Algorithms in Genome Analysis Spring 2023 – Lecture 1

LECTURES: VELI MÄKINEN
EXERCISES: NICOLA RIZZO

TEXTBOOK: WWW.GENOME-SCALE.INFO

COURSE PAGE: HTTPS://STUDIES.HELSINKI.FI/COURSES/CUR/HY-OPT-CUR-2223-A870D69D-ABBD-4598-990B-A51CB98992F4/LSI31007/ALGORITHMS_IN_GENOME_ANALYSIS_LECTURES

Course introduction

WHAT YOU MIGHT LEARN

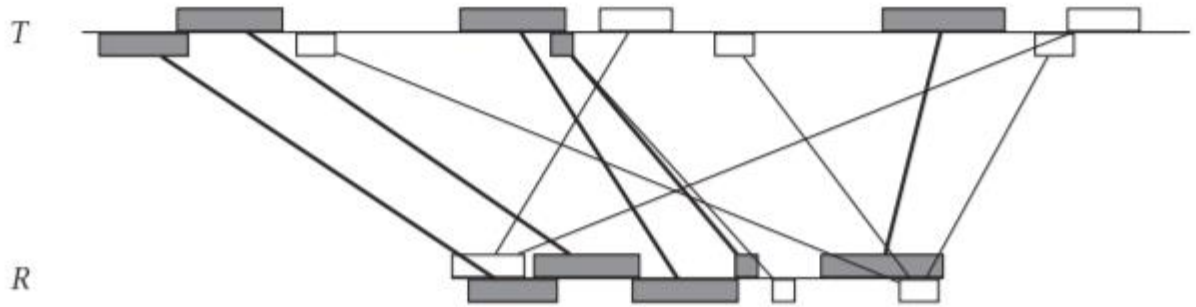
Prerequisites

- Some algorithms and basic statistics are assumed as background (e.g., *Elements of Bioinformatics, Design and Analysis of Algorithms)*
- Elective course: Bioinformatics and Systems Medicine study track / Life Science Informatics Master's Programme; Applied algorithms / Algorithms study track / Master's Programme in Computer Science
- Suitable for non-CS students also, but you'd probably need to have taken some algorithms course beforehand
- The required minimal knowledge of molecular biology is covered mostly during this first week
- *Python* used in some of the exercises (see also project associated with *Data Analysis with Python* course)
- The focus is on *algorithms* in genome analysis, but we follow the probabilistic notions common in *bioinformatics*
- This course used to be called *Biological Sequence Analysis*.

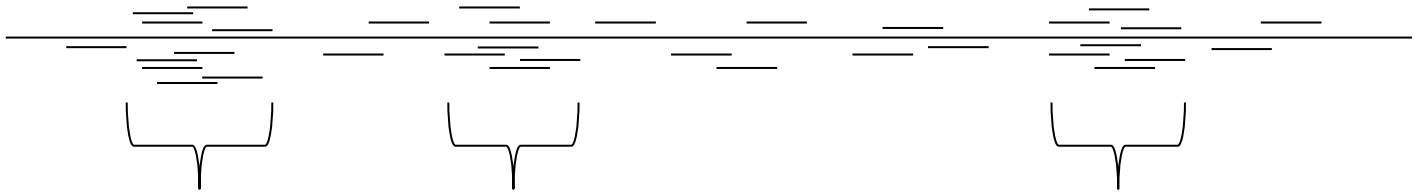
3

Topics / techniques / applications

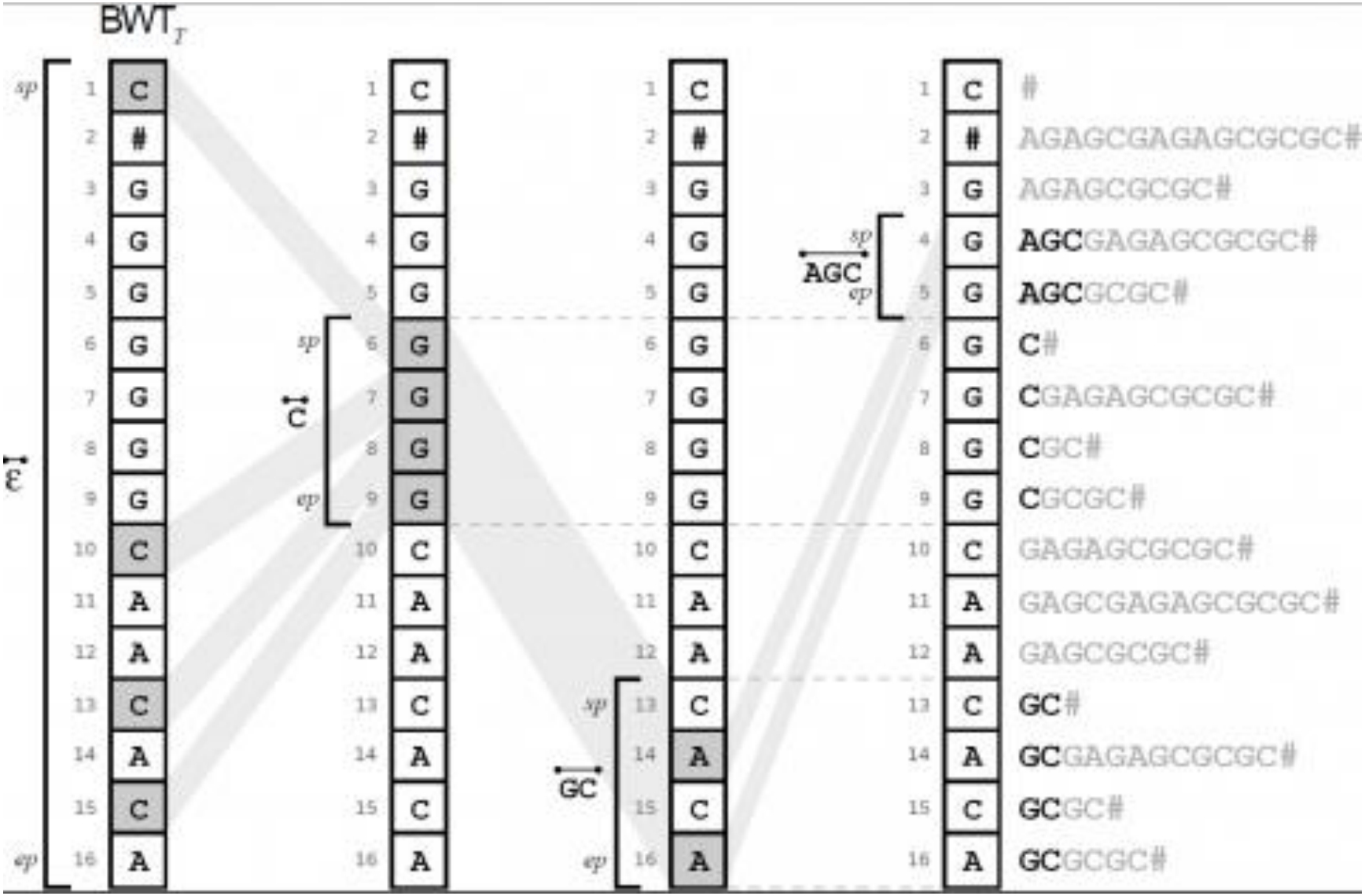
- Alignments / optimized dynamic programming / alternative splicing, phylogenetics
- Hidden Markov Models / dynamic programming, machine learning / gene prediction, peak detection
- High-throughput sequence analysis / text indexing, Burrows-Wheeler transform, minimizers / read alignment, variant calling



Spliced alignment by chaining local matches



Peak detection through segmentation with HMMs



Fast read alignment using BWT indexing

Motivation

- Topics not too far from current research

***De novo* assembly and genotyping of variants using colored de Bruijn graphs**

Zamin Iqbal, Mario Caccamo, Isaac Turner, Paul Flicek & Gil McVean

Nature Genetics **44**, 226–232(2012) | Cite this article

Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype

Daehwan Kim, Joseph M. Paggi, Chanhee Park, Christopher Bennett & Steven L. Salzberg

Nature Biotechnology **37**, 907–915(2019) | Cite this article

Genotyping structural variants in pangenome graphs using the vg toolkit

Glenn Hickey, David Heller, Jean Monlong, Jonas A. Sibbesen, Jouni Sirén, Jordan Eizenga, Eric T. Dawson, Erik Garrison, Adam M. Novak & Benedict Paten

Genome Biology **21**, Article number: 35 (2020) | Cite this article

Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms (SODA)

< Previous Chapter | Table of Contents | Next Chapter >

Abstract | PDF

Regular Languages meet Prefix Sorting*

Jarno Alanko, Giovanna D'Agostino, Alberto Policriti and Nicola Prezza

This Paper Appears in

Minimap2: pairwise alignment for nucleotide sequences FREE

Heng Li

Bioinformatics, Volume 34, Issue 18, 15 September 2018, Pages 3094–3100, https://doi.org/10.1093/bioinformatics/bty191

PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

High-Accuracy HLA Type Inference from Whole-Genome Sequencing Data Using Population Reference Graphs

Alexander T. Dilthey, Pierre-Antoine Gourraud, Alexander J. Mentzer, Nezih Cereb, Zamin Iqbal, Gil McVean

Sparse Dynamic Programming on DAGs with Small Width

Authors: Veli Mäkinen, Alexandru I. Tomescu, Anna Kuosmanen, Topi Paavilainen, Travis Gagie, Rayan Chikhi Authors Info & Affiliations

Publication: ACM Transactions on Algorithms • February 2019 • Article No.: 29 • https://doi.org/10.1145/3301312

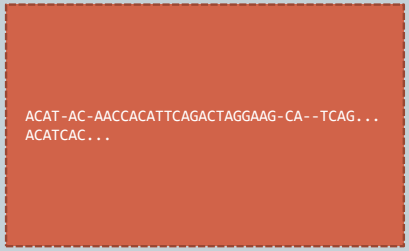
Bit-parallel sequence-to-graph alignment

Mikko Rautiainen, Veli Mäkinen, Tobias Marschall

Bioinformatics, Volume 35, Issue 19, 1 October 2019, Pages 3599–3607, https://doi.org/10.1093/bioinformatics/btz162

Example of ongoing research: Covid-19

- 30794 strains of Covid-19 were downloaded and their multiple sequence alignment (MSA) was constructed
- This MSA is of size 30794 x 29003 bases



- One can represent the MSA as a small graph that can be indexed for exact string search: such index takes only 2,2% of the size of the MSA
- Such graph / index is a so-called *pangenome* representation, useful e.g., for future vaccine development, to design antibodies for many strains (current and plausible recombinations)

Conduct of the course

NO EXAM – WEEKLY EXERCISES DETERMINE THE GRADE

Typical week

- Tuesday and Wednesday lectures 10-12
  - Some are replaced by prerecorded lecture videos
- Wednesday exercise sessions 12-14:
  - Group work for draft solutions
- Monday evening deadline to submit finalized solutions in Moodle
- See Moodle for detailed programme
- First exercise is a quiz, starting already this week!

Grading

- 8 weeks with 5 assigments each
  = 40 assignments / points
- You can postpone max 5 to the exam week
- 20 p -> grade 1
- 24 p -> grade 2
- 27 p -> grade 3
- 31 p -> grade 4
- 34 p -> grade 5

Course practices

- Course book (see course web page) contains most of the things we cover
- Some prerecorded lectures are linked in Moodle
- Most lectures are Powerpoint presentations given in the lecture room (pdf's available afterwards)
- Model solutions are not provided
- Communication by email or by Moodle discussion forum
- Course contains some new material from the 2nd edition of the course book (in press)