

Genome-Scale Algorithm Design (PhD reading course, 7.5hp)

Course Description:

Combinatorial algorithms and discrete data structures are put to the test by genomic data, which routinely reaches gigabytes to terabytes in scale. This is a reading course based on the book *Genome-Scale Algorithm Design*, covering advanced algorithms and data structures that form the foundation of modern bioinformatics: suffix trees and wavelet trees, succinct and compressed data structures, and graph-based approaches to sequence alignment, genome assembly, and pangenomics.

We will study both the theoretical guarantees behind these algorithms and the engineering techniques needed to make them work on real-world genomic data. Emphasis is placed on connecting core algorithmic ideas in string matching, graph algorithms, and space-efficient data structures to their applications in high-throughput sequencing problems such as read alignment, variant calling, and genome assembly.

Course coordinator: Kristoffer Sahlin (ksahlin@math.su.se)

Meeting plan: Course pace 25%. Course period covers parts of HT 2026 - VT 2027.

Meetings are once a week. Each meeting lasts 2-2.5 hours (except the short intro meeting below).

- Short course intro meeting: **Wednesday September 16 at 11:00-11:30.**
- Tentative schedule: Wednesdays 9:30-12:00 (starting Sept 23 with some gap weeks, will be discussed at intro meeting)

Place: TBD

Learning goals:

1. Learn a set of advanced algorithms and data structures (e.g., dynamic range minimum queries, rank-select operations, Wavelet trees, MSA, HMMs, BWT variations), and be able to reason about their correctness and time and space complexity.
2. Learn the core set of problems in bioinformatic sequence analysis and some of their solutions, focusing on scalability to large genomic data sets.
3. Present and summarize in a clear and structured way the acquired knowledge by delivering lecture-style presentations of the reading material.

Prerequisites:

To qualify for the course, you should have a basic knowledge in discrete mathematics and probability theory, as well as intermediate knowledge in Computer Science with a focus on algorithms, data structures, and complexity analysis.

This means knowledge roughly equivalent to the following courses at SU: Mathematics I, 30 ECTS (MM2001), Programming techniques, 7.5 ECTS (DA2005), Probability theory 1, 7.5 ECTS (MT3001), Computer Science for Mathematicians, 7.5 ECTS (DA3018), Algorithms and Complexity, 7.5 ECTS (DA4005), Data structures and algorithms, 7.5 ECTS (DA4006).

Assessments:

1. Hold formal presentations based on the material presented in the book.

2. Prepare questions/insights/remarks on the material and presentation when not presenting.
3. Write a short 1-2 page proposal for how at least one of the algorithms/data structures/techniques presented in the course could help your research.

Literature: Mäkinen V, Belazzougui D, Cunial F, Tomescu AI. *Genome-Scale Algorithm Design: Bioinformatics in the Era of High-Throughput Sequencing*. 2nd ed. Cambridge University Press; 2023.

Reading Course Schedule

Meeting	Contents
1	Chapter 1 and 2: Molecular Biology and High-Throughput Sequencing Algorithm Design
2	Chapter 3: Data structures
3	Chapter 4: Graphs
4	Chapter 5: Network flows
5	Chapter 6 (sections 6.1–6.3): Alignments
6	Chapter 6 (sections 6.4–6.9): Alignments ctd.
7	Chapter 7: HMMs
8	Chapter 8: Classical Indexes
9	Chapter 9: Burrows-Wheeler indexes
10	Chapter 10: Alignment-based genome analysis
11	Chapter 11: Alignment-free genome analysis and comparison
12	Chapter 12: Compression of genomes
13	Chapter 13: Fragment assembly
14	Chapter 14: Haplotype analysis
15	Chapter 15: Pangenomics
16	Chapter 16: Transcriptomics
17	Chapter 17: Metagenomics