UNIVERSITY OF TORONTO
FACULTY OF ARTS & SCIENCE | Department of Mathematics

Departmental PhD Thesis Exam

Thursday, June 13, 2024 at 11:00 a.m. (sharp)
via Zoom / BA6183

PhD Candidate : Jim Shaw

Supervisor : Yun William Yu

Thesis title : Practical and theoretical problems in biological sequence comparison

Abstract

DNA sequencing technologies have revolutionized the study of biology, allowing unprecedented access to the blueprint of life – our genomes. As these technologies have matured, massive amounts of DNA sequences are now becoming available for computers to analyze. The increasing throughput of these technologies has rendered old algorithms unusable. In this thesis, we first build a principled mathematical foundation for approximate DNA string matching, also called sequence alignment. We then show that using theoretically-backed approaches can result in faster and better software implementations, resulting in useful new tools for biologists.

In Chapter 3, we rigorously analyze how to subsample DNA strings to speed up alignment while retaining sensitivity. We show how to determine a sampling algorithm's conservation, which is a measure of how sensitively it can match strings. Surprisingly, different sampling methods have vastly different conservation even when retaining the same amount of "information". We show that modifying existing software to use better subsampling algorithms gives more sensitive results.

In Chapter 4, we provide the first non-trivial runtime and accuracy bounds on a widely-used DNA alignment algorithm called seed-chain-extend. We break the worst-case quadratic runtime barrier of sequence alignment by performing an average-case analysis under a probabilistic evolutionary model of DNA sequence. Our results are concordant with algorithmic results on real data and provide new insights into the rigorous analysis of sequence alignment.

In Chapter 5, we utilize the subsampling and the seed-chain-extend approaches analyzed in Chapters 3 and 4 to build a new genome-genome comparison method and software called skani. skani can estimate the evolutionary divergence between two genomes $>$ 25 times faster than previous algorithms. We show that skani can compare hundreds of thousands of genomes in seconds on a standard desktop computer, enabling large-scale comparisons not possible before.