EDAM: EDIT DISTANCE TOLERANT APPROXIMATE MATCHING CONTENT ADDRESSABLE MEMORY

Robert Hanhan¹, Esteban Garzón²,³, Zuher Jahshan¹, Adam Teman², Marco Lanuzza³, and Leonid Yavits²
WHY SEQUENCING AND ANALYSIS OF VIRAL SAMPLES ARE CRITICAL

- Covid pandemic exposed the scientific infrastructures to an unpresented stress test
- A conventional (PCR, antigen) testing does not allow precise tracking of virus mutations
  - With PCR testing, SARS-CoV-2 mutations such as Delta or Omicron would go unnoticed
- Genome sequencing and analysis (detection, classification, phylogenetic analysis) are the ultimate pandemic control tools
WHY ACCELERATION OF VIRAL GENOME ANALYSIS IS CRITICAL

- To understand the pandemic, each new SARS-CoV-2 genome needs to be analyzed in a pangenomic context.

- Problem: such an analysis in a 100k-genome context takes more than 1 day on a high-performance server\(^1\)
  - To date, there are >11 million SARS-CoV-2 genomes sequenced\(^2\)

- Imagine the following scenario:
  - It’s 2030; the world is plagued by a viral pandemic.
  - 100M people are tested worldwide, 10M test positive, 2M samples are sequenced daily.

---

\(^1\) UShER, Turakhia et al., Nature Gen, 2021
\(^2\) GISAID
PROBLEM 1: CLASSIFICATION

- **Step 1 – Take a sample**
  (human patient or environmental)

- **Step 2 – sequence the sample**

- **Step 3 – classify the pathogens using**
  - Exact matching software tools like Kraken and Clark
  - EDAM targets the classification step
WHY EXACT MATCHING ISN’T WORKING

- Exact matching fails in the presence of
  - Sequencing errors (the problem is alleviated by high coverage)
  - Mutations

- As a result, detection by exact matching may miss a new lineage
  - If DNA fragments carrying the mutations are discarded
WHY TOLERATING A SMALL HAMMING DISTANCE ISN’T WORKING EITHER

Sequence 1  A T A T A T A T A T A T   HD=0
Sequence 2  A T A T A T A T A T A T   ED=0

Sequence 1  A T A T A T A T A T A T   HD=12
Sequence 2  C A T A T A T A T A T A   ED=1
EDAM – THE GENERAL IDEA

A Content Addressable Memory that tolerates edit distance to overcome the sequencing errors

EDAM evaluates the co-located basepair and its left and right neighbors:

Query Data:  
Stored Data:

Match (T):

Mismatch (T):
EDAM ARCHITECTURE

EDAM as a Virus Detection and Classification Platform
For $F_{op} = 667 \text{Mhz}$ and $k = 64$, EDAM classification throughput is $2561 \frac{\text{Giga Bp}}{\text{Min}}$, achieving the speedup of $1,214 \times$ over Kraken2* on Intel core I9-10900X at 3.70GHz with 32GB of memory.

* Sensitivity $= \frac{TP}{TP + FN}$
  
  * Precision $= \frac{TP}{TP + FP}$
EDAM: EDIT DISTANCE TOLERANT APPROXIMATE MATCHING CONTENT ADDRESSABLE MEMORY

Robert Hanhan\textsuperscript{1}, Esteban Garzón\textsuperscript{2,3}, Zuher Jahshan\textsuperscript{1}, Adam Teman\textsuperscript{2}, Marco Lanuzza\textsuperscript{3}, and Leonid Yavits\textsuperscript{2}