Projects & Seminars
Mobile Genomics
Genome Sequencing on Mobile Devices

Prof. Onur Mutlu
Dr. Mohammed Alser
ETH Zürich
Fall 2020
29 September 2020
The Role of This Course
Projects & Seminars: Mobile Genomics

- We will cover the **basics** of **genome analysis** to understand the **speed-accuracy tradeoff** in using computationally-lightweight heuristics versus accurate computationally-expensive algorithms.

- Students will **experimentally** evaluate different heuristic **algorithms** and observe their effect on **the end results**.

- This evaluation will give the students the chance to carry out a **hands-on project** to implement one or more of these heuristic algorithms in **their smartphones** and help the society by enabling on-site analysis of genomic data.
Key Objectives

- Multiple components that are aimed at improving students’
  - Basic knowledge in genome analysis (dry lab)
  - Technical skills in genome analysis and computer architecture
  - Critical thinking and analysis
  - Familiarity with key research directions
  - Technical presentation of your project
Key Goal

(Learn how to) efficiently implement one of the key steps in genome analysis on portable devices
Prerequisites of the Course

- No prior knowledge in bioinformatics or genome analysis is required.
- A good knowledge in C programming language and programming is required.
- Interest in making things efficient and solving problems
Course Info: Who Are We?

- **Onur Mutlu**
  - Full Professor @ ETH Zurich ITET (INFK), since September 2015
  - Strecker Professor @ Carnegie Mellon University ECE/CS, 2009-2016, 2016-...
  - PhD from UT-Austin, worked at Google, VMware, Microsoft Research, Intel, AMD
  - [https://people.inf.ethz.ch/omutlu/](https://people.inf.ethz.ch/omutlu/)
  - [omutlu@gmail.com](mailto:omutlu@gmail.com) (Best way to reach me)
  - [https://people.inf.ethz.ch/omutlu/projects.htm](https://people.inf.ethz.ch/omutlu/projects.htm)

- **Research and Teaching in:**
  - Computer architecture, computer systems, hardware security, bioinformatics
  - Memory and storage systems
  - Hardware security, safety, predictability
  - Fault tolerance
  - Hardware/software cooperation
  - Architectures for bioinformatics, health, medicine
  - ...
Course Info: Who Are We?

- Lead Supervisor:
  - Dr. Mohammed Alser

- Supervisors:
  - Dr. Juan Gomez Luna
  - Jeremie Kim
  - Can Firtina

- Get to know them and their research
  - [https://safari.ethz.ch/safari-group/](https://safari.ethz.ch/safari-group/)
Course Requirements and Expectations

- Attendance required for all meetings

- Study the learning materials

- Each student will carry out a hands-on project
  - Build, implement, code, and design with close engagement from the supervisors

- Participation
  - Ask questions, contribute thoughts/ideas
  - Read relevant papers

We will help the projects with good progress to get published in good venues!
Course Website

- [https://safari.ethz.ch/projects_and_seminars/doku.php?id=genome_seq_mobile](https://safari.ethz.ch/projects_and_seminars/doku.php?id=genome_seq_mobile)

- Useful information for the course

- Check your email frequently for announcements

- We will also have Piazza for Q&A, announcements, ..
Next Meetings

- We will announce the projects and their descriptions next week.

- We will give you a chance to select a project,

- Then, we will have 1-1 meetings to match your interests, skills, and background with a suitable project.

- It is important that you study the learning materials before our next meeting!
WHAT IS GENOME ANALYSIS?
Our goal is to find the complete sequence of A, C, G, T’s in DNA (or RNA).

NO machine can read the entire content of a genome
1990-2003: The Human Genome Project (HGP) provides a complete and accurate sequence of all DNA base pairs that make up the human genome and finds 20,000 to 25,000 human genes.
Vast Improvement in Sequencing

Read Mapping

---

1x10^{12} bases*

44 hours*

<1000 $

* NovaSeq 6000
High-Throughput Sequencers

Illumina MiSeq

Pacific Biosciences Sequel II

Illumina NovaSeq 6000

Pacific Biosciences RS II

Oxford Nanopore MinION

Oxford Nanopore PromethION

Oxford Nanopore SmidgION

... and more! All produce data with different properties.
Reads lack information about their order and location (which part of genome they are originated from)
HTS Sequencing Output

Small pieces of a broken vase
short reads

Large pieces of a broken vase
long reads

Which sequencing technology is the best?

- 50-300 bp
- low error rate (~0.1%)

- 10K-100K bp
- high error rate (~15%)
Building up the Donor’s Genome
Map reads to a known reference genome with some minor differences allowed.
Metagenomics Analysis

Reads from different unknown donors at sequencing time are mapped to many known reference genomes.

Genetic material recovered directly from environmental samples are converted into "text format" and compared against a reference database.

Reference Database
Challenges in Read Mapping

- Need to find many mappings of each read
- Need to tolerate small variances/errors in each read
- Need to map each read very fast (i.e., performance is important, life critical in some cases)
Read Mapping: A Brute Force Algorithm

Reference

Read

Very Expensive!

\[ O(m^2kn) \]

\( m \): read length
\( k \): no. of reads
\( n \): reference genome length
Bottlenecked in Read Alignment!!

378 Million bases/minute
Read Sequencing**

150x slower

2 Million bases/minute
Read Mapping*

* BWA-MEM
** NovaSeq 6000, MinION
ACCELERATING GENOME ANALYSIS
What Makes Read Mapper Slow?

Key Observation # 1

70-90% of the read mapper’s execution time is spent in read alignment.

Alser et al, Bioinformatics (2017)
What Makes Read Mapper Slow? (cont’d)

Key Observation # 2

98% of candidate locations have high dissimilarity with a given read.

Xin et al, BMC genomics (2013)
What Makes Read Mapper Slow? (cont’d)

Key Observation # 3

- **Quadratic-time** dynamic-programming algorithm  
  **WHY?!**  
  Enumerating all possible prefixes

- NETHERLANDS x SWITZERLAND
- NETHERLANDS x S
- NETHERLANDS x SW
- NETHERLANDS x SWI
- NETHERLANDS x SWIT
- NETHERLANDS x SWITZ
- NETHERLANDS x SWITZE
- NETHERLANDS x SWITZER
- NETHERLANDS x SWITZERL
- NETHERLANDS x SWITZERLAN
- NETHERLANDS x SWITZERLAND
What Makes Read Mapper Slow? (cont’d)

Key Observation # 3

- **Quadratic-time** dynamic-programming algorithm
  - Enumerating all possible prefixes

- **Data dependencies** limit the computation parallelism
  - Processing row (or column) after another

- **Entire matrix** is computed even though strings can be dissimilar.
  - Number of differences is computed only at the backtracking step.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>E</th>
<th>T</th>
<th>H</th>
<th>E</th>
<th>R</th>
<th>L</th>
<th>A</th>
<th>N</th>
<th>D</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>S</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>W</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>I</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>T</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>Z</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>E</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>R</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>L</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>A</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>N</td>
<td>10</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>D</td>
<td>11</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>
Finding SNPs Associated with Complex Trait

<table>
<thead>
<tr>
<th>SNP1</th>
<th>SNP2</th>
<th>Blood Pressure</th>
</tr>
</thead>
<tbody>
<tr>
<td>...ACATG<strong>CCGACATTTTCA</strong>TAGGCC...</td>
<td><strong>180</strong></td>
<td></td>
</tr>
<tr>
<td>...ACATG<strong>CCGACATTTTTCATA</strong>AGGCC...</td>
<td><strong>175</strong></td>
<td></td>
</tr>
<tr>
<td>...ACATG<strong>CCGACATTTTTCATA</strong>AGGCC...</td>
<td><strong>170</strong></td>
<td></td>
</tr>
<tr>
<td>...ACATG<strong>CCGACATTTTTCATA</strong>AGGCC...</td>
<td><strong>165</strong></td>
<td></td>
</tr>
<tr>
<td>...ACATG<strong>CCGACATTTTTCATA</strong>AGGCC...</td>
<td><strong>160</strong></td>
<td></td>
</tr>
<tr>
<td>...ACATG<strong>CCGACATTTTTCATA</strong>AGGCC...</td>
<td><strong>145</strong></td>
<td></td>
</tr>
<tr>
<td>...ACATG<strong>CCGACATTTTTCATA</strong>AGGCC...</td>
<td><strong>140</strong></td>
<td></td>
</tr>
<tr>
<td>...ACATG<strong>CCGACATTTTTCATA</strong>AGGCC...</td>
<td><strong>130</strong></td>
<td></td>
</tr>
<tr>
<td>...ACATG<strong>TCGACATTTTTCATA</strong>AGGCC...</td>
<td><strong>120</strong></td>
<td></td>
</tr>
<tr>
<td>...ACATG<strong>TCGACATTTTTCATA</strong>AGGCC...</td>
<td><strong>120</strong></td>
<td></td>
</tr>
<tr>
<td>...ACATG<strong>TCGACATTTTTCATA</strong>AGGCC...</td>
<td><strong>115</strong></td>
<td></td>
</tr>
<tr>
<td>...ACATG<strong>TCGACATTTTTCATA</strong>AGGCC...</td>
<td><strong>110</strong></td>
<td></td>
</tr>
<tr>
<td>...ACATG<strong>TCGACATTTTTCATA</strong>AGGCC...</td>
<td><strong>110</strong></td>
<td></td>
</tr>
<tr>
<td>...ACATG<strong>TCGACATTTTTCATA</strong>AGGCC...</td>
<td><strong>110</strong></td>
<td></td>
</tr>
<tr>
<td>...ACATG<strong>TCGACATTTTTCATA</strong>AGGCC...</td>
<td><strong>105</strong></td>
<td></td>
</tr>
<tr>
<td>...ACATG<strong>TCGACATTTTTCATA</strong>AGGCC...</td>
<td><strong>100</strong></td>
<td></td>
</tr>
</tbody>
</table>

Different individuals

Eleazar Eskin: Discovering the Causal Variants Involved in GWAS Studies, CGSI 2018, UCLA
Mirror Phenotypes of 593 Kb CNVs

AUTISM
Deletion of 593 kb

OBESITY
Walters, *Nature* 2010
Deletion of 593 kb

SCHIZOPHRENIA
McCarthy, *Nat Genet* 2009
Duplication of 593 kb

UNDERWEIGHT
Duplication of 593 kb

Deletion in the short arm of chromosome 16 (16p11.2)

Duplication in the short arm of chromosome 16 (16p11.2)
City-Scale Microbiome Profiling

City-Scale Microbiome Profiling (cont’d)

Afshinnekoo+, "Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics", Cell Systems, 2015
Plague in New York Subway System?

Plague (Yersinia Pestis)

What Is It?

Plague is caused by Yersinia pestis bacteria. It can be a life-threatening infection if not treated promptly. Plague has caused several major epidemics in Europe and Asia over the last 2,000 years. Plague has most famously been called "the Black Death" because it can cause skin sores that form black scabs. A plague epidemic in the 14th century killed more than one-third of the population of Europe within a few years. In some cities, up to 75% of the population died within days, with fever and swollen skin sores.
Plague in New York Subway System?

Plague (Yersinia Pestis)

What Is It?

Published: December, 2018

Plague is caused by Yersinia Pestis. Plague has lasted 2,000 years. Plague causes skin sores that form before the infection spreads throughout the body. More than one-third of the population died within 2 weeks.

The findings of Yersinia Pestis in the subway received wide coverage in the lay press, causing some alarm among New York residents.

The New York Times

Bubonic Plague in the Subway System? Don’t Worry About It

In October, riders were not deterred after reports that an Ebola-infected man had ridden the subway just before he fell ill. Robert Stolarik for The New York Times

Living in a microbial world

Charles Schmidt


https://www.nature.com/articles/nbt.3868
There is a critical need for fast and accurate genome analysis.
Open Questions

How and where to enable fast, accurate, cheap, privacy-preserving, and exabyte scale analysis of genomic data?
Pushing Towards New Architectures

Microprocessor  Main Memory  Storage (SSD/HDD)

Single memory request consumes
>160x-800x more energy compared to
performing a complex add operation

Sequencing Machine
Processing Genomic Data Where it Makes Sense

Modern systems

FPGAs

Hybrid Main Memory

Heterogeneous Processors and Accelerators

(General Purpose) GPUs

Sequencing Machine

Persistent Memory/Storage
Key Takeaways

Most speedup comes from parallelism enabled by novel architectures and algorithms
Projects & Seminars
Mobile Genomics
Genome Sequencing on Mobile Devices

Prof. Onur Mutlu
Dr. Mohammed Alser
ETH Zürich
Fall 2020
29 September 2020