2022 IEEE International Parallel and Distributed Processing Symposium
Workshops (IPDPSW)

21st IEEE International Workshop on High Performance Computational Biology (HiCOMB 2022)

May 30, 2022

Virtual Workshop

The size and complexity of data sets in bioinformatics continues to grow at an increasing rate, and the analysis of these complex and error-prone data sets demands efficient algorithms and hardware. Recently a new approach, namely pangenomics, started to gain attraction, which aims to analyze collections of genomes jointly or use them as reference, replacing the 20+ years practice of utilizing a single reference genome. This paradigm shift will require better algorithms and hardware to enable faster computation and better memory management. Hence high-performance computing (HPC) has become an integral part of research and development in bioinformatics, computational biology, and medical and health informatics. The goal of this workshop is to provide a forum for discussion of latest research in developing HPC solutions to data- and compute-intensive problems arising from all areas of computational life sciences.

This year's program will feature a keynote talk by Christina Boucher from University of Florida and an invited talk by Yatish Turakhia from University of California, San Diego. We received 21 submissions to the workshop. Each submission was reviewed by three program committee members, and we accepted seven submissions as full papers and five submissions as extended abstracts. The authors of the accepted papers will also present at the workshop. We thank the authors, the program committee members, and the keynote and invited speakers for contributing to this year's high-quality technical program.

Workshop General Chairs:
Alba Cristina M. A. de Melo, University of Brasilia
Ananth Kalyanaraman, Washington State University

Program Chair:
Can Alkan, Bilkent University

Program Committee:
Mohammed Alser, ETH Zurich
Rolf Backofen, University of Freiburg
Irem Boybat, IBM Zurich
Sarah Bruningk, ETH Zurich
Somali Chaterji, Purdue University
Ercument Cicek, Bilkent University
Priyanka Ghosh, NIH NCBI
Zam Iqbal, EMBL EBI
Benjamin Langmead, Johns Hopkins University
Ryan Layer, University of Colorado Boulder

Serghei Mangul, University of Southern California
Ibrahim Numanagic, University Victoria
Pierre Peterlongo, University of Rennes
Knut Reinert, Freie Universitat Berlin
Jared Simpson, Ontario Institute for Cancer Research
Ewa Szczurek, University of Warsaw
Yatish Turakhia, UCSD
Leonid Yavits, Bar-Ilan University and Technion
Federico Zambelli, University of Milan

HiCOMB 2022 Keynote Speaker

Building scalable indexes that can be efficiently queried

Christina Boucher, University of Florida

Abstract: Recently, Gagie et al. proposed a version of the FM-index, called the r-index, that can store thousands of human genomes on a commodity computer. We later showed how to build the r-index efficiently via a technique called prefix-free parsing (PFP) and demonstrated its effectiveness for exact pattern matching. Exact pattern matching can be leveraged to support approximate pattern matching but the r-index itself cannot support efficiently popular and important queries such as finding maximal exact matches (MEMs). To address this shortcoming, Bannai et al. introduced the concept of thresholds, and showed that storing them together with the r-index enables efficient MEM finding --- but they did not say how to find those thresholds. We present another novel algorithm that applies PFP to build the r-index and find the thresholds simultaneously and in linear time and space with respect to the size of the prefix-free parse. Our implementation can rapidly find MEMs between reads and large sequence collections of highly repetitive sequences. Compared to existing methods, ours used 2 to 11 times less memory and was 2 to 32 times faster for index construction. Moreover, our method was less than one thousandth the size of competing indexes for large collections of human chromosomes.

Biography: Dr. Boucher is an Associate Professor in the Department of Computer and Information Science and Engineering at the University of Florida. She has over 85 publications in bioinformatics, with over two dozen of them in succinct data structures and/or alignment. Considering this, she was a keynote speaker for IABD 2019, FAB 2018, RECOMB-SEQ 2016 and the ECCB 2016 Workshop on Pan-Genomics. She is a recipient of an ESA 2016 Best Paper Award. She oversees the development and maintenance of several software methods, including MEGARes and AMRPlusPlus, METAMarc, Kohdista, Vari, VariMerge — and most recently, Moni. In addition, she has built a team of collaborators in various biomedical sciences including microbiology, veterinarian medicine, epidemiology, public health, and clinical sciences.

In addition, she actively works on increasing the diversity in bioinformatics education. Her efforts include being a member of the University of Florida's Implicit Bias committee, being a panellist for the NSF-funded ACM BCB 2015 Women in Bioinformatics meeting, serving as a faculty advisor for an ACM-W chapter, and being an active member of the Diversity Committee for over three years. She also received a fellowship from The Institute for Learning and Teaching (TILT) for her course redevelopment and served on the advisory committee for an NSF Research Traineeships Program.

She has been the PC chair for several conferences, including SPIRE 2020, RECOMB-Seq 2019, and ACM-BCB 2018. Most recently, she was nominated to serve on the NIH BDMA Study Section as a Standing Member, and a Board Member of SIG BIO.

HiCOMB 2022 Invited Speaker

Pandemic-scale Phylogenetics

Yatish Turakhia, University of California, San Diego

Abstract: Phylogenetics has been central to the genomic surveillance, epidemiology and contact tracing efforts during the COVD-19 pandemic. But the massive scale of genomic sequencing has rendered the pre-pandemic tools quite inadequate for comprehensive phylogenetic analyses. In this talk, I will discuss a high-performance computing (HPC) phylogenetic package that we developed to address the needs imposed by this pandemic. Orders of magnitude gains were achieved by this package through several domain-specific optimization and parallelization techniques. The package comprises four programs: UShER, matOptimize, RIPPLES and matUtils. Using high-performance computing, UShER and matOptimize maintain and refine daily a massive mutation-annotated phylogenetic tree consisting of all (>9M currently) SARS-CoV-2 sequences available on online repositories. With UShER and RIPPLES, individual labs – even with modest compute resources – incorporate newly-sequenced SARS-CoV-2 genomes on this phylogeny and discover evidence for recombination in real-time. With matUtils, they rapidly query and visualize massive SARS-CoV-2 phylogenies. This has empowered scientists worldwide to study the SARS-CoV-2 evolutionary and transmission dynamics at an unprecedented scale, resolution and speed. This has laid the groundwork for future genomic surveillance of MOST infectious pathogens.

Biography: Dr. Turakhia is an Assistant Professor in the Department of Electrical and Computer Engineering at the University of California San Diego since July 2021. He is also affiliated with the Department of Computer Science and Engineering and the Bioinformatics and Systems Biology graduate program at UCSD. His lab is also affiliated with the Center of Machine-Integrated Computing and Security and Center of Microbiome Innovation at UCSD. Previously, he was a postdoctoral scholar at the Genomics Institute at UC Santa Cruz. Dr. Turakhia obtained his Ph.D. in Electrical Engineering in 2019 from Stanford University and his bachelor's and master's degrees in Electrical Engineering from the Indian Institute of Technology (IIT), Bombay in 2014.