

MBI, 25.06.2025, czas: 60 min

Imię i nazwisko:

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Proszę wpisywać odpowiedzi w ramkach

Zadanie 1 (3 pkt)

Badamy podobieństwo lokalne (algorytm Smitha - Watermana) dla sekwencji ABABA i BBBA. Stosujemy liniową karę za przerwę $d = -n$. Macierz podobieństwa pokazano obok, n jest liczbą liter Twojego nazwiska.

	A	B
A	$2 * n$	$-2 * n$
B	$-2 * n$	$2 * n$

Uzupełnij macierz podobieństwa, macierz pomocniczą. Podaj najlepsze dopasowanie, jeżeli jest kilka, podaj jedno z nich.

	A	B
A		
B		

$d =$

--

	A	B	A	B	A
B					
B					
B					
B					
A					

Rozwiązanie:

Zadanie 2 (3 pkt)

Sekwencje AAAAA, ABABAB, BBBB, AAAB oraz BABBB użyto do budowy multi-grafu de Bruijna, wagi w krawędziach pokazują, ile razy krawędź ma być użyta przy generowaniu rozwiązania. Zbuduj graf 4 rzędu (w wierzchołkach są sekwencje o długości 3 symboli).

Narysuj graf 4 rzędu.

Czy istnieje rozwiązanie?

--

Sekwencja wyjściowa obliczona z grafu 4 rzędu to:

Zadanie 3 (3 pkt)

Znajdź motyw o długości $l = 3$, w 4 sekwencjach przedstawionych poniżej.

$$\begin{aligned}s_0 &= A & A & T & T, \\s_1 &= T & A & A & T, \\s_2 &= T & A & T & A, \\s_3 &= A & A & A & T.\end{aligned}$$

Motyw:

Pozycje występowania motywu, indeksujemy od 0:

$$P(s_0) = \boxed{}, P(s_1) = \boxed{}, P(s_2) = \boxed{}, P(s_3) = \boxed{}.$$

Zadanie 4 (3 pkt)

Posługujemy się monetami A i B, obserwując sekwencje rzutów $RRRR$ (orły i reszki). Podaj najbardziej prawdopodobną sekwencję użytych monet, zakładając, że to doświadczenie może być opisane ukrytym modelem Markowa pokazanym niżej.

$$\begin{aligned}Q &= \{A, B\}, \\V &= \{O, R\}, \\P_A &= 0, P_B = 1.\end{aligned}$$

	A	B
A	$\frac{2}{3}$	$\frac{1}{3}$
B	$\frac{1}{5}$	$\frac{4}{5}$

	O	R
A	0	1
B	$\frac{1}{2}$	$\frac{1}{2}$

Uzupełnij macierz pomocniczą w algorytmie Viterbiego

	R	R	R	R
A				
B				

Podaj najbardziej prawdopodobną sekwencję użytych monet:

Pytania (2 pkt)

Jakie zagadnienie / zagadnienia warto byłoby dodać do MBI

Ile godzin poświęciłeś na MBI (sumarycznie wykłady, ćwiczenia, praca własna):

W jakim obszarze, Twoim zdaniem, przedstawione na wykładzie metody mogą przynieść największą korzyść.

Notatki lub uwagi do R. Nowaka:

Imię i nazwisko:

Zadanie 5 (3 pkt)

Poniżej podano fragment sekwencji referencyjnej oraz pliku SAM (kolumny POS i SEQ, odpowiadające pozycji oraz sekwencji) dla czterech kolejnych odczytów.

>chr1:101-124
accacatggtgttccgctttatg

101 accacatggtgttccgctt
102 ccacatggtgttccgcttt
103 cacatggtgttccgctttt
104 acatggtgttccgctttta

- Określ sekwencję aminokwasów, które są kodowane przez tą sekwencję zakładając, że na pozycji 1 zaczyna się nowy kodon.

- Dla pozycji 1-24 określ głębokość pokrycia odczytami. Na których pozycjach ta głębokość jest największa a na których najmniejsza?

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24

- Czy dane z pliku SAM pozwalają wnioskować o wystąpieniu wariantu(ów)? tak / nie ☐

Jeżeli tak:

- a) Określ typ wariantu(ów) (SNV / Indel ?).
- b) Jaki jest ich efekt na białko? (np. missense / stopgain / frameshift)?
- c) Jaka jest prawdopodobna zygotyczność wariantu (homozygotyczny czy heterozygotyczny)?

Tablica kodonów:

		Second base					
		U	C	A	G		
First base	U	UUU } Phenylalanine F UUC } UUA } Leucine L UUG }	UCU } Serine S UCC } UCA } UCG }	UAU } Tyrosine Y UAC } UAA } Stop codon UAG } Stop codon	UGU } Cysteine C UGC } UGA } Stop codon UGG } Tryptophan W	Third base	U
	C	CUU } Leucine L CUC } CUA } CUG }	CCU } Proline P CCC } CCA } CCG }	CAU } Histidine H CAC } CAA } Glutamine Q CAG }	CGU } Arginine R CGC } CGA } CGG }		C
	A	AUU } Isoleucine I AUC } AUA } AUG } Methionine start codon M	ACU } Threonine T ACC } ACA } ACG }	AAU } Asparagine N AAC } AAA } Lysine K AAG }	AGU } Serine S AGC } AGA } Arginine R AGG }		A
	G	GUU } Valine V GUC } GUA } GUG }	GCU } Alanine A GCC } GCA } GCG }	GAU } Aspartic acid D GAC } GAA } Glutamic acid E GAG }	GGU } Glycine G GGC } GGA } GGG }		G

Notatki lub uwagi do Tomasza Gambina:

MBI, 25.06.2025, czas: 60 min

Imię i nazwisko:

[illegible]

Zadanie 6 (3 pkt)

Zaproponuj ile przestrzeni dyskowej należy przygotować na nieskompresowany plik z odczytami DNA w formacie FASTQ.

- sekwencjonowanie Whole Genome Sequencing;
- długość genomu referencyjnego: 2 Gbp;
- pokrycie odczytami: 30×;
- długość odczytu 100 bp.

W szacowaniu pomiń znaki nowych linii, przyjmij że każdy identyfikator sekwencji to 50 B.

Notatki lub uwagi do W. Kuśmirka: