SIGBio Record

Newsletter of the SIGBio

ACM Special Interest Group

Volume 7, Issue 2, Spet 2017 ISSN 2159-1210

Notice to Contributing Authors to SIG Newsletters

Contents

Editor's Notes

The SIGBio community is growing and the research interests are broadening. We presented the structure of the SIGBio Record during the SIGBio meeting at ACM BCB Conference 2017. Consequently, we hope the SIGBio Record will contain in future more articles and perspective covering novel arguments. We continbute the Record with contributed articles and discussions, workshop and conference reports, scientist profiles, all topics interesting for SIGBio community.

The current issue presents a contributed article, a Phd Thesis abstract, and highlights from young researchers.

Also, a short report of the ACM BCB 2017 held in Boston, MA. The conference had high number of participants.

We thank contributors for this issue and hope that readers will find interesting references to their work in Bioinformatics and Health Informatics area.

Pietro Hiram Guzzi, Pierangelo Veltri - SIGBio Record Editors

SIGBIO Record - Submission Guidelines

Submission categories

Submissions to the newsletter can be either on a special issue topic or on topics of general interest to the SIGBIO community.

These can be in any one of the following categories:

- Survey/tutorial articles (short) on important topics.
- Topical articles on problems and challenges
- Well-articulated position papers.
- Review articles of technical books, products and .
- Reviews/summaries from conferences, panels and special meetings within 1 to 4 pages [1500-2500 words]
- Book reviews and reports on relevant published technical books
- PhD dissertation abstracts not exceeding 10 pages
- Calls and announcements for conferences and journals not exceeding 1 page
- News items on the order of 1-3 paragraphs

Brief announcements Announcements not exceeding 5 lines in length can simply be sent as ASCII text to the
editors by e-mail. SIGBIO Record publishes announcements that are submitted as is without review.

Announcements cannot be advertisements and should be of general interest to the wider community. The Editor reserves the right to reject any requests for announcements at his discretion.

Authors are invited to submit original research papers or review papers in all areas of bioinformatics and computational biology. The papers published in SIGBioinformatics Record will be archived in ACM Digital Library. Papers should follow the ACM format, and there is no page limitation.

http://www.acm.org/sigs/publications/proceedings-templates

Submissions should be made via email to the editors Pierangelo Veltri and Pietro Hiram Guzzi *(University Magna Graecia of Catanzaro, Italy at acmsigbio@gmail.com, veltri@unicz.it, hguzzi@unicz.it),*

ASSOCIATIVE PATTERN RECOGNITION FOR BIOLOGICAL REGULATION DATA

Yiou Xiao
B.E. Nanjing University, 2009
M.S. Syracuse University, 2011

Phd DISSERTATION
Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Computer and Information Science and Engineering

Abstract

The value of data analysis has become unprecedentedly recognized in the last decade. Nowadays, the great potential of data is appreciated by people from various backgrounds. From public relation experts to strategy makers; from big companies in silicon valley to scientists in fundamental sciences, people are devoting tremendous amount of attention, money and time to exploring its value. According to EMBL the total size of bioinformatics databases has grown to 70 Petabytes in 2015 (10 9 MBytes); Twitter users generate 250 million posts every day.

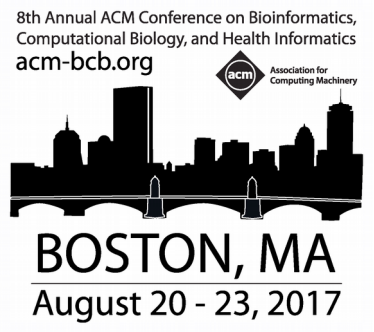
The difficulty of storing, processing and analyzing big data has been recognized for a long time. However, the term "big data" gained great attention in recent years because of the big advancement in technologies. New infrastructures such as MapReduce, HDFS, Hadoop, NoSQL databases and GPU computation as well as deep neural networks algorithms such as LSTM and CNN rekindled the enthusiasm.

Associative patterns between sets of objects are of interest in many disciplines such as social networks, economics and biology. The goal is to discover the interactions or relations between sets of objects. Although many approaches have been proposed, most focus on interactions between single objects, considered using similar characteristics of objects. In this dissertation, we focus on associative patterns recognition in bioinformatics area.

Bioinformatics is the ensemble of computational approaches to large-scale information analysis in biological data. It is now considered to be a self-contained branch of molecular biology, and helps researchers to better understand life systems; invent new diagnosis or treatment procedures; and design highly efficient medicines in target based therapies using data-centric techniques. Bioinformatics research accelerates the development of fundamental advances in biological hypothesis generation, data analysis and modeling, and provides tools for pharmaceutical, biomedical, chemical and even insurance companies. It encompasses a wide spectrum of topics that address questions about biological composition, structure, function and evolution of molecules, cells, tissues and organisms by computational methods that include mathematical modeling, machine learning and data mining. Biological regulation is defined as any process which modulates the frequency, rate or extent of biological processes, where computational approaches for recognizing interactions between objects (i.e., genes, RNAs, promoters, transcription factors and histone modifiers) are crucially important in hypotheses generation and experiment design.

In protein-DNA associative pattern recognition, we introduce an efficient algorithm for affinity test by searching for over-represented DNA sequences using a hash function and modulo addition calculation. This substantially improves the efficiency of next generation sequencing data analysis. In gene regulatory network inference, we propose a framework for refining weak networks based on transcription factor binding sites, thus improved the precision of predicted edges by up to 52%. In histone modification code analysis, we propose an approach to genome-wide combinatorial pattern recognition for "histone code to function" associative pattern recognition, and achieved improvement by up to 38.1%. We also propose a novel shape based modification pattern analysis approach, using this to successfully predict sub-classes of genes in flowering-time category. We also propose a "combination to combination" associative pattern recognition, and achieved better performance compared against multi-label classification and bidirectional associative memory methods. Our proposed approaches recognize associative patterns from different types of data efficiently, and provides a useful toolbox for biological regulation analysis. This dissertation presents a road-map to associative patterns recognition at genome wide level.

Program and General Chairs' Report on ACM-BCB 2017



The 8th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB) took place in Boston, Massachusetts on August 20th - 23rd. ACM-BCB is the premier dissemination forum for interdisciplinary research linking computer science, mathematics, statistics, biology, bioinformatics, biomedical informatics, and health informatics. Building on a steadily growing community and quality of research since its inception in 2010, ACM-BCB 2017 showcased leading-edge research on new technologies and techniques around gathering, processing, analyzing, and modeling of data and information for a variety of scientific, clinical, and healthcare applications, from bench to bedside.

The main call for original research submissions for both long and short papers resulted in 132 submissions coming from 35 different Countries around the world, to confirm the international and high relevance of ACM-BCB 2017 in attracting the interest of several research groups worldwide.

Furthermore, the conference received 15 submissions in the "Highlight" category, 5 submissions in the "Tutorial" category, 7 submissions in the "Workshop" category and 62 submissions in the "Poster" category.

The 132 submissions were grouped into 16 tracks: Advancing Algorithms and Methods, Automated Diagnosis and Prediction, Applications to Healthcare Processes, Applications to Microbes and Imaging Genetics, Big Data in Bioinformatics, Biological Modeling, Clinical Databases and Information Systems, Cancer Genomics, Genomic Variations and Disease, Inferring Phylogenies and Haplotypes, Knowledge Representation Applications, Protein and RNA Analysis, Protein Structure and Dynamics, Sequence Analysis and Genome Assembly, Systems Biology, Text Mining and Classification.

Every track had 3 track chairs, which by the help of many program committee members, performed the task of carefully reviewing every submission. Every submission received at least three reviews, evaluating quality, originality, novelty, and relevance of the submission. The Program Chairs, in a strict and very fruitful cooperation with the General Chairs, considered the reviewers' recommendations on every submission, resulting in 42 submissions accepted as long papers, and 28 submissions accepted as short papers.

The papers and the Highlight papers accepted as oral presentations were then scheduled in three parallel sessions, spanning over the three days of the conference, while accepted tutorials and accepted workshops were scheduled on the first day of the conference. The conference had a fourth parallel session, welcoming the Workshop on Algorithms in Bioinformatics (WABI), as it does in every odd year.

The conference had a full program and featured 3 keynote speakers, Shawn Murphy of Partners in Healthcare, Dagmar Ringe of Brandeis University, and Tandy Warnow of the University of Illininois at Urbana-Champaign. The program was diverse, including an industry panel, with a keynote from Enoch Huang, the Head of Computational Sciences at Pfizer R&D, a Women in Bioinformatics Panel, which is now a regular feature of every ACM-BCB conference, an NSF-sponsored student research forum, that

was highly instructive to student presenters and highly entertaining for attendees, as well as a demos and exhibits session.

The quality of papers was very high, and three presenters were recognized for their work. Laraib Iqbal and Robert Patro received the best paper award for their “Rich chromatin structure prediction from Hi-C data” paper. Alan Cleary, Thiruvarangan Ramaraj, Indika Kahanda, Joan Mudge, and Brendan Mumey received the best student paper award for their “Exploring frequented regions in pan-genomic graphs” paper. The best poster award went to Ali Foroughi Pour and Lori A. Dalton for their “Integrating Prior Information with Bayesian Feature Selection” poster. The best sponsored poster award went to Fatima Zare, Sardar Ansari, Kayvan Najarian, and Sheida Nabavi for their "Bias and Noise Cancellation for Robust Copy Number Variation Detection" poster. The best WABI paper award went to Jens Quedenfeld and Sven Rahmann for their "Analysis of min-hashing for variant tolerant DNA read mapping" paper.

The Program Chairs and the General Chairs welcomed all 307 registered participants and other attendees. Thank you to all for making ACM-BCB 2017 a great conference, and we hope to see all of you, and more of your colleagues and students in Washington, D.C. in 2018.

Amarda Shehu
George Mason University
Fairfax, VA

Tamer Kahveci
University Of Florida
Gainesville, FL

Giuseppe Pozzi
Politecnico di Milano
Milano, Italy

Nurit Haspel
University of Massachusetts
Boston, MA

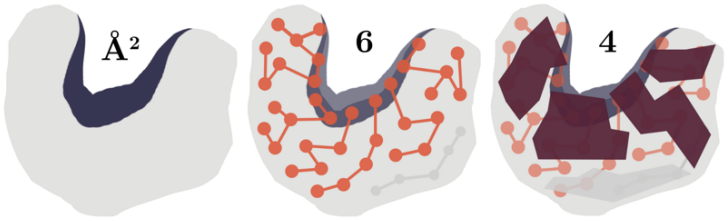
Lenore Cowen
Tufts University
Boston, MA

ACM-SIGBIO Undergraduate Research Highlight

At ACM-BCB 2017, several contributions in the form of paper and poster presentations were given by undergraduate students. A solicitation then went out to the ACM-SIGBIO community to highlight undergraduate research in the ACM-SIBGIO newsletter. What follows below are four contributions that showcase computational biology and bioinformatics research performed by undergraduate students.

Stephanie Mason*, Filip Jagodzinski*, and Brian Chen‡
Western Washington University* and Lehigh University‡

Stephanie Mason is an undergraduate student at Western Washington University majoring in Biology/Mathematics and minoring in Computer Science and Chemistry. Stephanie was always interested in research in computational biology and bioinformatics, and in Fall of 2016 she began working on a project involving a collaboration between Dr. Jagodzisnki of Western Washington University and Dr. Brian Chen of Lehigh University. Her work culminated in a workshop paper, titled “Investigating Rigidity Properties of Protein Cavities”, which she presented at the Computational Structural Biology Workshop at the ACM-BCB 2017 conference in Boston, Massachusetts.



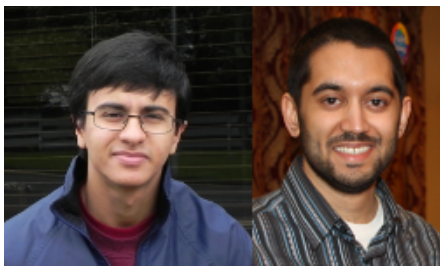
At CSBW, Stephanie presented a computation pipeline for investigating rigidity properties of protein cavities, which has relevance to drug design and ligand binding studies. The approach relies on rigidity analysis to determine the flexible and rigid regions of surface cavities that are identified using Dr. Chen's cavity detection software. Hundreds of thousands of cavities from among thousands of protein structures in the PDB have been analyzed. Stephanie's Python and Bash scripts gather a variety of structural metrics from the cavity, rigidity, and PDB data files. She aggregates that information into a single, large dataset, and mines the data using custom R scripts for relationships among a variety of cavity and structural metrics, including cavity surface area, quantities of rigid clusters in a cavity, and the sizes of rigid clusters that are members of large and small cavities. Some emerging relationships were presented at CSBW. More information about Stephanie’s work can be found in the ACM proceedings of ACM-BCB 2017. This summer, Stephanie is continuing this line of work as part of an REU at Lehigh University.

Michael Nissenson and Dong Si
University of Washington Bothell

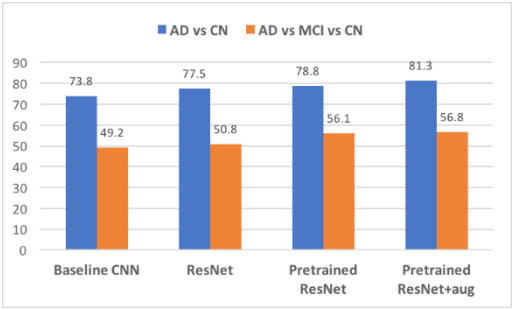
Michael Nissenson is a Computer Science and Software Engineering undergraduate student that joined Dr. Dong Si's research group in the department of computer science in the School of Science, Technology, Engineering & Mathematics (STEM) at the University of Washington Bothell in October 2016. Michael found out about the research group after attending a research fair on campus. In Dr. Si's lab, Michael received training on analysis of cryo-EM data. His work culminated in a workshop paper, titled "Automated Protein Chain Isolation from 3D Cryo-EM Data and Volume Comparison Tool", which he presented at the Computational Structural Biology Workshop at the ACM-BCB 2017 conference in Boston, Massachusetts.

At CSBW, Michael presented two tools that he developed in Dr. Si's lab and that can be used the by the larger cryo-EM community. Both tools have been implemented as add-ons that can be used inside UCSF Chimera, which is the current most popular cryo-EM visualization tool. The first tool allows users to quickly generate experimental training data to be fed into machine learning algorithms, reducing the time it takes to create this type of data from hours to seconds. The second tool provides a rough estimate of the quality of experimental cryo-EM data, allowing them to find out if some data is missing from the experimental data. More information about Michael's work can be found in the ACM proceedings of ACM-BCB 2017. Michael is continuing this line of work in Dr. Si's work. As he states, "It's really exciting to be able to present my work to others, knowing it might one day help to save lives."

Aly Valliani and Ameet Soni
Swarthmore College

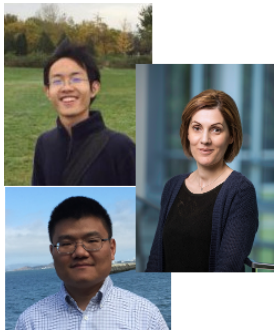


Aly Valliani is a Computer Science undergraduate student working in Dr. Soni's laboratory in the department of computer science at Swarthmore College. In Dr. Soni's lab, Aly focused on machine learning tools of relevance for Alzheimer's. His work culminated in a poster, titled "Deep Residual Nets for Improved Alzheimer's Diagnosies", which he presented at the ACM-BCB 2017 conference in Boston, Massachusetts.

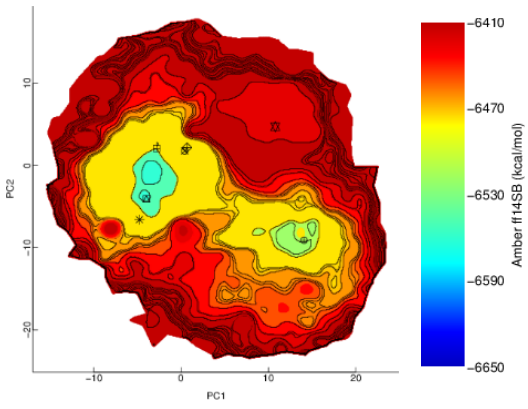


At ACM-BCB, Aly ResNet, a deep residual network, for the task of predicting Alzheimer's Disease from brain images. The work was based on the hypothesis that deep, pretrained convoluted neural networks learn cross-domain features that improve low-level interpretation of images. Aly evaluated this hypothesis using MRI brain images from the Alzheimer's Disease Neuroimaging Initiative (ADNI) with each patient being diagnosed as having Alzheimer's Disease (AD), Cognitively Normal (CN), or demonstrating Mild Cognitive Impairment (MCI). The results (some of which are highlighted on the right) confirm the hypothesis, as a ResNet pretrained on millions of natural images outperforms a randomly initialized ResNet on both 2-way (AD vs CN) and the harder 3-way (AD vs. MCI vs. CN) prediction tasks. Furthermore, the results show the importance of depth (ResNet structures outperform a Baseline CNN) and data augmentation to prevent overfitting. More information about Aly's work can be found in the ACM proceedings of ACM-BCB 2017.

Xiaowen Fang, Wanli Qiao, and Amarda Shehu
George Mason University



Xiaowen Fang is an undergraduate student in the department of Computer Science at George Mason University. He is part of the China 1-2-1 program that allows Chinese undergraduate students to attend an American university for two years. Xiaowen joined the computer science laboratory of Prof. Shehu and the statistics laboratory of Dr. Qiao to conduct research on a project funded by the Jeffress Memorial Trust Award in Interdisciplinary Sciences, which exclusively promotes undergraduate research. Xiaowen's research has already resulted in a conference manuscript that is currently under review.



Xiaowen's research is on a novel framework to analyze molecular energy landscapes with tools from spatial data analytics and high-dimensional geometry. At the moment, Xiaowen has put together two tools, one that automatically detects basins in a landscape, and another that automatically detects saddles. A two-dimensional projection of a landscape with automatically-detected basins is shown on the right. This research is highly interdisciplinary and serves to reveal the organization of molecular energy landscapes that is crucial to understanding the structure and thermodynamic basis of human diseases that are proteinopathies. Xiaowen is continuing this line of work, and a journal manuscript with him as first author is currently in the works.