TECHNION Israel Institute of Technology

ANDREW AND ERNA VITERBI FACULTY OF ELECTRICAL ENGINEERING

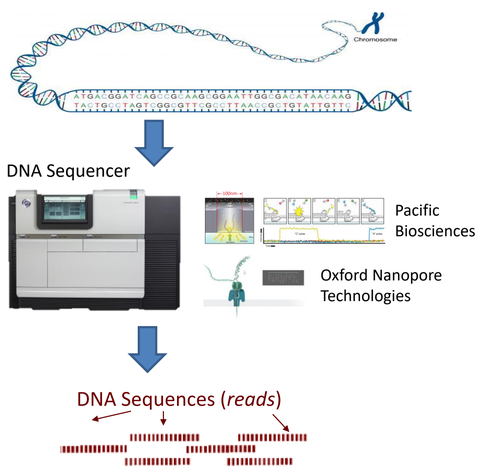
Accelerating DNA Long Read Mapping with Emerging Technologies

Roman Kaplan

Background and Problem: DNA Long Read Mapping

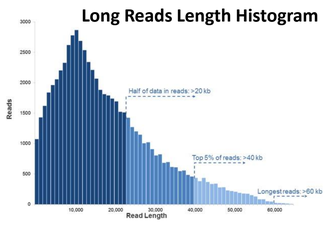
3rd Generation Sequencing: Real-Time Single-Molecule

- DNA is composed of 4 nucleotides: 'A', 'G', 'C', 'T' (base pairs, bps)
- Reading the content of an entire DNA strand at once (e.g., Chromosome) isn't possible
- DNA sequencers output strands of the DNA (*reads*)
- 3rd generation sequencers vs. 2nd generation:
  - Produce outputs faster (hours vs. days)
  - Reads are longer (10k+ bps vs. 100-300 bps)
    ➔ called *long reads*
  - Simpler preparation (less lab work)
- The downsides: many errors, up to 15%
  - Errors can be insertions, deletions and substitutions

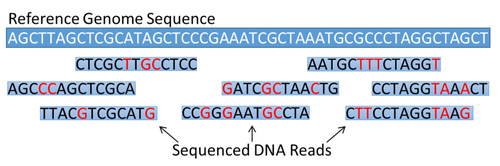


DNA Long Read Mapping

- The problem: "stitching" together all the DNA reads
- When an organism was previously sequenced (e.g., human), this sequence is used as a reference to construct the new organism genome
- Requires mapping 1M+ reads against a reference sequence (e.g., human is 3Gbps)
- Existing read mapping tools use technology-specific heuristics, complex data structures and large memory requirements
- Hardware solutions only addressed short reads with few errors (from 2nd generation)
- Long reads contain many errors and pose a challenge for an effective hardware acceleration solution



DNA Read Mapping

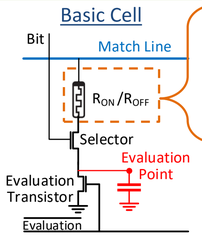


The Solution

RASSA: Resistive Approximate Similarity Search Accelerator

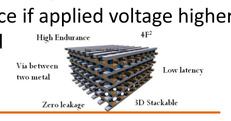
Processing-in-Memory Using Memristors

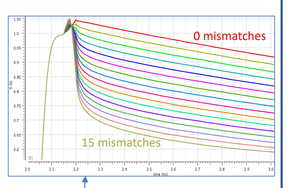
- Basic cell: 1 **memristor**, 2 transistors (2T1R)
- The memristor serves as a programmable non-volatile switch
  - Low resistance (R_{on}) – open switch
  - High resistance (R_{off}) – closed switch
- One-hot encoding to store DNA base pairs
  - 4 cells encode one base pair



Memristors (Memory Resistor)

- Change resistance if applied voltage higher than a threshold
- High density
- Non-volatile

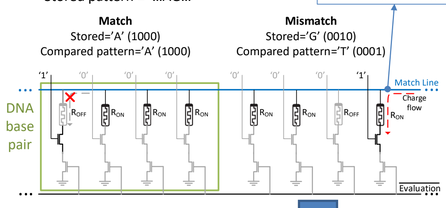




Comparing a pattern

- Compared pattern is applied on all bitcells
- R_{on} allows charge to flow
  - **Match**: no charge flow → N.C. in Match line
  - **Mismatch**: charge flows through R_{on} memristor → Match line voltage drops
- Charge flow through one bitcell allows to quantify the drop in match line
- Match line voltage level translates to number of matches

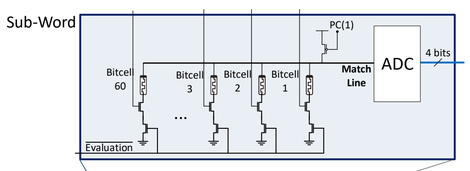
Example
Compared pattern = '…AT…'
Stored pattern = '…AG…'



The RASSA Bottom-Up Architecture

Sub-Word

- Contains 60 bitcells ➔ encodes 15 DNA base pairs (0 through 15 matches = 4 bit)
- Analog-to-Digital converter translates match line voltage level to number of matches

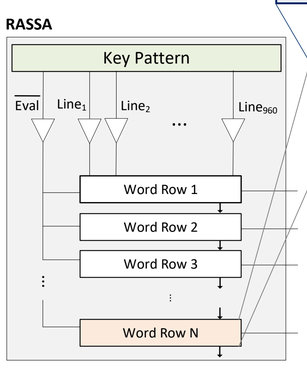


RASSA

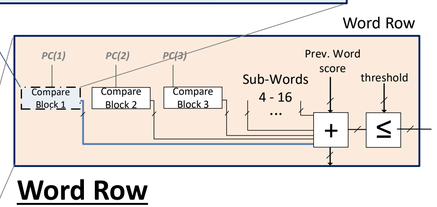
- Holds multiple Word Rows
- A compare pattern is applied on all Word Rows
- The comparison takes place in all Word Rows ***in-situ*** and **in parallel**

(Massively parallel processing-in-memory)

- Scores below a threshold indicate a potential mapping location



Word Row



Word Row

- Contains 16 Sub-Words (240 bps)
- Connected to top and bottom Word Rows
- All mismatch values are summed *(mismatch score)*
- Two options for a mismatch score:
  1) Add mismatch score from top Word Row
  2) Compare mismatch score to a threshold

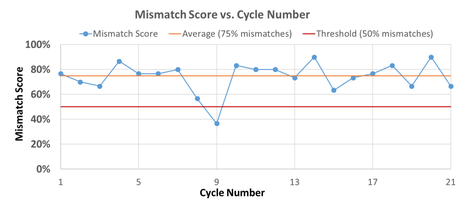
DNA Read Mapping with RASSA

The Main Idea: High-Similarity Regions → Mapping Locations

- A fixed-size chunk is compared against a long sequence
- Counting mismatch score in every position approximates the correlation between the chunk and sequence
- In the example, the read chunk is mismatch score is calculated for every position against the reference sequence
- The mapping location for the chunk is the position of a mismatch score below the threshold

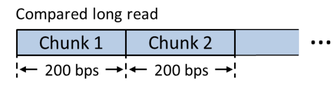
Simplified Example

RASSA

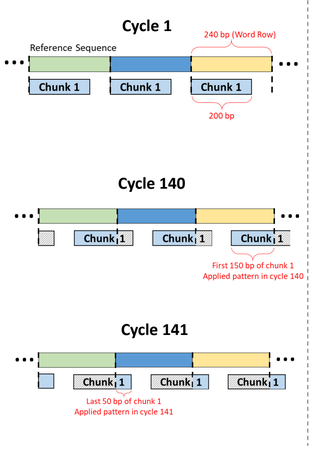


Find Long Read Mapping Locations with RASSA

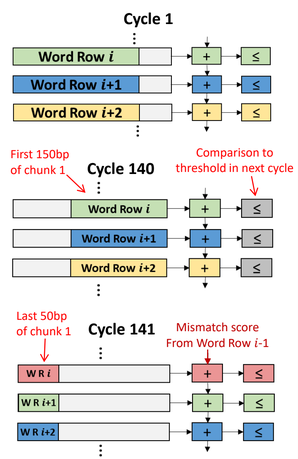
- Reads are divided to fixed-size chunks
  For example: 100bps / 200bps
- The threshold is set at 40-50% of the chunk length (determined empirically)
- Every chunk is compared against the entire reference sequence (stored in RASSA)
- Chunk with mismatch score below a threshold signals a mapping location
- When chunk is split between two Word Rows, mismatch score for every part is found in separate cycles
  - Two consecutive Word Rows are needed
  - Mismatch score from top Word Row is transmitted to the bottom Word Row
  - Bottom Word Row sums and compares to threshold



Long Read Chunk Compared with a Reference Sequence



How Comparison is Performed in RASSA



Evaluations

Chip Parameters, Performance and Accuracy

Chip Parameters

- A Sub-Word circuit was designed, placed and routed using the Global Foundries 28nm CMOS High-k Metal Gate library for:
  - Transistor sizing
  - Timing
  - Power analysis
- Spectre simulations for the FF and SS corners at 70^0C and nominal voltage

Parameter	Value
DNA bps per row (bits)	240 (960)
Words per chip	$131k$ (2^{17})
Memory size (DNA bps)	$31.5M$
Frequency	$1GHz$
Single chip power	235W
Single chip area	$209mm^2$

Performance and Accuracy Comparison with minimap2

- Two organism reference sequences were used: e.coli and yeast
- Input sequences from Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT)
  - For PacBio: regular and high-quality Circular Consensus Sequencing (CCS) reads
- Accuracy and performance compared to minimap2, a state-of-art long read mapper
  - **Executing platform:** 16-core Intel Xeon E5-2650 @2GHz w/ 64GB of RAM
  - Minimap2 is was executed active SIMD extensions and multi-threading enabled
  - Sensitivity: % of reads found by RASSA from those found by minimap2

Reference Sequences
E.coli K-12 MG1655: 4.6Mbp
S.cerevisiae (Yeast) W303: 11.7Mbp

Datasets: #reads, avg len
e.coli PacBio: 100k, 5.2kbp
e.Coli CCS: 260k, 940bp
e.Coli ONT: 165k, 9kbp
Yeast PacBio: 100k, 6.3kbp
Yeast ONT: 30k, 11.3kbp

		200bp chunk			100bp chunk		
		Sensitivity	False Positives	Speedup	Sensitivity	False Positives	Speedup
e.coli	PacBio	79.3%	13.4%	25×	83.2%	13.6%	16×
	PacBio CCS	96.3%	8.9%	43×	96.2%	6.9%	24×
	ONT	88.8%	10.5%	48×	87.6%	12.4%	31×
Yeast	PacBio	69.8%	8.7%	77×	72%	11.8%	51×
	ONT*	85.9%	34.9%	31×	85.1%	39.2%	49×

* minimap2 mapped only 20% of all reads, with 50% of mappings with lower quality score than 60 (indicates a high-confidence mapping). RASSA sensitivity = % of reads mapped from the entire dataset. False positives = % of mapped reads with two or more mapping locations.

Performance Comparison with FPGA

- Gatekeeper [1], a pre-alignment FPGA accelerator
  - Counts number of mismatches between short reads and a reference sequence
  - Implemented in a Virtex-7 FPGA using Xilinx VC709 board, running @250MHz
  - Host machine uses 3.6GHz Intel i7-3820 CPU w/ 8GB of RAM
- Comparison of RASSA vs. GateKeeper throughput
  - Throughput measured in Billion Evaluated Mappings Locations per second (BEML/s)
  - GateKeeper results were taken from [1], RASSA results are normalized to 250MHz

RASSA vs. GateKeeper Throughput Comparison

Read Lengths	GateKeeper	RASSA @250MHz
100bp	1.7 BEML/s	226.8 BEML/s
200bp	-	175.2 BEML/s
300bp	0.2 BEML/s	142.8 BEML/s

[1] Alser, M., Hassan, H., Xin, H., Ergin, O., Mutlu, O. and Alkan, C. "GateKeeper: a new hardware architecture for accelerating pre-alignment in DNA short read mapping." Bioinformatics, vol. 33, no. 21, pp. 3355-3363, 2017.