

Exercise 4

Problem 10

Idea: full enumeration of $\alpha' \in \Sigma^w$

Inputs: • alphabet Σ of size c [constant]

• searched sequence $P \in \Sigma^m$

• seed length $w \in \mathbb{N}$ [constant]

• alignment scoring $\delta: \Sigma^2 \rightarrow \mathbb{Q}$ [here: goal $\delta = \min$]

• seed similarity threshold $s \in \mathbb{Q}$

The output: $A = \{ \alpha' \mid \delta(\alpha', \alpha) \leq s \wedge |\alpha| = |\alpha'| = w$
 $\wedge \alpha \text{ is a substring of } P \}$

for each $\alpha' \in \Sigma^w$.

for $j = 1, \dots, m - w + 1$:

• Compute $t := \delta(P_{j, j-w+1}, \alpha')$

• if $t \leq s$ then $A = A \cup \{ \alpha' \}$; break;

Correctness:

• It terminates, since Σ^w is finite

• we consider all pairs of (α', α) where α is a substring of length w of P

• we add α' only if it suits a threshold s

Runtime:

$$\leq c^w \cdot (m-w+1) \cdot (w+d) \stackrel{c^w \text{ constant}}{=} O(m)$$

↑ ↑ ↑
strings # positions score computation
in Σ^w in \mathcal{P} + 'rest'

Justification for DNA/RNA:

$$|\Sigma| = c = 4$$

$$m \approx 10^6$$

$$w \approx 10$$

Problem 11

Relevant aspects:

- which datasets?
- which algorithm/parameters?
- scores (for comparison/justification)

⇒ robust results/claim

Seq 1: Cryptosporidium

Seq 2: Plasmodium falciparum

Seq 3: Tubulin delta protein (Plasmodium)

Seq 4: DNA fragmentation factor, beta polypeptide

Problem 12

symbolic method

$$E = Z_0 + Z_{\uparrow} * W * Z_{\downarrow}$$

go up

go down

excursion
of length 0

$$W = \epsilon + Z_{\uparrow} * W * Z_{\downarrow} * W$$

empty

$$\leadsto F(z) = (1-p)z^0 + p(1-p)z^2 W(z)$$

$$W(z) = 1 + p(1-p)z^2 W(z)^2$$

$$\leadsto F(z) = (1-p) + \frac{1 - \sqrt{1 - 4p(1-p)z^2}}{2}$$

$$\leadsto \mathbb{E}[L] = F'(1) = p \left(1 + \frac{1}{1-2p} \right) \quad \left[p < \frac{1}{2} \right]$$