Your Presentation Title

Subtitle or Conference Name

Author Name
Institution or Affiliation
February 2026

Outline

1. Background & Motivation
2. Problem Formulation
3. Methodology
4. Results
5. Conclusion

Background

- Way to index massive bacterial datasets
- Each genome is a color. Each k-mer is associated with a set of colors.
- The color set of a k-mer is the set of color associated with it.
- Interesting object: the set of distinct color sets
- Key to compressing the data structure

Color matrix

- Color matrix: rows are k-mers, column are colors
- So we want to build to distinct rows
- It's easy to build the matrix column by column.
- But the query is row by row.

Color matrix

	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8
`ACGTA`	1	0	1	1	0	0	1	0
`CGTAC`	1	1	0	0	1	0	0	1
`GTACG`	0	1	1	0	0	1	0	0
`TACGT`	1	0	1	1	0	0	1	0
`ACGTG`	0	0	1	0	1	1	0	1
`TGCAA`	1	1	0	1	0	0	0	0
`GCAAC`	0	1	1	0	0	1	0	0
`CAACT`	0	0	0	1	1	0	1	1
`TTGCA`	1	0	1	1	0	0	1	0
`AACGT`	0	1	1	0	1	0	1	0
`CGTAT`	0	0	1	0	1	1	0	1
`GTATC`	0	0	1	0	0	1	1	0
`TATCG`	1	1	0	0	1	0	0	0
`ATCGA`	0	1	1	1	0	0	1	1
`CGAAC`	0	0	1	0	1	1	0	1

Color matrix

	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8
ACGTA	1	0	1	1	0	0	1	0
CGTAC	1	1	0	0	1	0	0	1
GTACG	0	1	1	0	0	1	0	0
TACGT	1	0	1	1	0	0	1	0
ACGTG	0	0	1	0	1	1	0	1
TGCAA	1	1	0	1	0	0	0	0
GCAAC	0	1	1	0	0	1	0	0
CAACT	0	0	0	1	1	0	1	1
TTGCA	1	0	1	1	0	0	1	0
AACGT	0	1	1	0	1	0	1	0
CGTAT	0	0	1	0	1	1	0	1
GTATC	0	0	1	0	0	1	1	0
TATCG	1	1	0	0	1	0	0	0
ATCGA	0	1	1	1	0	0	1	1
CGAAC	0	0	1	0	1	1	0	1

	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	**Fingerprint**
ACGTA	1	0	1	1	0	0	1	0	**1011010011**
CGTAC	1	1	0	0	1	0	0	1	**0110100101**
GTACG	0	1	1	0	0	1	0	0	**1100011110**
TACGT	1	0	1	1	0	0	1	0	**1011010011**
ACGTG	0	0	1	0	1	1	0	1	**0101110100**
TGCAA	1	1	0	1	0	0	0	0	**1001001010**
GCAAC	0	1	1	0	0	1	0	0	**1100011110**
CAACT	0	0	0	1	1	0	1	1	**0010111001**
TTGCA	1	0	1	1	0	0	1	0	**1011010011**
AACGT	0	1	1	0	1	0	1	0	**1110001100**
CGTAT	0	0	1	0	1	1	0	1	**0101110100**
GTATC	0	0	1	0	0	1	1	0	**0011000111**
TATCG	1	1	0	0	1	0	0	0	**1000110010**
ATCGA	0	1	1	1	0	0	1	1	**0100101011**
CGAAC	0	0	1	0	1	1	0	1	**0101110100**

k-mer	Fingerprint
ACGTA	**1011010011**
CGTAC	**0110100101**
GTACG	**1100011110**
TACGT	**1011010011**
ACGTG	**0101110100**
TGCAA	**1001001010**
GCAAC	**1100011110**
CAACT	**0010111001**
TTGCA	**1011010011**
AACGT	**1110001100**
CGTAT	**0101110100**
GTATC	**0011000111**
TATCG	**1000110010**
ATCGA	**0100101011**
CGAAC	**0101110100**

Color-set covering subset of k-mers

k-mer	Fingerprint
CAACT	**0010111001**
GTATC	**0011000111**
ATCGA	**0100101011**
ACGTG	**0101110100**
CGTAC	**0110100101**
TATCG	**1000110010**
TGCAA	**1001001010**
ACGTA	**1011010011**
GTACG	**1100011110**
AACGT	**1110001100**

Requirements for the fingerprint function F

- F takes in a fingerprint and a color, and adds the color to the fingerprint.
- Given set $\{c_1, c_2, c_3\}$ the fingerprint is $F(F(F(0, c_1), c_2), c_3)$
- Commutative: $F(F(F(0, c_1), c_2), c_3) = F(F(F(0, c_3), c_2), c_1)$
- Atomically updateable: $x \leftarrow F(x, c)$ is an atomic CPU operation
- Collision-resistant

Fingerprinting scheme

- **Initialization**: For each color, pick an l-bit fingerprint uniformly at random. Denote the fingerprint of color c with $f(c)$.
- **Fingerprinting**: The fingerprint of a *set* $A = \{c_1, c_2, ..., c_m\}$ is $F(A) = c_1 \oplus c_2 \oplus ... \oplus c_m$, where $\oplus$ is bitwise xor.
- **Wishlist**: Incremental ✓, Commutative ✓, Atomically updatable ✓, Collision-resistant: ?

Collision analysis

The fingerprint function F is **universal hash family** over the single-color fingerprint picks $f(c)$. By the union bound:

Lemma 2. *Given a set of distinct sets $A_0, ..., A_{N-1}$, the probability that there exists two sets $A_i \neq A_j$ such that $F(A_i) = F\left(A_j\right)$ is at most $\frac{N^2}{2^{\ell+1}}$, where l is the length of a fingerprint.*

For example, for $\ell = 128$ and $N = 10^9$, we have a collision probability of at most 10^{18} / $2^{129} \approx 1.47 \cdot 10^{-21}$.