The k-mismatch problem revisited

R. Clifford[1], A. Fontaine[1], E. Porat[2], B.Sach[1] and T. Starikovskaya[1]
[1]University of Bristol, Department of Computer Science, Bristol, U.K. (@bristolcstheory)
[2]Bar-Ilan University, Department of Computer Science, Israel
allyx.fontaine@bristol.ac.uk

The k-mismatch problem

The Hamming distance

The Hamming distance between two strings of the same length is the number of mismatches between them.

Let $Ham(P, T)[i]$ be the Hamming distance between a m-length pattern P and a subtext $T[i-m, i]$.

The k-mismatch problem

Input: $\begin{cases} \text{A pattern } P \text{ of length } m \\ \text{A streaming text } T \text{ of length } n \end{cases}$

Output: For all positions $m \leq i < n$,
$\begin{cases} Ham(P, T[i]) & \text{if } Ham(P, T[i]) \leq k \\ \text{No} & \text{otherwise} \end{cases}$

3-*mismatch example*

T b b a c c b c a b a c ...
P a b c a b
a b c a b
a b c a b
a b c a b
a b c a b
a b c a b

x: 3-mismatch position - x: No 3-mismatch - x: Mismatch

Faster solutions using less space

Space complexity

Problem	Previous	Ours
Deterministic online	$\widetilde{\Theta}(m)$	–
Randomised online	–	$\widetilde{O}(k^2)$
$(1+\varepsilon)$-approximation	$\widetilde{O}(m/\varepsilon^2)$	$\widetilde{O}(k^2/\varepsilon^2)$
Randomised online worst case	$\widetilde{O}(k^3)$	$\widetilde{O}(k^2)$

Time complexity

Problems	Previous	Ours
Deterministic offline	$\widetilde{O}(nk^3/m)$	$\widetilde{O}(nk^2/m)$
Randomised online	–	$\widetilde{O}(nk^2/m)$
$(1+\varepsilon)$-approximation	–	$\widetilde{O}(1/\varepsilon^2)$
Randomised online worst case	$\widetilde{O}(k^2)$	$\widetilde{O}(\sqrt{k})$

Algorithm

Approximate period

The $3k/2$-period of P is the smallest integer ℓ, such that $\text{HAM}(P[\ell, m-1], P[0, m-1-\ell]) \leq 3k/2$.

3k/2-period example

Let $P = abcabaacabcaccacca$ and $k = 4$.
The $3k/2$-period of P is $\ell = 3$.

1 shift
a b c a b a a c a b c a c c a c c a
a b c a b a a c a b c a c c a c c

2 shifts
a b c a b a a c a b c a c c a c c a
a b c a b a a c a b c a c c a c

3 shifts
a b c a b a a c a b c a c c a c c a
a b c a b a a c a b c a c c a

Case 1: Small approximate period ($\ell \leq k$)

1. Identify a compressible region of the text which contains all the k-mismatches.
2. Partition this region into $O(k)$ subtexts and the pattern into $O(k)$ subpatterns.
3. Run length encode all the subpatterns and subtexts.
4. Compute run length encoded Hamming distances for each subpattern/subtext pair.
5. Sum the Hamming distances from Step 4.

Run-length encoding example

Let $P = abcabaacabcaccacca$.

Partition	Encoding
$P^0 = aaabcc$	$(a,3)(b,1)(c,2)$
$P^1 = bbcccc$	$(b,2)(c,4)$
$P^2 = caaaaa$	$(c,1)(a,5)$

All required information from P and T encoded in only $O(k)$ space!

Case 2: Large approximate period ($\ell > k$)

1. Filter out all alignments of the pattern and text with Hamming distance greater than $3k/2$.
2. Verify whether the Hamming distance is at most k at those positions.

University of BRISTOL
EPSRC Engineering and Physical Sciences Research Council