NSF Sponsored Student Research Forum

Location: Room 225

Co-Chairs
Professor May D. Wang, Georgia Tech and Emory University
Professor Jaroslaw Zola, University at Buffalo, SUNY

Student Co-Chair
Ying Sha, Georgia Tech

Schedule

<table>
<thead>
<tr>
<th>Time</th>
<th>Presentations</th>
</tr>
</thead>
<tbody>
<tr>
<td>4:00–4:05 pm</td>
<td>NSF Travel Fund Awardee Forum Chairs and NSF Program Directors</td>
</tr>
<tr>
<td>4:05–4:09 pm</td>
<td>Md. Ashad Alam Robust Kernel Canonical Correlation Analysis Toward Imaging Genetics Data</td>
</tr>
<tr>
<td>4:09–4:13 pm</td>
<td>Minghan Chen Hybrid Method for Stochastic Modeling and Simulation of Reaction-Diffusion System</td>
</tr>
<tr>
<td>4:13–4:17 pm</td>
<td>Olivia Choudhury HAPI-Gen: Highly Accurate Phasing and Imputation of Genotype Data</td>
</tr>
<tr>
<td>4:17–4:21 pm</td>
<td>Thomas Hahn Advanced Feature-Driven Disease Named Entity Recognition Using Conditional Random Fields</td>
</tr>
<tr>
<td>4:21–4:25 pm</td>
<td>Edward W Huang PaReCat: Patient Record Subcategorization for Precision Traditional Chinese Medicine</td>
</tr>
<tr>
<td>4:25–4:29 pm</td>
<td>Priyanka Ghosh A Fast Sketch-based Assembler for Genomes</td>
</tr>
<tr>
<td>4:29–4:33 pm</td>
<td>Kishlay Jha Mining Novel Knowledge from Biomedical Literature using Statistical Measures and Domain Knowledge</td>
</tr>
<tr>
<td>4:37–4:41 pm</td>
<td>Jonathan Kho Disease Similarity using Large Graphs</td>
</tr>
<tr>
<td>4:41–4:45 pm</td>
<td>Eunji Kim Model-based study of the Effectiveness of Reporting Lists of Small Feature Sets using RNA-Seq Data</td>
</tr>
<tr>
<td>4:45–4:49 pm</td>
<td>Alexey Markin Manhattan Path-Difference Median Trees</td>
</tr>
<tr>
<td>4:49–4:53 pm</td>
<td>Tatiana Maximova Sample-based Models of Protein Structural Transitions</td>
</tr>
<tr>
<td>4:53–4:57 pm</td>
<td>Jucheon Moon Robinson-Foulds Median Trees: A Clique-based Heuristic</td>
</tr>
<tr>
<td>4:57–5:01 pm</td>
<td>Rahul Nihalani Utilizing paired reads to improve Genome Assembly quality</td>
</tr>
<tr>
<td>5:01–5:05 pm</td>
<td>Paola Gabriela Pesantez-Cabrera Detecting Communities in Biological Bipartite Networks</td>
</tr>
<tr>
<td>5:05–5:09 pm</td>
<td>Qiuling Suo Risk Factor Analysis Based on Deep Learning Models</td>
</tr>
<tr>
<td>5:13–5:17 pm</td>
<td>Subrata Saha Genome-wide Association Study</td>
</tr>
<tr>
<td>5:17–5:21 pm</td>
<td>Aisharjya Sarkar Counting independent motifs in probabilistic networks</td>
</tr>
<tr>
<td>5:21–5:25 pm</td>
<td>Ying Sha A Novel Temporal Similarity Measure for Patients Based on Irregularly Measured Data in Electronic Health Records</td>
</tr>
<tr>
<td>5:25–5:29 pm</td>
<td>Michael Wyatt Development of a Scalable Method for Creating Food Groups Using the NHANES Dataset and MapReduce</td>
</tr>
<tr>
<td>5:29–5:33 pm</td>
<td>Haotian Xu Text Classification with Topic-based Word Embedding and Convolutional Neural Networks</td>
</tr>
<tr>
<td>5:33–5:37 pm</td>
<td>Hasini Yatawatte Automated Capture of Naturalistic Child Vocalizations for Health Research</td>
</tr>
</tbody>
</table>
Robust Kernel Canonical Correlation Analysis
Toward Imaging Genetics Data
Md. Ashad Alam, PhD.
Postdoc Fellow
Department of Biomedical Engineering
Tulane University, New Orleans, LA 70118, USA
Cellphone: 504-505-9198
Email: malam@tulane.edu

ABSTRACT

Methods based on the gene are more effective than the methods based on a single SNP. Recent years, the kernel canonical correlation analysis (Classical kernel CCA) based U statistic (KCCU) has been proposed to detect the nonlinear relationship between genes. To estimate the variance in KCCU, they have used resampling based methods which are highly computationally intensive. In addition, classical kernel CCA is not robust to contaminated data. We, therefore, propose a method based on influence function to estimate the variance of the KCCU and a nonparametric robust KCCU method based on robust kernel CCA, which is designed for contaminated data and less sensitive to noise than classical kernel CCA. We investigate the proposed methods to synthesized data and imaging genetic data set. Based on gene ontology and pathway analysis, the synthesized and genetics analysis demonstrate that the proposed robust method shows the superior performance over the state-of-the-art methods.

BIOGRAPHY

Dr. Md. Ashad Alam is a postdoc fellow at Department of Biomedical Engineering in Tulane university, where he has developed noble robust kernel methods for multi-view data analysis, especially imaging genetics data. For his work on statistical machine learning, received his PhD. in Statistical Science under the supervision of Professor Dr. Kenji Fukumizu in September 2014 at the Institute of Statistical Mathematics (ISM), Tokyo, Japan. He achieved his B.Sc. (Hons) and M.Sc. degrees in Statistics in 2004 and 2005, respectively from the Department of Statistics, University of Rajshahi, Bangladesh. He was a lecturer from September 2005 to January 2008, an assistant professor from February 2008 to August 2013 and he was promoted to Associate Professor in September 2013 at the Department of Statistics, Hajee Mohammad Danesh Sciences and Technology University, Bangladesh. His research interests are: theoretical and computational aspects of statistical machine learning, and robust statistics for multi-view and topological data analysis in Biomedical.
Hybrid Method for Stochastic Modeling and Simulation of Reaction-Diffusion System

Minghan Chen
Virginia Tech
208 Turner St NW, Blacksburg, VA
540-200-7960
cmhshirl@vt.edu

ABSTRACT

Stochastic simulation of reaction-diffusion systems presents great challenges for spatiotemporal biological modeling and simulation. One widely used framework for stochastic simulation of reaction-diffusion systems is reaction diffusion master equation (RDME). Previous studies have discovered that for the RDME, when discretization size approaches zero, reaction time for bimolecular reactions in high dimensional domains tends to infinity (Hellander, 2012). In the study (Chen, 2016), we demonstrate that in the 1D domain, highly nonlinear reaction dynamics given by Hill function may also have dramatic change when discretization size is smaller than a critical value. Our analysis reveals that the switch-like Hill dynamics reduces to a linear function of discretization size when the discretization size is small enough. Moreover, we proposed several methods (smoothing over space, fixed length smoothing over space and a hybrid method), to correctly (under certain precision) simulate Hill function dynamics in the microscopic RDME system.

BIOGRAPHY

My name is Minghan Chen, and I'm a third 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 deterministic, stochastic 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 hybrid ODE/SSA modeling on reaction-diffusion systems.
In my previous study, I 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. 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.

System parameters, such as reaction rates and initial conditions, play an important role in biology models. I finished a project on parameter optimization on a budding yeast cell cycle model. Currently, I’m more focused on the hybrid ODE/SSA method for reaction-diffusion systems due to the inefficiency of stochastic simulation. Also, hybrid method can correct the nonlinear reactions dynamics in spatial models.
HAPI-Gen: Highly Accurate Phasing and Imputation of Genotype Data

Olivia Choudhury

Ph.D. Candidate,
Department of Computer Science and Engineering,
University of Notre Dame, IN 46556, USA

247a FitzPatrick Hall of Engineering,
University of Notre Dame, IN 46556, USA

765 409 7837
ochoudhu@nd.edu

ABSTRACT

High-throughput sequencing is continuously generating large volumes of genotype data. Although haplotype phasing is an effective means to deal with large numbers of variants in analyses such as genome-wide association studies (GWAS), the quality of phasing is degraded by missing genotypes. Most existing imputation tools [1, 2] rely on reference genotype panels and a linear order of markers in the form of a physical or genetic map. A large number of genomics projects, however, do not have access to these resources. In this work, we introduce HAPI-Gen, a Highly Accurate Phasing and Imputation tool for Genotype data that does not require reference genotype panels or the global order of markers, thereby filling an important need for improving phasing in less studied organisms. Other major advantages of HAPI-Gen include low runtime, reduced memory footprint, and easy parallelization. We test HAPI-Gen on the malaria parasite Plasmodium falciparum and three plant datasets of grape, apple, and maize. For varying data sizes and proportions of missing genotypes, HAPI-Gen consistently performs better than three of the leading tools in terms of accuracy, runtime, and memory usage.
References:


BIOGRAPHY

Olivia Choudhury is a Ph.D. candidate at the Department of Computer Science and Engineering, University of Notre Dame, USA. Her research areas include bioinformatics, cloud computing, predictive modeling, and reinforcement learning. Her thesis focuses on solving the two key challenges in bioinformatics - (i) expediting the rate of large-scale genome data analysis, and (ii) improving the fidelity of emerging genome data. To address the first problem, she has developed parallel and high-throughput data analysis frameworks that allow optimized utilization of computational resources. For the second problem, she is interested in exploring hidden Markov model and reinforcement learning to impute missing values and correct erroneous data in applications like haplotype phasing and third generation sequencing (PacBio sequencing), respectively.

As a research intern at the Broad Institute of MIT and Harvard in summer 2016, she developed analysis tools to study the genomes of the mosquito *Aedes aegypti*, responsible for Zika outbreak, and the malarial parasite *Plasmodium falciparum*. This study will help in understanding their role in disease transmission and determining potential measures of disease control. As a research intern at IBM Watson in summer 2015, she designed automated cloud-based infrastructures for big data analysis.

She received the EIGH fellowship from the ECK Institute for Global Health for the years 2015 – 2017. Currently, she serves as the President of Graduate Society of Women Engineers (SWE) and co-Vice President of the Graduate Student Union (GSU) at the University of Notre Dame.
PaReCat: Patient Record Subcategorization for Precision Traditional Chinese Medicine

Edward W Huang
University of Illinois at Urbana-Champaign
1107 W. Green Street
Apartment 122
Urbana, Illinois 61801
714-261-6687
ewhuang3@illinois.edu

ABSTRACT

Traditional Chinese medicine (TCM), a style of medicine widely used in China for thousands of years, can complement modern western medicine by taking personalization as the core principle of clinical practice. A fundamental task in TCM, particularly important for achieving effective precision medicine, is to subcategorize patients with a general disease into groups corresponding to variations of that disease. In this paper, we conduct the first study of the problem of subcategorizing electronic patient records in TCM. While the general problem of subcategorization can be solved using basic clustering algorithms, accommodating variations in symptoms and herb prescriptions of TCM patient records when computing patient similarity is a major technical challenge that has yet to be addressed. To tackle this problem, we propose to learn inexact matchings of both symptoms and herbs from a TCM dictionary of herb functions by using an embedding algorithm. Our hypothesis is that the prior knowledge of herb-symptom associations in the TCM dictionary can be used to discover latent relationships among comorbid symptoms and functionally similar herbs, thereby improving the quality of subcategorization. We performed extensive experiments on large-scale real-world datasets. As expected, our approach leads to more accurate matchings between patient records than baseline approaches, and thus better subcategorization results. We also show that the proposed algorithm can be used immediately in multiple clinical applications, such as retrieving similar patients as well as discovering two special TCM cases: similar symptoms treated by different herbs and different symptoms treated by similar herbs.

Subcategorizations can help inexperienced doctors learn pattern analysis by providing specific medical cases. Most importantly, patient record subcategorization can help doctors view all possible diagnoses for their patients, reducing the chance of misdiagnosis. Misdiagnoses across all fields not only present ongoing risks to the health and safety of patients, but also cost the United States roughly $750 billion.
annually. Furthermore, doctors can cross-reference their records with existing databases to obtain comparisons helpful in determining trends in prescriptions and treatment.

**BIOGRAPHY**

I am currently a second year PhD student at the University of Illinois at Urbana-Champaign. My advisor is ChengXiang Zhai in the database and information systems group. I work primarily in data mining and information retrieval for the purpose of medical data analysis. I graduated in 2015 from the California Institute of Technology with a B.S. with honor in Computer Science. Previously, I have done research in computational chemistry, internal clock synchronization, system awareness platforms, and information retrieval. Current projects in addition to the published work include drug response prediction, gene clustering, and drug-pathway correlation analysis. I am an NSF GRFP recipient and hold a NASA copyright for my work on Precision Time Protocol.
Advanced Feature-Driven Disease Named Entity Recognition Using Conditional Random Fields

Thomas Hahn
University of Arkansas at Little Rock
Little Rock, AR, 72204
(318) 243 3940
TFHahn@UALR.edu

ABSTRACT

Named entity recognition (NER) is crucial for biomedical text mining because it enhances information retrieval and knowledge discovery. Extensive research has been conducted using semantic type categories, such as “DNA”, “RNA”, “proteins”, “genes” and other named entities. However, especially human disease NER has not been sufficiently explored yet partly because traditional machine learning approaches still lack the precision needed for integrating orthographic, contextual, linguistic, sentence and token level features. We propose a new method for human disease NER based on sentence and token level features using Conditional Random Field (CRF). Using our new feature set consisting of orthographic, contextual, affixes, bigrams, part of speech and stem based features we could raise the maximum F-score for the training set to 94% when applying 10-fold cross validation for semantic labeling of the NCBI disease corpus. For the testing and development corpus our model could raise the F score to 88% and 85% respectively.

BIOGRAPHY

Thomas Hahn earned two bachelor degrees, one in biology and the other in chemistry and a master’s degree in biology from Louisiana Tech University in Ruston, Louisiana. Since his visual disability prevented him from conducting wet-lab research within the time constraints of his biomedical PhD program he had to switch his career path. His vocational rehabilitation agency enrolled him in a ten months vocational training program, where he learned the basics of programming. This gave him the foundation to earn a master’s degree in bioinformatics at the University of Arkansas at Little Rock,
where he is now pursuing his doctorate degree in bioinformatics. His research focuses on improving our understanding of the mechanisms underlying and driving the aging process to increase our chances for delaying, stopping and hopefully eventually reversing the adverse and detrimental consequences of aging because otherwise they'll inevitably keep destroying everyone's existence eventually. Thomas Hahn hopes to carry on this kind of research professionally after graduation.
A Fast Sketch-based Assembler for Genomes

Priyanka Ghosh
Washington State University
1590 NE Northwood Dr. Apt A303, Pullman WA 99163
832-421-0549
pghosh@eecs.wsu.edu

ABSTRACT

Genome assembly, a classical problem in bioinformatics, with the goal to reconstruct an unknown long genome (DNA sequence) from very short fragments (“DNA reads”) obtained from it, is a constantly evolving research topic since most Next-Generation-Sequencing technologies generating the reads have become orders of magnitude faster and less expensive. Digressing from methods adopted by most state-of-the-art assemblers [1,2], our efforts are devoted towards designing and developing scalable parallel algorithms in order to accomplish large-scale plant, animal and microbial genome assemblies on supercomputers, focusing not only on performance (runtime and memory) but also on problem and resource scalability. Our paper A Fast Sketch-based Assembler for Genomes accepted to ACM-BCB 2016 addresses this issue and presents a fast multi-threaded algorithm incorporating probabilistic data structures namely the Count-min sketch [3] in a novel fashion to build a De Bruijn graph with a significantly reduced memory and time footprint. More specifically, FastEtch constructs an approximate de Bruijn graph and stores, with high probability, only those vertices corresponding to highly frequent k-mers, thereby achieving significant savings in space and time. Furthermore, our algorithm incorporates user-defined heuristic parameters offering users a way to control the inherent quality-performance (time, memory) trade-off in the assembly output. We compared FastEtch [3] with four other state-of-the-art assemblers — viz. SOAPdenovo, Velvet, ABySS, and Minia. We observed that all three variants of FastEtch performs consistently the best in terms of runtime performance. With respect to memory usage, FastEtch was second only to Minia, which also uses the disk. More specifically, FastEtch performed between 11.6% to 30.6% faster than the second fastest algorithm, across all the experiments. Also, on an average FastEtch consumes 20-30% less memory in comparison to the traditional de Bruijn graph based assemblers. In terms of assembly accuracy, FastEtch surpasses Velvet and SOAPdenovo by a large margin both in terms of N50 and % genome covered. In summation, although the FastEtch approach relies on approximation, we are able to deliver longer contigs and high assembly accuracy, in less time and memory compared to most other short-
read assemblers. Parallel tools for conducting genome assembly are gradually gaining popularity and show promise towards meeting the demands of processing such huge amounts of data. However the widening gap between between massively parallel sequencing technologies (orders of magnitude faster and less expensive) and the ability to analyze and assemble such genomes remains a challenge in terms of finding a scalable solution that is both accurate and efficient with respect to both memory (and resource) consumption and runtime performance. FastEtch provides a parallel multi-threaded approach to performing genome assembly. With our current efforts dedicated towards achieving large-scale performance, the cost of genome assembly can be reduced. This is needed in order to target treatment for rare and common genetic diseases, particularly cancer. Cancer arises fundamentally from genomic defects in tumor cells, and the identification and characterization of these mutations can not only guide the treatment, but also transform and personalize medicine. Apart from cancer treatment, genome assembly may also lend itself to applications in the field of genome mapping.

References:

List of my publications:

BIOGRAPHY

I am a 2nd year Computer Science PhD student at Washington State University, working with Dr. Ananth Kalyanaraman. My research interests lie broadly in designing efficient scalable algorithms for solving various combinatorial problems arising in genome / metagenome assembly, that can be further extended to incorporate parallelism (on shared and distributed memory systems) in order to target a solution at scale. To that effect, my primary research goal builds upon ideas from parallel algorithms, optimization techniques and memory-efficient data-structures to more specifically address an active problem in the field of genomics, namely De novo genome assembly.
I have earned my Master's degree in Computer Science from University of Houston, USA in 2012, under the supervision of Dr. Barbara Chapman, and a Bachelor's degree in Information Technology from the West Bengal University of Technology in 2007. I have a background in high performance computing, parallel programming models, and as part of my Master's thesis I have implemented extensions creating a prototype to allow support for explicit point-to-point synchronizations among OpenMP tasks in the OpenUH OpenMP runtime library. These extensions introduced in the OpenMP 4.0 standard provide support for a data flow model within OpenMP.
Mining Novel Knowledge from Biomedical Literature using Statistical Measures and Domain Knowledge

Kishlay Jha *

University at Buffalo, The State University of New York

341 Davis Hall
Buffalo, NY 14260-2500
(701) 205-5783
kishlayj@buffalo.edu

ABSTRACT

Recent decades have witnessed a massive increase in scientific reporting. MEDLINE, a preeminent bibliographic database has more than 23 million references to journal articles in life sciences. This overloaded textual resource (scientific knowledge) presents us with challenges and opportunities to apply sophisticated text mining techniques and gain deeper insights. The technique of generating meaningful knowledge (previously unknown) from existing knowledge base is known as Literature Based Discovery (LBD) [1, 2]. In simple words, the problem statement of our work is: Can we generate novel knowledge from large knowledge bases (example: MEDLINE) by rationally connecting complementary but non-interactive set of articles? A crucial component in this research problem is to discover related concepts. To find related concepts, researchers have proposed several approaches. Some of them are distributional approaches, graph based approaches, latent semantic indexing, and so forth. In our work, we proposed a new approach which combines statistical information measures and semantic relatedness in an intelligent way. Also, we layout an effective “preprocessing” strategy which raises precision of results by eliminating outputs that are bogus, spurious or semantically unrelated. The preliminary experiments demonstrate that our proposed methods are capable of generating results with higher precisions. We believe our results [Publications] will help biomedical scientists develop a cognitive perspective and investigate further for potential scientific discoveries. The central idea of this work is to generate plausible scientific hypotheses worthy of further experimentation. Thus, if the experimentation on generated hypotheses corroborates then it leads to creation of new knowledge –
highlights upon the prolific leverage this work endures. In past, researchers working in related problems have successfully developed application useful to society at large. Some of the examples include: identification of viruses as bio-weapons [3], finding functional connection between genes [4], drug-disease association [5]. Therefore, it is imperative to assume that conclusions of this research problem shall be very intriguing with a broader impact on overall humanity.

**BIOGRAPHY**

Kishlay Jha is a PhD Student in the Department of Computer Science and Engineering at State University of New York at Buffalo. Prior to that, he received his MS degree in Computer Science from North Dakota State University in 2016. His research interests are in the area of data/text mining, machine learning and bioinformatics.

Key references:


Relevant Publications:

2. Vishrawas Gopalakrishnan, **Kishlay Jha**, Aidong Zhang and Wei Jin, “Generating Hypothesis: Using Global and Local Features in Graph to Discover New Knowledge from Medical Literature”, 8th International Conference on Bioinformatics and Computational Biology (BICoB) Las Vegas, Nevada, USA, April 4-6, 2016 (In conjunction with CATA-2016)
Computational Framework for in-Silico Study of Virtual Cell Biology via Process Simulation and Multiscale Modeling

Hanyu Jiang

Stevens Institute of Technology

Burchard Building, 524 River Street #713, Hoboken, NJ 07030
(+1) 917-680-5999
hjiang5@stevens.edu

ABSTRACT

Biological systems encompass complexity that far surpasses many artificial systems. Modeling and simulation of large and complex biochemical pathways is a computationally intensive challenge. Traditional tools, such as ordinary differential equations and partial differential equations are all limited either by their modeling fidelity or computational efficiency or both. In this work, we present a scalable computational framework based on modeling biochemical reactions in explicit 3D space, which is designed to exploit parallelism and scalability offered by commodity massively parallel processors (i.e., GPUs), and it is also able to offer a deeper insight into various biological processes within the cell. By using a novel method of domain partition as well as particle data exchange, experimental results collected on various types of GPUs have shown that both strong and weak scaling are achieved and considered to be linear for our system. Besides, the accuracy and error analysis are also studied to validate the proposed framework. Overall, this work is first of its kind to study biological pathways in such fine-grained spatial and temporal resolution with higher accuracy, and it also pushes the boundaries of size and scale of systems that can be efficiently simulated with commodity massively parallel processors.

BIOGRAPHY

Hanyu Jiang obtained his MEng degree of Computer Engineering from Stevens Institute of Technology. After that, he began working toward the PhD degree in Computer Engineering at Stevens and supervised by Dr.Narayan Ganesan. His current research interests include Computational System Biology and Bioinformatics, Heterogeneous and Parallel Computing, Multi/Many-core Processor
Architecture as well as accelerated approach on Big Data processing.
Disease Similarity using Large Graphs

Jonathan Kho

Georgia Institute of Technology, Department of Computational Science and Engineering

Atlanta, GA
914-318-1701
jkho3@gatech.edu

ABSTRACT

In this project graph databases and graph algorithms will be leveraged and analyzed to characterize diseases prone to recent outbreak such as Zika. This Zika outbreak has shown the importance of not only predicting the spread of disease but also identifying known medication and biological markers to focus research on in search of similar treatment for the spreading disease. This project aims to use graph databases and properties and develop scalable graph pattern matching algorithms to create an objective measure of disease similarity and assist medical researchers in finding important biological markers to investigate.

BIOGRAPHY

Jonathan Kho is a Ph.D. student in the Department of Computational Science and Engineering at the Georgia Institute of Technology. He earned his B.S. in Electrical and Computer Engineering in 2014 from the University of Rochester. His research interests include distributed systems and graph databases. His current application interest is bioinformatics.
Model-based study of the Effectiveness of Reporting Lists of Small Feature Sets using RNA-Seq Data

Eunji Kim

Department of Electrical Engineering and Computer Engineering, Texas A&M University
223B Wisenbaker Engineering Building, College Station, TX
(734) 353-6531
eunjikim@tamu.edu

ABSTRACT

Next Generation Sequencing (NGS) technologies have recently emerged as popular tools to quantify gene transcripts. However, NGS technologies pose new computational challenges and a recent study by N. Ghaffari et al. showed that a NGS pipeline could lead to transformation degradation in classifier performance. In this study, we address another important question: the effects of the non-linear transformation induced by the sequencing machine and the choice of error estimators on the feature sets ranking. Two different types of synthetic data are generated for simulation experiments: actual gene expression concentration, modeled by a hybrid multivariate Gaussian distribution (MVN), and (2) Poisson transformed MVN data, emulating NGS reads. Comparison of the ranking power results obtained from two different types of data demonstrates that the data become less discriminative when gene concentrations are transformed by the sequencing. Moreover, the consistency between the ranked lists of feature sets based on the MVN and the NGS data is poor, which is yet another indicator of unreliable classification performance in the case of RNA-Seq data.

BIOGRAPHY

Eunji Kim is a third year Ph.D student in the Electrical and Computer Engineering at Texas A&M University. She received her M.S in Electrical and Computer Engineering at the University of Michigan, Ann Arbor, and her B.S in Electrical and Computer Engineering at Ewha Womans University. She joined the Genomic Signal Processing Laboratory in 2014 and during 2014 - 2015, she worked as a research assistant in the Center for Bioinformatics and Genomic Systems Engineering. Since 2015, she has been conducting collaborative research with the NIH-sponsored Center for Environmental...
Translational Health Research at Texas A&M University and the Fred Hutchinson Cancer Research Center in Seattle. Her research interests include (i) the ranking of gene sets for phenotype classification, error estimators in the designed classifiers, (ii) mathematical representation of dynamics of canalizing gene in gene regulatory networks, regulation power of canalizing genes (iii) the effects of diet and microbial metabolites on the human sigmoid colon and rectal tissue. In addition, she is drawing comparisons between data generated from biopsied tissue (invasive sampling) and stool-derived exfoliated cells (noninvasive sampling).
Manhattan Path-Difference Median Trees

Alexey Markin

Iowa State University, Computer Science department

3732 Tripp street, Ames, Iowa 50014
(515)444-14-70
amarkin@iastate.edu

ABSTRACT

The Median Tree approach arises in computational evolutionary biology as a very powerful tool for inference of large-scale phylogenetic trees. A median tree problem seeks an optimal phylogenetic tree given an input set of discordant gene trees and a problem-specific cost function. This work investigates into a well-established and extensively studied tree metric, Manhattan path-difference distance (Williams and Clifford, 1971), as applied to the median tree problem. Our work shows that the corresponding Manhattan median tree problem is NP-hard and approaches it using a standard local search heuristic. In order to enable the heuristic to compute median trees for large-scale phylogenetic datasets we devised a time-efficient algorithm that significantly improves on the complexity of the previously best-known naïve approach. As for exact solutions to the Manhattan median tree problem, we developed a novel ILP formulation.

Intellectual Merit: As the performed experimental evaluation shows, the developed algorithm for the Manhattan median tree problem represents a valuable alternative to the mainstream supertree methods, such as MRP, and can bring new insights in understanding of evolution. Broader Impact: Computationally efficient median tree and supertree methods bring the scale of the phylogenetic analysis to the unprecedented scale. The analysis and predictive power of large-scale evolutionary trees represent a great value not only to the evolutionary biology, but also immensely benefits such areas as agronomy, medical sciences, epidemiology, and the whole society at large.
BIOGRAPHY

Alexey Markin received a B.S. degree in Computer Science from Higher School of Economics (one of the leading universities in Russia) in 2015. His interest in computational biology led him to seek a doctoral degree at the Computer Science department at Iowa State University, where he works with Prof. Oliver Eulenstein on problems in computational evolutionary biology. Alexey has received a "Robert Stewart Early Research Recognition Award" in regards to his work on median trees in phylogenetics. His research interests include graph theory and computational biology.
ABSTRACT

Modeling structural transitions of a protein at equilibrium is central to understanding function modulation but challenging due to the disparate spatio-temporal scales involved. Of particular interest are sampling-based methods that embed sampled structures in discrete, graph-based models of dynamics to answer path queries. These methods have to balance between further exploiting low-energy regions and exploring unpopulated, possibly high-energy regions needed for a transition. We recently presented a strategy that leverages experimentally-known structures to improve sampling. The intellectual merit of the work presented here is that we demonstrate how such structures can further be leveraged to improve both exploitation and exploration and obtain paths of very high granularity. We show that such improvement is key to obtain accurate sample-based models of structural transitions. The broader impact of this work is improvement of understanding of the current capabilities and limitations of sampling-based methods. Proposing strategies to address some of these limitations is a first step towards sampling-based methods becoming reliable tools for modeling protein structural transitions and being more broadly adopted by the computational and molecular biology community at large.

BIOGRAPHY

Tatiana Maximova is a postdoctoral fellow in the department of Computer Science at George Mason University. She obtained her Ph.D. from Odessa National Polytechnic University, Ukraine and was a researcher at the Ben Gurion University in Israel. At Ben Gurion, Dr. Maximova was one of the developers of MESHI, an object-oriented open-source software for protein structure modeling and prediction. MESHI was integral to the performance of the "Keasar" group in CASP10, ranking 2nd and 3rd among single model methods in per-target assessment, rank 2nd in the best-model group.
performance in CASP10. Dr. Maximova's research interests include computational structural biology, protein biophysics, structure prediction, force-fields development, and sampling-based algorithms with a focus on transition modeling. At George Mason University, Dr. Maximova's work focuses on structure-function related studies on proteins central to cell function and human health. As part of this work, Dr. Maximova has developed computational methods that lie at the interface between computer science, biophysics, proteomics, and statistics. In particular, some of these methods are shown capable of mapping energy landscapes and structural transitions of wildtype and pathogenic variants in signaling proteins and enzymes of central importance to human diseases, such as Ras GTPases, calmodulin, and superoxidase dismutase.
Robinson-Foulds Median Trees: A Clique-based Heuristic

Jucheol Moon

Iowa State University

B26B Atanasoff Hall, Ames, Iowa 50011, USA
+1 (614) 556-7289
moon@iastate.edu

ABSTRACT

Species trees are branching diagrams depicting the evolutionary relationships among a set of species, and species trees are inferred from gene trees that are describing the evolutionary history of genes [1]. Unfortunately, for many genes the corresponding gene trees have discordant topologies when compared to the actual species tree, and to each other. Solving Robinson-Foulds (RF) median tree problems is a common approach to reconcile discordance in gene trees in order to infer a species tree [2]. The RF median tree problem seeks a median tree for a given collection of gene trees under the RF distance. While this problem is NP-hard, it has been addressed by standard local search heuristics. We introduce a graph-theoretic formulation of the RF median tree problem where optimal trees relate to minimum vertex weight cliques in a compatibility graph. The intellectual merit of this work is that our clique-based heuristic improves significantly on the RF median tree estimates resulting from the standard RF heuristics. Investigating how the clique-based heuristic and standard local search heuristics may relate, we found that the clique based heuristic can be equivalently expressed as a standard local search heuristic albeit using a distinct tree edit operation that we introduce. In terms of the broader impact, RF median tree problem is applied cancer research to reconstruct processes of evolution in tumor progression. By estimating the evolution history of tumors more accurately, the disease could be detected and treated preemptively.

References.
BIOGRAPHY

Jucheol Moon is a Ph.D. student of Computer Science at ISU since 2012. He received his M.S. degree in Computer Science from South Dakota State University in 2012 and B.S. degree in Physics from Korea University in 2004. His research interest is in algorithmic and theoretical properties of problems and machine learning in Computational Biology and Bioinformatics. He has published 2 journal papers and 7 conference papers in these areas and received distinguished awards including Best Paper Award Finalist at the 8th International Conference on Bioinformatics and Computational Biology and a Teaching Excellence Award in 2015. He is a member of the ACM Special Interest Group on Bioinformatics, Computational Biology, and Biomedical Informatics (SIGBio).

JOURNALS

CONFERENCES
Utilizing paired reads to improve Genome Assembly quality

Rahul Nihalani
PhD candidate, Georgia Tech
266 Ferst Drive NW, Atlanta, Georgia
515-708-6289
rahuln@gatech.edu

ABSTRACT

Genome assembly has been a problem of interest for over two decades. A quality solution to the problem will have a high impact in various areas such as personalized medicine, cancer research, studying new species and many more. Numerous assemblers have been developed till date to solve the problem. There still remains, however, plenty of scope to improve the quality of assembly. Some sequencers like Illumina can produce reads in pairs, where an approximate genomic distance between the pairs is known. This information mostly remains underutilized in most assemblers. Our research focuses on utilizing this paired-read information to improve the quality of genome assembly. We have developed a framework to embed the paired-read information into de-Bruijn graphs, a popular framework for approaching genome assembly. We utilize this information to produce paths in the graph to form long contigs. We produced some preliminary results that show a great promise in the approach. The research is published in ACM-BCB 2016 [1].

BIOGRAPHY

I obtained my bachelor's degree in communications and computer engineering from LNMIIT Jaipur, India, and masters in computer science and engineering from IIT Bombay, India. After obtained my masters, I joined Iowa State University as a PhD student in the dept of ECpE and transferred to Georgia Tech in the second year. Currently I am a PhD candidate in the dept. of CSE at Georgia Tech, advised by Dr. Srinivas Aluru.
Detecting Communities in Biological Bipartite Networks

Paola Gabriela Pesantez-Cabrera
Washington State University
1401 NE Merman Drive APT 904, Pullman, WA 99163
(509)339-4109
p.pesantezcabrera@wsu.edu

ABSTRACT

Community detection is a widely used operation in graph theory. The goal is to partition the vertex set of an input graph into tightly-knit groups or “communities”, such that vertices that are assigned to a community have a higher density of edges among them than to the vertices in the rest of the network. Numerous efficient algorithms have been developed for community detection in unipartite networks [S. Fortunato, 2010]. However, its treatment in bipartite networks has been sparse [M. J. Barber, 2007, R. Guimerà et al., 2007, X. Liu and T. Murata, 2010]. In our research we visit the problem of community detection in bipartite networks which has a critical impact in the analysis of complex biological data where there are two different types of interacting entities (e.g., genes and diseases, drugs and protein complexes, plants and pollinators, hosts and pathogens). For instance, identifying a group of genes that have been implicated across a set of diseases could possibly reveal hidden links among seemingly different diseases or disease conditions, and in the process help identify new drugs and therapies. Similarly, identifying active gene clusters across different subsets of brain regions could provide new insights into brain function. Thus, toward detecting communities in such bipartite networks, we make the following contributions: i) we define a variant of the bipartite modularity function defined by Murata to overcome one of its limitations; ii) we present an algorithm (biLouvain), building on an efficient heuristic that was originally developed for unipartite networks; and iii) we present a thorough experimental evaluation of our algorithm compared to other state-of-the-art methods to identify communities on bipartite networks. Experimental results show that our biLouvain algorithm identifies communities that have a comparable or better quality (bipartite modularity) than existing methods, while significantly reducing the time-to-solution between one and three orders of magnitude. Although the
idea of modularity optimization to solve the community detection problem is not new, to the best of our knowledge our proposed approach hasn't been tried before and it lead us to a better understanding of the structural properties of the bipartite networks as well as to improve modularity results and performance. Once the communities have been obtained the next steps are analysis and interpretation of them. These domain specific tasks have direct impact on other studies for example: in the case of the gene-drug network we could identify groups of drugs that might inhibit or otherwise modulate groups of genes. We also could detect groups of genes that are suitable for drug targeting but may not currently have a drug targeting them. Similarly, for plant-pollinator networks we could identify groups of pollinators that limit or promote plant species establishment and persistence.

**BIOGRAPHY**

I was born in Cuenca, Ecuador, in 1985. Since I was a child I was fascinated by computers and how they can be used to research and solve problems. Thus, I received the B.S. degree in Computer Science from University of Cuenca, Cuenca, Ecuador, in 2009, and the MSc. degree in Computer Science from Washington State University, Washington, in 2013.

Since Fall 2013, I have been a Computer Science Ph.D. student at Washington State University under the research supervision of Dr. Ananth Kalyanaraman. I just finished my third year as part of the program, completed all course work, and passed my preliminary examination.

I am interested in developing large-scale algorithms and software for cluster-based analysis of heterogeneous networks originating from the life science domain. My goal is to incorporate complex, heterogeneous information available into cluster analytics. In normal life, I am passionate about promoting engineering, science, and math among young women.

I received the “Benigno Malo” prize in recognition of academic merit and I graduated with the highest honor from University of Cuenca. I was one of the recipients of the Fulbright Scholarship to study my master degree in the United States. I was honored with the EECS Outstanding Teaching Assistant award from Washington State University in 2014. I was awarded a travel grant to attend the ACM International Workshop on Big Data in Life Sciences (BigLS) in 2014, and I have been honor with the NSF award to attend ACM-BCB in 2015 and 2016.
Risk Factor Analysis Based on Deep Learning Models

Qiuling Suo

University at Buffalo
341 Davis Hall, University at Buffalo
814770286
qiulings@buffalo.edu

ABSTRACT

Risk factor (RF) analysis on healthcare data aims to mine the interaction relationships among various factors, evaluate their effect on the target disease, and identify the causing factors to the disease. This research focuses on disentangle informative features from the high-dimensional, diverse and heterogeneous healthcare data, in order to improve disease diagnosis and prediction. We investigate the application of deep learning techniques on healthcare datasets, as deep learning is powerful at extracting high-level abstract representations. Our method is based on two unsupervised feature learning techniques, restricted Boltzmann machine (RBM) and denoising autoencoder. We first investigated and interpreted the latent integrated features by calculating the importance score of each input RF; then we compared prediction performance with other classifiers. The importance score is measured by calculating the gradient of a latent feature in terms of each input RF, as larger gradient means larger impact. The experimental results show that pre-trained neural network models can learn meaningful latent concepts, which aggregate correlated informative RFs and reduce the effect from other noisy RFs. The high-level abstract features can achieve better prediction accuracy than raw risk factors, which indicates that deep learning is promising in the healthcare area. This work is published in ACM-BCB 2016. Our work propose novel ways to understand how neural network models learn and represent abstract features for healthcare data. We show that deep learning has the advantage to analyze risk factors. We hope that this work will promote a better understanding of neural network models on healthcare domain, and offer some guidelines on how to apply these models in various healthcare applications. With the understanding of the relations among disease attributes, patient can be advised to avoid unnecessary tests, and change their modifiable attributes (such as quit smoking, reduce alcohol) for disease control.
**BIOGRAPHY**

I am currently working as a PhD student at Robin Li Data Mining& Machine Learning Laboratory with supervision by Prof. Aidong Zhang. I received my master’s degree from Penn State University in 2013. My research interests are in the area of data mining, machine learning and health informatics. I am currently focuses on understanding interactions among diverse risk factors, and building models to better predict the risk of developing the target disease.
Detecting Anomalies in Alert Firing within Clinical Decision Support Systems using Anomaly/Outlier Detection Techniques

Soumi Ray
Northeastern University and Brigham and Women’s Hospital
360 Huntington Ave
Boston, MA 02115
443-765-5333
ray.so@husky.neu.edu

ABSTRACT

Clinical Decision Support (CDS) systems play an integral role in the improvement of health care quality and safety. Alert malfunctions within CDS are a common problem and these greatly limit its usability [1]. Often these go undetected for years with adverse consequences. Anomaly detection is a novel approach to identify malfunctioning within CDS systems. Once an anomaly in alert firing is detected, it can be used to rectify and repair CDS system to make it robust and reliable. We introduce and apply four anomaly detection algorithms namely Poisson changepoint detection, ARIMA (Autoregressive Integrated Moving Average), Seasonal Hybrid Extreme Studentized Deviate algorithm (S-H-ESD) and E-Divisive with Median (EDM) to estimate the dates when malfunctions and/or changes in alert firing occur. These algorithms were applied to 4 commonly utilized rule alerts and at least one of the four algorithms was able to successfully determine the date of an anomaly in the alert firing. Preliminary results demonstrate successful detection of anomaly occurrences within the alert data [2]. The use of these algorithms in detecting alert trigger malfunctions in novel and the results are promising. The growing acceptance of electronic health records by healthcare system demands robustness and minimization of such errors or anomalies. The techniques we have demonstrated can be broadly applied for early detection of anomalies in the CDS systems and subsequent root cause analysis.
Reference:


BIOGRAPHY

I am a full-time Masters’ student of Health Informatics at Northeastern University. After completing my Bachelors in Computer Science, I enrolled at University of Maryland Baltimore County (UMBC) for a Masters degree in the same specialty. I developed a keen interest in artificial intelligence and machine learning and stayed back at UMBC to pursue my PhD. During my PhD I focused on developing models for discovering and characterizing hidden variables in different problem domains. I did an internship at National Library of Medicine, National Institutes of Health (NIH) where I worked on exploring feature selection for identifying image categories for improved content-based image retrieval. While working on this project I developed a strong interest in the application of my technical knowledge in the healthcare domain. Upon completion of my PhD I joined the National Library of Medicine, NIH as a postdoctoral fellow and worked on preservation and access of unstructured text by automated metadata extraction (AME) and semantic search. I branched out to Bioinformatics in an effort to apply my experience in Computer Science and Artificial Intelligence in the medical domain. My current area of research involves the application of machine learning algorithms to improve robustness of Clinical Decision Support systems.
Genome-wide Association Study

Subrata Saha
University of Connecticut, Storrs
Department of Computer Science and Engineering
Phone Number: 860-208-0449
Email: subrata.saha@engr.uconn.edu

ABSTRACT

Advances made in sequencing technology have resulted in the sequencing of thousands of genomes. Novel analysis tools are needed to process these data and extract useful information. Such tools could aid in personalized medicine. As an example, we could identify the causes for a disease by comparing the genomes of people who have the disease and those who do not have this disease. Given that human variability happens due to single nucleotide polymorphisms (SNPs), we could focus our attention on these SNPs. Investigations that try to understand human variability using SNPs fall under genome-wide association study (GWAS). A crucial step in GWAS is the identification of the correlation between genotypes (SNPs) and phenotypes (i.e., characteristics such as the presence of a disease). This step can be modeled as the $k$-locus problem (where $k$ is any integer). A number of algorithms have been proposed in the literature for this problem. In this research we present an algorithm for solving the 2-locus problem that is up to two orders of magnitude faster than the previous best known algorithms. The $k$-locus problem can be thought of as a special case of the closest pair problem (CPP). CPP is one of the well-studied and fundamental problems in computing. Given a set of points in a metric space, the problem is to identify the pair of closest points. There are numerous applications where this problem finds a place. Examples include computational biology, computational finance, share market analysis, weather prediction, entomology, electro cardiograph, N-body simulations, molecular simulations, etc. As a result, any improvements made in solving CPP will have immediate implications for the solution of numerous problems in these domains. A naive deterministic algorithm can solve CPP in quadratic time. Quadratic time may be too much given that we live in an era of big data. Speeding up data processing algorithms is thus much more essential now than ever before. In this research we present algorithms for CPP that improve (in theory and/or practice) the best-known algorithms reported in the literature for CPP.
**BIOGRAPHY**

Mr. Subrata Saha is a PhD student in Computer Science at University of Connecticut (UConn), Storrs. He completed his BSc in Computer Science and Engineering at Bangladesh University of Engineering and Technology (BUET), Dhaka. 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 data mining and bioinformatics. His research interest in bioinformatics includes genome-wide association study, 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’s research articles have been published in top-tier journals and conferences (e.g., ICDM, CIKM, Bioinformatics, BMC Bioinformatics, BMC Genomics, ACM BCB, ADMA). He has also served as a reviewer in several 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).
Counting independent motifs in probabilistic networks

Aisharjya Sarkar

Computer & Information Science & Engineering, University of Florida, Gainesville, Florida

3800 SW 34th ST.
Apt# BB273, Stoneridge Apartments
Gainesville, FL - 32608
518-322-1501
aisharjya@ufl.edu

ABSTRACT

We propose a novel algorithm for counting independent copies of a given motif topology in probabilistic biological networks. Network motifs which are frequent topological patterns, are key structures through which biological networks operate. Counting independent (i.e. non-overlapping) copies of a given motif however remains to be a computationally hard problem. Motif counting problem becomes computationally even harder for biological networks as biological interactions are uncertain events. The central challenge behind this problem is that different embeddings of a given motif in a network can share edges. Such edges can create complex computational dependencies between different instances of the given motif. Our algorithm first finds all instances of a motif in the network (i.e. overlapping instances). It then assigns a priority value to each of those instances. The priority value of each instance depends on the existence of that instance as well as the existence of all overlapping instances with this instance. The algorithm iteratively picks the instance with the highest priority value to include in the output. It then removes that instance along with its overlapping instances. Our algorithm continues this process until no instances remain. We tested our algorithm on synthetic as well as real datasets. In order to test the robustness of our method, we tested based on three parameters: network size, probability model and network topology model. Among the three parameters that we test, the performance of our method mainly depends on topology. The size of the network also has great effect, while the probability model has limited effect. We tested our algorithm on seven real networks with a maximum node size of 159 and 127 edges. The real network experimental results concur with those of synthetic networks, where network topology plays the key impact. The proposed method provides a mathematical formulation to precisely calculate the priority value for an embedding to exist. First it calculates a gain value, which represents the probability that the corresponding embedding will exist. It then calculates a loss value which represents the number of overlapped neighbors of that embedding given that embedding exists. The priority value of that embedding is expressed as a function of these two values. This algorithm can be applied to any real-life applications, not just probabilistic biological
networks, wherever the network topology is uncertain in nature and requires counting motifs that can help in determining the key topological patterns.

Key references:

**BIOGRAPHY**

Aisharjya Sarkar is a second year PhD student in the department of Computer & Information Science & Engineering at University of Florida, USA. She is working in The Bioinformatics Lab under the supervision of Dr. Tamer Kahveci since January 2016 at UF. She graduated with a Masters of Engineering in the department of Information Technology from Indian Institute of Engineering, Science & Technology, Shibpur, India in July 2010. She has worked as a software developer at Tata Consultancy Services Ltd. until July 2015. Her current research interests include the following areas: Computational Systems Biology, Biological Network Modeling and Analysis and Predictive Modeling and Analytics in Healthcare.
A Novel Temporal Similarity Measure for Patients Based on Irregularly Measured Data in Electronic Health Records

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

ABSTRACT

Patient similarity measurement is an important tool for cohort identification in clinical decision support applications. A reliable similarity metric can also be used for deriving diagnostic or prognostic information about a target patient using other patients with similar trajectories of health-care events. However, the measure of similar care trajectories is challenged by the irregularity of measurements, inherent in health care. To address this challenge, we propose a novel temporal similarity measure for patients based on irregularly measured laboratory test data from the Multiparameter Intelligent Monitoring in Intensive Care database and the pediatric Intensive Care Unit (ICU) database of Children’s Healthcare of Atlanta. This similarity measure, which is modified from the Smith Waterman algorithm, identifies patients that share sequentially similar laboratory results separated by time intervals of similar length. We demonstrate the predictive power of our method; that is, patients with higher similarity in their previous histories will most likely have higher similarity in their later histories. In addition, compared with other non-temporal measures, our method is stronger at predicting mortality in ICU patients diagnosed with acute kidney injury and sepsis.

Intellectual Merit: This study demonstrates the advantages of applying one of the well-known sequence alignment algorithms, Smith-Waterman algorithm, into patient similarity analysis in healthcare analytics. The proposed technique addresses the irregularity and noisiness of measurements that are inherent in healthcare data.

Broader Impact: This study proposes a useful application that allows clinicians to find similar patients to a target patient based on healthcare histories stored in electronic health records, which demonstrates the power of personalized medicine.
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 second-year doctoral student in the school of biology, where she is conducting research pertaining to temporal data mining using intensive care unit (ICU) data under supervision of Professor May D. Wang. During her one year 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.
Development of a Scalable Method for Creating Food Groups Using the NHANES Dataset and MapReduce

Michael Wyatt
University of Delaware
Newark, DE
19709
mwyatt@udel.edu

ABSTRACT
We tackle the need for meaningful food group classifications in dietary datasets such as the National Health and Nutrition Examination Survey (NHANES). Our method for identifying food groups provides a new standard to group food items for use in studies which deal with dietary impact on health and wellbeing. This standard will provide a basis for defining food groups, where typically food groups are defined by experts or informed by dietary trends [Reedy, 2010][Niclis, 2015]. We perform extensive analysis and preprocessing of the NHANES dataset. Additionally, we provide a data-driven approach for creating food groups based on micro- and macro-nutrient values using unsupervised learning algorithms. Our methods are parallelized to benefit from the scalable MapReduce paradigm. Results show that our food groups are denser and better separated with respect to nutrient content than expert-informed food groups [Wyatt, 2016].

BIOGRAPHY
During his undergraduate career at the University of Delaware, Michael Wyatt studied Biochemistry. At the University of New Mexico, Michael earned a MS in Computer Science. He later returned to the University of Delaware to earn his PhD in Computer Science. With his knowledge of biology, chemistry, and computer science, Michael strives to perform interdisciplinary research, where problems
related to biology can be solved using the tools and techniques learned through the study of computer science. Michael’s current work focuses on the National Health and Nutrition Examination Survey dataset. He is working to solve problems related to diet and its impact on health using advanced analysis techniques. Much of the work that Michael does involves machine learning and building scalable frameworks for analysis of big data. While finishing his PhD, Michael will continue to perform interdisciplinary research and expand his data analysis skill set.
Text Classification with Topic-based Word Embedding and Convolutional Neural Networks

Haotian Xu
Wayne State University
42 W Warren Ave, Detroit, MI 48202
267-283-7342
htxu@wayne.edu

ABSTRACT

(1) **Background.** Recently, distributed word embeddings trained by neural language models are commonly used for text classification with Convolutional Neural Networks (CNNs) (Kim) (Rios and Kavuluru).

(2) **Approach.** We propose a novel neural language model, Topic-based Skip-gram, to learn topic-based word embeddings for biomedical literature indexing with CNNs. Topic-based Skip-gram leverages textual content with topic models, e.g., Latent Dirichlet Allocation (LDA), to capture precise topic-based word relationship and then integrate it into distributed word embedding learning. We then describe two multimodal CNN architectures, which are able to employ different kinds of word embeddings at the same time for text classification.

(3) **Key preliminary results and discussions.** Through extensive experiments conducted on several real-world datasets, we demonstrate that combination of our Topic-based Skip-gram and multimodal CNN architectures outperforms state-of-the-art methods in biomedical literature indexing, clinical note annotation and general textual benchmark dataset classification. This work is published in (Xu, Dong and Zhu).

(4) **Intellectual Merit.** Our framework first leverages the whole text corpus with topic models to capture semantic relationship between words and then take it as the input for word representation learning using Topic-based Skip-gram with a novel objective function. Then, these topic-based word representations are used together with other state-of-the-art word embeddings for text classification in multimodal CNN models.

(5) **Broader Impact.** As a general text classification approach, our framework works well on indexing of biomedical articles, annotation of clinical text fragments with behavior codes, and
classification of news groups and it can also be utilized for emotion analysis, questioning & answering, etc. Furthermore, our framework is also extendable for image/audio classification.

**BIOGRAPHY**

I am currently a PhD student of Machine Vision and Pattern Recognition Laboratory (MVPRL), part of the Center for Visual Informatics and Intelligence at Wayne State University since August 2013. I received M.S. degree in Computer and Information Sciences at Temple University in Summer 2013. I received B.S. degree in Electronic Engineering at University of Science and Technology of China (USTC) in Summer 2012.

My areas of research include feature learning and data mining.

1) Biomedical Literature and Clinical Notes Classification with Topic Models and CNNs
   I proposed a topic-based natural language model to learn semantic word embedding for biomedical literature and clinical notes classification with multimodal convolutional neural networks. This is the work that I will present in ACM-BCB 2016.

2) Communication Research with Topic Models
   I used unsupervised topic models to identify the content of adolescent and caregiver speech in an approach akin to qualitative thematic analysis. This is the main technical content of the poster presentation, “Accelerating the Pace of Qualitative Communication Research with Computational Technology”, in the 37th Annual Meeting and Scientific Sessions of the Society of Behavioral Medicine.

3) Clinical Transcript Labeling with Semi-supervised Topic Model
   I proposed a semi-supervised hierarchical Bayesian topic model which utilizes only a few labeled clinical transcripts and much more unlabeled transcripts, and obtained a much better classification performance than famous supervised models (SVM and Naive Bayes).
Automated Capture of Naturalistic Child Vocalizations for Health Research

Hasini Yatawatte

University of Notre Dame

Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN
404-7860691
hyatawat@nd.edu

ABSTRACT

Child phonetic production and pre-speech vocalizations are extensively being collected and investigated to identify biomarkers and progression markers to develop diagnostic and assessment tools for various neuro-developmental conditions and to develop new child monitoring and health diagnostic tools. Most prior studies rely on infrequent data collections (e.g. in clinical or laboratory settings) due to the limitations inherent to collection and analysis of naturalistic recordings from children [1-3]. However, data collection in naturalistic settings are preferred and can yield data that (a) more accurately provide insights into the links between health and speech and (b) provide the foundation for accurate and reliable speech-based assessment and diagnostic tools. This paper describes the design of a data collection process using a specially designed voice recorder application that is running on an iOS-based device to collect naturalistic child utterances. The usability of the application was evaluated through a small-scale pilot run, where the captured data was processed to automatically classify child utterances. A classifier to distinguish different types of child utterances (such as babbling, crying, screaming, and laughing) as well as the type of emotion associated with the voice was built with a good accuracy using statistical properties of MFCC features of the samples alone The primary goal of the work was to address the challenges in naturalistic speech data collections, thereby making it more feasible to use such data collections in future research. The experience gained through the study and the voice capturing application will be used in our future studies to investigate early detection of neuro-developmental disorders such as autism in children by analyzing patterns that such disorders leave in speech.

References


**BIOGRAPHY**

Hasini Yatawatte is a third year PhD student from University of Notre Dame currently researching the feasibility of using speech as an early marker and classifier of Autism Spectrum Disorder (ASD). The intention of the research is to build portable and low cost diagnosis and assessment tools for Autism and similar neurodevelopmental disorders. She pursued her Master's in Computer Science from University of Colombo, Sri Lanka and BSc Eng.(Hons) in Computer Science and Engineering from University of Moratuwa, Sri Lanka. Her research interests involve identifying novel biomarkers and progress markers for various neurodevelopmental, neurological and neurodegenerative disorders. In her spare time, Hasini likes to expand her horizons through reading books, travelling, and improving life skills.
May Dongmei Wang, Ph.D. is a professor of BME and ECE, a Georgia Cancer Coalition Scholar, a Kavli Fellow, an AIMBE Fellow, ACTSI Georgia Tech Co-Director of Biomedical Informatics Program, and Core Director of Bioinformatics and Bicomputing Core in Emroy-GT Cancer Nanotechnology Center, and Georgia Tech Center for Imaging Mass Spectrometry. Her research interest is primarily in Biomedical Big Data Analytics, focusing on Biomedical and Health Informatics (BHI) for personalized, predictive, and precision health. Her research group develops intelligent analytics for data quality control, multi-modal data integration, causality modeling, and real-time data streaming. Major clinical applications include omic data mining for precision medicine, bio-nanoinformatics for cancer therapeutics and surgery, biomedical imaging informatics for clinical decision support, critical/chronic health informatics for personalized patient care, and predictive systems modeling for healthcare delivery. She is the corresponding/co-corresponding author for articles published in Journal of American Medical Informatics Association (JAMIA), Journal of Biomedical and Health Informatics (JBHI), IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 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, and Circulation Genetics etc.

Dr. Wang has led next generation sequencing data analytics studies in a FDA-organized international consortium and has served as an Emerging Area Editor for Proceedings of the National Academy of Sciences, Senior Editor for IEEE Journal of BHI, an Associate Editor for IEEE Trans. on BME, and invited panelists for NIH, NSF, European Union ERC, Brain Canada, and UK MRC. She was also a Distinguished Lecturer of the IEEE Engineering in Medicine and Biology Society (EMBS 2014-2015). Dr. Wang has delivered over 200 invited talks, keynotes, and plenary lectures at professional conferences, academic institutions, health systems, government agencies, and industry. She has served in the IEEE Big Data Initiative (Steering Comm), ACM Bioinformatics, Comp Biology and Health Informatics Conference (Steering Comm Co-Chair, Conference Co-Chair), the AIMBE Nomination Comm on Biomedical Informatics, the EMBS BHI Technical Comm (Chair), the IEEE International Conference on BHI (Steering Comm Chair, Conference Co-Chair), and the 2016 IEEE EMBS Annual Conference (Co-Chair). Recently, Dr. Wang was elected Vice Chair of the Gordon Research Conference (GRC) on Advanced Health Informatics. She is the recipient of an 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.