Advanced Algorithms in Bioinformatics (P4)

Sequence and Structure Analysis

Freie Universität Berlin, Institut für Informatik
David Weese, Sandro Andreotti

Sommersemester 2011

1. Review, 2011-05-11

Name, Vorname	Matrikelnummer

Zur Bearbeitung des Reviews stehen Ihnen 50 Minuten zur Verfügung. Jeder Punkt entspricht in etwa einer Minute.

Geben Sie auf diesem Titelblatt und auf allen eventuell zusätzlich abgegebenen Blättern ihren Namen und ihre Immatrikulationsnummer an.

Schreiben Sie ihre Lösungen direkt auf die entsprechenden Aufgabenbögen. Sollte dort der Platz nicht ausreichen und Sie weitere Blätter benötigen, vermerken Sie dies bitte, damit wir auch den Rest ihrer Antwort finden und bei der Bewertung berücksichtigen können. Am Ende des Reviews sind sämtliche Aufgabenblätter wieder abzugeben.

Ergebnis:

Aufgabe	maximal	erreicht
1	12	
2	15	
3	10	
4	13	
$\sum$	50	

12/12

Exercise 1. 6 + 2 + 4 = 12 Pts

1. For the pattern **AGATA** and text **AGATACGATATATAC** apply the Horspool algorithm and explain the single steps.
2. What is the worst case runtime (number of comparisons) when searching a pattern of length m in a text of length n?
3. Give an example of a text T of length ≥ 20 and pattern P of length 5 where the number of character comparisons equals the worst case and P does not occur in T.

1) I. Preprocessing

A G A T A
4 3 2 1

$m = 5$
$n = 15$

shift table:

	G	T	A	*
	3	1	2	5

✓

II. Searching

pos = 0

A G A T A C G A T A T A T A C
A G A T A

j = 0 => pattern occurs at position 1

A G A T A

d [A] = 2 => shift 2

G A T A

G ≠ A, d [G] = 3

G A T A

C ≠ A, d [A] = 2

T ≠ G, d [A] = 2

T ≠ G, d [A] = 2

pos is not ≤ n − m => algorithm ends

✓ 6/6

→ output: pattern occurs at position 1

2) $O(m \cdot n)$ ✓ 2/2

3) T = TTTTTTTTTTTTTTTTTTTT

P = ATTTT ✓ 4/4

15/15

Exercise 2. 6 + 6 + 3 = 15 Pts

The Myers Bitvector algorithm uses binary encoding of the dynamic programming matrix.

1. Use the bitvectors to fill out the dynamic programming matrix

$$\begin{aligned} VN_1 &= 000000 \\ VP_1 &= 111110 \\ D0_2 &= 111110 \\ HN_3 &= 111100 \\ HP_3 &= 000010 \end{aligned}$$

		t_1	t_2	t_3
	0	0	0	0
p_1	1	0	1	1
p_2	2	1	0	1
p_3	3	2	1	0
p_4	4	3	2	1
p_5	5	4	3	2
p_6	6	5	4	3

✓ 6/6

2. Below you find the pseudocode of the Myers Bitvector algorithm where the variables are renamed to $\alpha, \beta, \gamma, \delta, \theta$. Map these identifiers back to the original names $D0, HN, HP, VN, VP$.

3. How can you modify the algorithm to compute edit distance (global alignment) instead of the semi-global alignment?

// Preprocessing

for $c \in \Sigma$ do $B[c] = 0^m$ od

for $j \in 1 \ldots m$ do $B[p_j] = B[p_j] \mid 0^{m-j}10^{j-1}$ od

$\delta = 1^m$; $\gamma = 0^m$;

$score = m$;

// Searching

for $pos \in 1 \ldots n$ do

  $X = B[t_{pos}] \mid \gamma$;

  $\alpha = ((\delta + (X \;\&\; \delta)) \wedge \delta) \mid X$;

  $\theta = \delta \;\&\; \alpha$;

  $\beta = \gamma \mid \sim (\delta \mid \alpha)$;

  $X = \beta \ll 1$;

  $\gamma = X \;\&\; \alpha$;

  $\delta = (\theta \ll 1) \mid \sim (X \mid \alpha)$;

  // Scoring and output

  if $\beta \;\&\; 10^{m-1} \neq 0^m$

   then $score$ += 1;

   else if $\theta \;\&\; 10^{m-1} \neq 0^m$

    then $score$ −= 1;

   fi

  fi

  if $score \leq k$ report occurrence at pos fi;

od

---

2) α = D0 ✓

β = HP ✓

γ = VN ✓

6/6

δ = ~~[illegible]~~ VP ✓

θ = HN ✓

3) ~~mod~~ adding $| \; 0^{m-1} 1$ to the second definition of X

$\Rightarrow X = HP \ll 1 \mid 0^{m-1}1$

✓ 3/3

5/10

Exercise 3. 6 + 4 = 10 Pts

1. State the q-gram Lemma for contiguous shapes and prove it.
2. Is the threshold we compute by that Lemma tight (if ≥ 0)? Justify your answer.

1) q-gram: a short subsequence of length q

Let P and S be two strings of length w with at most K differences

Then the number of possible q-grams amounts to $w + 1 - q$ and the number of q-grams that can be ~~at~~ destroyed at most amounts to kq

why?

5/6

threshold:

$t = w + 1 - q - qk = w + 1 - q(k+1)$

P and S share at least t common q grams

2) Yes.

why?

1/4

5/13

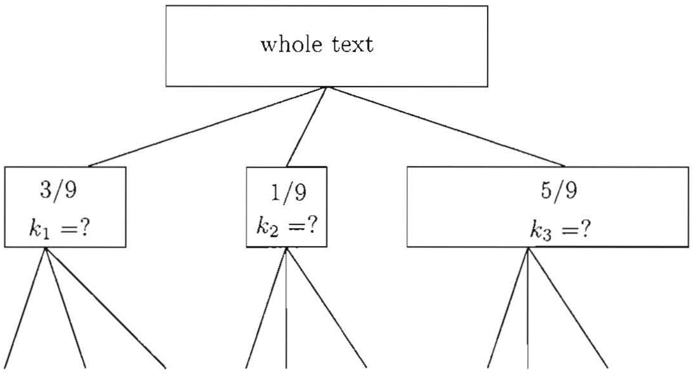
Exercise 4. 6 + 4 + 3 = 13 Pts

1. Prove the pigeonhole Lemma:

   Lemma 1. *Let Occ match P with k errors, $P = p^1, \ldots, p^j$ be a concatenation of subpatterns, and $a_1, \ldots, a_j$ be nonnegative integers such that $A = \sum_{i=1}^{j} a_i$. Then, for some $i \in 1, \ldots, j$, Occ includes a substring that matches p^i with $\lfloor a_i k / A \rfloor$ errors.*

2. Anne, Paul und Peter want to use hierarchical filtering. They are asked for error bounds (k_1, k_2, k_3) for hierarchical verification. The pattern shall be searched with 5 errors and is split in three parts.

   Anne suggests $(1, 0, 2)$, Paul $(1, 0, 1)$ and Peter $(1, 1, 2)$.



   (a) Are all three error bounds valid? Justify your answer.

   (b) Who suggested the best bounds? Justify your answer.

1) Let k_i be the number of errors in p^i

then $k = \sum_{i=1}^{j} k_i$

according to the Lemma $\exists i : \left\lfloor \frac{a_i k}{A} \right\rfloor \geq k_i$

Proof by contradiction

assume $\forall i \left\lfloor \frac{a_i k}{A} \right\rfloor < k_i$

Then applying chaining rules:

$k_i \geq \left\lfloor \frac{a_i k}{A} \right\rfloor + 1 > \frac{a_i k}{A}$

now it is easy to derive the contradiction

$k = \sum_{i=1}^{j} k_i \geq \sum_{i=1}^{j} \left(\left\lfloor \frac{a_i k}{A} \right\rfloor + 1 \right) > \sum_{i=1}^{j} \frac{a_i k}{A} = k$

5/6

a, b

$\Rightarrow k > k$ ↯