Computer Architecture

Lecture 26b: RawHash

Can Firtina

ETH Zurich
Fall 2024
13 December 2024

Brief Self Introduction

- **Can Firtina**
  - Senior researcher in the SAFARI Research Group and a lecturer at ETH Zurich (PhD thesis defended in November 2024)

- **Research interests:** Bioinformatics & Computer Architecture
  - Real-time genome analysis
  - Similarity search in a large space of genomic data
  - Hardware-Algorithm co-design to accelerate genome analysis
  - Genome editing
  - Error correction

- Get to know **our group and our research: https://safari.ethz.ch**
  - Contact me: canfirtina@gmail.com
  - Website: https://canfirtina.com
  - Twitter (aka X): https://twitter.com/FirtinaC
  - Bluesky: https://bsky.app/profile/firtinac.bsky.social

Recall: Key Applications of Genome Analysis

Uncovering and treating diseases linked to genomic variations

Altering genomes to solve fundamental challenges of life

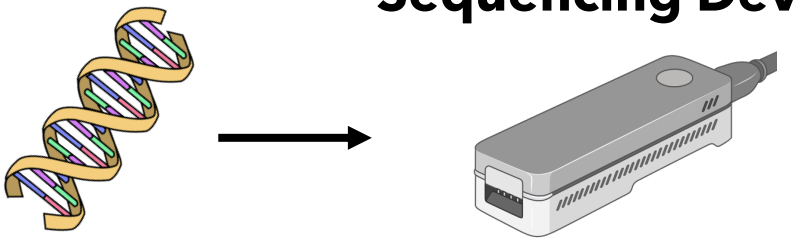
Detecting **pathogens** in the environment

Rapid surveillance of **disease outbreaks**

Recall: Genome Sequencing Data Generation

Sequencing process **converts biological molecules**
into **digital nucleotide sequences called reads**

Sequencing Device

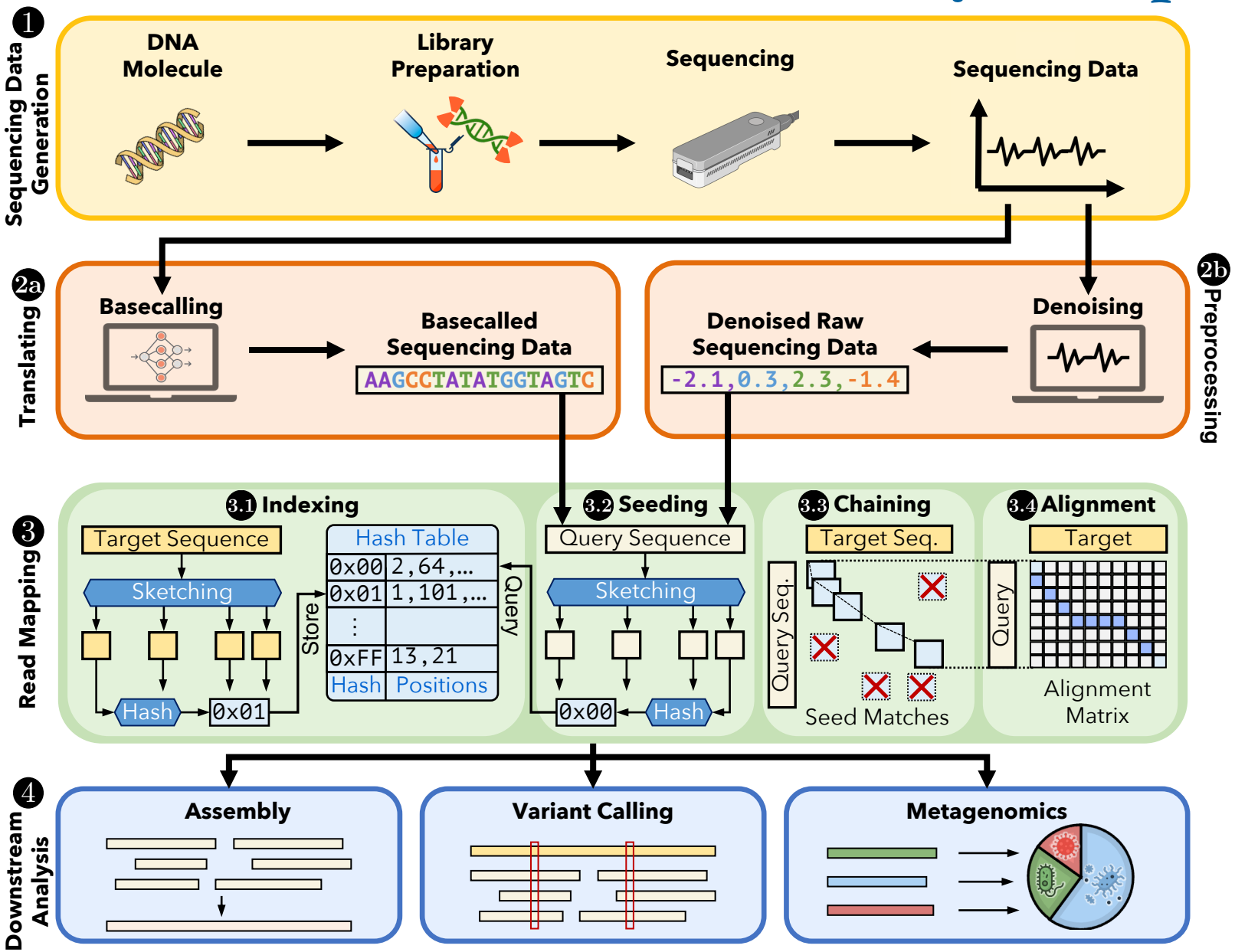


Biological Molecule (e.g., DNA)

Challenge: Unknown origins

Challenge: Large volume of data to analyze

Recall: A Common Genome Analysis Pipeline



Agenda for Today

- Real-Time and Raw Sequencing Data Analysis
  - RawHash [in ISMB/ECCB 2023]
  - RawHash2 [Bioinformatics 2024]
  - Rawsamble [arXiv 2024]

RawHash

Enabling Fast and Accurate Real-Time Analysis of Raw Nanopore Signals for Large Genomes

Can Firtina

Nika Mansouri Ghiasi Joel Lindegger Gagandeep Singh

Meryem Banu Cavlak Haiyu Mao Onur Mutlu

RawHash

RawHash2

Code

SAFARI

ETH zürich

Outline

Background

RawHash

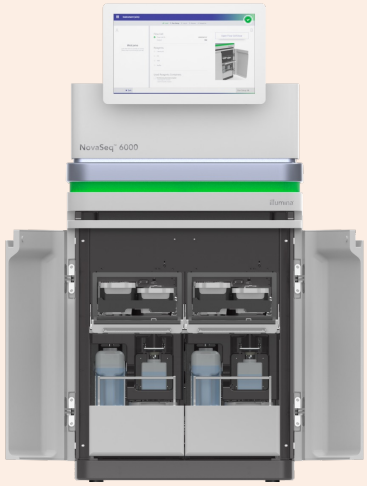
RawHash2

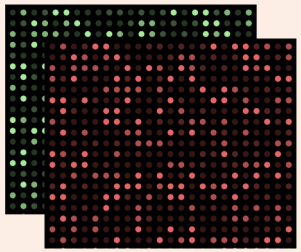
Evaluation

Conclusion

Recall: Different Raw Sequencing Data

Illumina

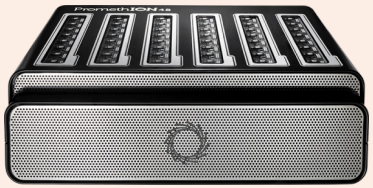


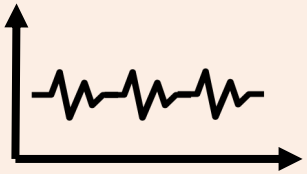


Multiple images

.BCL/.CBCL

Nanopore

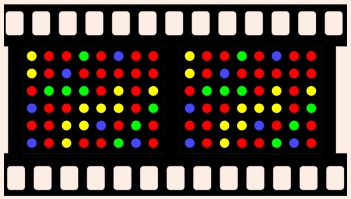




Electrical Signal

.POD5

PacBio



30-hour movie

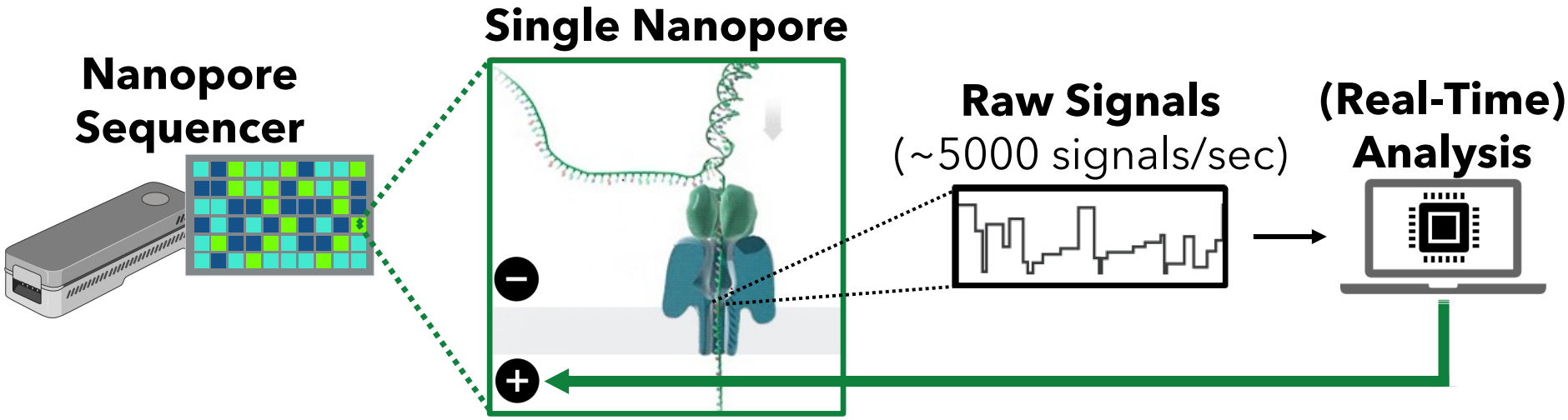
.BAM

Nanopore Sequencing

Nanopore Sequencing: a widely used sequencing technology

- Can sequence large fragments of nucleic acid molecules
- Offers high throughput
- Cost-effective
- Enables **real-time and portable genome analysis**

Nanopore Sequencing – How it Works



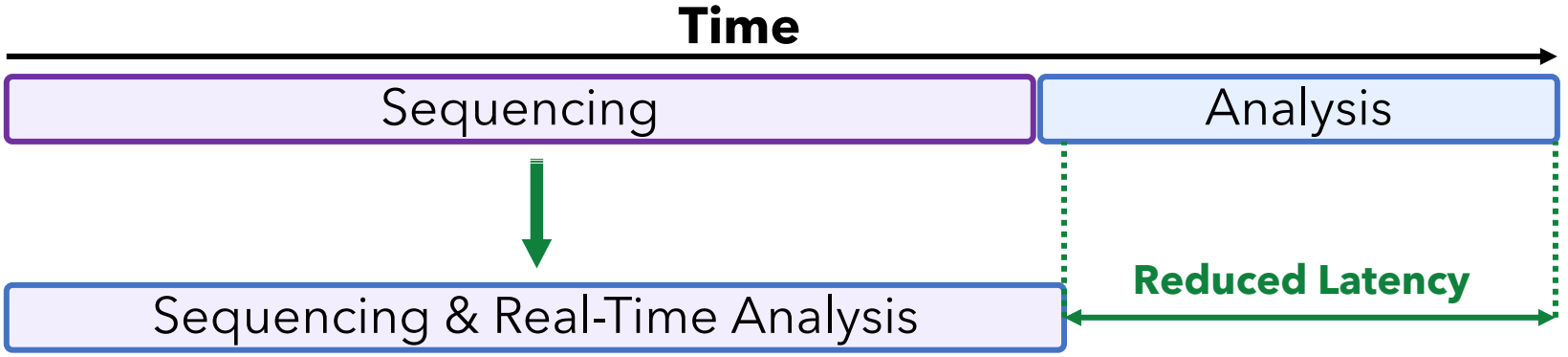
Raw Signals: Ionic current measurements generated at a certain **throughput**

(Real-Time) Analysis: Analyzing raw signals **instantly as they are generated**

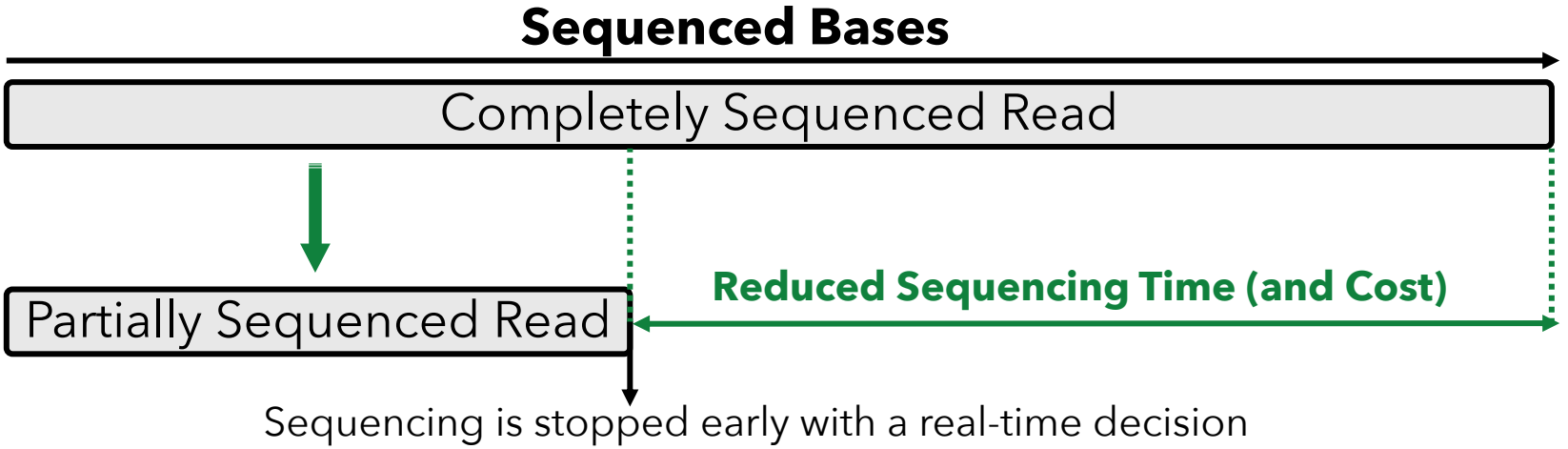
Real-Time Decisions: Stopping sequencing **early** based on real-time analysis

Benefits of Real-Time Analysis

Reducing latency by overlapping analysis with sequencing



Reducing sequencing time and cost by stopping sequencing early



Challenges in Real-Time Analysis

Rapid analysis to match the nanopore sequencer throughput

Timely decisions to stop sequencing as early as possible

Accurate analysis from noisy raw signal data

Power-efficient computation for scalability and portability

Outline

Background

RawHash

RawHash2

Evaluation

Conclusion

Executive Summary

Problem: Real-time analysis of nanopore raw signals is **inaccurate** and **inefficient for large genomes**

Goal: Enable **fast** and **accurate** real-time analysis of raw nanopore signals

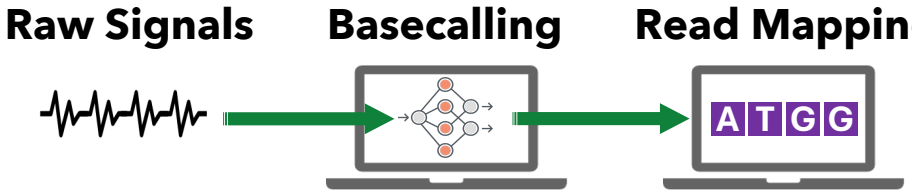
Key Contributions:
1) The **first hash-based mechanism** for mapping raw nanopore signals

2) The novel **Sequence Until** technique can accurately and **dynamically stop the entire sequencing of all reads at once** if further sequencing is not necessary

Key Results: Across 3 use cases and 5 genomes of varying sizes
- **27× 19×, and 4× better average throughput** compared to the state-of-the-art works
- **Most accurate raw signal mapper for all datasets**
- Sequence Until **reduces the sequencing time and cost by 15×**

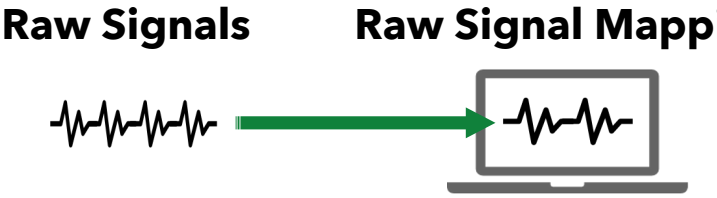
Analyzing Raw Nanopore Signals

Traditional: Translating (**basecalling**) signals to bases **before** analysis



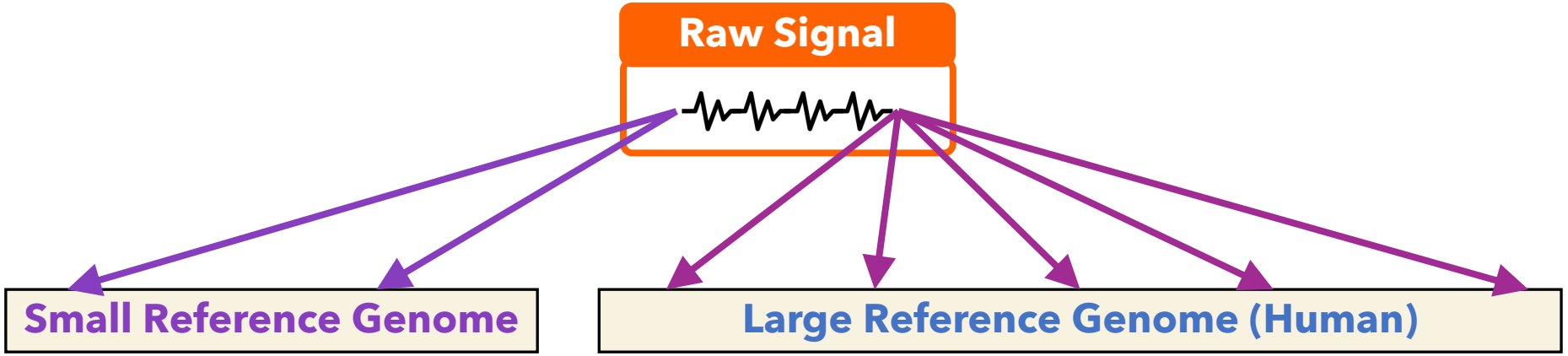
- Basecalled sequences are **less noisy** than raw signals
- **Many analysis tools** use basecalled sequences
- **Costly and power-hungry** computational requirements

Recent Works: Directly analyzing signals **without basecalling**



- **Efficient analysis** with better scalability and portability
- Raw signals retain **more information** than just bases

The Problem with Raw Signal Mapping



Fewer candidate regions in **small genomes**

Substantially **larger number of regions** to check **per read** as the genome size increases

Accurate mapping

Problem: Probabilistic mechanisms on **many regions → inaccurate mapping**

High throughput

Problem: Distance calculation on **many regions → reduced throughput**

The Problem with Raw Signal Mapping

Existing solutions are
inaccurate or inefficient
for large genomes

Accurate mapping

on **many regions → inaccurate mapping**

High throughput

Problem: Distance calculation
on **many regions → reduced throughput**

Goal

Enable **fast and accurate real-time analysis** of raw nanopore signals **for large genomes**

RawHash

The **first hash-based search mechanism**
to quickly and accurately map raw nanopore signals
to reference genomes

Sequence Until can accurately and **dynamically stop**
the entire sequencing run at once
if further sequencing is unnecessary

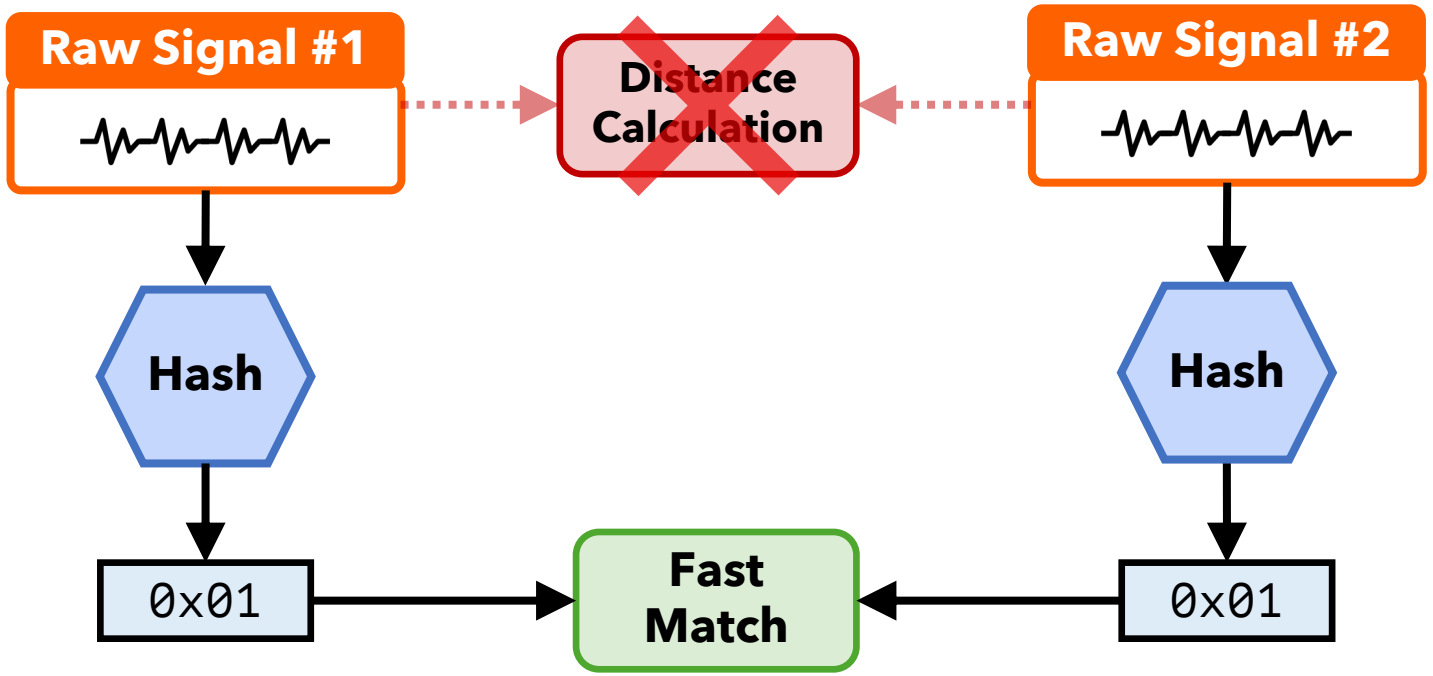
RawHash

The **first hash-based search mechanism**
to quickly and accurately map raw nanopore signals
to reference genomes

Sequence Until can accurately and **dynamically stop**
the entire sequencing run at once
if further sequencing is unnecessary

RawHash – Key Idea

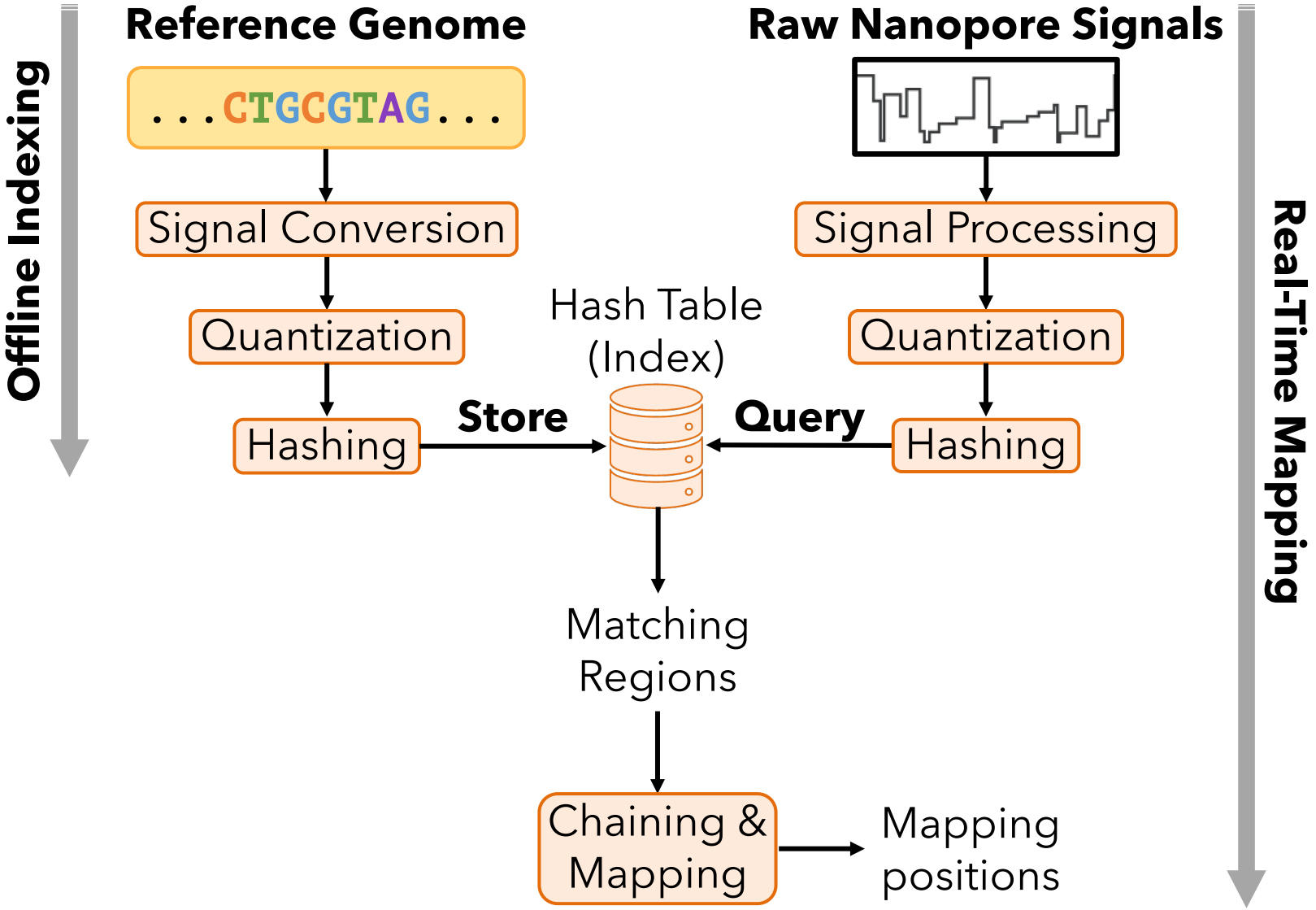
Key Observation: **Identical** nucleotides generate **similar** raw signals



Challenge #1: Generating the **same** hash value for **similar enough** signals

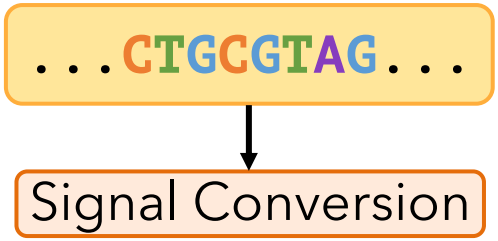
Challenge #2: **Accurately** finding as **few** similar regions as possible

RawHash Overview

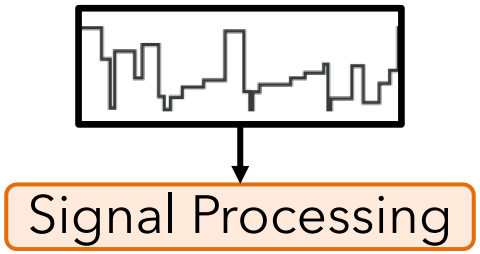


RawHash Overview

Reference Genome

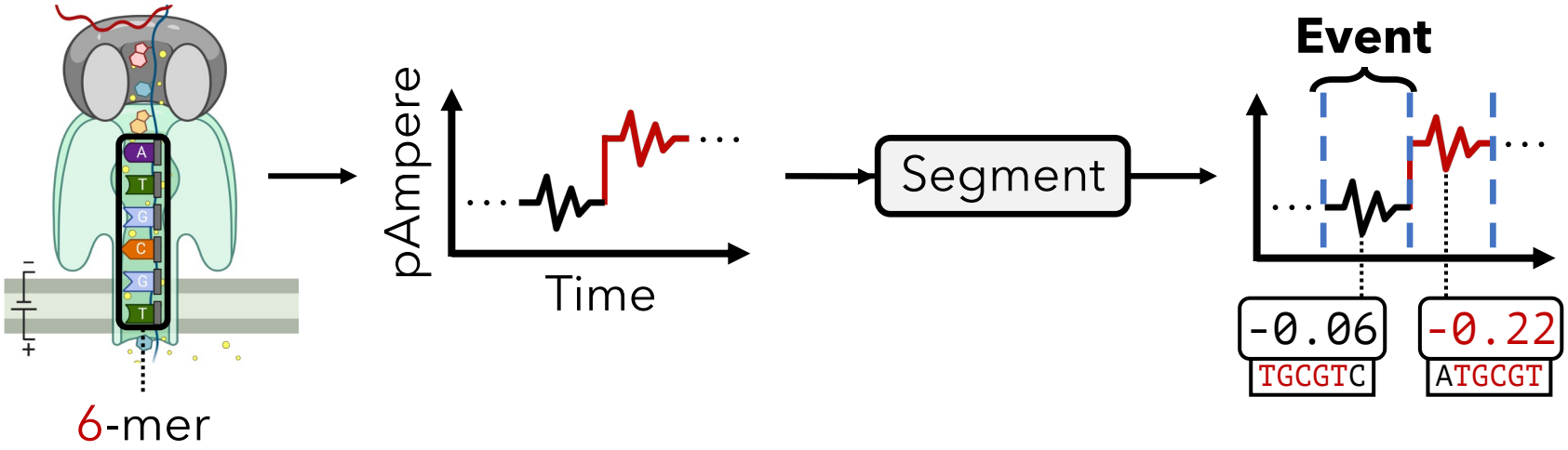


Raw Nanopore Signals



Processing Raw Nanopore Signals

- K many nucleotides (k-mers) sequenced at a time
- **Observation:** Abrupt change in the signal as DNA moves inside a nanopore (e.g., when sequencing a new k-mer)
- **Goal:** Identify **raw signal segments** corresponding each sequenced k-mer
  - **Event:** A raw signal **segment** corresponding to a **particular k-mer**
  - **Statistical tests (segmentation)** to find these events by identifying abrupt changes

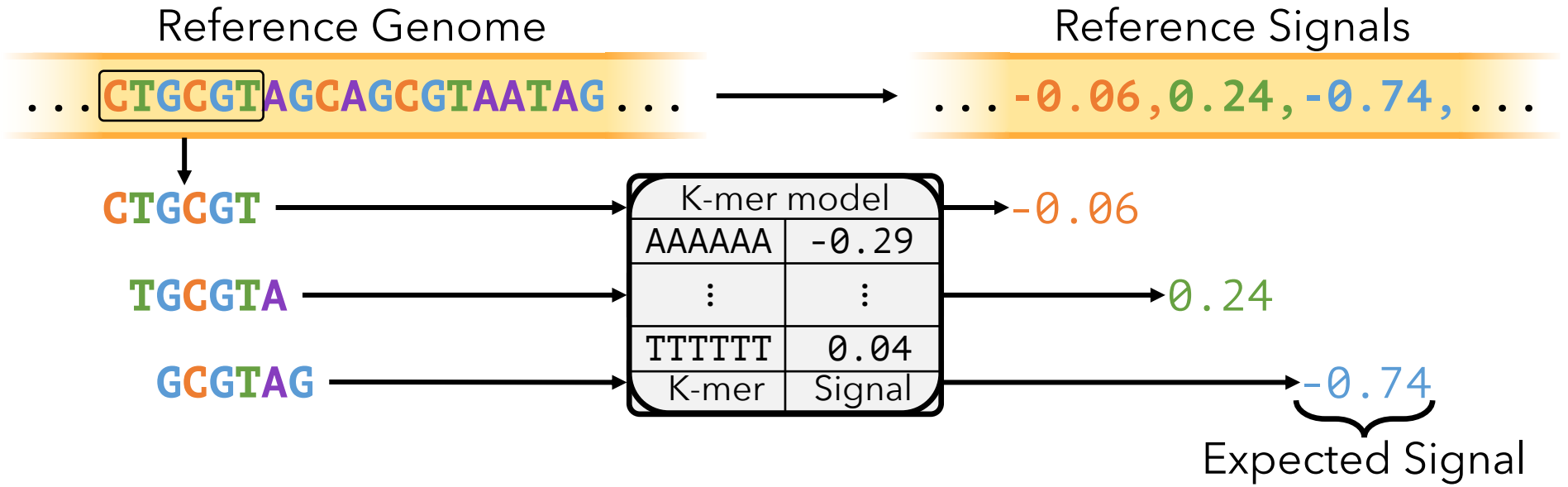


Reference-to-Signal Conversion

- **Goal:** Enable direct comparison to raw signals by converting reference genome into its synthetic signal (one-time task)

- **K-mer model:** Provides **expected** signal values **for every possible k-mer**
  - A lookup table preconstructed based on nanopore's characteristics

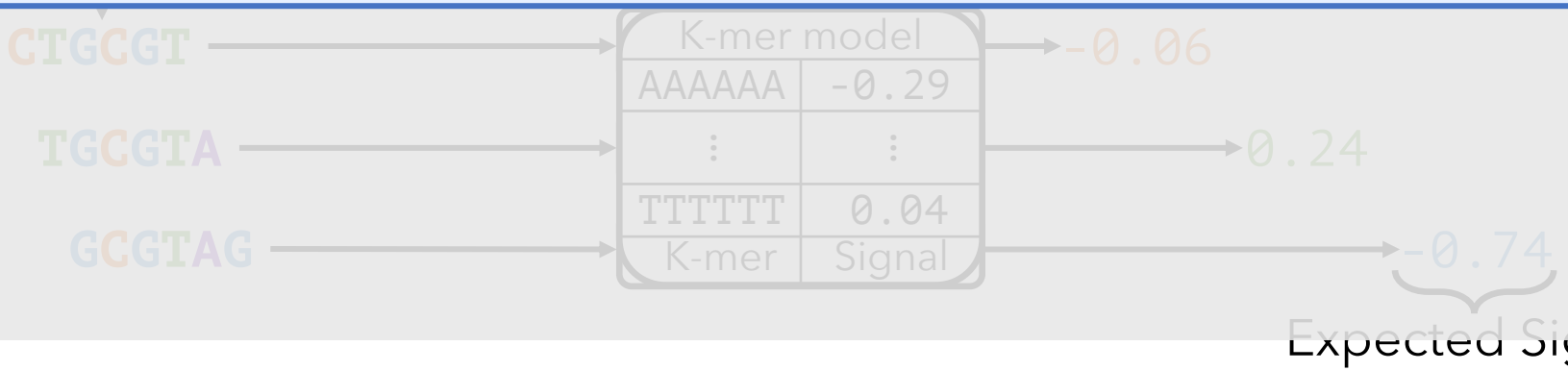
- Use the **k-mer model** to convert **all k-mers** of a reference genome to their **expected** signal values



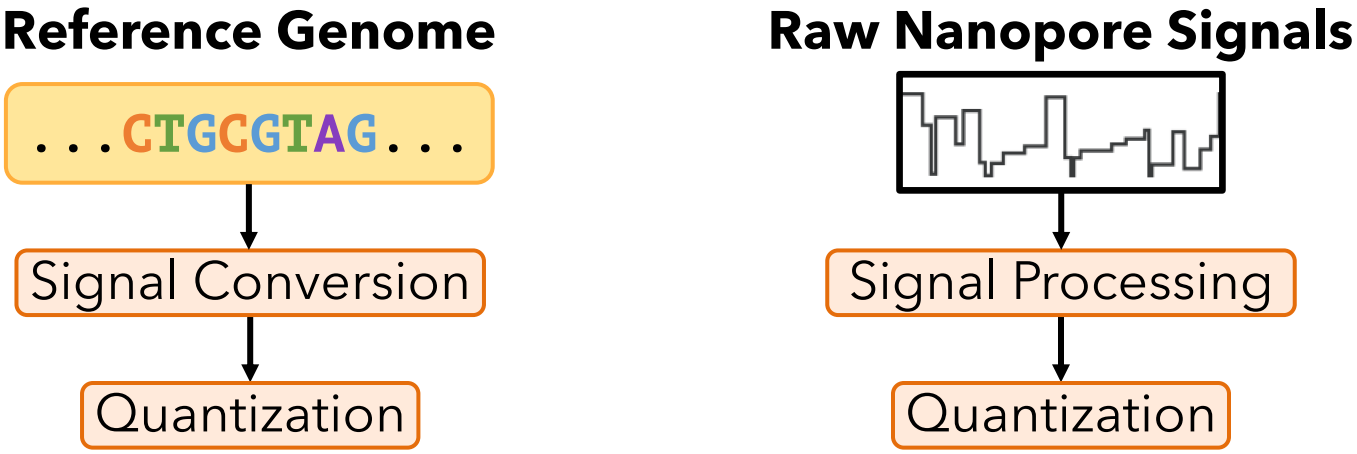
Reference-to-Signal Conversion

- **Goal:** Enable direct comparison to raw signals by converting reference genome into its synthetic signal (one-time task)
- **K-mer model:** Provides **expected** signal values **for every possible k-mer**

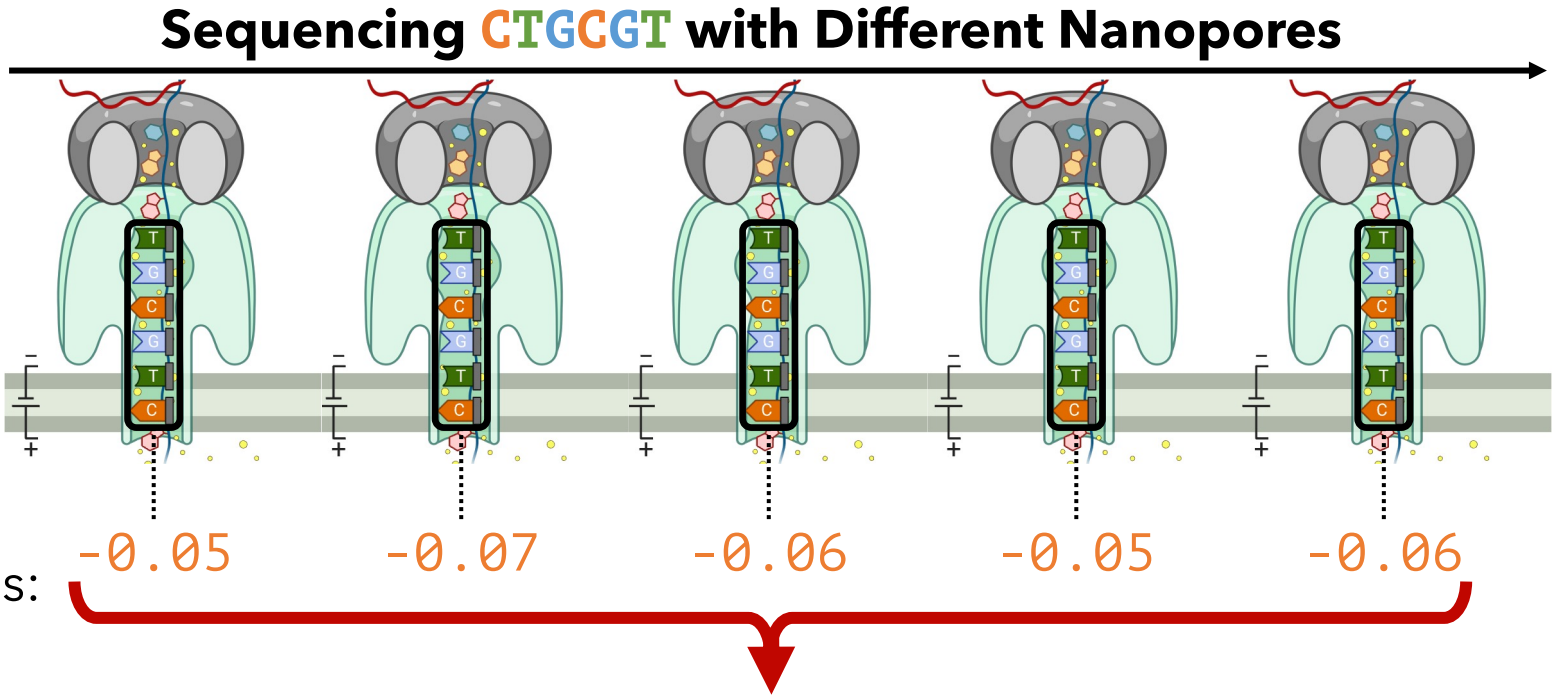
Can we directly match signals to each other?



RawHash Overview



Noise in Raw Signal Analysis

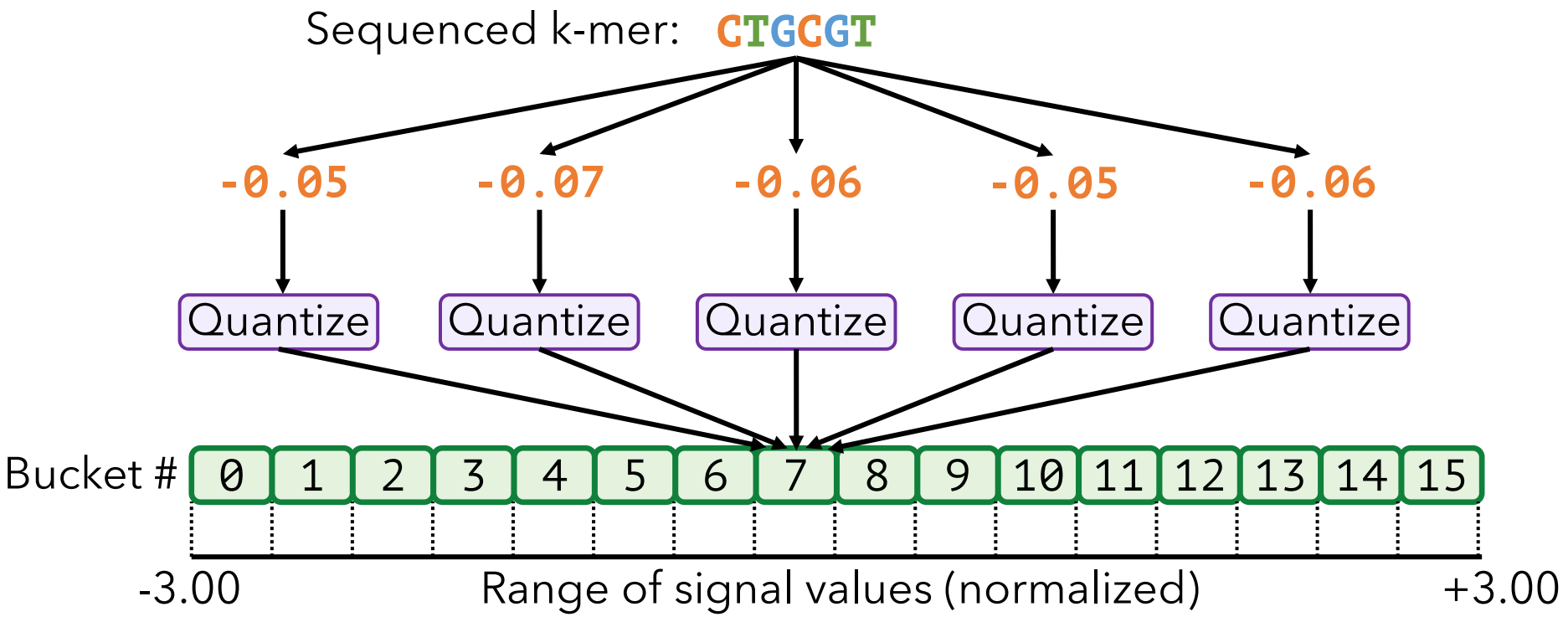


Noise causes slight differences in raw signals from **the same k-mer**

Challenge: Directly matching raw signals is not feasible

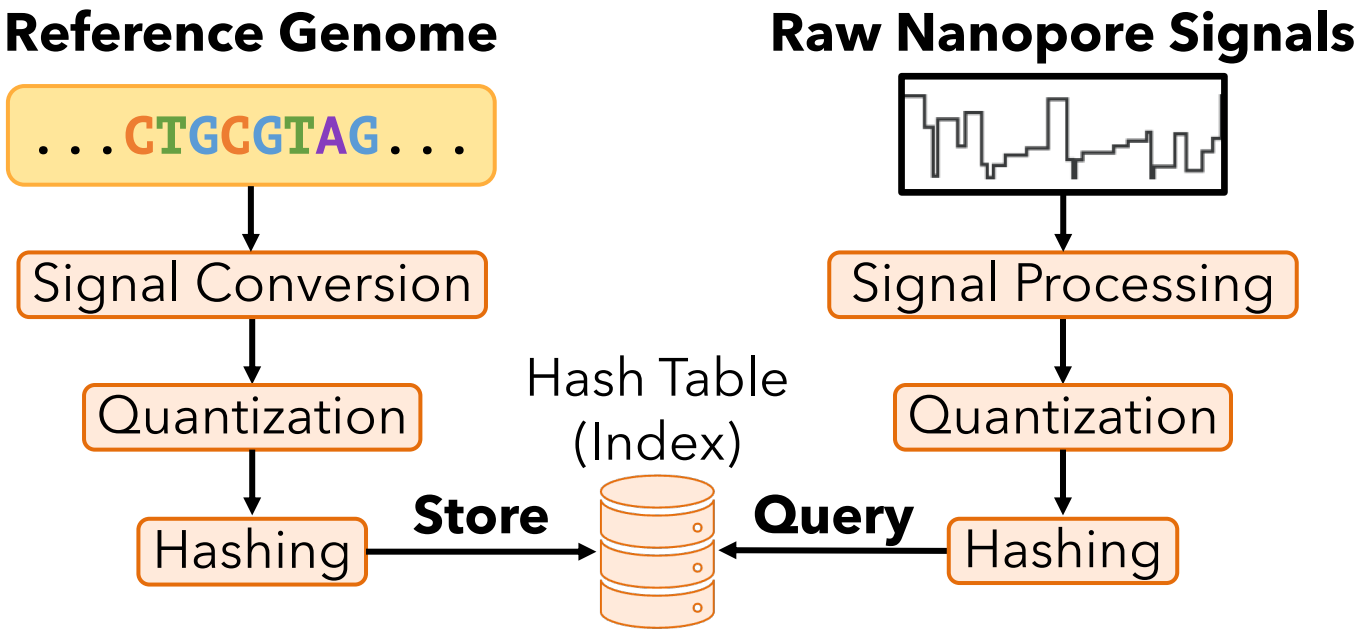
Challenge: A single k-mer is too short for accurate matching

RawHash Key Idea – Quantization

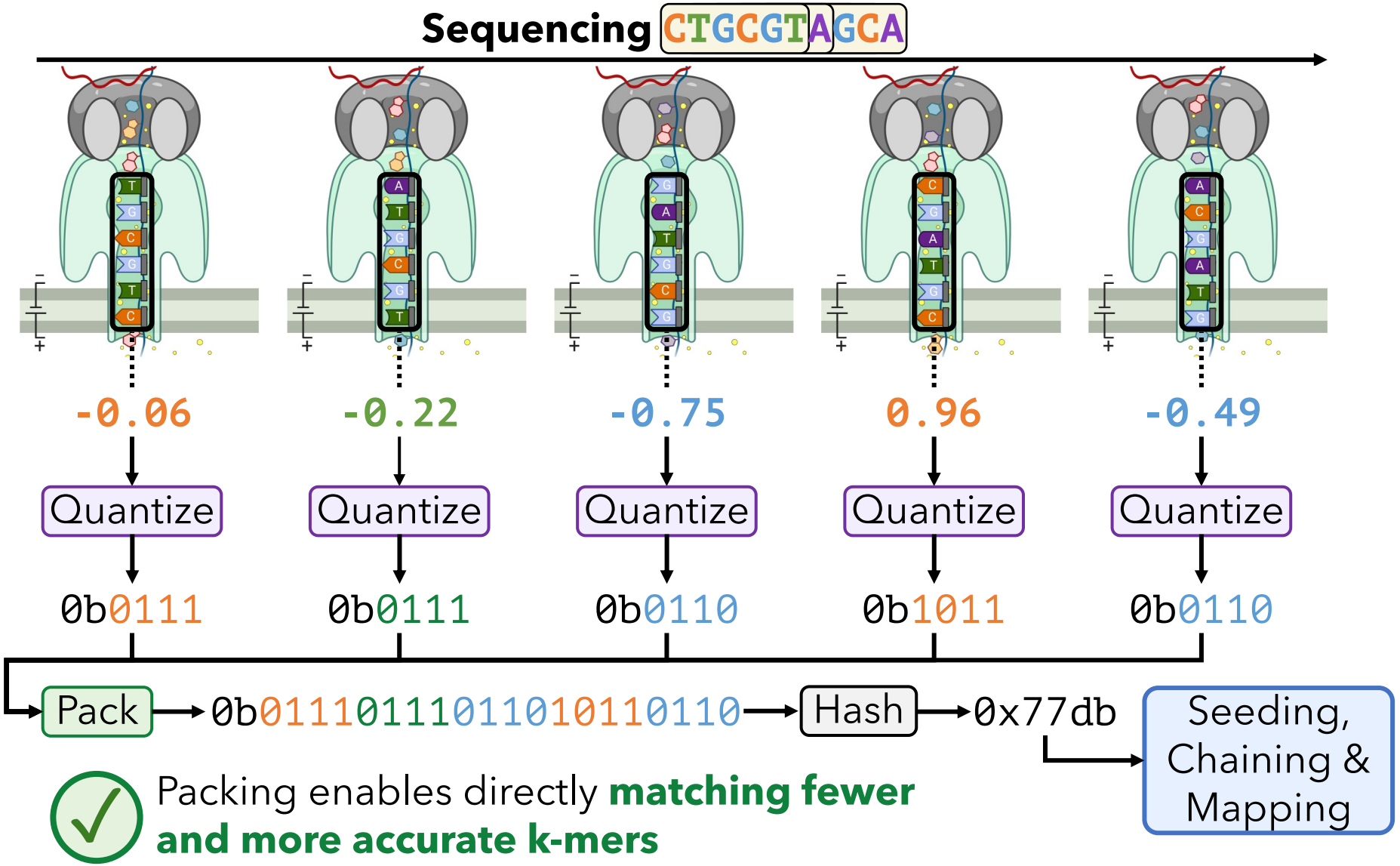


- **Reducing noise** by **quantizing** raw signals into equal-width buckets
- **Enables matching raw signals** by eliminating slight differences

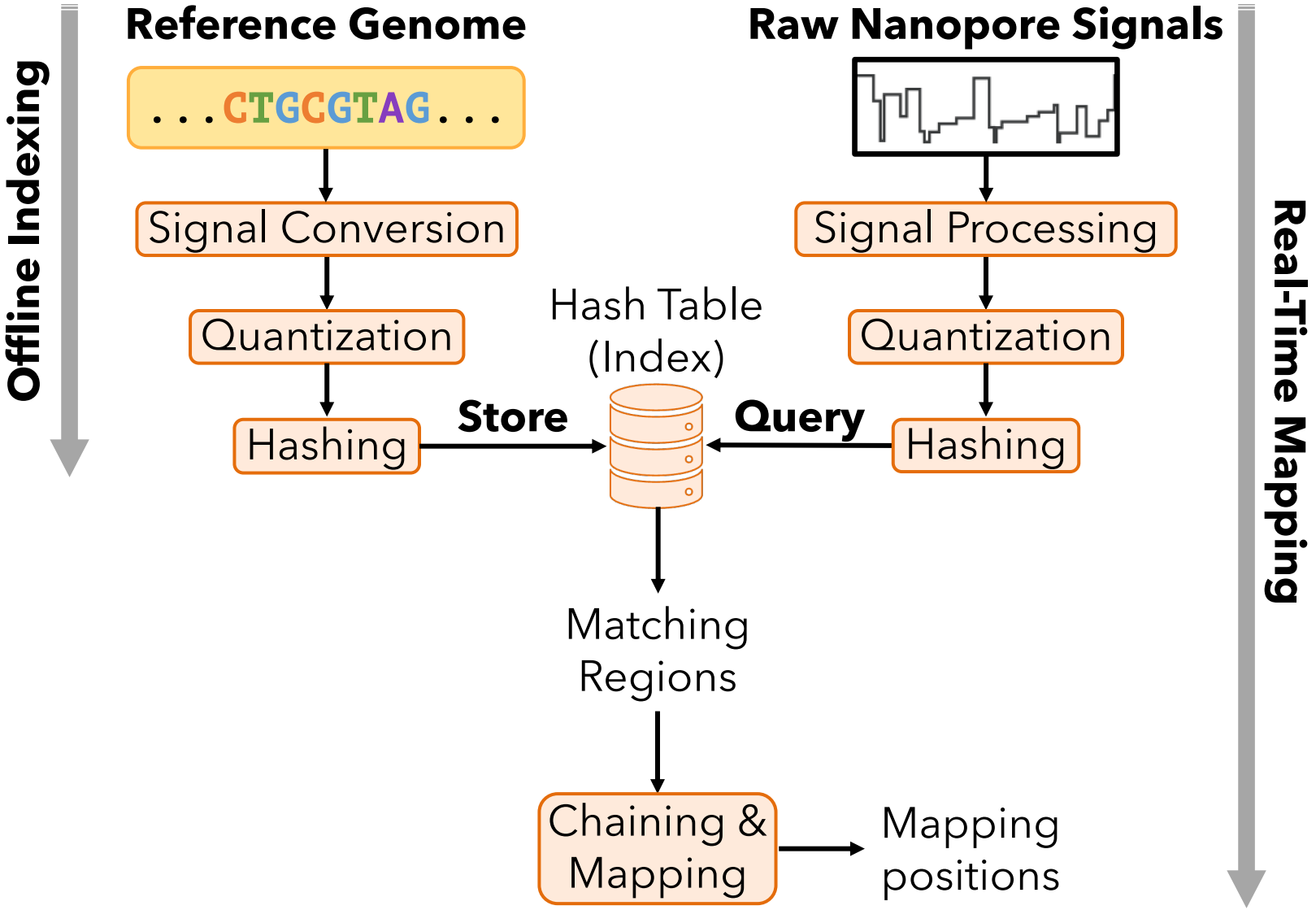
RawHash Overview



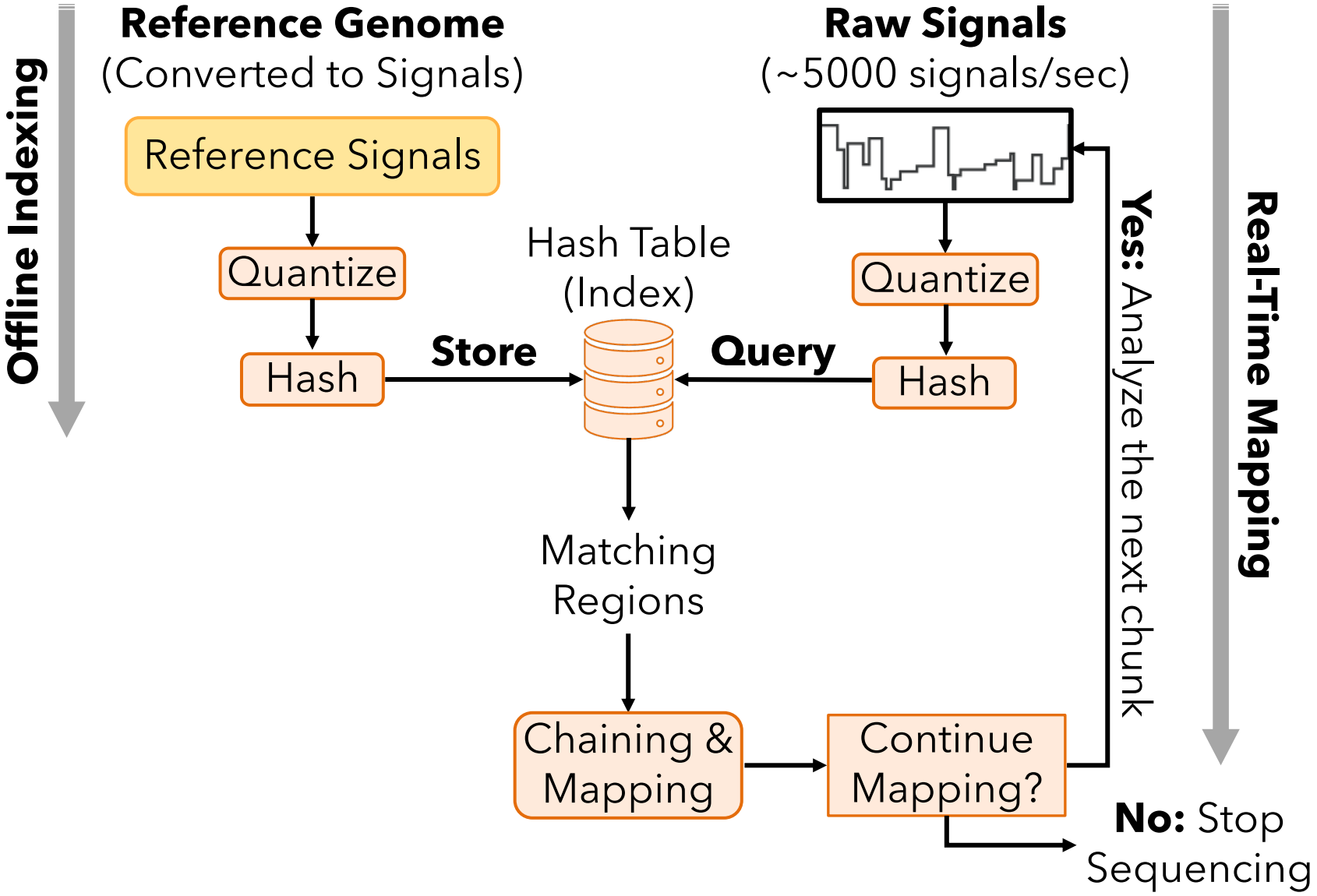
RawHash Key Idea – Hash-based Seeding



RawHash Overview



Real-Time Mapping with RawHash



RawHash

The **first hash-based search mechanism**
to quickly and accurately map raw nanopore signals
to reference genomes

Sequence Until can accurately and **dynamically stop**
the entire sequencing run at once
if further sequencing is unnecessary

RawHash

The **first hash-based search mechanism**
to quickly and accurately map raw nanopore signals
to reference genomes

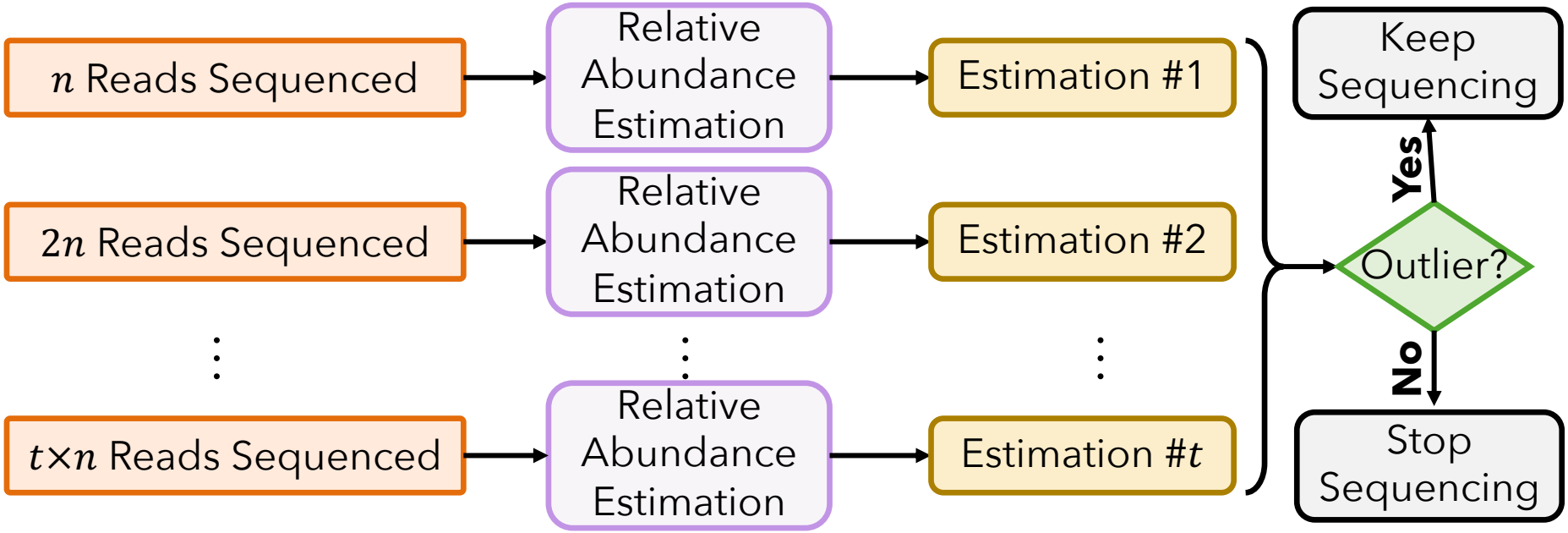
Sequence Until can accurately and **dynamically stop**
the entire sequencing run at once
if further sequencing is unnecessary

The Sequence Until Mechanism

- **Problem:**
  - Unnecessary sequencing waste time, power and money
- **Key Idea:**
  - **Dynamically** decide if further sequencing of the entire sample is necessary to achieve high accuracy
  - Stop sequencing early without sacrificing accuracy
- **Potential Benefits:**
  - Significant **reduction in sequencing time and cost**
- Example real-time genome analysis use case:
  - **Relative abundance estimation**

The Sequence Until Mechanism

- **Key Steps:**
  1. Continuously generate relative abundance estimation after every n reads
  2. Keep the last t estimation results
  3. **Detect outliers** in the results via **cross-correlation** of the recent t results
  4. Absence of outliers indicates **consistent results**
     - Further sequencing **is likely** to generate consistent results ➔ Stop the sequencing



Outline

Background

RawHash

RawHash2

Evaluation

Conclusion

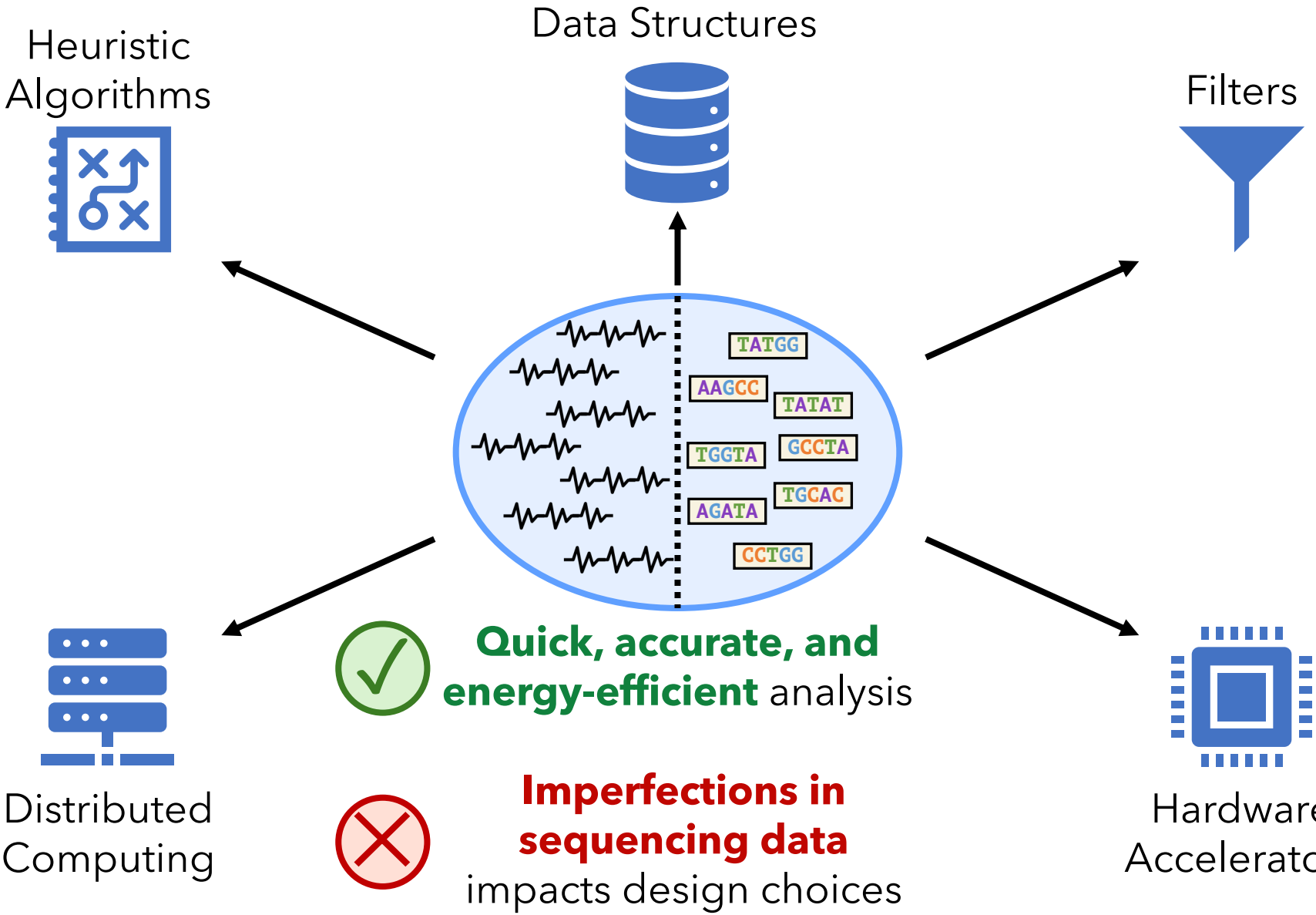
Computer Architecture

Lecture 26b: RawHash

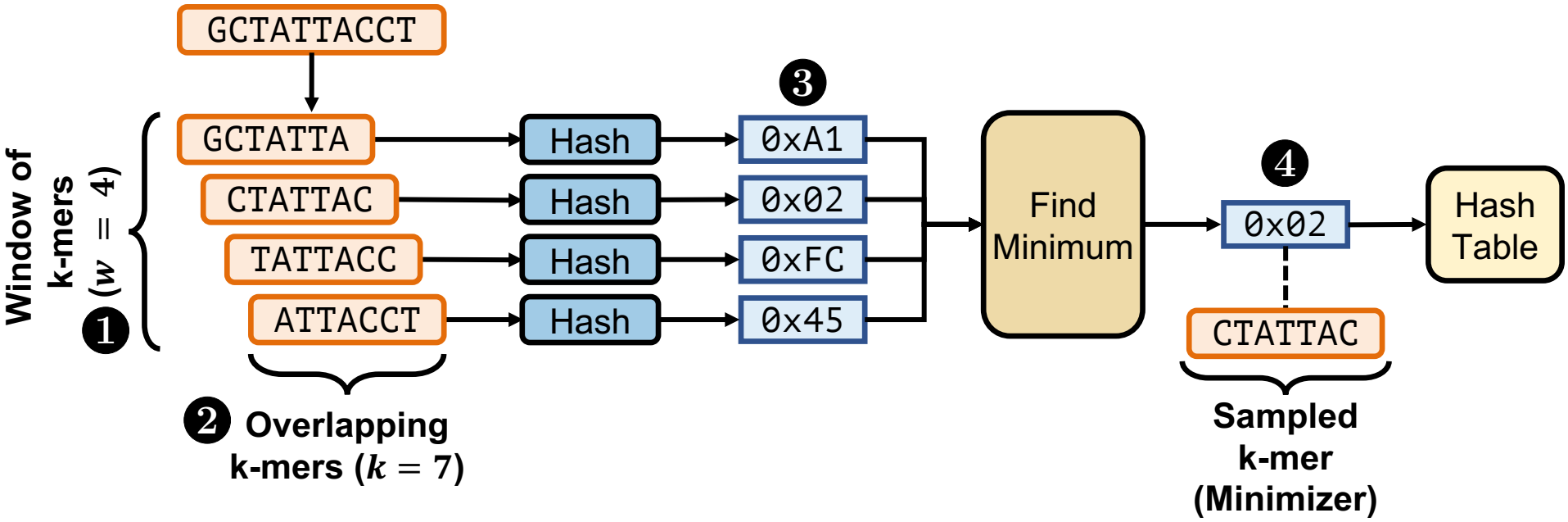
Can Firtina

ETH Zurich
Fall 2024
13 December 2024

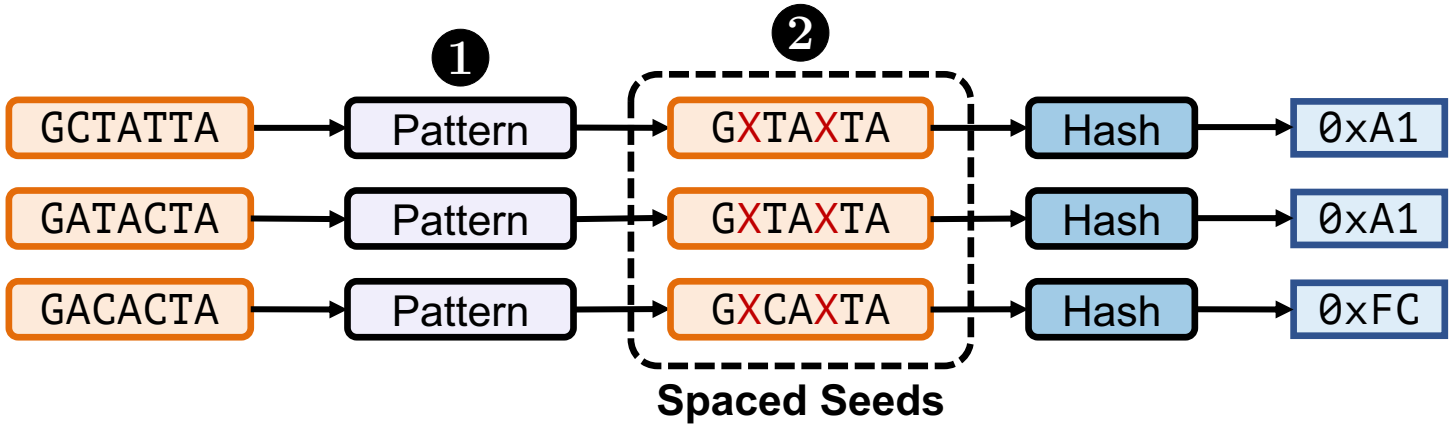
Sequencing Data Analysis



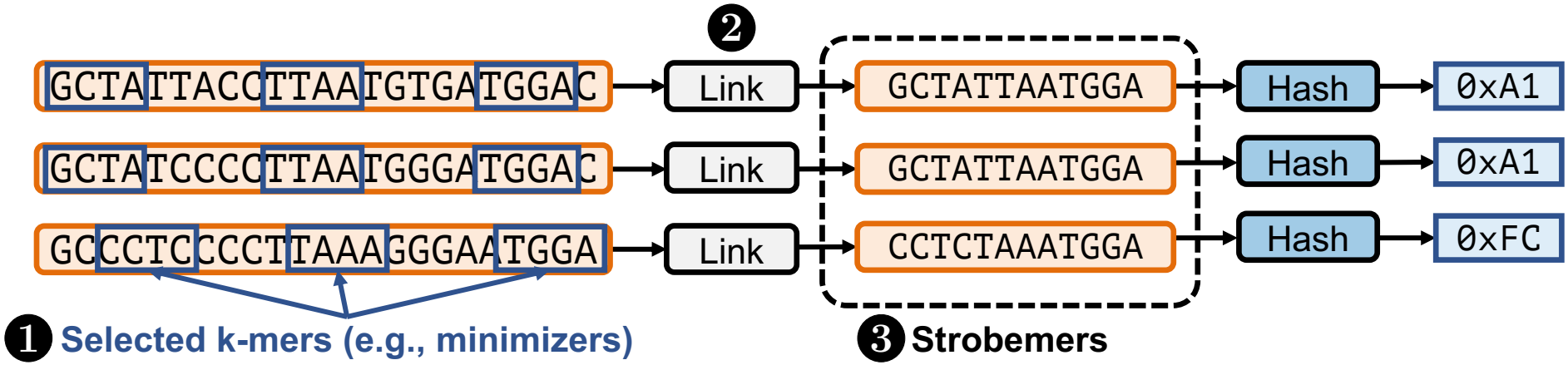
Minimizer Sketching



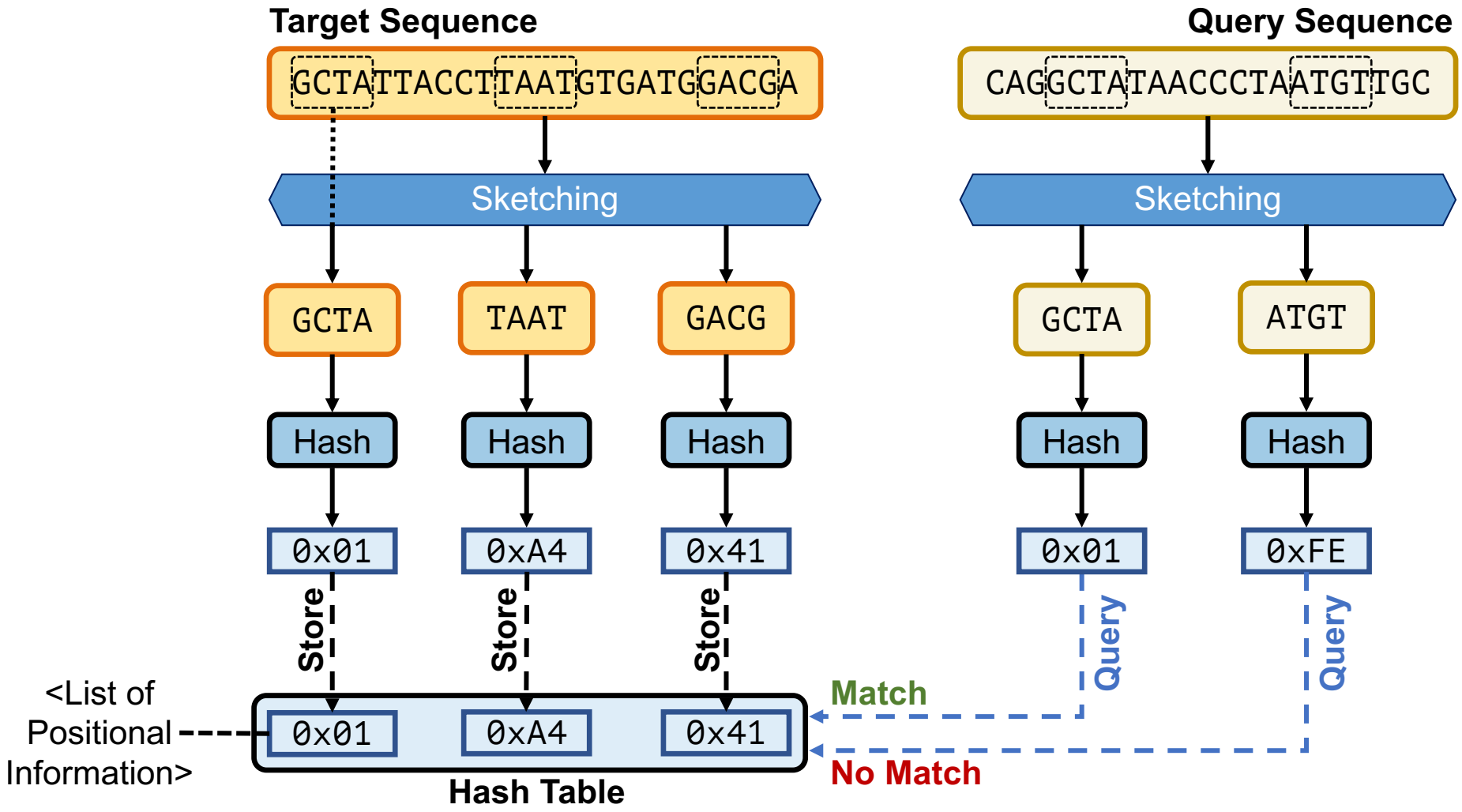
Spaced Seeding



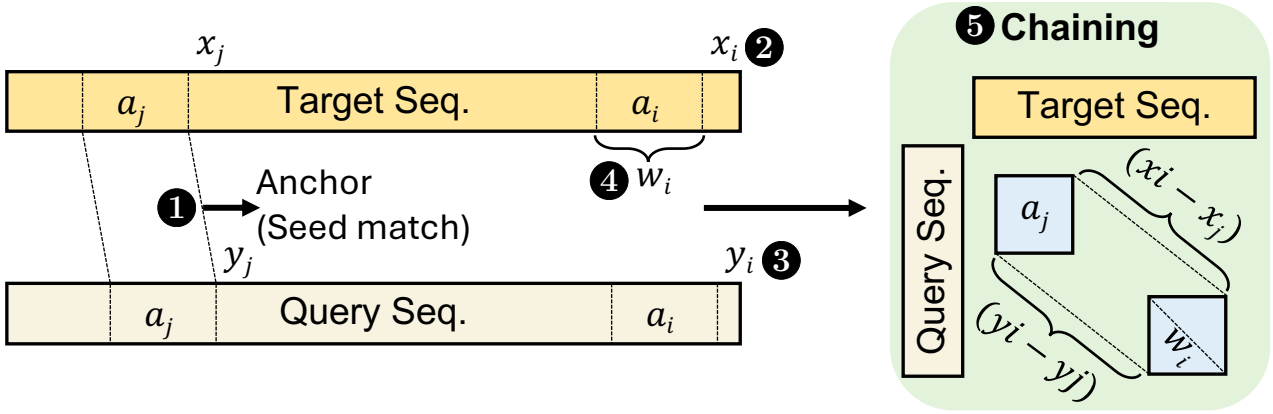
Strobemer Sketches



Hash-Based Sketching and Seed Matching



Chaining (Two Points)

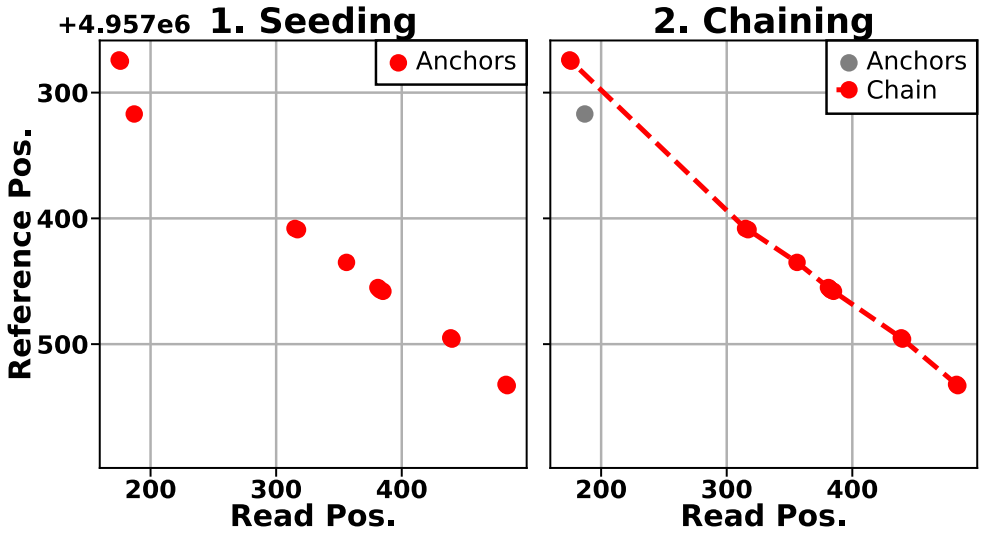


Chaining (Multiple Points)

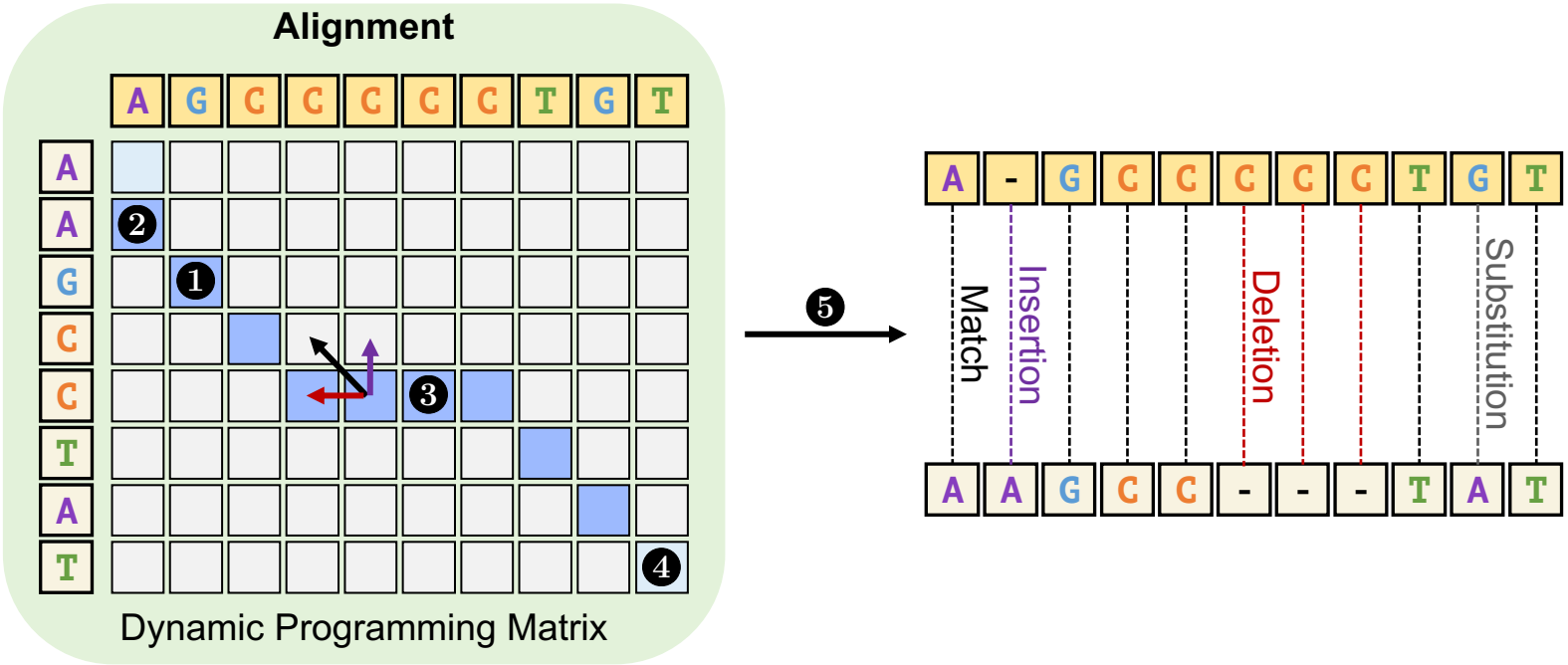
- **Exact hash value matches:** Needed for finding matching regions between a reference genome and a read

- What if there are mutations or errors?
  - **No hash (seed) match** will occur in such positions

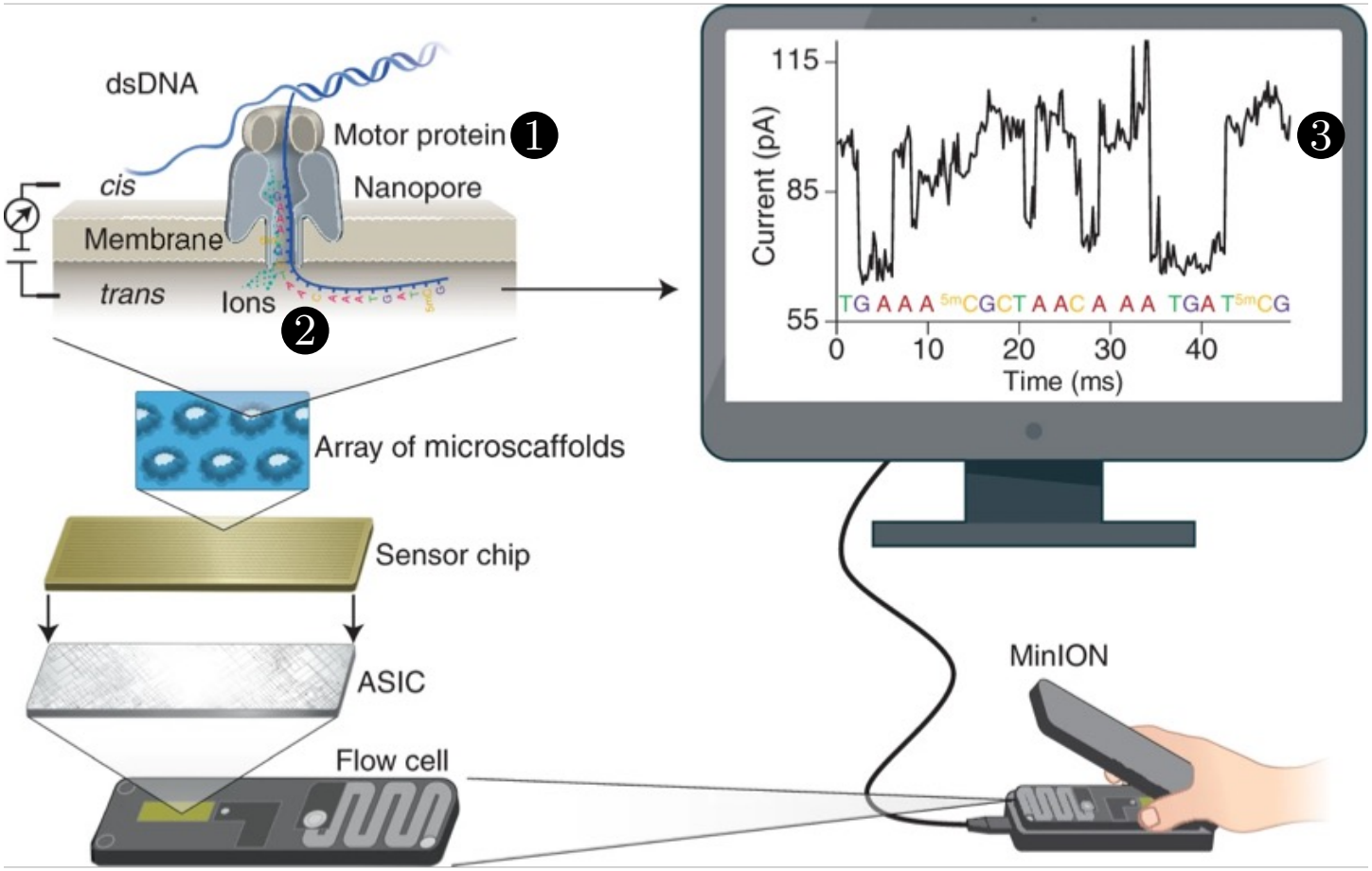
- The chaining algorithm links **exact matches in a proximity** even though there are gaps (no seed matches) between them



Sequence Alignment



Nanopore Sequencing

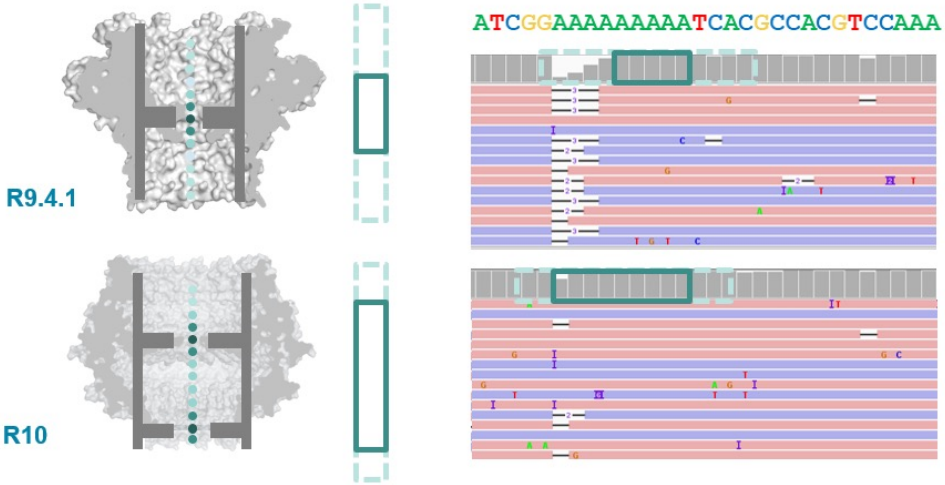


Source of Noise in Nanopore Sequencing

- **Stochastic thermal fluctuations in the ionic current**
  - Random ionic movement due to inherent thermal energy (Brownian motion)
- **Variations in the translocation speed**
  - Mainly due to the motor protein
- **Environmental factors**
  - **Temperature:** Affecting enzymes including the motor protein
  - **pH levels:** Affecting charge and the shape of molecules
- **Maybe: Aging & material-related noise between nanopores**
  - Their effects potentially can be minimized with normalization techniques

R9 vs. R10 Chemistries

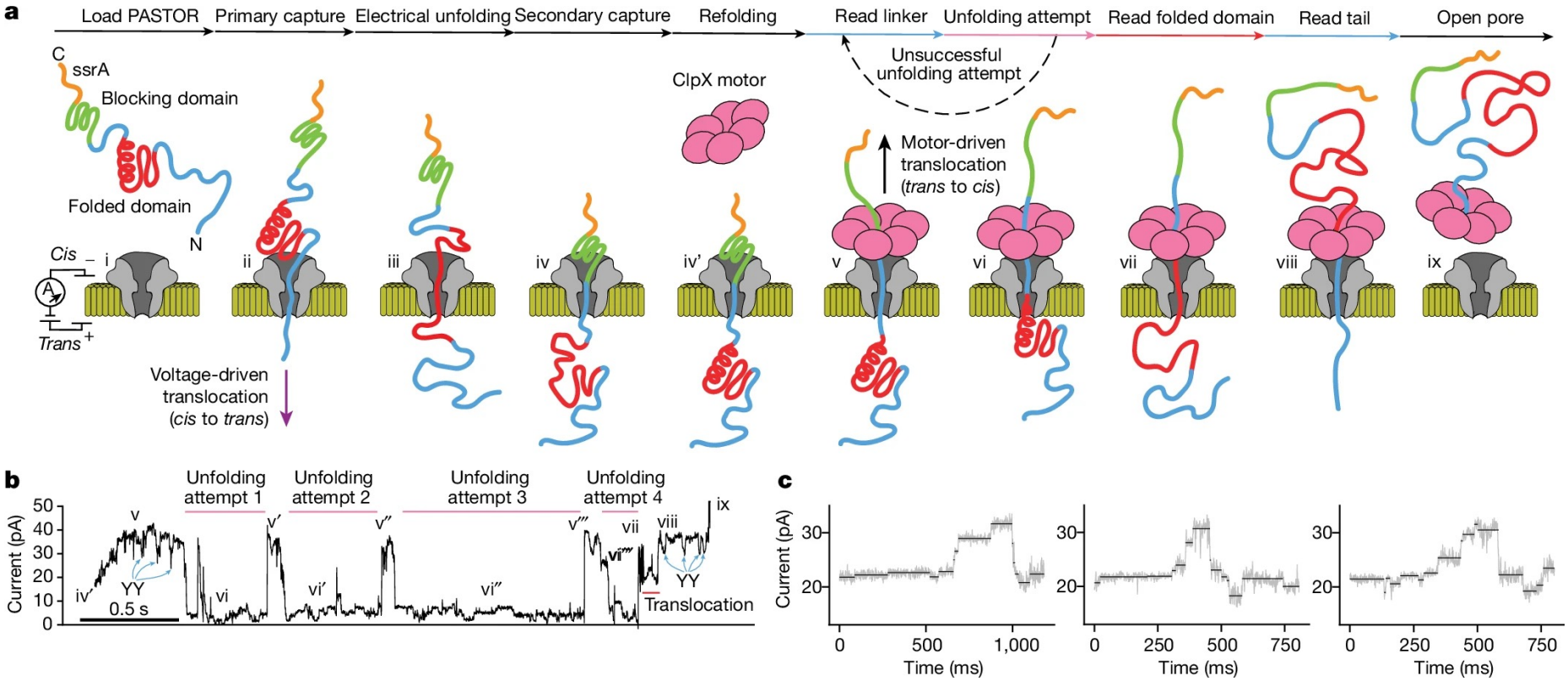
- **Dual reader head**



- **Motor protein** with more consistent translocation speed in R10

- **Duplex sequencing** in R10

Proteomics with Nanopores



Motone+, "Multi-pass, single-molecule nanopore reading of long protein strands", Nature 2024.

Applications of Read Until

Depletion: Reads mapping to a particular reference genome is ejected

- Microbiome studies by removing host DNA
- Eliminating known residual DNA or RNA (e.g., mitochondrial DNA)
- High abundance genome removal

Enrichment: Reads **not** mapping to a particular reference genome is ejected

- Removing contaminated organisms
- Targeted sequencing (e.g., to a particular region of interest in the genome)
- Low abundance genome enrichment

Applications of Run Until & Sequence Until

Run Until: Stopping the entire sequencing run

- Stopping when reads reach to a particular depth of coverage
- Stopping when the abundance of all genomes reach a particular threshold

Sequence Until: Run Until with accuracy-aware decision making

- Stopping when relative abundance estimations do not change substantially (for high-abundance genomes)
- Stopping when finding that the sample is contaminated with a particular set of genomes
- …

In Vitro (e.g., PCR) vs. In Silico

- **Polymerase Chain Reaction (PCR)** as a way of in vitro "analysis"
  - Can increase the quantity of DNA in a sample
  - **Non-dynamic** targeted sequencing (e.g., low abundance *known* targets)
  - **Requires additional resources:** Time and money for preparation and execution of PCR

- **Adaptive sampling** as a way of in silico (i.e., computational) analysis
  - **Cannot** increase the existing quantity of DNA in a sample
  - **Dynamic targeted sequencing:** Decisions can be made based on real-time analysis (e.g., Sequence Until)
  - Minimal additional resources
    - ***Almost* no additional resources** for preparation and execution
    - **Simultaneous** enrichment and depletion is possible
    - Better suited for rapid whole genome sequencing
  - *Beauty* of computational analysis (e.g., high flexibility – no need for primers)

- PCR and adaptive sampling can be combined depending on the analysis type

Finding Mapping Positions

- Useful for **any application** that requires exact genomic position
  - Variant calling in downstream analysis
  - Specifically: Identifying rare variants in cancer genomics
  - Methylation profiling

- Accurate and flexible **depth of coverage estimation**
  - **Alternative: DNA quantification** (without computational analysis)
    - DNA quantification is challenging for metagenomics analysis
  - **Computational method:** We can map to almost entire set of known reference genomes to accurately estimate the coverage of a metagenomics sample

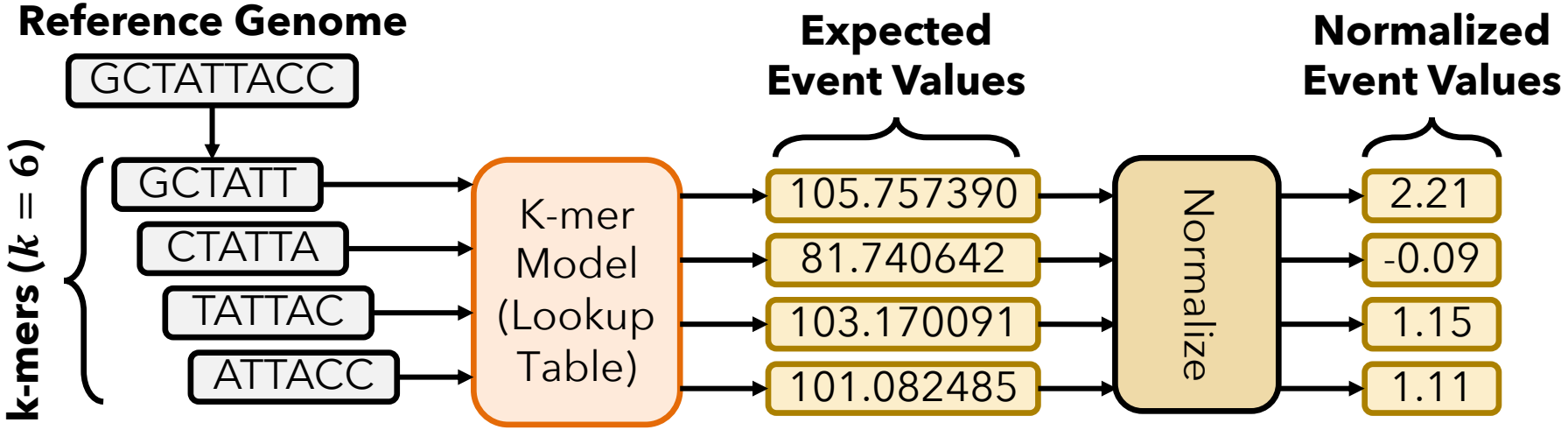
- **Transcriptome analysis**
  - Accurately quantifying expression levels & alternative splicing

- **Better resolution** (i.e., more sensitive analysis) for any other application that does not specifically require mapping positions

Reference-to-Event Conversion

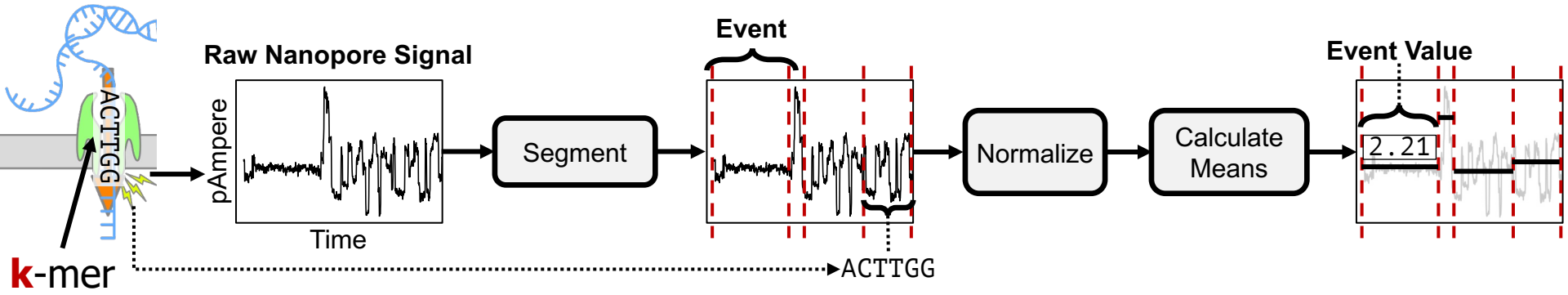
- **K-mer model:** Provides **expected** event values **for each k-mer**
  - Preconstructed based on nanopore sequencer characteristics

- Use the **k-mer model** to convert **all k-mers** of a reference genome to their **expected** event values



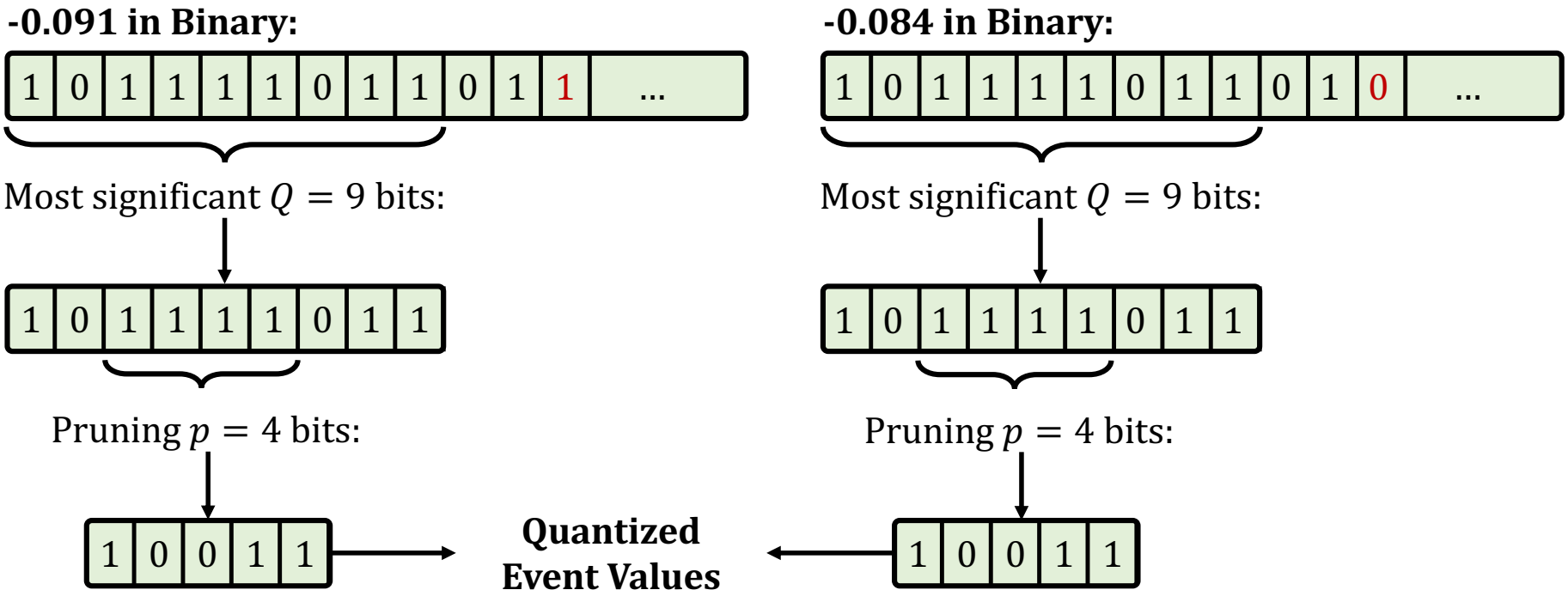
Enabling Analysis From Electrical Signals

- K many nucleotides (k-mers) sequenced at a time
- **Event:** A **segment** of the raw signal
  - Corresponds to a **particular** k-mer

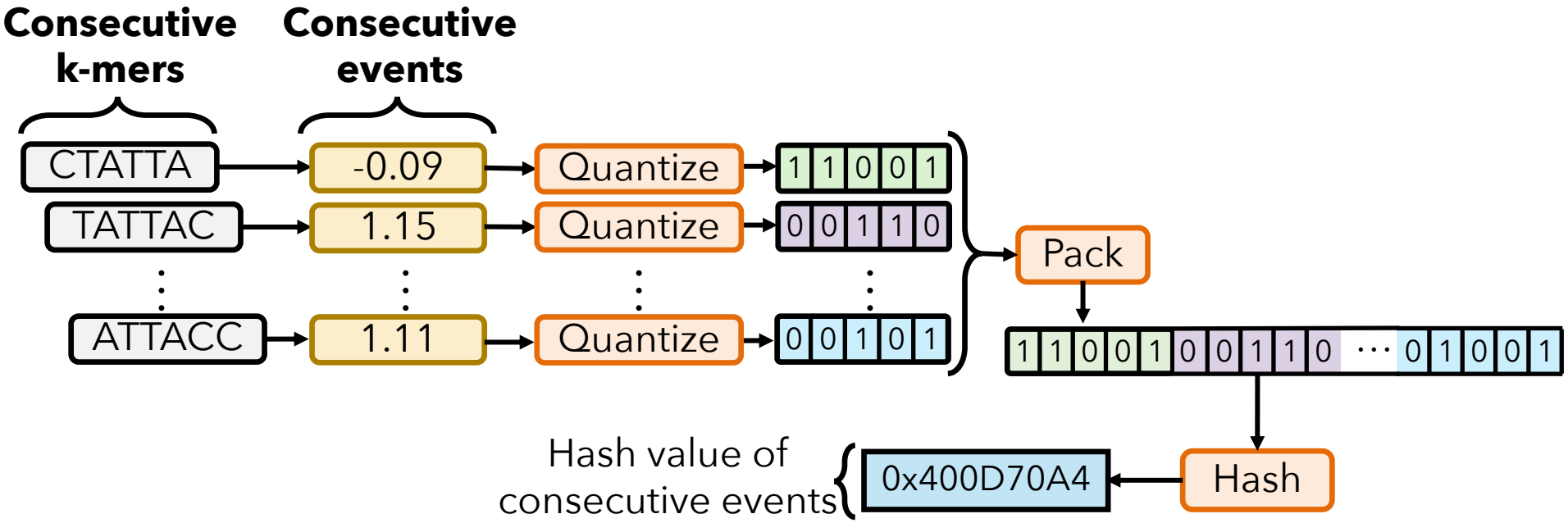


- **Observation:** Event values generated after sequencing **the same k-mer** are **similar** in value (not necessarily the same)

Quantization -- RawHash



Packing and Hashing



Sequence Until – RawHash & UNCALLED

Tool	Estimated Relative Abundance Ratios					
	SARS-CoV-2	***E. coli***	***Yeast***	***Green Algae***	***Human***	**Distance**
Ground Truth	0.0929	0.4365	0.0698	0.1179	0.2828	N/A
UNCALLED (25%)	0.0026	0.5890	0.0613	0.1332	0.2139	0.1910
RawHash (25%)	0.0271	0.4853	0.0920	0.0786	0.3170	**0.0995**
UNCALLED (10%)	0.0026	0.5906	0.0611	0.1316	0.2141	0.1920
RawHash (10%)	0.0273	0.4869	0.0963	0.0772	0.3124	**0.1004**
UNCALLED (1%)	0.0026	0.5750	0.0616	0.1506	0.2103	0.1836
RawHash (1%)	0.0259	0.4783	0.0987	0.0882	0.3088	**0.0928**
UNCALLED (0.1%)	0.0040	0.4565	0.0380	0.1910	0.3105	0.1242
RawHash (0.1%)	0.0212	0.5045	0.1120	0.0810	0.2814	**0.1136**
UNCALLED (0.01%)	0.0000	0.5551	0.0000	0.0000	0.4449	0.2602
RawHash (0.01%)	0.0906	0.6122	0.0000	0.0000	0.2972	**0.2232**

Sequence Until – RawHash

	Estimated Relative Abundance Ratios in 50,000 Random Reads					
Tool	***SARS-CoV-2***	***E. coli***	***Yeast***	***Green Algae***	***Human***	**Distance**
RawHash (100%)	0.0270	0.3636	0.3062	0.1951	0.1081	N/A
RawHash + *Sequence Until* (7%)	0.0283	0.3539	0.3100	0.1946	0.1133	0.0118

Presets

Preset (-x)	**Corresponding parameters**	Usage
viral	-e 5 -q 9 -l 3	Viral genomes
sensitive	-e 6 -q 9 -l 3	Small genomes (i.e., < 50M bases)
fast	-e 7 -q 9 -l 3	Large genomes (i.e., > 50M bases)

Versions – RawHash

Tool	Version
RawHash	0.9
UNCALLED	2.2
Sigmap	0.1
Minimap2	2.24