Chengze Shen

Phone: (858) 291-2443 | Email: chengze5@illinois.edu | Personal Page | LinkedIn | Google Scholar | GitHub

EDUCATION

University of Illinois, Urbana-Champaign, IL (current GPA: 3.85/4.00) August 2020 - present
Ph.D. in Computer Science
QUAL - passed | PRELIM - passed | Expected Graduation - May 2024

Carnegie Mellon University, Pittsburgh, PA (GPA: 3.91/4.00) August 2018 - May 2020
M.S. in Computational Biology

University of California, San Diego, CA (GPA: 3.89/4.00, Magna Cum Laude**)** Sept 2015 - June 2018
B.S. in Bioengineering: Bioinformatics (minor in Cognitive Science)

Related coursework

UIUC (Ph.D.):	Advanced Bioinformatics	Cancer Genomics	**Parallel Programming**
Advanced Algorithms	**GPU programming**	Genomic Biology	Network and Clustering
CMU (M.S.):	**Machine Learning**	**Deep Learning**	Modeling and Simulation
Bioimage Informatics	Computational Genomics	Cell Molecular Biology	**Algorithms**

TECHNICAL SKILLS

Research Expertise: Machine learning, deep learning, GPU computing, sequence data analysis, large-scale analysis, multiple sequence alignment, phylogenetic tree estimation, phylogenetic placement, taxonomic identification

Programming Languages: Python, Rust, C++, C, Cuda C, shell/bash scripts, Java

Frameworks: PyTorch, TensorFlow, Git, AWS EC2, S3 bucket, Slurm, HPC computing, UNIX/Linux environment

Spoken Languages: Mandarin (native), English (fluent), Japanese (basic read/write/listen/speak)

WORK EXPERIENCE

Machine Learning Intern in R&D team, **Human Longevity Inc.** May - August 2019

- Worked with a mentor in developing a new machine learning pipeline to study risk factors of human aging.
- Procured large datasets from UK Biobank, NHANES, and GEN3 biomarkers and transformed data for model training
- Optimized both prediction quality and computations by training and evaluating various machine learning models: gradient/ada boosting machine, random forest, elastic net, and deep neural network.
- Compiled and built a reusable pipeline with Pytorch for publication and company use cases.

RESEARCH & TEACHING EXPERIENCE

Graduate Research Assistant at Warnow Lab, **UIUC** August 2021 - present

- Worked as a Ph.D. student to study multiple sequence alignment (MSA).
- Identified low accuracy issues of current MSA methods when inputs have sequence length heterogeneity and high rates of evolution.
- Developed novel MSA methods that beat state-of-the-art methods in accuracy and can scale to ultra-large data.
- Published the new methods (see Software WITCH and EMMA) in reputable journals and GitHub.
- (Current) Developing TIPP3, a new marker-gene-based abundance profiling tool for accurate and scalable microbiome analysis on next-gen sequencing data.

Graduate Research Assistant at Dr. Hosein Mohimani Lab, **Carnegie Mellon University** Dec 2018 - Sept 2020

- Worked on new methods for fast and efficient read mapping to other reads or genomes.
- Developed a new method using Distributed Sensitive Bucketing, Hidden Markov Models, and prefix/suffix trees.
- Our new method, DSB, can map reads more sensitively than Minimap2, BlasR, and MMSeqs2 on distant homologs without losing accuracy.

- Published our new method on Nature Scientific Report and GitHub (see Software DSB).

Teaching Assistant for CS125: Introduction to Computer Science, **UIUC** August 2020 - May 2021

- Organized discussion sections to review homework questions and prepare for exams.
- Led the development of open-access quiz review materials for all students (URL).

PUBLICATIONS

1. Chengze Shen, Paul Zaharias, and Tandy Warnow. **MAGUS+eHMMs: Improved Multiple Sequence Alignment Accuracy for Fragmentary Sequences**. Bioinformatics, Volume 38, Issue 4, 15 February 2022, pp. 918–924 (URL).
2. Park, Minhyuk, Stefan Ivanovic, Gillian Chu, Chengze Shen, and Tandy Warnow. **UPP2: Fast and Accurate Alignment of Datasets with Fragmentary Sequences.** Bioinformatics 39, no. 1 (January 1, 2023) (URL).
3. Chengze Shen, Minhyuk Park, and Tandy Warnow. **WITCH: Improved Multiple Sequence Alignment Through Weighted Consensus Hidden Markov Model Alignment.** Journal of Computational Biology Special Issue for Michael Waterman, 2022 29:8, 782-801 (URL).
4. Chengze Shen, Baqiao Liu, Kelly P. Williams, and Tandy Warnow. **Computing Multiple Sequence Alignments given a Constraint Subset Alignment using EMMA.** Workshop in Algorithms for Bioinformatics 2023 (URL). Also submitted for publication in Algorithms for Molecular Biology (invited).
5. Eleanor Wedell, Chengze Shen, and Tandy Warnow. **BATCH-SCAMPP: Scaling phylogenetic placement methods to place many sequences.** Workshop in Algorithms for Bioinformatics 2023 (URL).
6. Mihir Mongia*, Chengze Shen*, Arash Gholami Davoodi, Guillaume Marcais, and Hosein Mohimani. **Large Scale Sequence Alignment via Efficient Inference in Generative Models**, Sci Rep 13, 7285 (2023) (URL).

COURSE PROJECT HIGHLIGHTS

GPU-BLAST-plus (Ph.D., GitHub) Jan - May 2023

- Implemented block-queuing kernels and constant memory, added thread number, and reduced synchronization.
- Achieved 4x speedup compared to baseline GPU-BLAST.

Deep Learning Projects (M.S., GitHub) Jan - May 2019

- Implemented ConvNet layers from scratch for image classification tasks: fully-connected layer, convolution layer, activations, batch normalization, dropout, pooling layers, and loss functions.
- Implemented a translation recurrent neural network with Attention module and Beam Search algorithm.
- Applied Convolution neural network to solve ordinary and partial differential equations.

AWARDS

- Professional Honors/Industry Track - CMU May 2020
- Provost Honors - UCSD Fall 2015, F/W/S 2016, F/W/S 2017, W/S 2018

SOFTWARE

VireTap - https://github.com/c5shen/VireTap
- A bash script framework for onco-viral transcriptome detection.

WITCH - https://github.com/c5shen/WITCH
- A multiple sequence alignment tool for large-scale data with sequence length heterogeneity based on HMMs.

EMMA - https://github.com/c5shen/EMMA
- A multiple sequence alignment tool for large-scale data with sequence length heterogeneity based on MAFFT.

DSB - https://github.com/mohimanilab/DistributionSensitiveBucketing
- An efficient sequences/reads mapping tool with high recall and precision even for distant homologs.

GPU-BLAST-plus - https://github.com/c5shen/GPU-BLAST-plus
- A faster version of the GPU version of the BLAST algorithm.

PEOPLE OF REFERRAL

> Kelly P. Williams	Sandia National Laboratories @ Livermore	kpwilli@sandia.gov
> Tandy Warnow	University of Illinois, Urbana-Champaign	warnow@illinois.edu