Chengze Shen

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

I am a computer scientist with strong programming skills and a broad background in bioinformatics.
I am looking for a permanent position in bioinformatics industry. I want to develop scalable solutions and software for challenging problems in biology where advanced computer science is needed.

EDUCATION

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

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

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

RELEVANT COURSES

UIUC (Ph.D.): Algorithmic Genomic Biology, Advanced Bioinformatics, Computational Cancer Genomics, Applied Regression and Design, Applied Parallel Programming Advanced Social Network, Algorithms, Bioinformatics and Computation, Computational Scientometrics, Manycore Parallel Algorithms.

CMU (M.S.): Machine Learning, Deep Learning, Applied Cell Molecular Biology, Convex Optimization, Biology Modeling and Simulation, Bioimage Informatics, Algorithm/Advanced Data Structure, Automation, Computational Genomics, Computer System, Computational Medicine.

RESEARCH INTERESTS

Bioinformatics, Computational Biology, Microbial Analysis, Metagenomics, Phylogenetics, Phylogenomics.

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. **SALMA: Scalable ALignment using MAFFT-Add.** Available on Biorxiv (URL).
5. Mihir Mongia*, Chengze Shen*, Arash Gholami Davoodi, Guillaume Marcais, and Hosein Mohimani. **Large Scale Sequence Alignment via Efficient Inference in Generative Models**. Under revision at Nature Scientific Report.

RESEARCH

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

- Developing novel multiple sequence alignment methods to address sequence length heterogeneity and ultra-large-scale data (see papers 1-4 above).
- Developing TIPP3, a new marker-gene-based abundance profiling tool for accurate and scalable microbiome analysis.

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

- Used Distributed Sensitive Bucketing and Hidden Markov Model algorithms for efficient sequences and reads mapping in large-scale data (see paper 5 above).

Chengze Shen

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

TEACHING

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

PROJECTS

VireTap (M.S., GitHub) Jan – March 2019

- Led the software development for VireTap, an onco-viral transcriptome detection tool in human cancer models using Tophat, Trinity, and BLAST.
- The software has been maintained and published on GitHub for public use.

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 language translation recurrent neural network with Gated Recurrent Unit, Attention module, and Beam Search algorithm.
- Applied Convolution neural network to solve ordinary and partial differential equations such as Laplace Equation.

INTERNSHIP

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

- Explored and implemented machine learning models to learn human age with data sets from UK Biobank, NHANES, and GEN3 biomarkers.
- Trained and benchmarked gradient/ada boosting machine, random forest, elastic net, and deep neural network.
- Improved deep neural network accuracy with optimized network architecture and hyper-parameters random search.
- Built a reusable PyTorch deep neural network training pipeline with GPU support.

AWARDS

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

SKILLS

Programming Languages: Python, Rust, C++, C, shell scripts, Java, MATLAB, Swift, PHP, JavaScript
Frameworks: PyTorch, TensorFlow, Bootstrap, Django, MySQL, Git, AWS EC2, SageMaker, S3 bucket, Slurm
Languages: Mandarin (native), English (fluent), Japanese (basic read/write/listen/speak)

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.

SALMA - https://github.com/c5shen/SALMA
- 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.