UC San Diego
Electrical and Computer Engineering
JACOBS SCHOOL OF ENGINEERING

TURAKHIA LAB

TALCO: Tiling Genome Sequence Alignment using Convergence of Traceback Pointers

Sumit Walia, Cheng Ye, Arkid Bera, Dhruvi Lodhavia and Yatish Turakhia
University of California San Diego

Outline

- Emergence of **Long Genome Sequence Alignment** (LGSA)
- Current LGSA **algorithms**, **accelerators** and their **limitations**
- **TALCO**: A tiling technique based on convergence of traceback pointers for long genome sequence alignment
- **Key Contributions** and **Results**
- **Conclusion**

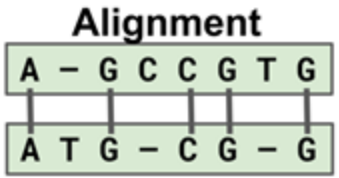
Outline

- Emergence of **Long Genome Sequence Alignment** (LGSA)
- Current LGSA **algorithms**, **accelerators** and their **limitations**
- **TALCO**: A tiling technique based on convergence of traceback pointers for long genome sequence alignment
- **Key Contributions** and **Results**
- **Conclusion**

Genome Sequence Alignment

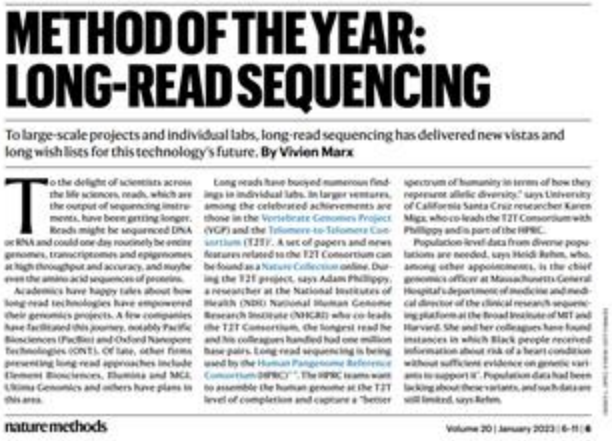
- **Sequence alignment:** Given two sequences, Reference (R) and Query (Q), assign gaps ("**–**") in R and Q to produce a valid alignment that maximizes the alignment score

R: AGCCGTG
Q: ATGCGG



Long Genome Sequence Alignment: Applications

- **Third-generation sequencing technologies** (produce reads of length 10 kb - 4 Mb), leads to major breakthroughs in recent past:



METHOD OF THE YEAR: LONG-READ SEQUENCING

To large-scale projects and individual labs, long-read sequencing has delivered new vistas and long wish lists for this technology's future. By Vivien Marx

To the delight of scientists across the life sciences, reads, which are the output of sequencing instruments, have been getting longer. Reads might be sequenced DNA or RNA and could one day routinely be entire genomes, transcriptomes and epigenomes at high throughput and accuracy, and maybe even the amino acid sequences of proteins.

Academics have happy tales about how long-read technologies have empowered their genomics projects. A few companies have facilitated this journey, notably Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). Of late, other firms presenting long-read approaches include Element Biosciences, Illumina and MGI. Ultima Genomics and others have plans in this area.

Long reads have buoyed numerous findings in individual labs. In larger ventures, among the celebrated achievements are those in the Vertebrate Genomes Project (VGP) and the Telomere-to-Telomere Consortium (T2T)[1]. A set of papers and news features related to the T2T Consortium can be found as a Nature Collection online. During the T2T project, says Adam Phillippy, a researcher at the National Institutes of Health (NIH) National Human Genome Research Institute (NHGRI) who co-leads the T2T Consortium, the longest read he and his colleagues handled had one million base pairs. Long read sequencing is being used by the Human Pangenome Reference Consortium (HPRC)[2]. The HPRC teams want to assemble the human genome at the T2T level of completion and capture a "better spectrum of humanity in terms of how they represent allelic diversity," says University of California Santa Cruz researcher Karen Miga, who co-leads the T2T Consortium with Phillippy and is part of the HPRC.

Population-level data from diverse populations are needed, says Heidi Rehm, who, among other appointments, is the chief genomics officer at Massachusetts General Hospital's department of medicine and medical director of the clinical research sequencing platform at the Broad Institute of MIT and Harvard. She and her colleagues have found instances in which Black people received information about risk of a heart condition without sufficient evidence on genetic variants to support it[3]. Population data had been lacking about these variants, and such data are still limited, says Rehm.

nature methods

Volume 20 | January 2023 | 6–11 | 6

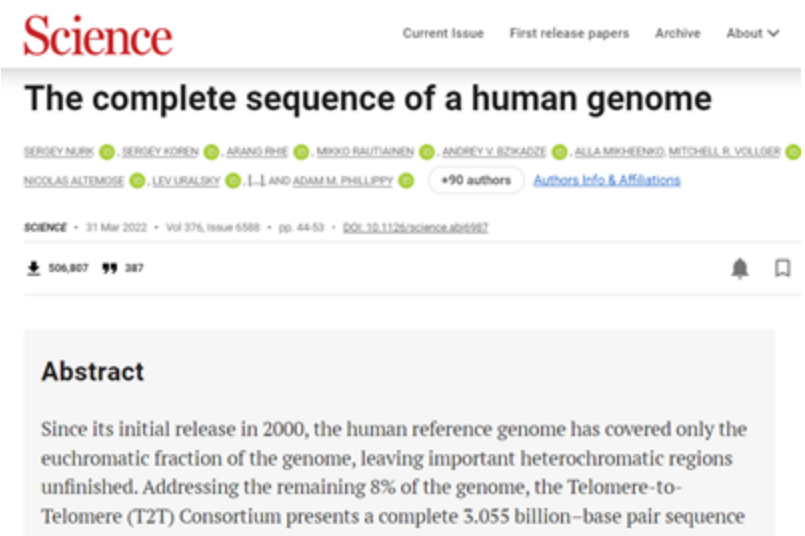
Oxford Nanopore Technologies (ONT)

Pacific Biosciences (PacBio)

Marx, Vivien. "Method of the year: long-read sequencing." Nature Methods 20.1 (2023): 6-11.

Long Genome Sequence Alignment: Applications

- **Third-generation sequencing technologies** (produce reads of length 10 kb - 4 Mb), leads to major breakthroughs in recent past:
  - **Human Genome Assembly**



Science

Current Issue | First release papers | Archive | About

The complete sequence of a human genome

SERGEY NURK, SERGEY KOREN, ARANG RHIE, MIKKO RAUTIAINEN, ANDREY V. BZIKADZE, ALLA MIKHEENKO, MITCHELL R. VOLLGER, NICOLAS ALTEMOSE, LEV URALSKY, [...] AND ADAM M. PHILLIPPY +90 authors Authors Info & Affiliations

SCIENCE • 31 Mar 2022 • Vol 376, Issue 6588 • pp. 44-53 • DOI: 10.1126/science.abj6987

506,807 387

Abstract

Since its initial release in 2000, the human reference genome has covered only the euchromatic fraction of the genome, leaving important heterochromatic regions unfinished. Addressing the remaining 8% of the genome, the Telomere-to-Telomere (T2T) Consortium presents a complete 3.055 billion–base pair sequence

Nurk, Sergey, et al. "The complete sequence of a human genome." Science 376.6588 (2022): 44-53.

Long Genome Sequence Alignment: Applications

- **Third-generation sequencing technologies** (produce reads of length 10 kb - 4 Mb), leads to major breakthroughs in recent past:
  - **Human Genome Assembly**
  - **Rapid genetic diagnosis**

Fastest DNA sequencing technique helps undiagnosed patients find answers in mere hours

A research effort led by Stanford scientists set the first Guinness World Record for the fastest DNA sequencing technique, which was used to sequence a human genome in just 5 hours and 2 minutes.

January 12, 2022 - By Hanae Armitage

Researchers were able to quickly determine that Matthew Kunzman's **heart failure** was the result of a **genetic condition** — a finding that cleared the way for him to be placed on a heart transplant list immediately. *Courtesy of Jenny Kunzman*

Armitage, Hamae. "Fastest DNA sequencing technique helps undiagnosed patients find answers in mere hours." News Center 12 (2022).

Long Genome Sequence Alignment: Applications

- Third-generation sequencing technologies (produce reads of length 10 kb - 4 Mb), leads to major breakthroughs in recent past:
  - Human Genome Assembly
  - Rapid genetic diagnosis
  - **Characterize structural variations and complex regions**

Article

Patterns of somatic structural variation in human cancer genomes

https://doi.org/10.1038/s41586-019-1913-9
Received: 22 September 2017
Accepted: 18 November 2019
Published online: 5 February 2020

Yilong Li, Nicola D. Roberts, Jeremiah A. Wala, Ofer Shapira, Steven E. Schumacher, Kiran Kumar, Ekta Khurana, Sebastian Waszak, Jan O. Korbel, James E. Haber, Marcin Imielinski, PCAWG Structural Variation Working Group, Joachim Weischenfeldt, Rameen Beroukhim, Peter J. Campbell & PCAWG Consortium

ARTICLE

Structural Variation of Chromosomes in Autism Spectrum Disorder

Christian R. Marshall,[1] Abdul Noor,[2] John B. Vincent,[2] Anath C. Lionel,[1] Lars Feuk,[1] Jennifer Skaug,[1] Mary Shago,[3] Rainald Moessner,[1] Dalila Pinto,[1] Yan Ren,[1] Bhooma Thiruvahindrapduram,[1] Andreas Fiebig,[6] Stefan Schreiber,[6] Jan Friedman,[7] Cees E.J. Ketelaars,[8] Yvonne J. Vos,[8] Can Ficicioglu,[9] Susan Kirkpatrick,[10] Rob Nicolson,[11] Leon Sloman,[2] Anne Summers,[12] Clare A. Gibbons,[12] Ahmad Teebi,[4] David Chitayat,[4] Rosanna Weksberg,[4] Ann Thompson,[13] Cathy Vardy,[14] Vicki Crosbie,[14] Sandra Luscombe,[14] Rebecca Baatjes,[1] Lonnie Zwaigenbaum,[15] Wendy Roberts,[5,16] Bridget Fernandez,[14] Peter Szatmari,[13] and Stephen W. Scherer[1,*]

Circulation: Genomic and Precision Medicine
Volume 14, Issue 4, August 2021
https://doi.org/10.1161/CIRCGEN.120.003223

American Heart Association.

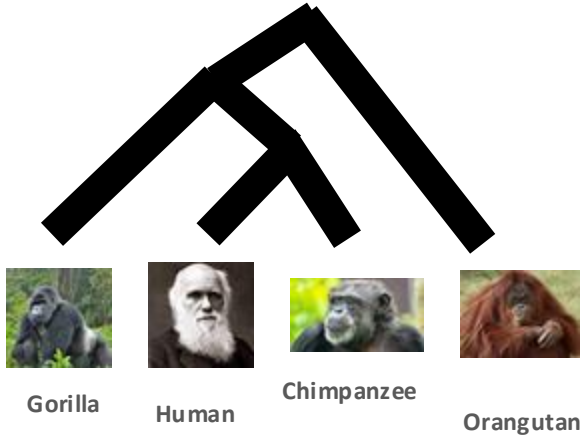
RESEARCH LETTERS

Long-Read Sequence Confirmed a Large Deletion Including *MYH6* and *MYH7* in an Infant of Atrial Septal Defect and Atrial Arrhythmias

Long Genome Sequence Alignment: Applications

- Third-generation sequencing technologies (produce reads of length 10 kb - 4 Mb), leads to major breakthroughs in recent past:
  - Human Genome Assembly
  - Rapid genetic diagnosis
  - Characterize structural variations and complex regions

- **New insights into the evolution of different species through whole genome analysis**



Long Genome Sequence Alignment: Applications

- Third-generation sequencing technologies (produce reads of length 10 kb - 4 Mb), leads to major breakthroughs in recent past:
  - Human Genome Assembly
  - Rapid genetic diagnosis
  - Characterize structural variations and complex regions

- New insights into the evolution of different species through whole genome analysis

Bottlenecked by LGSA

Outline

- Emergence of **Long Genome Sequence Alignment**
- Current LGSA **algorithms**, **accelerators** and their **limitations**
- **TALCO**: A tiling technique based on convergence of traceback pointers for long genome sequence alignment
- **Key Contributions** and **Results**
- **Conclusion**

Broad Classification of Alignment Algorithms

Classical Dynamic Programming (DP) Algorithms	Non-Classical Algorithms
Ex: *Needleman-Wunsch, Smith-Waterman*	Ex: *WFA, O(ND)*

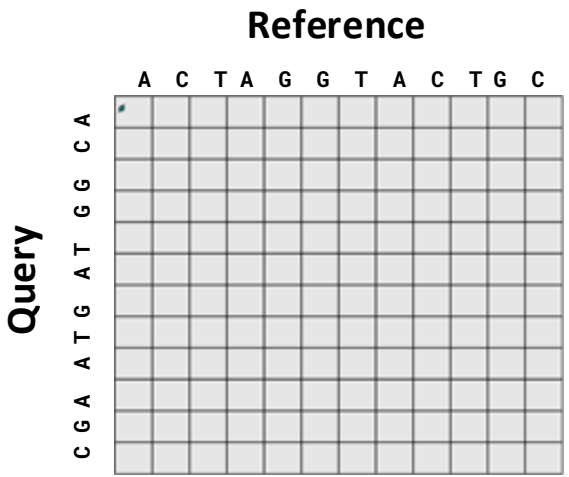
Classical DP based Alignment Algorithms

Classical Dynamic Programming (DP) Algorithms

Ex: *Needleman-Wunsch, Smith-Waterman*

Non-Classical Algorithms

Ex: *WFA, O(ND)*



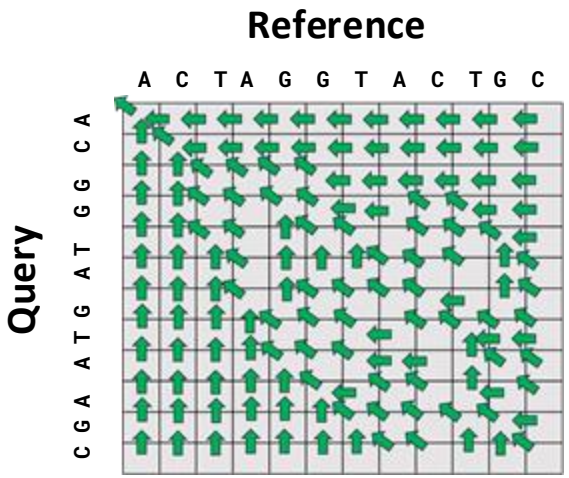
Classical DP based Alignment Algorithms

Classical Dynamic Programming (DP) Algorithms

Ex: *Needleman-Wunsch, Smith-Waterman*

Non-Classical Algorithms

Ex: *WFA, O(ND)*



1. **Matrix Fill (Store traceback pointers)**

Classical DP based Alignment Algorithms

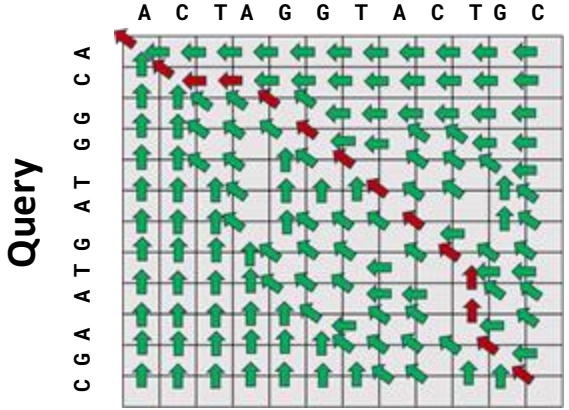
Classical Dynamic Programming (DP) Algorithms

Ex: *Needleman-Wunsch, Smith-Waterman*

Non-Classical Algorithms

Ex: *WFA, O(ND)*

Reference




```
A C T A G G T A C T - - G C
A C - - G G T A G T A A G C
```

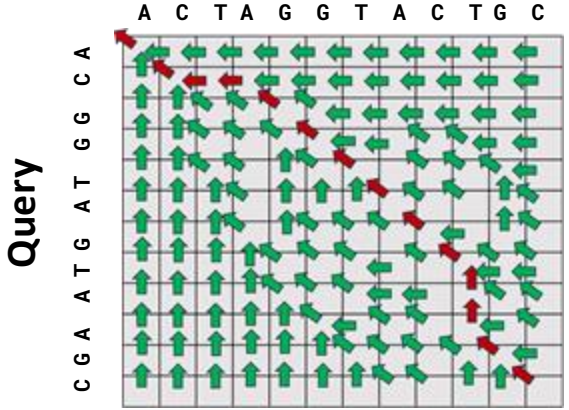
1. Matrix Fill (Store traceback pointers)
2. **Optimal traceback path**

Non-Classical Alignment Algorithms

Classical Dynamic Programming (DP) Algorithms

Ex: *Needleman-Wunsch, Smith-Waterman*

Reference

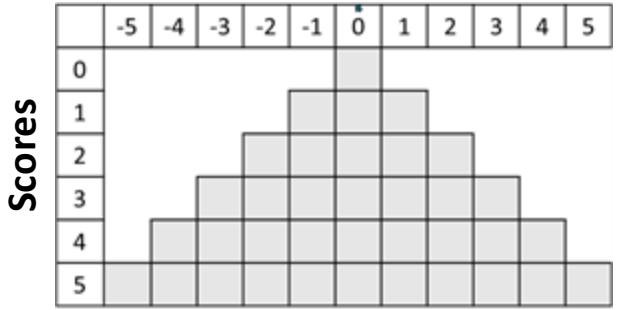


1. Matrix Fill (Store traceback pointers)
2. Optimal traceback path

Non-Classical Algorithms

Ex: *WFA, O(ND)*

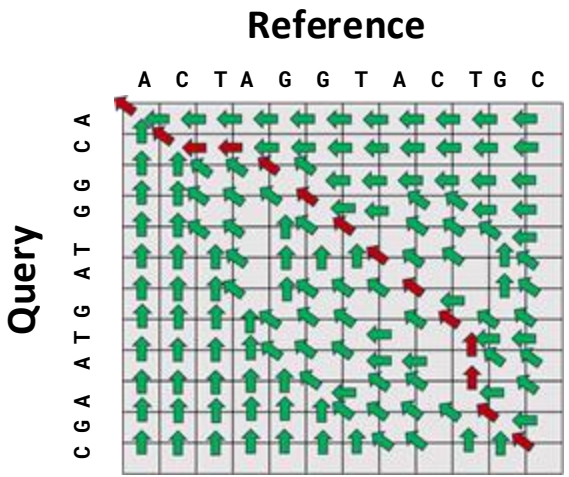
Diagonals



Non-Classical Alignment Algorithms

Classical Dynamic Programming (DP) Algorithms

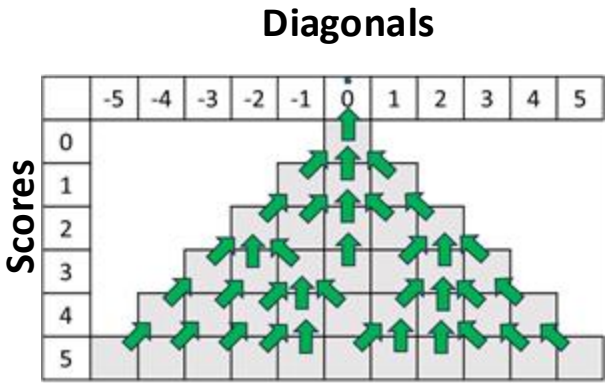
Ex: *Needleman-Wunsch, Smith-Waterman*



1. Matrix Fill (Store traceback pointers)
2. Optimal traceback path

Non-Classical Algorithms

Ex: *WFA, O(ND)*

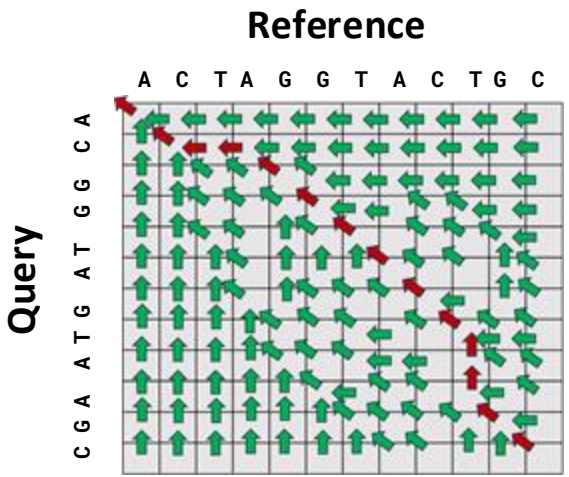


1. **Matrix Fill (Store traceback pointers)**

Non-Classical Alignment Algorithms

Classical Dynamic Programming (DP) Algorithms

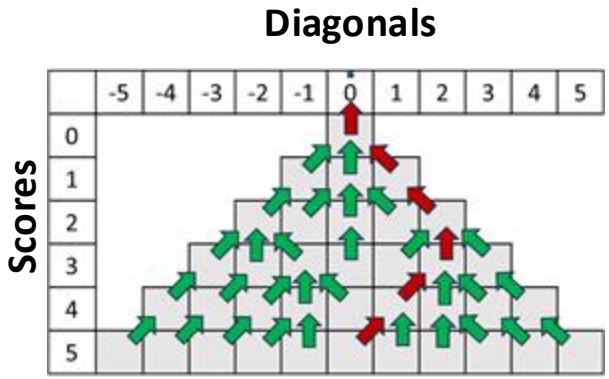
Ex: *Needleman-Wunsch, Smith-Waterman*



1. Matrix Fill (Store traceback pointers)
2. Optimal traceback path

Non-Classical Algorithms

Ex: *WFA, O(ND)*



1. Matrix Fill (Store traceback pointers)
2. **Optimal traceback path**

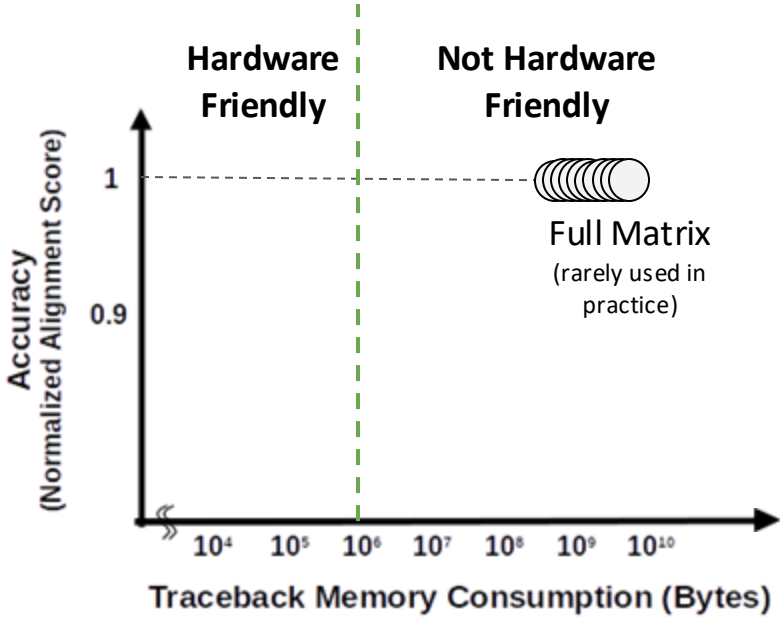
Comparison: Classical-DP and Non-Classical

Classical Dynamic Programming (DP) Algorithms	Non-Classical Algorithms
Ex: *Needleman-Wunsch, Smith-Waterman*	Ex: *WFA, O(ND)*

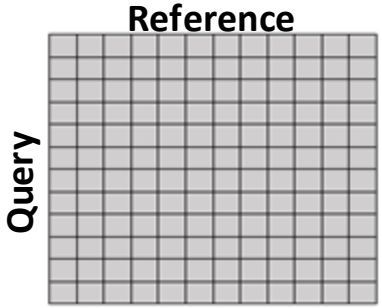
Both categories of algorithms produce optimal alignments

Uniform dependencies **Easier to accelerate**	Non-Uniform dependencies **Harder to accelerate**
More popular	Very Fast for similar sequences

Full Matrix Sequence Alignment Algorithms



Classical-DP



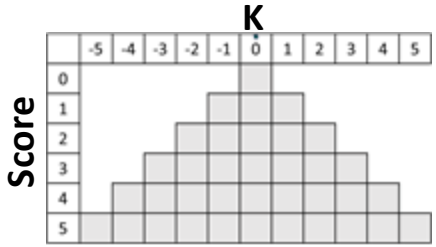
Computed cells

Uncomputed cells

Space Complexity O(NM)

Ex: *Needleman-Wunsch*

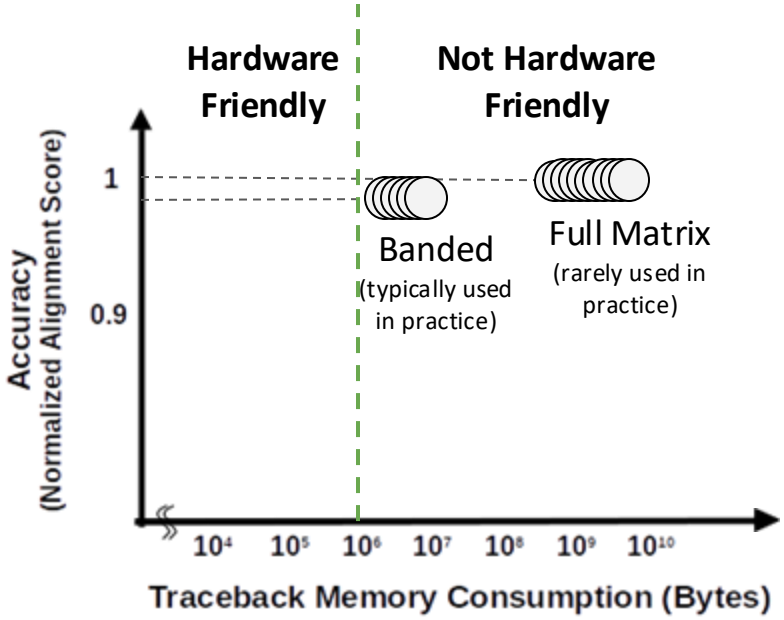
Non Classical



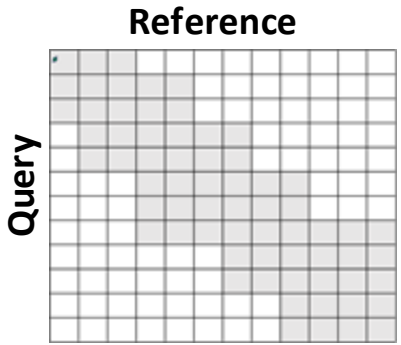
Space Complexity $O(S^2)$

Ex: *WFA*

Banded Sequence Alignment Algorithms



Classical-DP



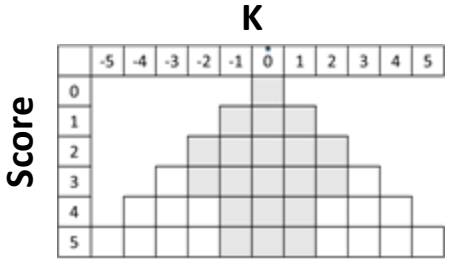
Computed cells

Uncomputed cells

Space Complexity O(ND)

Ex: X-Drop

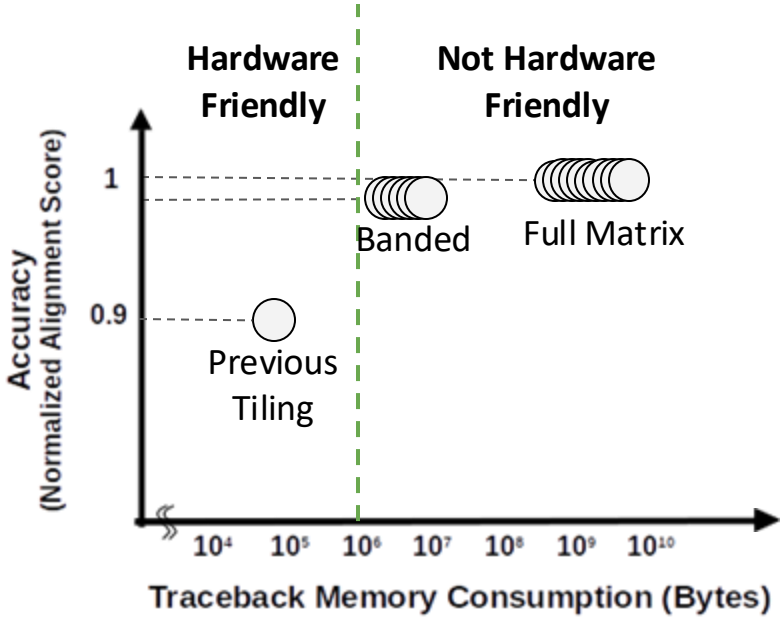
Non Classical

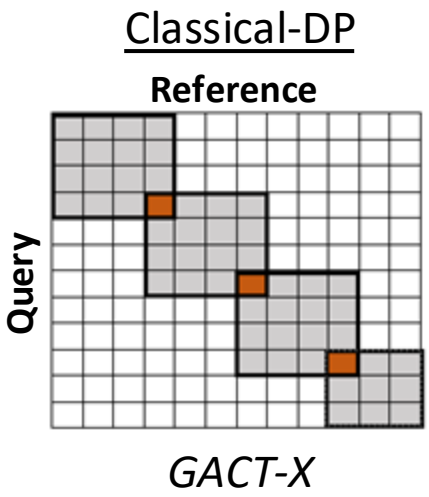


Space Complexity O(SD)

Ex: WFA-Adapt

Tiling heuristic





T: Tile size

Space Complexity $O(T^2)$

Independent of sequence length!

Non Classical

Tiling is **never applied** to **non-classical sequence alignment algorithm**

Architecture Papers Adopting Tiling Heuristics

- **GACT** – Darwin: A Genomics Co-processor Provides up to 15,000X Acceleration on Long Read Assembly (**ASPLOS 2018 Best Paper Award**)
- **GACT-X** – Darwin-WGA: A Co-processor Provides Increased Sensitivity in Whole Genome Alignments with High Speedup

Lower accuracy imposes challenges for tiling-based accelerators to be adopted in critical real-world applications (e.g. medical diagnoses)

Van der Auwera, Geraldine A., et al. "From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline." Current protocols in bioinformatics 43.1 (2013

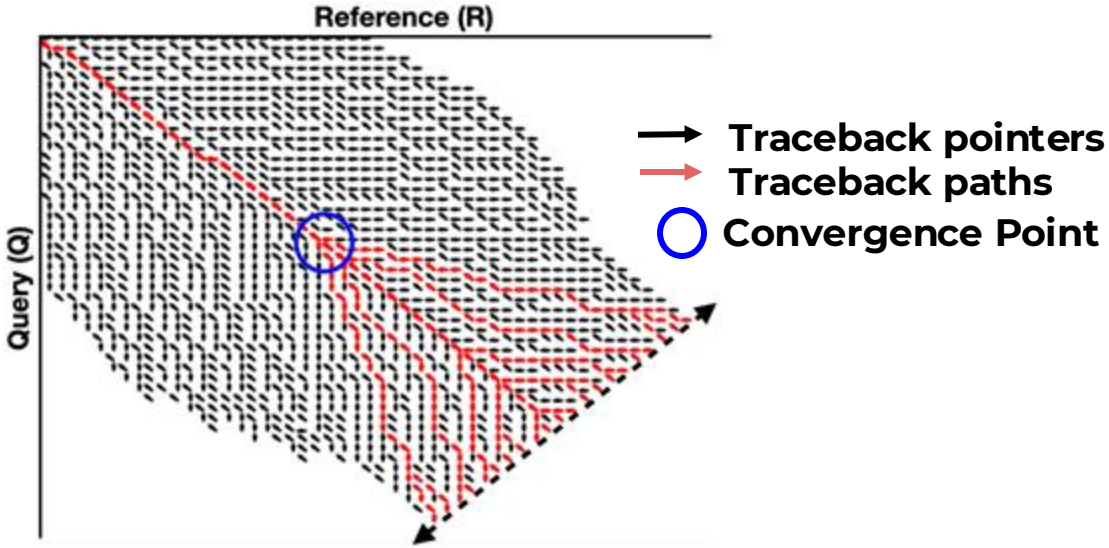
- **RAPIDx**: High-performance ReRAM processing in-memory accelerator for sequence alignment (**TCAD 2023**)
- **GMX: I**nstruction Set Extensions for Fast, Scalable, and Efficient Genome Sequence Alignment **(MICRO 2023)**
- **Scrooge:** a fast and memory-frugal genomic sequence aligner for CPUs, GPUs, and ASICs **(Bioinformatics 2023)**

Outline

- Emergence of **Long Genome Sequence Alignment**
- Current LGSA algorithms, accelerators and their limitations
- **TALCO**: A tiling technique based on convergence of traceback pointers for long genome sequence alignment
- **Key Contributions** and **Results**
- **Conclusion**

Key Insight: Convergence of Traceback Paths

TALCO (Tiling Long Genome Alignment using Convergence of Traceback Pointers) is based on the following observation:



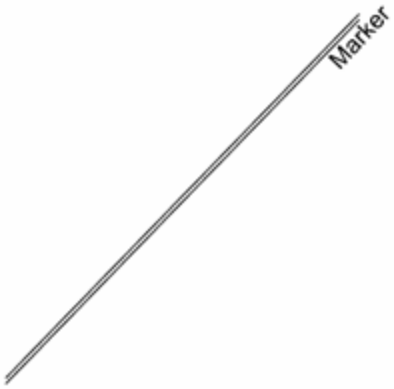
Experiment: Pairwise sequence alignment using Needleman-Wunsch with X-Drop banding

TALCO: Tiling technique for long genome alignment

TALCO algorithm has **two** phases:

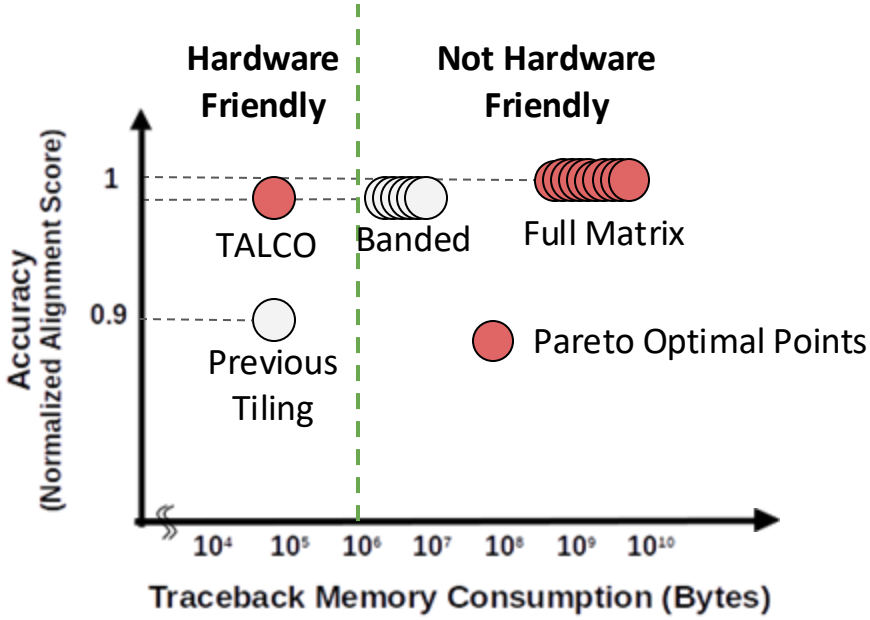
1. Stores traceback pointers till the **Marker**

1. Find point of convergence of traceback pointers using **pointer-redirection**



TALCO applied to X-Drop Algorithm

TALCO is on the Pareto Optimal Frontier



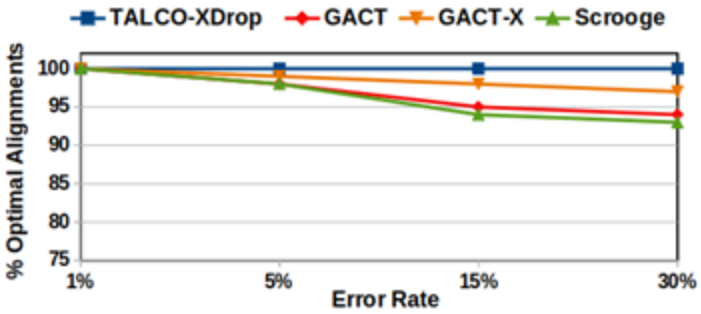
- Constant Space complexity
- Guarantees optimality under banding constraints

Outline

- Emergence of **Long Genome Sequence Alignment**
- Current LGSA algorithms, accelerators and their limitations
- **TALCO**: A tiling technique based on convergence of traceback pointers for long genome sequence alignment
- **Key Contributions** and **Results**
- **Conclusion**

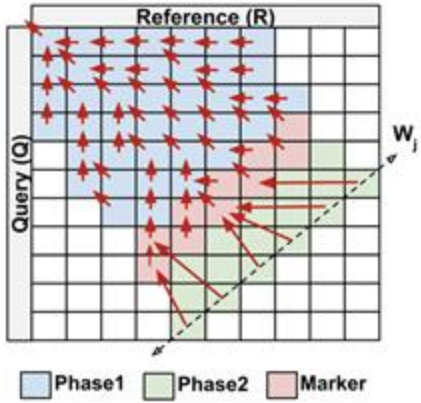
Key Contributions and Results

1. TALCO, **guarantees optimality** under banding constraints

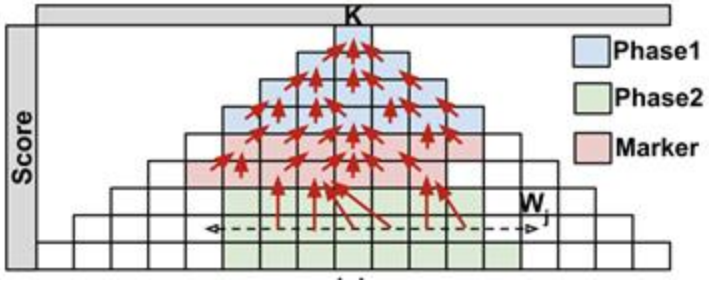


Key Contributions and Results

1. TALCO, **guarantees optimality** under banding constraints
2. We applied TALCO to X-Drop (**TALCO- XDrop**) and WFA-Adapt (**TALCO-WFAA**)



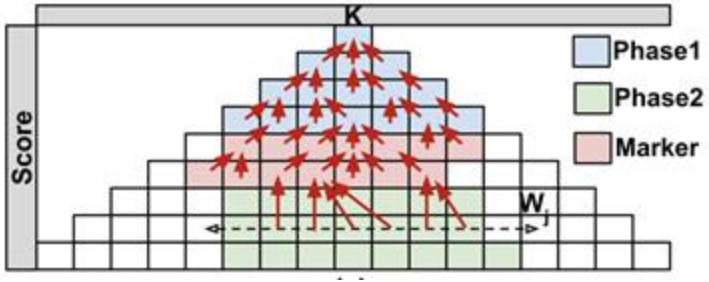
TALCO-XDrop



TALCO-WFAA

Key Contributions and Results

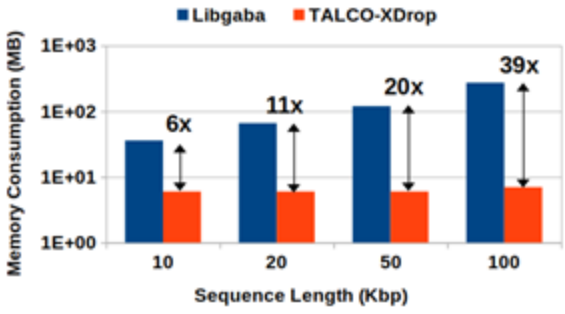
1. TALCO, **guarantees optimality** under banding constraints
2. We applied TALCO to X-Drop (**TALCO- XDrop**) and WFA-Adapt (**TALCO-WFAA)**
3. TALCO-WFAA is the **first accelerator** based on the **WFA-Adapt algorithm** capable of performing arbitrary long sequence alignments

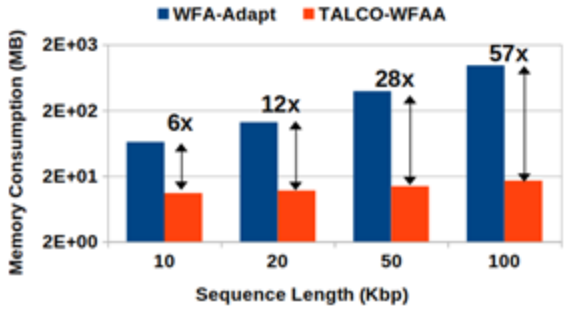


TALCO-WFAA

Key Contributions and Results

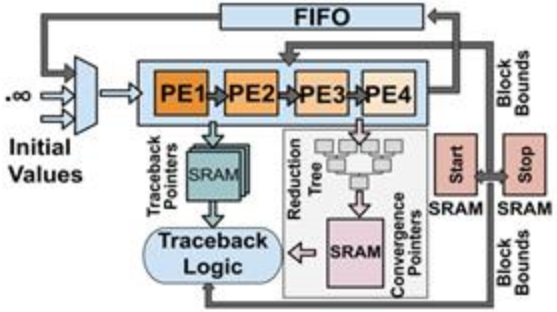
1. TALCO, **guarantees optimality** under banding constraints
2. We applied TALCO to X-Drop (**TALCO- XDrop**) and WFA-Adapt (**TALCO-WFAA)**
3. TALCO-WFAA is the **first accelerator** based on the **WFA-Adapt algorithm** capable of performing arbitrary long sequence alignments
4. TALCO-XDrop and TALCO-WFAA (**software**) achieves up to **39x** and **57x** improvement in **memory footprint**, respectively, compared to software baselines



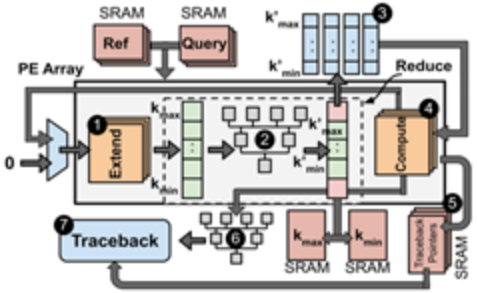


Key Contributions and Results

1. TALCO, **guarantees optimality** under banding constraints
2. We applied TALCO to X-Drop (**TALCO- XDrop**) and WFA-Adapt (**TALCO-WFAA)**
3. TALCO-WFAA is the **first accelerator** based on the **WFA-Adapt algorithm** capable of performing arbitrary long sequence alignments
4. TALCO-XDrop and TALCO-WFAA (**software**) achieves up to **39x** and **57x** improvement in memory footprint, respectively, compared to software baselines
5. **Designed hardware accelerators** for TALCO- XDrop and TALCO-WFAA



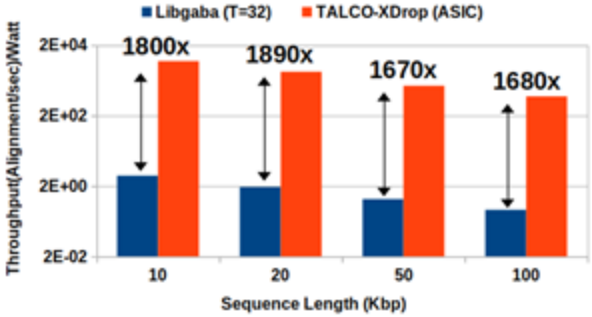
TALCO-XDrop hardware design

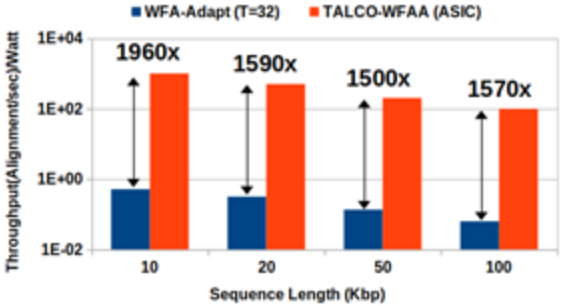


TALCO-WFAA hardware design

Key Contributions and Results

1. TALCO, **guarantees optimality** under banding constraints
2. We applied TALCO to X-Drop (**TALCO- XDrop**) and WFA-Adapt (**TALCO-WFAA)**
3. TALCO-WFAA is the **first accelerator** based on the **WFA-Adapt algorithm** capable of performing arbitrary long sequence alignments
4. TALCO-XDrop and TALCO-WFAA (**software**) achieves up to **39x** and **57x** improvement in memory footprint, respectively, compared to software baselines
5. **Designed hardware accelerators** for TALCO- XDrop and TALCO-WFAA
6. TALCO-XDrop and TALCO-WFAA (**ASIC**) achieves up to **~1,900X** and **~2,000X**, respectively, improvement in **alignment throughput/watt** over software baselines





Key Contributions and Results

1. TALCO, **guarantees optimality** under banding constraints
2. We applied TALCO to X-Drop (**TALCO- XDrop**) and WFA-Adapt (**TALCO-WFAA)**
3. TALCO-WFAA is the **first accelerator** based on the **WFA-Adapt algorithm** capable of performing arbitrary long sequence alignments
4. TALCO-XDrop and TALCO-WFAA (**software**) achieves up to **39x** and **57x** improvement in memory footprint, respectively, compared to software baselines
5. **Designed hardware accelerators** for TALCO- XDrop and TALCO-WFAA
6. TALCO-XDrop and TALCO-WFAA (**ASIC**) achieves up to **~1,900X** and **~2,000X**, respectively, improvement in **alignment throughput/watt** over software baselines
7. We **synthesized** TALCO-XDrop and TALCO-WFAA for **FPGAs** available on the **Amazon EC2 FPGA instances**

https://github.com/TurakhiaLab/TALCO/blob/main/hardware/README.md

Building on AWS EC2 F1 instance

Follow the below instructions to execute TALCO-XDrop and TALCO-WFAA on the AWS EC2 F1 instance, f1.2xlarge.

- Clone aws-fpga repository

```
git clone https://github.com/aws/aws-fpga
cd aws-fpga
source vitis_runtime_setup.sh
```

- Clone TALCO repository

```
git clone https://github.com/TurakhiaLab/TALCO.git
export TALCO_DIR=$PWD/TALCO
cd TALCO/hardware/TALCO-XDrop
```

- Steps for running on the EC2 F1 instance, f1.2xlarge (MODE-hw)

```
source $TALCO_DIR/hardware/scripts/run.sh
$TALCO_DIR/dataset/sequence.fa TALCO_XDrop.awsxclbin
```

Key Contributions and Results

1. TALCO, **guarantees optimality** under banding constraints
2. We applied TALCO to X-Drop (**TALCO- XDrop**) and WFA-Adapt (**TALCO-WFAA)**
3. TALCO-WFAA is the **first accelerator** based on the **WFA-Adapt algorithm** capable of performing arbitrary long sequence alignments
4. TALCO-XDrop and TALCO-WFAA (**software**) achieves up to **39x** and **57x** improvement in memory footprint, respectively, compared to software baselines
5. **Designed hardware accelerators** for TALCO- XDrop and TALCO-WFAA
6. TALCO-XDrop and TALCO-WFAA (**ASIC**) achieves up to **~1,900X** and **~2,000X**, respectively, improvement in **alignment throughput/watt** over software baselines
7. We **synthesized** TALCO-XDrop and TALCO-WFAA for **FPGAs** available on the **Amazon EC2 FPGA instances**

https://github.com/TurakhiaLab/TALCO/

HPCA Artifact Evaluation

Outline

- Emergence of **Long Genome Sequence Alignment**
- Current techniques and their **limitations** to **hardware accelerate** long genome sequence alignment
- **TALCO**: A tiling technique based on convergence of traceback pointers for long genome sequence alignment
- **Key Contributions** and **Results**
- **Conclusion**

Conclusion

- We present TALCO, a novel tiling technique for long genome sequence alignment
  - **Maintains a constant memory footprint**
  - Ensures **optimal alignments** under banding constraints

- We applied TALCO to X-Drop (**TALCO-XDrop**) and WFA-Adapt (**TALCO-WFAA**)

- TALCO-XDrop (TALCO-WFAA) software achieve up to **39X** (**57X**) improvement in **memory footprint** for long alignments compared to software baselines

- We present **hardware accelerator designs** for **TALCO-XDrop** and **TALCO-WFAA**

- TALCO-XDrop (TALCO-WFAA) ASIC achieves up to **1,900X** (**2,000X**) improvement in **alignment throughput/watt** over software baselines implementing the same algorithm

UC San Diego
Electrical and Computer Engineering
JACOBS SCHOOL OF ENGINEERING

TURAKHIA LAB

TALCO: Tiling Genome Sequence Alignment using Convergence of Traceback Pointers

Sumit Walia
Ph.D student

Cheng Ye
MS student

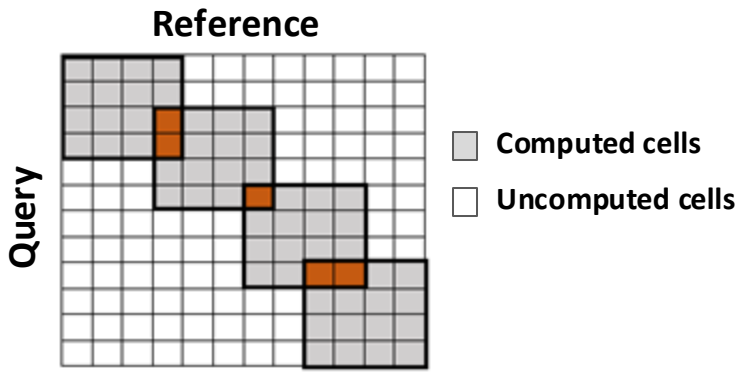
Arkid Bera
MS student

Dhruvi L.
MS student

Yatish Turakhia
Assistant Professor, UCSD

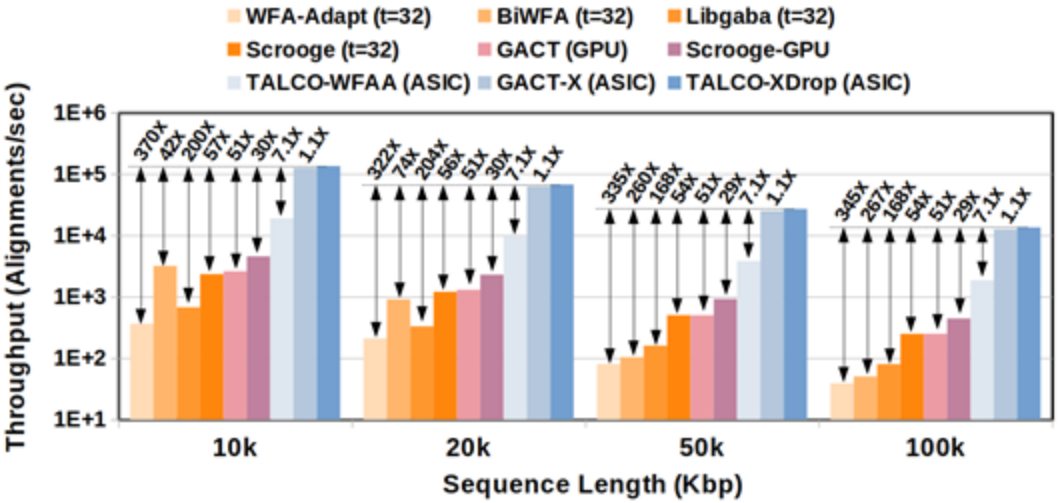
Thank you!

Additional Slides

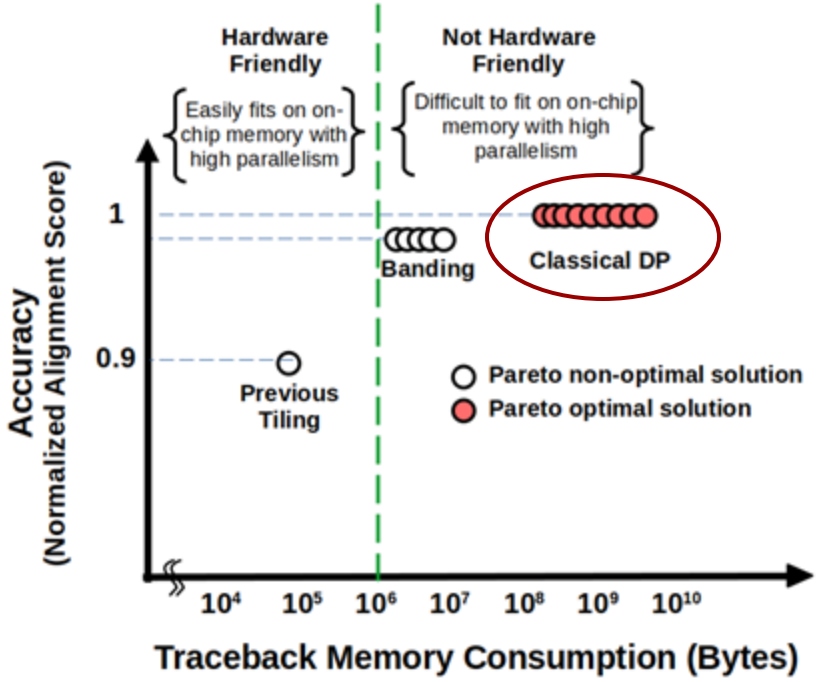

Reference
Query
Computed cells
Uncomputed cells

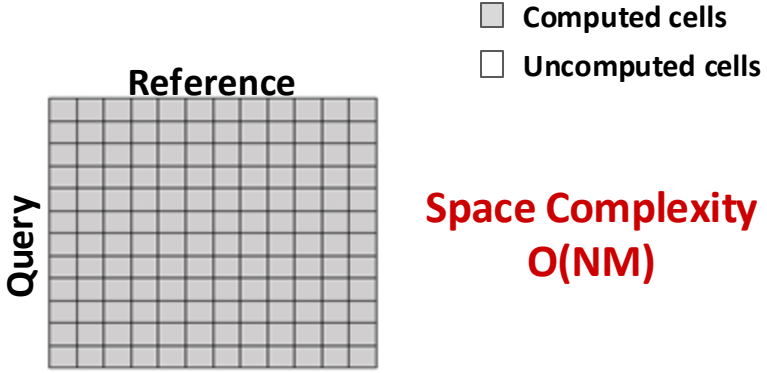
Key Results

3. TALCO improves the alignment throughput over state-of-the-art GPU and ASIC baselines that implement tiling heuristics by over **50X** and **1.1X**, respectively.

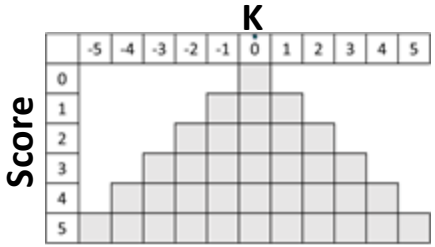


Full Matrix Sequence Alignment Algorithms



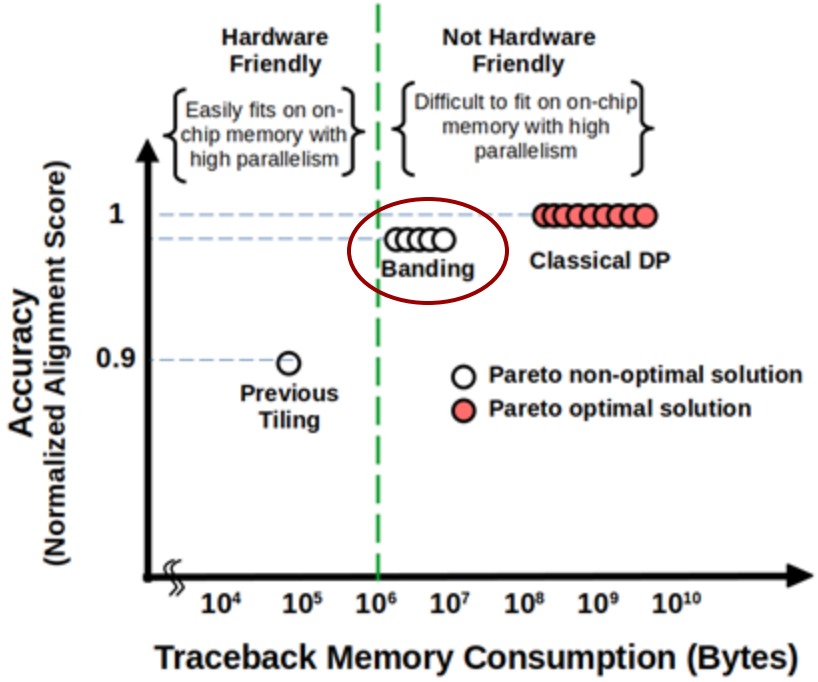


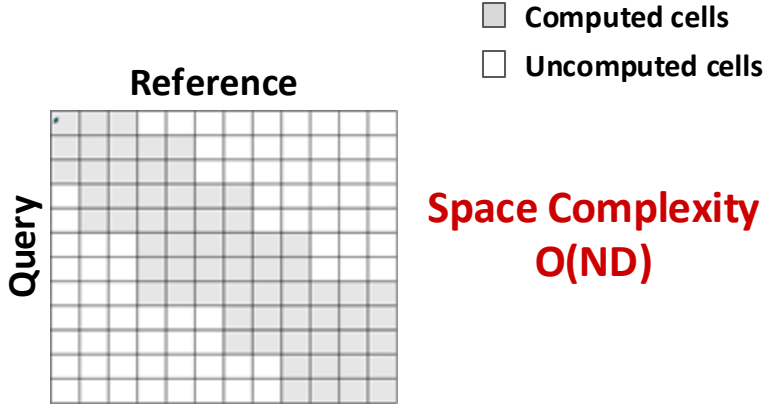
Space Complexity O(NM)



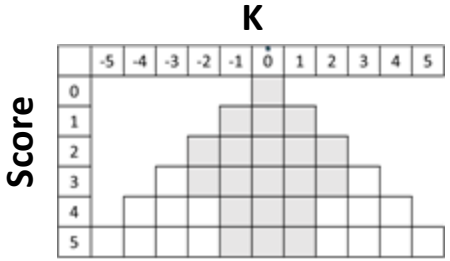
Space Complexity $O(S^2)$

Banded Sequence Alignment Algorithms



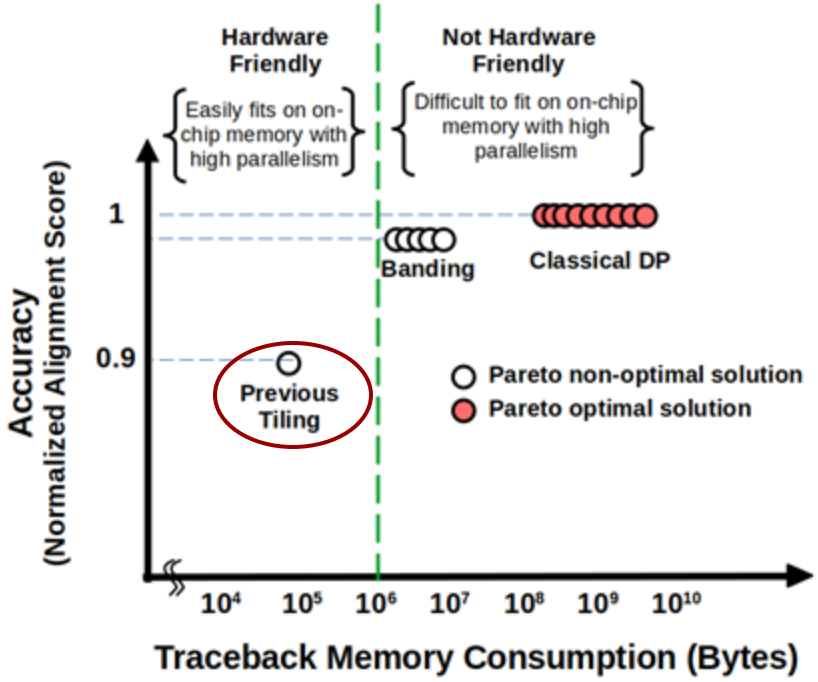


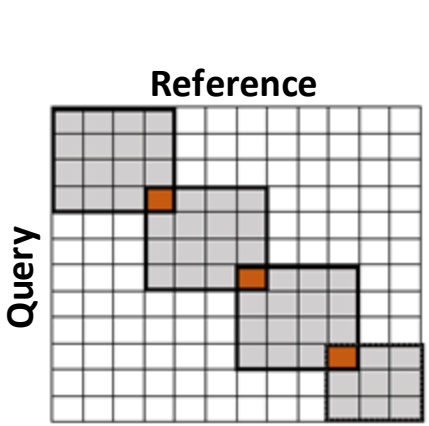
Space Complexity O(ND)



Space Complexity O(SD)

Tiling heuristic

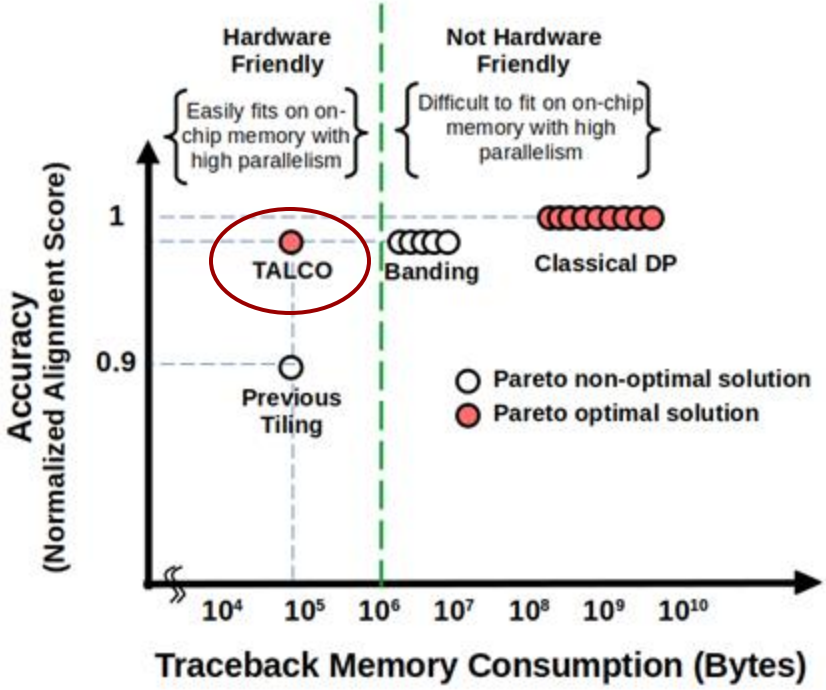




Space Complexity $O(T^2)$

Tiling is **never applied** to **non-classical sequence alignment algorithm**

TALCO on Pareto Plot



- **Constant Space complexity**
- **Guarantees optimality under banding constraints** (dynamic overlap between consecutive tiles)

Key Contributions

3. We implemented TALCO-XDrop and TALCO-WFAA in software, and achieves up to **39x** and **57x** improvement in **memory footprint**, respectively, compared to software baselines