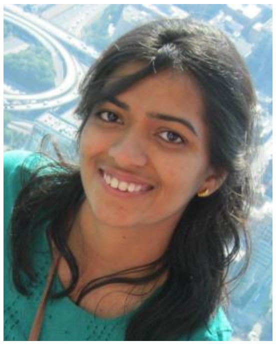
Fort Collins, CO 80526, USA
512-517-8009
tkind94@gmail.com

ABSTRACT

Biological sequence data for one to many individuals from thousands of species has been generated or is currently being generated, resulting in enormous amounts of next generation sequence (NGS) data that must be stored and managed. It has thus become imperative that tailored compression algorithms be developed that exploit the particular redundancy present in NGS data in order to reduce storage costs. We present two LZ-style compression algorithms for short read compression: Faust and Afin. Both methods work without requiring a reference genome and take only the sequence reads as input. We compare our new techniques to state-of-the-art methods using a collection of one billion human sequence reads that were sequenced from a well-studied African male. Our experiments demonstrate that Afin and Faust provide powerful compression while having superior time and memory usage to alternative methods during compression and decompression. Both Faust and Afin are available at https://github.com/tkind94/afin---faust.

BIOGRAPHY

I am a Texas native at Colorado State University studying as a senior in computer science with a minor in math. I currently work with Christina Boucher at Colorado State University as an undergraduate research assistant working in bioinformatics. I have strong interests towards data sciences and its practical applications in other scientific fields. My goals are to pursue graduate school and continue studying data analytics.



Scaling Genomic Sequence Search with Fine Grained Parallelization

Kanak Mahadik

Purdue University

465 Northwestern Avenue, West Lafayette, IN - 47906
765-426-9097
kmahadik@purdue.edu

ABSTRACT

Gene sequencing instruments are producing huge volumes of data, straining the capabilities of current computational algorithms and hindering researchers interested in analyzing the data to improve the human condition. We propose a fine-grained parallelism technique that is suited to the current and emerging workflows for doing alignment of genomics sequences. It achieves higher parallelism (hence speedup, of more than one order of magnitude) and load balancing, while maintaining 100% accuracy.

BIOGRAPHY

Kanak Mahadik is a PhD student in the School of Electrical and Computer Engineering (ECE) at Purdue University, West Lafayette, Indiana. She completed her MS from Purdue in 2012 and her BS from Pune University, India in 2010. After working on performance optimizations at Salesforce.com, she resumed graduate school in 2013 as a Ross Graduate Fellow. Her current research is on developing computational genomics applications and making them run on very large datasets and at very large scales. She is passionate about taking computational achievements and making a difference to the human condition, in close collaboration with biologists and clinicians.

Scalable Algorithms for Building and Clustering Heterogeneous Biological Networks

Paola Gabriela Pesantez Cabrera

ACM, WWCode

1401 NE Merman Drive APT 904, Pullman, WA 99163
509-339-4109
p.pesantezcabrera@wsu.edu

ABSTRACT

The collection of complex experimental data from natural and human-made systems, that continue to evolve, has resulted in large-scale complex networks. How these networks can be constructed, inferred, understood, and controlled has become a really interesting topic of research. Real situations that could be explained using networks include but are not limited to: spreading of viruses, evolution of diseases, biomedical citation networks, social networks, and molecular interaction networks. Additionally, these vast amounts of data come from innumerable, not always compatible, sources; therefore, data mining and clustering from heterogeneous data provide a powerful function to extract information, especially in bioinformatics. Consider the example of a disease-gene-protein tripartite network, containing three types of vertices {diseases, genes, proteins} and up to three types of inter-type edges: disease-gene if a mutation in that gene is implicated for that disease, gene-protein connecting genes to their protein products, and disease-protein edges for proteins implicated for diseases. Here, a heterogeneous “cluster” can help link a subset of genes to a subset of proteins through their implication on a common disease, or alternatively, link two seemingly unrelated diseases by common gene or protein contributors. Our current goals are: a) to build a k-partite network representation using biomedical data from varied sources (e.g., proteins vs. domains, literature corpus vs. keyword/entities), and b) to develop methods for efficient and scalable clustering of these data sets. We are currently exploring the applicability and extension of graph-theoretic community detection techniques and related metrics (e.g., modularity) that would more readily suit the annotation of heterogeneous biological data.

BIOGRAPHY

I was born in Cuenca, Ecuador, in 1985. I received the B.S. degree in Computer Science from the Universidad de Cuenca, Cuenca, Ecuador, in 2009, and the MSc. degree in Computer Science from Washington State University, Washington, in 2013.

Currently, I am part of the Computer Science Ph.D. program at Washington State University under the research supervision of Dr. Ananth Kalyanaraman. I have just completed my second year, finishing my course work. The aim of my dissertation research is to develop large-scale algorithms and software for cluster-based analysis of heterogeneous networks originating from the life science domain. The goal is to incorporate complex, heterogeneous information available into cluster analytics.

I am a member of WSU Society of Women Engineers to promote engineering, science, and math among young women. I am also a volunteer tutor at WSU Louis Stokes for participation Alliance Minority Program to help computer science students to reach their goals.

I received the “Benigno Malo” prize in recognition of academic merit and I graduated with the highest honor from Universidad de Cuenca. I was one of the recipients of the Fulbright Scholarship to study my master degree in the United States. I was awarded a travel grant to attend the ACM International Workshop on Big Data in Life Sciences (BigLS 2014) from University at Buffalo, and I was honored with the EECS Outstanding Teaching Assistant award from Washington State University.

Super Deduper, fast PCR duplicate removal in fastq files.

Kristen R. Petersen

University of Idaho

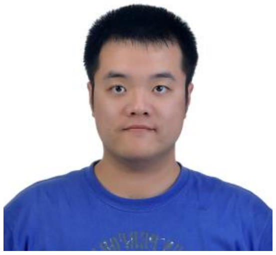
875 Perimeter Drive, Moscow, ID 83844
208-885-2929
pete7861@vandals.uidaho.edu

ABSTRACT

Our goal was to explore the accuracy and utility of identifying and removing polymerase chain reaction (PCR) duplicates from high throughput sequencing (HTS) data using Super Deduper. Super Deduper is a pre-alignment, sequence read based technique developed at the University of Idaho, which examines and uses only a small portion of each read’s sequence in order to identify and remove PCR and/or optical duplicates. Through comparisons with well-known pre- and post-alignment techniques, Super Deduper’s parameters were optimized and its performance assessed. The results conclude that Super Deduper is a viable pre-alignment alternative to post-alignment techniques. Super Deduper is both independent of a reference genome and choice in alignment application, allowing for its use in a greater variety of HTS applications. Super Deduper is an open source application and can be found at https://github.com/dstreett/Super-Deduper.

BIOGRAPHY

Kristen Petersen is a second year graduate student at the University of Idaho (UI). She received her B.S. in Mathematics from Point Loma Nazarene University in San Diego, CA, and is currently pursuing a M.S. in Statistics and Ph.D. in Bioinformatics and Computational Biology at UI. Aware of her interest in applying her analytical skills to solve biology related problems, her undergraduate mentor Dr. Ryan Botts presented her with a summer project that involved the development of a technique to identify novel plasmids based on gene clusters found within known plasmid sequences, which she turned into her honor’s thesis. This past year Kristen validated Super Deduper, an application that her faculty advisor, Dr. Matthew Settles and collaborators, developed at UI. She is currently performing community profiling on soil samples taken from the Priest River Experimental Forest in Northern Idaho to determine microbial and fungal interactions. Outside of research, Kristen enjoys hiking, skiing, wakeboarding, and spending time with friends and family.



Integrated Querying of Disparate Association and Interaction Data in Biomedical Applications

Shi Qiao

Case Western Reserve University

Cleveland, OH, 44106
sxq18@case.edu

ABSTRACT

In biomedical applications, network models are commonly used to represent interactions and higher-level associations among biological entities. Integrated analyses of these interaction and association data has proven useful in extracting knowledge, and generating novel hypotheses for biomedical research. For example, integrated mining of clinical similarity among diseases, known disease-gene associations, and molecular interactions among proteins provide insight on prioritizing candidate disease genes. However, since most datasets provide their own schema and query interface, opportunities for exploratory and integrative querying of disparate data are currently limited. In this study, we capitalize on RDF-based representations of biomedical interaction and association data to develop a querying framework that enables efficient processing and flexible specification of graph template matching queries. The proposed framework enables integrative querying of biomedical databases to discover complex patterns of associations among a diverse range of biological entities, including biomolecules, biological processes, organisms, and phenotypes. Our experimental results on the UniProt dataset show the proposed framework can be used to efficiently process complex queries, and identify biologically relevant patterns of associations that cannot be readily obtained by querying each dataset independently.

BIOGRAPHY

Shi Qiao is a Ph.D. candidate in the Department of Computer Science at Case Western Reserve University. He received his bachelor degree in Computer Science from Nanjing University, China. His current research focuses on Query Processing and Optimization, RDF and Bioinformatics. Meanwhile, Shi's research has been published in leading academic conferences including SIGMOD, VLDB, and BCB.

Genome Compression

Subrata Saha

University of Connecticut, Storrs
860-208-0449
subrata.saha@engr.uconn.edu

ABSTRACT

Nowadays genome sequencing is becoming faster and more affordable. Consequently, the number of complete genomic sequences ranging from human to microscopic organisms is increasing rapidly. As a result, the cost to store, process, analyze and transmit the data is becoming a bottleneck for research and future medical applications. So, the need for devising efficient data compression and data reduction techniques targeting only to compress biological sequencing data is growing by the day. Although there exists a number of standard data compression algorithms, they are not efficient to compress biological data. These algorithms are often criticized to be extravagant as they do not consider some inherent properties of the sequencing data while compressing. To exploit statistical and information-theoretic properties of genomic sequences, we need specialized compression algorithms to effectively compress biological sequencing data.

BIOGRAPHY

Mr. Subrata Saha is a PhD student in Computer Science at University of Connecticut (UConn), Storrs. He completed his B.Sc. in Computer Science and Engineering at Bangladesh University of Engineering and Technology (BUET). Before joining UConn Mr. Saha was a Senior Software Engineer at TigerIT Bangladesh Limited. He arrived at UConn in August 2011 and joined the applied algorithms lab under Professor Sanguthevar Rajasekaran. Mr. Saha mainly focuses on designing and developing efficient algorithms in the fields of bioinformatics and computational biology. His research interest in bioinformatics includes biological data compression, error correction for short reads, spliced reads mapping, structural variation detection and sequence assembly. His research interest also includes machine learning and data mining. Mr. Saha has served as a reviewer in many journals and conferences including BMC Bioinformatics, European Conference on Computational, Biology (ECCB), Parallel Processing Letters (PPL) and IEEE Engineering in Medicine and Biology Society (EMBS). He was also a TPC member of IEEE Symposium on Computers and Communications (ISCC).

A systems biology approach for the identification of significantly perturbed genes

Adib Shafi

Wayne State University

4762 2nd Ave, Apt 204, Detroit, MI 48201
313-421-5590
fj9079@wayne.edu

ABSTRACT

Identifying the list of genes that are involved in the mechanisms differentiating two phenotypes is a crucial step in the analysis high-throughput gene expression experiments. Although in the last decade several approaches have been developed in order to address this challenge, these approaches share a number of important limitations. Even the most widely used approaches fail to incorporate information about known interactions among genes, and they often fail to yield reproducible results across similar experiments, both in terms of the list of genes and in terms of the list of mechanisms that are related to those genes. Here we propose a novel systems biology approach able to i) identify the genes that are involved in a biological mechanism relevant to the condition in analysis and ii) yield reproducible results across multiple data sets related to the same condition, by using gene expression levels and existing knowledge of the interactions among genes. We apply our method on four data sets describing two conditions, and we compare our approach with the classical approach of identifying relevant genes based on their differential expression and p-value. The results show that our method is better at identifying genes that are involved in the mechanisms relevant to the phenotype in analysis, as well as producing more consistent results across data sets describing the same biological condition.

BIOGRAPHY

Adib Shafi is currently a second year PhD student at the Department of Computer Science at Wayne State University in Detroit, Michigan. He is a member of Intelligent Systems and Bioinformatics Laboratory (ISBL). His research is focused on building new methodologies to find biomarkers, signaling pathway analysis, data integration, finding mechanism for newly repurposed drugs, etc. Before stating his PhD, he worked as a Software Developer for a company in Bangladesh for around 2 years. He earned his Bachelor’s degree in Computer Science from Military Institute of Science and Technology (MIST), Bangladesh in 2011.

Scalable Clustering in Genetic Mapping

Veronika Strnadová-Neeley

LBNL, UCSB

Harold Frank Hall, Rm 5110, University of California, Santa Barbara, CA 93106-5110
veronika@cs.ucsb.edu

ABSTRACT

Next generation sequencing technologies have produced a flood of information invaluable to genetic mapping research. However, the rapid growth of this data has outpaced our ability to analyze it effectively. In this talk, I will present an overview of my recent work, which highlights the utility of scalable clustering algorithms in accurately and efficiently analyzing large-scale genetic mapping data. This work has not only contributed to the scalability of genetic mapping software and the ability to assemble challenging genomes such as wheat, but to new theoretical approaches for clustering and reducing large-scale data to a more manageable size. I will explain the ideas which have proven successful in handling the large-scale and characteristically noisy and incomplete genetic mapping data. I will also mention possible avenues for future research, extending the methods developed for large-scale genetic mapping to other application domains.

BIOGRAPHY

I am a Ph.D. Candidate with a Computational Science and Engineering emphasis at UC Santa Barbara, working with adviser John R. Gilbert in the Combinatorial Scientific Computing Lab. For the past few years I have been collaborating with researchers at Lawrence Berkeley National Lab, UC Berkeley and the Joint Genome Institute to design scalable algorithms for genetic mapping. Broadly, my research interests include scalable clustering algorithms, bioinformatics, graph algorithms, linear algebra and scientific computing. I completed my BS in applied mathematics at the University of New Mexico.

Engineering High-performance Parallel Algorithms with Applications to Bioinformatics

Jesmin Jahan Tithi

Stony Brook University

700 Health Sciences Drive, Chapin C 1034AY (#350), Stony Brook 11790
631-428-7669
jtithi@cs.stonybrook.edu

ABSTRACT

Bioinformatics is considered to be the next Google industry and predicted to be a $13 billion market by 2020. Bioinformatics is currently seeing a paradigm shift due to recent changes in parallel architectures and programming platforms, smart mobile and wearable devices, faster generation of massive data, big-data analytics, and data-driven design and discovery (e.g., drug). This paradigm shift brings new challenges, making bioinformatics a field of opportunities for algorithm research.

Recent developments in computer architecture have emphasized parallelism over increased clockspeed. Taking advantage of these developments requires designing algorithms that parallelize well on diverse parallel architectures. In my research, I show how to take advantage of several algorithm design techniques and data structures to harness modern heterogeneous parallel architectures to efficiently solve algorithms used in Bioinformatics. The main goal while designing algorithms is to achieve high-performance in terms of runtime and scalability. Other desirable goals include energy-efficiency, portability, and adaptivity. My current research focuses on designing highly efficient parallel algorithms to solve several dynamic programming (DP) and graph problems with applications to bioinformatics, and the problem of computing various molecular energetics terms required for molecular dynamics simulations, targeting a range of available parallel architectures including multicores, manycores, and special purpose accelerators.

BIOGRAPHY

Jesmin Jahan Tithi is a Ph.D. candidate at the State University of New York at Stony Brook. Her broad research focus is enabling efficient algorithms on heterogeneous parallel platforms in terms of runtime, scalability, portability, and energy-consumption. Jesmin's dissertation particularly focuses on engineering high-performance parallel algorithms with applications to bioinformatics, targeting multicores, manycores, special-purpose accelerators, and clusters of multicores. Her research has been published in IEEE/ACM conferences such as IPDPS, ISPASS, PPoPP, and ACM-BCB, and her paper “Exploiting Spatial Architectures for Edit Distance Algorithms” was included in the best paper session at ISPASS-2014.

Incorporating Known SNPs in Short Read Aligners

Saima Sultana Tithi

Virginia Tech

KWII, 2202 Kraft Dr., Blacksburg, VA 24060
540-750-6312
saima5@vt.edu

ABSTRACT

Today's high-throughput sequencing technologies like Next Generation Sequencing (NGS) produce a large number of short reads (short DNA sequences) from random locations of the genome. Aligning these short reads to the appropriate position in the reference genome with less error is a computationally significant challenge. Some of the existing alignment tools include Bowtie, Bowtie2, BWA, BLAST, and SNAP which have been proved to be efficient. However, none of these tools consider known genomic variants while sequencing the reads. To address this problem, recently we developed SNPwise, a short read aligner that takes into account known SNP variations provided by the database or the user while aligning the reads. By considering known SNPs during the alignment, the accuracy of the alignment is improved as well as the total number of mapped reads is increased. One paper and one poster based on this research work have been published in BICoB’15 and ISBRA’15 respectively.

BIOGRAPHY

I, Saima Sultana Tithi, am an enthusiastic PhD student with a goal to pursue a research based career. I received Bachelor of Science in Computer Science and Engineering from Bangladesh University of Engineering and Technology (BUET), the top ranked engineering school in Bangladesh. Currently I am working in the “Computational Biology and Bioinformatics Lab” at Virginia Tech under the supervision of Dr. Liqing Zhang. My research interests include Evolutionary and Comparative Genomics, Bioinformatics, and Computational Biology.

I have just started 2nd year of my PhD. Currently, I am working on short read mapping of human genome. We developed SNPwise, a short read aligner that while aligning the reads, takes into account known SNP variations documented in dbSNP. We have published one paper and one poster on this project in BICoB’15 and ISBRA’15 respectively. Besides, I have also worked on the mathematical modeling of Ebola outbreak and this work has been published in the ACM-BCB’15. Moreover, one of my current projects includes mapping of bisulfite-treated reads. Additionally, I am working on different types of assemblers for the metagenomic sequencing data as well.

I am from Bangladesh, a beautiful and picturesque country located in South Asia. I am the only child of my parents. My father is a journalist and my mother is government employee. My husband, Mohammad Shabbir Hasan, is a PhD student in the Department of Computer Science at Virginia Tech. My hobbies are reading literature and watching movies.

Computational Inference of Expression and Alternative Splicing using High Throughput Data

Kun Wang

University of Maryland

3104E Biomolecular Sciences Bldg
301-405-8218
kunwang0331@gmail.com

ABSTRACT

Gene expression is the process in which the genetic information is used and synthesized into a functional product. It determinates cell’s phenotype to a large extent. In addition, most genes function in complex regulatory networks and exhibit correlated expression patterns. Currently, RNA-seq deep sequencing has become a widely used approach to quantify transcripts expression level due to its higher resolution and more precise estimation. Alternative splicing is a regulated process by which genes can encode multiple mature mRNAs, which would contribute to both gene expression and protein complexity. As we know over 95% genes with multiple exons will undergo alternative splicing. In many cases, mis-regulation of splicing will increase the probability to generate diseases, such as HGPS (Hutchinson-Gilford progeria syndrome) which is a rare and devastating accelerated aging disease. Our overall goal is to develop methods for the studying coordinated gene expression changes in diseases and mechanisms of alternative splicing, with special emphasis to HGPS.

BIOGRAPHY

Kun Wang received the BSc degree in computer science from Shandong University in China in 2008 and the MSc degree in computer science from Arkansas State University in 2011. Currently, he is working towards the PhD degree in computational biology and bioinformatics program in the University of Maryland at College Park. His research interests include comparative genomics, gene regulation, splicing regulation, and normal aging.

Using Motion Planning to Rank Ligand Binding Affinity

Hsin-Yi (Cindy) Yeh

Texas A&M University

3112 TAMU College Station TX 77843
979-847-8835
hyeh@cse.tamu.edu

ABSTRACT

In the drug discovery process, pharmaceutical companies have to screen many drug (or ligand) candidates to find the most promising ones for trial. This process is very costly and attention as turned to computational approaches to predict binding affinity to the desired target protein. In this work, we develop a computational tool for ranking ligand binding affinity that uniformly samples ligand conformations over the target protein's surface and analyzes the resulting set to compute an affinity ranking. Experiments on one target protein show that our method is able to correctly rank different ligands for the target protein as determined by experimental data. Our method is a promising technique and potential cost-saving tool for pharmaceutical companies to narrow the search for good drug candidate.

BIOGRAPHY

Hsin-Yi (Cindy) Yeh is a Ph.D. candidate in the Department of Computer Science and Engineering at Texas A&M University working with Dr. Nancy M. Amato in Parasol Laboratory. Her research focus is on a uniform sampling framework for randomized motion planning algorithms and their application to problems in computational biology, particularly in protein folding and ligand binding.

Simultaneous Optimization of Both Node and Edge Conservation in Network Alignment via WAVE

Joseph Crawford

University of Notre Dame

Notre Dame, IN 46556
678-650-5062
jcrawfo7@nd.edu

ABSTRACT

Network alignment can be used to transfer functional knowledge between conserved regions of different networks. Existing methods use a node cost function (NCF) to compare nodes across networks and an alignment strategy (AS) to find high-scoring alignments with respect to total NCF over all aligned nodes (or node conservation). Then, they evaluate alignments via a measure that is different than node conservation used to guide alignment construction. Typically, one measures edge conservation, but only after alignments are produced. Hence, we recently directly maximized edge conservation while constructing alignments, which improved their quality. Here, we aim to maximize both node and edge conservation during alignment construction to further improve quality. We design a novel measure of edge conservation that (unlike existing measures that treat each conserved edge the same) weighs conserved edges to favor edges with highly NCF-similar end-nodes. As a result, we introduce a novel AS, Weighted Alignment VotEr (WAVE), which can optimize any measures of node and edge conservation. Using WAVE on top of well-established NCFs improves alignments compared to existing methods that optimize only node or edge conservation or treat each conserved edge the same.

BIOGRAPHY

My name is Joseph Crawford. I am a recipient of the Deans' Fellowship and I am currently a third year Ph.D student at the University of Notre Dame under the Department of Computer Science and Engineering. I am a member of the Complex Networks (CoNe) Lab and I am currently doing research under Dr. Tijana Milenkovic. In this lab, we develop graph theoretic and computational approaches for network mining, and apply these methods to real-world networks.

Machine Learning for Data Analysis in Social Health Networks

Hao Wang

University of Oregon

1202 University of Oregon, Eugene, Oregon, 97403, USA
541-346-1380
csehao@cs.uoregon.edu

ABSTRACT

We introduce an ontology-based Restricted Boltzmann Machine (ORBM) model for human behavior prediction in health social networks. A bottom up algorithm is proposed to design the deep learning architecture from the SMASH ontology. Such deep learning architecture is used later to incorporate self-motivation, social influences, and environmental events together in a human behavior prediction model. Experiments conducted on both real and synthetic data from health social networks have shown the tremendous effectiveness of our approach compared with conventional methods.

BIOGRAPHY

I'm a fourth year PHD student working with Professor Dejing Dou at Computer and Information Science, University of Oregon since 2011. I obtained BS. and MS. degree in Electrical Engineering from Fudan University in 2006 and 2009 respectively. My current research interest includes ontology-based data-mining and ontology based deep learning. In the past four years, I have been actively involved in projects with knowledge engineering, data mining and health informatics. My past publications have revealed crucial association and correlation results in the health social network project, the SMASH project. I have involved in and made major contributions in the design of the SMAHS ontologies. The SMASH ontologies have been submitted to the National Center for Biomedical Ontology (NCBO) in March 2015. I'm also the co-author of the paper titled "Ontology-based Deep Learning for Human Behavior Prediction in Health Social Networks" which has been accepted to the ACM-BCB 2015 as a full paper.

Medical information retrieval: Challenges and Solutions

Yanshan Wang

Mayo Clinic

200 1st ST SW, Rochester, MN 55905
507-293-1382
Wang.Yanshan@mayo.edu

ABSTRACT

With the rapid adoption of electronic health records (EHRs) as a result of HITECH act of American Recovery and Reinvestment Act (ARRA), medical information retrieval (MIR) has shown promise for retrieving valuable information. In order to retrieve ad hoc cohorts from clinical text for physicians, we are funded by NLM to develop a layered language model. We also developed a topical information retrieval model to retrieve research articles for a systematic review. In addition to clinical text, there is growing number of healthcare forums and websites such as WebMD, HealthBoards, etc. with valuable contents on patients’ healthcare concerns. Thus, we studied healthcare question answer systems by retrieving semantically similar questions that were previously posted and answered on a healthcare forum. An ensemble model is proposed to combine and rank the results of keyword-based and topic-based search. The practical experiments have indicated the effectiveness of our approaches for users.

BIOGRAPHY

Yanshan Wang, Ph.D. is a postdoctoral research fellow in the Department of Health Sciences Research at Mayo Clinic under the supervision of Dr. Hongfang Liu. His work is centered on developing and integrating innovative machine learning approaches to clinical information retrieval. Recently, he has been working on an R01 NLM-funded project, which aims to accurately retrieve ad hoc cohorts from clinical data repository. In addition to the R01 project, he also leverages statistical models to help physicians systematically review healthcare disparities from the literature. Prior to joining Mayo Clinic in 2015, Yanshan received his Ph.D. in Management Engineering from Korea University. His doctoral thesis emphasized on theoretical analysis of indexing and ranking methods in information retrieval. He discovered that an ensemble model combining with a number of ranking methods could enhance the retrieval performance considerably. This phenomenon, though being commonly observed in the task of classification, has rarely been known in IR. He has brought this work to fruition with a journal paper in the Journal of American Society for Information Science and Technology, a highly reputable journal in the area of IR. Recently, he has submitted a new result on this topic to the IEEE transactions on Knowledge and Data Engineering.

Discovering De Facto Diagnosis Specialties

Aston Zhang

University of Illinois at Urbana-Champaign

201 N Goodwin Ave, Urbana, IL
217-979-6030
lzhang74@illinois.edu

ABSTRACT

In health care institutions, medical specialty information may be lacking or inaccurate, in part because there is no official code to express such specialties. Diagnosis histories offer information on which medical specialties may exist in practice, regardless of whether they have official codes. We refer to such specialties that are predicted with high certainty by diagnosis histories de facto diagnosis specialties. The objective of our research is to discover de facto diagnosis specialties under a general discovery–evaluation framework. Specifically, we employ a semi-supervised learning model (based on heterogeneous information network analysis) and an unsupervised learning method (based on topic modeling) for discovery. We further employ four supervised learning models for evaluation. We use one year of diagnosis histories from a major medical center, which consists of two data sets. One is fine-grained and has diagnoses assigned to 41,603 patients that are accessed by 2,504 medical service providers. The other is general and has diagnoses assigned to 291,562 patients that are accessed by 3,269 medical service providers. The semi-supervised learning model discovers a specialty for Breast Cancer on the fine-grained data set; while the unsupervised learning method confirms this discovery and suggests another specialty for Obesity on the larger general data set. The evaluation results reinforce that these two specialties can be recognized accurately by supervised learning models in comparison with 12 common diagnosis specialties defined by the Health Care Provider Taxonomy Code Set.

BIOGRAPHY

Aston is a 3rd-year Ph.D. Candidate at University of Illinois at Urbana-Champaign, supervised by Carl A. Gunter and Jiawei Han. His research interests include data mining, machine learning, and privacy. He studies healthcare and query log data. He interned at Microsoft Research and Yahoo Labs.

Session Chair

May Dongmei Wang, Ph.D.

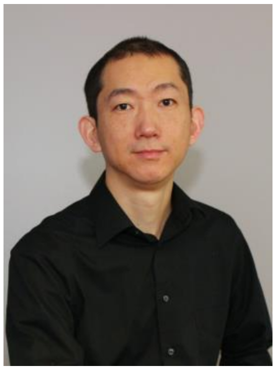
Associate Professor, The Wallace H. Coulter Joint Department of Biomedical Engineering
Kavli Fellow, Georgia Research Alliance Distinguished Cancer Scholar, Fellow of AIMBE
Georgia Institute of Technology and Emory University

UA Whitaker Bldg. Suite 4106, 313 Ferst Dr., Atlanta, GA 30332-0535, USA
404-385-2954
maywang@bme.gatech.edu

BIOGRAPHY

Dr. May D. Wang is an Associate Professor in the Joint Department of Biomedical Engineering of Georgia Tech and Emory. She is a Kavli Fellow, a Georgia Research Alliance Distinguished Cancer Scholar, and a Fellow of The American Institute for Biological and Medical Engineering (**FAIMBE**). She serves as Co-Director of Georgia Tech Biomedical Informatics Program in Atlanta Clinical and Translational Science Institute (ACTSI), Co-Director of Georgia-Tech Center for Bio-Imaging Mass Spectrometry, and Biocomputing and Bioinformatics Core Director in Emory-Georgia-Tech Cancer Nanotechnology Center. She is with Georgia Tech IBB, IPaT, and Emory Winship Cancer Institute. Prof. Wang’s research is in Biomedical Big Data analytics with a focus on Biomedical and Health Informatics (**BHI**) for Personalized/Predictive Health. Her research includes high throughput NGS and -omic data mining for clinical biomarker identification, bionanoinformatics, pathological imaging informatics for clinical diagnosis, critical and chronic care informatics for improving healthcare outcome, and systems modeling. Prof. Wang published 190+ peer-reviewed articles and is the corresponding/co-corresponding author for publications in IEEE/ACM Transactions on Computational Biology and Bioinformatics (**TCBB**), Journal of American Medical Informatics Association (**JAMIA**), Journal of Biomedical and Health Informatics (**J-BHI**), Briefings in Bioinformatics, BMC Bioinformatics, Journal of Pathology Informatics, Proceedings of The IEEE, Proceedings of National Academy of Sciences (**PNAS**), Annual Review of Medicine, Nature Protocols, Circulation Genetics, and Nanomedicine etc. She has led RNA-data analytics within FDA-led Sequencing Consortium (**SEQC**) of **MAQC-III.**

Prof. Wang serves as ACMBCB’2015 conf. co-chair, and has acquired NSF travel grant to sponsor students and young professionals from US-based institutions to attend ACMBCB’2015 in Atlanta, GA. She also serves as a Steering Committee member for **IEEE/ACM TCBB**, Senior Editor for **IEEE J-BHI**, an Associate Editor for IEEE Transactions on Biomedical Engineering (**TBME**), and an Emerging Area Editor for **PNAS**. Dr. Wang is an IEEE-EMBS 2014-2015 Distinguished Lecturer. In addition, Dr. Wang has devoted to the training of young generation of data scientists and engineers, and is a recipient of Georgia-Tech’s Outstanding Faculty Mentor for Undergraduate Research.



Session Chair

Zhaohui (Steve) Qin, Ph.D.

Associate Professor, Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University

1518 Clifton Rd, Atlanta, GA 30322, USA
404-712-9576
zhaohui.qin@emory.edu

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 a 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 from 2000 to 2003, 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. 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 and resequencing. Dr. Qin also actively collaborates with biomedical scientists and clinicians on various projects that utilize next-generation sequencing technologies to study cancer genomics. Dr. Qin has published more than 80 peer-reviewed research papers covering statistics, bioinformatics, statistical genetics and computational biology.

Session Chair

Po-Yen Leo Wu

Georgia Institute of Technology

313 Ferst Drive, Atlanta, GA 30332
404-385-5059
pwu33@gatech.edu

BIOGRAPHY

Po-Yen Wu earned his B.S. degree in electrical engineering from National Taiwan University in Taiwan in 2008 and his M.S. degree in electrical and computer engineering from the Georgia Institute of Technology in 2011. He is currently a graduate research assistant pursuing a Ph.D. degree in the School of Electrical and Computer Engineering at the Georgia Institute of Technology. His research interests include developing novel methodologies to improve the data analysis pipeline for RNA-sequencing data; exploring data-mining techniques for extracting features from electronic health records; and integrating knowledge from RNA-sequencing data and electronic health records to improve the prediction performance of clinical endpoints. For each semester from 2009 to 2013, he has consecutively received “President’s Fellowships,” which recognized his exemplary levels of scholarship and innovation. Moreover, he received the “Outstanding ECE Graduate Teaching Assistant Award” in 2011 in recognition of his excellent performance as a teaching assistant.