K. Jun **Gao**

+1 312-731-4398 / +1 416-881-0140 | kgao@cs.toronto.edu | www.cs.toronto.edu/~kgao | gaojunxuan | junx-gao

Undergraduate student in computer science and bioinformatics at the University of Toronto (2024 graduation). Interested in algorithms and data structure research with application in computational biology, especially in the area of space efficient genome analysis.

Education

University of Toronto *Toronto, ON*

Hon. B.Sc. in Computer Science and Computational Biology *Sep. 2020 - Jun. 2024*

- **Major** (specialist) in Computer Science, Bioinformatics and Computational Biology; Focus in Theory of Computation; Minor in Mathematics
- **Relevant coursework**: Software Design, Algorithm Design, Data Structures and Analysis, System Programming, Operating System, Database Management System, Parallel Programming, Machine Learning, System Biology, Applied Bioinformatics, Genetics, Molecular Biology
- **CGPA**: 3.94/4.00

Coursework

- **CSC265**: Enriched Data Structures and Analysis
- **CSC373**: Algorithms Design
- **CSC343**: Intro to Databases
- **CSC367**: Parallel Programming
- **CSC207**: Software Design
- **CSC209**: System Programming
- **CSC311**: Intro to Machine Learning
- **CSC2420**: Algorithm Design, Analysis and Theory: Online and Myopic Algorithms
- **CSC463**: Computational Complexity and Computability
- **CSC438**: Computability and Logic
- **CSC473**: Advanced Algorithms
- **MAT344**: Intro to Combinatorics
- **MAT332**: Intro to Graph Theory
- **MAT495**: Independent Reading in Math (Math for Massive Data Analysis)
- **BCB410**: Applied Bioinformatics
- **BCB420**: Computational Systems Biology

Skills

- **Languages**: C++, C, C#, Java, Python, JavaScript/TypeScript, R, Shell
- **Platforms & Frameworks**: Node.js, React, React Native, Flask, Celery, Django, Docker, Microsoft Azure, Universal Windows Platform, WinUI3 and Windows App SDK, .NET, PyTorch, Numpy, Scikit-learn, Snakemake, Workflow Description Language
- **General skills**: Object-oriented programming, software design, debugging and problem solving, database (MySQL, SQLite), web development, backend API development, operating systems, algorithm design, combinatorics and combinatorial optimization, string processing algorithms, graph theory, information and coding theory, bioinformatics.

Research Experience

A new approach for efficient storage and retrieval of homology data

KJ Gao, I Pilizota, F Martin, D Thybert *Sep. 2022*

We created a new data structure and index format for storing homology pair data, accommodating a large homology database. By leveraging gene tree hierarchy, we avoided storing all homologous relationships, reducing space complexity from $O(n^2)$ to $O(n \log n)$. We also implemented interval-based labeling to efficiently parse trees and extract homology information.

ChromMiniGraph: Data structure and algorithms for mapping sequencing reads to population references

J Shaw, KJ Gao, J Simpson, YW Yu *Ongoing*

We created ChromMiniGraph, a memory-efficient tool for constructing pangenome references that utilizes k-mer sampling and node coloring to reduce storage while maintaining accuracy. ChromMiniGraph maps reads efficiently and accurately using subsampling and colinear chaining on a linearized coordinate. ChromMiniGraph offers a streamlined workflow for pangenome references, read phasing, and structural variation identification.

Conference Presentations & Publications

ChromMiniGraph: Data structure and algorithms for mapping sequencing reads to population references

Lyon, France

J Shaw*, KJ Gao*, J Simpson, YW Yu (* co-first authorship) *July, 2023*

Accepted for poster presentation in HitSeq (high-throughput sequencing) track at ISMB/ECCB 2023 in Lyon, France

Work Experience

Ontario Institute for Cancer Research (OICR)

Toronto, ON

Research Student

Sep. 2022 - Present

- Worked on the development of new algorithms and data structures for **haplotype mapping**, advancing the potential of personalized medicine through individual genome analysis.
- Used chromatic **minimizer graph** and designed **dynamic programming** graph chaining algorithms and **topological linearization algorithm** to facilitate the construction of genome graphs for human pangenome and read mapping onto the pangenome. Analyzed the performance on HPC clusters.

European Bioinformatics Institute (EMBL-EBI)

Remote

Google Summer of Code Student

Jun. 2022 - Sep. 2022

- Devised scalable solution using **tree structures**, XML, and Newick format to replace map-based caches, **reducing storage requirement by over 90%**.
- Led the development of interval-based tree labeling algorithms for querying the new data format. Optimized for **batch query** using **tree labeling** and allowed for **30000+ homology queries per second** on gene trees with more than 1000 genes. Developed command-line tool in **C++**.

University of Toronto

Toronto, ON

Software Developer & Research Assistant

Sep. 2021 - Present

- Worked with graduate students from LMSE lab to develop and evaluate novel **machine learning algorithms** using **Sklearn and PyTorch** for generating **protein embedding** with high accuracy: **0.93 ROC-AUC score** in classification test and **R-value of 0.84** in regression test.
- Developed efficient and highly scalable **distributed backend** APIs on **high-performance computing clusters** using **Celery**, **Redis DB** for a web-based **ML prediction pipeline** used by a group of more than **20 researchers and students** in the Laboratory for Metabolic Systems Engineering at the University of Toronto.

Teaching

- **CSC165: Mathematical Expression and Reasoning for Computer Science**
  Introductory discrete math course
  In-class and marking teaching assistant — Winter 2022 and Winter 2023
- **CSC236: Intro to Theory of Computation**
  Second-year theory of computation course covering algorithm analysis, correctness proof, and formal language theory
  Tutorial and recitation teaching assistant — Fall 2022

Personal Projects

- **Rust-LSD (Rust)**
  Rust implementation of the direct superbubble detection graph algorithm as proposed by Gärtner et al. in their 2019 paper. Improved efficiency compared to the Python reference implementation and allowed for easy integration with existing bioinformatics software written in Rust.
- **JPDict (UWP and WinUI3, iOS/Android via React Native)**
  Popular Japanese language learning app on **Windows UWP** platform with a **userbase of 6000+ users**. The app includes features specially designed for language learners such as sentence structure analyzer and verb conjugation helper implemented using natural language processing techniques. Employed software engineering and design principles like MVVM with highly decoupled components for scalability. Developed and maintained the backend on **Azure Functions** serverless computing platform.
- **GradeTree (React Native)**
  Mobile app for the student and parent portal at Chicago Public Schools with responsive and accessible interface. Allowed student and parents to access academic and attendance records with ease. Data is obtained through web scraping and user data is stored locally with encryption. Gained recognition across the Chicago Public Schools district and inspired students in the district to learn to code.
- **Tokyo Planner (React Native, Flask, and C#)**
  Participated in the Tokyo Data Challenge and developed travel planning app using open data provided by the Tokyo Metropolitan Government and participating corporations. Created mobile app using React Native and highly customizable **routing algorithms** in **C#**. Finished product received **INIAD Special Award** in the contest.

Honors & Awards

2020 **Illinois State Scholar**, Illinois Student Assistance Commision
2020 **International Scholar**, University of Toronto
2021 **Dean's List Scholar**, University of Toronto, Faculty of Arts and Science
2021-2023 **Dean's List Scholar**, University of Toronto, Faculty of Arts and Science
2022 **St. Michael's College In-Course Scholarship**, University of Toronto