UNIVERSITÄT DES SAARLANDES

ZBI ZENTRUM FÜR BIOINFORMATIK

Research Topics of the Algorithmic Bioinformatics Group

Saarland Informatics Campus, Center for Bioinformatics

Prof. Dr. Sven Rahmann

14.11.2022

What is Bioinformatics ?

Definition (from the German Wikipedia page)

Bioinformatics is an interdisciplinary science
that solves problems from the life sciences
using theoretical and computer-based methods.

Main application areas (in Germany, according to FaBI)

- genes and genomes
- gene and protein expression and its regulation
- metabolic paths and networks
- structures of bio(macro)molecules, esp. DNA, RNA and proteins
- molecular interactions between DNA, RNA, proteins and chemical compounds
- molecular characterization of ecosystems

The Center for Bioinformatics at Saarland University

Modern Facilities

- Seminar rooms for small lectures and seminars
- PC pool for students
- Office space for research groups

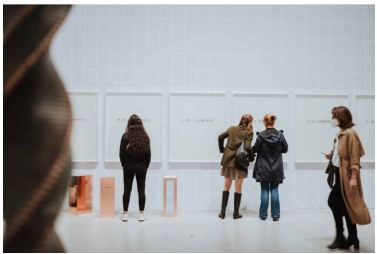
Research Groups in Bioinformatics

- Bioinformatics: Hans-Peter Lenhof
- Computational Biology: Volkhard Helms
- Clinical Bioinformatics: Andreas Keller
- Drug Bioinformatics: Olga Kalinina
- Integrative Cell Biology and Bioinformatics: Fabian Müller
- **Algorithmic Bioinformatics**: Sven Rahmann
- Spatial Transcriptomics: Fabian Kern
- Human-Microbe Systems Bioinformatics: Alexey Gurevich
- Data Driven Drug Development: Andrea Volkamer

Algorithmic Bioinformatics

Efficient algorithms for huge genomic datasets

- An individual genome can be printed on approx. 300 000 A4 pages.
- For genome-wide studies with hundreds of participants, one needs
  - a large compute cluster (and a petabyte of storage),
  - or a "modern gaming PC" and clever algorithms



DNA Sequencing: What, Why and How?

DNA sequencing: determining the sequence of nucleotides (ACGT) of each chromosome in the cell

Benefit: understanding variations in the human genome and relation to diseases

DNA Sequencing: What, Why and How?

DNA sequencing: determining the sequence of nucleotides (ACGT) of each chromosome in the cell

Benefit: understanding variations in the human genome and relation to diseases

(Illumina HiSeq 4000 Sequencer)

Second Generation Sequencers

- large device, up to 500 000 Euros
- DNA fragments of 100-300 bp
- extremely high throughput: up to 400 Gbp / day
- highly parallelized
- very accurate (error rate $< 0.1\%$)

DNA Sequencing: What, Why and How?

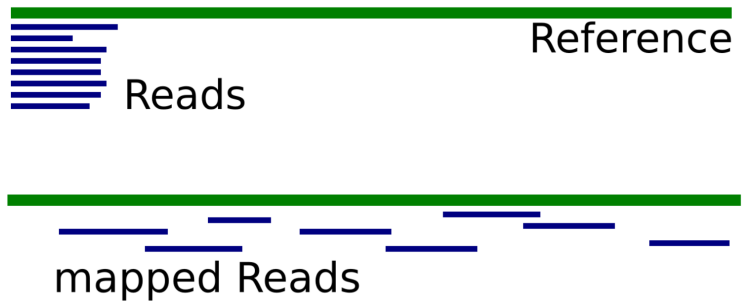
(Oxford Nanopore MinION)

Third Generation Sequencers

- sized like a USB stick (MinION)
- low initial investment
  (but running costs for chemicals)
- sequences long DNA fragments
  (10 000 bp and more)
- relatively low throughput,
  few fragments in parallel
- higher error rates (5% to 10%)

The Read Mapping Problem

Basic Question: Where is my sequenced DNA read coming from?



The Read Mapping Problem

Given

- a fragment of sequenced DNA (“read”)
- a database of known DNA (e.g., collection of **all** known genomes)

Sought

- most likely origin of the read
  - from which species ?
  - from which chromosome, at which position?
  - differences to the known reference sequence?

The Read Mapping Problem

Given

- a fragment of sequenced DNA ("read")
- a database of known DNA (e.g., collection of **all** known genomes)

Sought

- most likely origin of the read
  - from which species ?
  - from which chromosome, at which position?
  - differences to the known reference sequence?

About Uniqueness

- Is the found origin unique, or are the several plausible ones?
- How many likely places of origin are there?
- Enumerate all of them (in decreasing order of likeliness).

Read Mapping vs. Read Alignment

Mapping:

Read Mapping vs. Read Alignment

Mapping:

Alignment:

Reference sequence with aligned reads	CIGAR string	Explanation
C T G C A T G T T A G A T A A * * G A T A G C T G T G C T A		
A A G G A T A * C T G	1M2I4M1D3M	Insertion & Deletion
G A T A A * G G A T A	5M1P1I4M	Padding & Insertion
T G T T A ⟨skipped⟩ T G C T A	5M15N5M	Spliced read
a a a C A T G T T A G	3S8M	Soft clipping
A A A C A T G T T A G	3H8M	Hard clipping

Read Mapping vs. Read Alignment

Read mapping

Finding the (approximate) location of origin of a read

- e.g., only the species
- or only the chromosome, or an approximate position or interval

Read Mapping vs. Read Alignment

Read mapping

Finding the (approximate) location of origin of a read

- e.g., only the species
- or only the chromosome, or an approximate position or interval

Read alignment

comparison at DNA basepair resolution between read and genome

Read Mapping vs. Read Alignment

Read mapping

Finding the (approximate) location of origin of a read

- e.g., only the species
- or only the chromosome, or an approximate position or interval

Read alignment

comparison at DNA basepair resolution between read and genome

Differences between read mapping and alignment

- two distinct tasks, but often done together (by the same software)
- mapping is simpler (faster, more resource-efficient) than alignment
- **mapping is sufficient** for some applications,

Examples of Sequence Analysis Problems

- Genome assembly (3.1 Gbp) from millions of sequenced fragments (100 bp)

Examples of Sequence Analysis Problems

- Genome assembly (3.1 Gbp) from millions of sequenced fragments (100 bp)

- Determining the species composition of a sequenced sample:
  Which species are present (and how much)? Environmental metagenomics
- Comparative genomics of two related species:
  Common genes, missing/additional genes, evolution, common ancestor
- Construction of phylogenetic trees between species

Examples of Sequence Analysis Problems

- Genome assembly (3.1 Gbp) from millions of sequenced fragments (100 bp)

- Determining the species composition of a sequenced sample:
  Which species are present (and how much)? Environmental metagenomics
- Comparative genomics of two related species:
  Common genes, missing/additional genes, evolution, common ancestor
- Construction of phylogenetic trees between species

- Discovery of single nucleotide variants and short indels
  between a sequenced individual genome and the reference genome
- Discovery of copy number variants

Examples of Sequence Analysis Problems

- Genome assembly (3.1 Gbp) from millions of sequenced fragments (100 bp)

- Determining the species composition of a sequenced sample:
  Which species are present (and how much)? Environmental metagenomics
- Comparative genomics of two related species:
  Common genes, missing/additional genes, evolution, common ancestor
- Construction of phylogenetic trees between species

- Discovery of single nucleotide variants and short indels
  between a sequenced individual genome and the reference genome
- Discovery of copy number variants

- Quantification of transcriptional activity (“gene expression”, by RNA sequencing)
- Discovery of differential gene expression between samples

Our Research Program at Algorithmic Bioinformatics, UdS

For each of the problems on the previous slide (and more), we ask:

- What is the (alignment-based) state-of-the art?
- Can we do it **alignment-free** (by mapping only)?
- What is the best methodological approach?
- What are the savings in computing time, CPU work, energy?
- Long-term benefit: Make DNA analysis more green.

Our Research Program at Algorithmic Bioinformatics, UdS

For each of the problems on the previous slide (and more), we ask:

- What is the (alignment-based) state-of-the art?
- Can we do it **alignment-free** (by mapping only)?
- What is the best methodological approach?
- What are the savings in computing time, CPU work, energy?
- Long-term benefit: Make DNA analysis more green.