

Imię i nazwisko:

Proszę wpisywać odpowiedzi w ramkach

Zadanie 1 (3 pkt)

Badamy podobieństwo lokalne (algorytm Smitha-Watermana) dla sekwencji ABABB i ACBA. Stosujemy liniową karę za przerwę $d = -2 * n$. Macierz podobieństwa pokazano obok, n jest liczbą liter Twojego nazwiska.

	A	B	C
A	$2 * n$	$-n$	$-n$
B	$-n$	n	$n - 1$
C	$-n$	$n - 1$	n

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

	A	B	C
A			
B			
C			

d =

		A	B	A	B	B
A	<input style="width: 40px; height: 20px;" type="text"/>	<input style="width: 40px; height: 20px;" type="text"/>	<input style="width: 40px; height: 20px;" type="text"/>	<input style="width: 40px; height: 20px;" type="text"/>	<input style="width: 40px; height: 20px;" type="text"/>	<input style="width: 40px; height: 20px;" type="text"/>
A	<input style="width: 40px; height: 20px;" type="text"/>	<input style="width: 40px; height: 20px;" type="text"/>	<input style="width: 40px; height: 20px;" type="text"/>	<input style="width: 40px; height: 20px;" type="text"/>	<input style="width: 40px; height: 20px;" type="text"/>	<input style="width: 40px; height: 20px;" type="text"/>
C	<input style="width: 40px; height: 20px;" type="text"/>	<input style="width: 40px; height: 20px;" type="text"/>	<input style="width: 40px; height: 20px;" type="text"/>	<input style="width: 40px; height: 20px;" type="text"/>	<input style="width: 40px; height: 20px;" type="text"/>	<input style="width: 40px; height: 20px;" type="text"/>
B	<input style="width: 40px; height: 20px;" type="text"/>	<input style="width: 40px; height: 20px;" type="text"/>	<input style="width: 40px; height: 20px;" type="text"/>	<input style="width: 40px; height: 20px;" type="text"/>	<input style="width: 40px; height: 20px;" type="text"/>	<input style="width: 40px; height: 20px;" type="text"/>
A	<input style="width: 40px; height: 20px;" type="text"/>	<input style="width: 40px; height: 20px;" type="text"/>	<input style="width: 40px; height: 20px;" type="text"/>	<input style="width: 40px; height: 20px;" type="text"/>	<input style="width: 40px; height: 20px;" type="text"/>	<input style="width: 40px; height: 20px;" type="text"/>

Rozwiązanie:

Zadanie 2 (3 pkt)

Sekwencje CCCCCAAAAAB, BBBBCCC, AABCB oraz BCBCB użyto do budowy grafu de Bruijna, bez wag w krawędziach. Zbuduj graf 4 rzędu (w wierzchołkach są sekwencje o długości 3 symboli) oraz graf 3 rzędu (w wierzchołkach sekwencje o długości 2 symboli).

Narysuj graf 4 rzędu.

Sekwencja wyjściowa obliczona z grafu 4 rzędu to:

Narysuj graf 3 rzędu.

Sekwencja wyjściowa obliczona z grafu 3 rzędu to:

Zadanie 3 (3 pkt)

Zbuduj drzewo filogenetyczne metodą UPGMA (średnich połączeń) dla 5 taksonów. Macierz odległości podano obok.

	B	C	D	E
A	3	7	8	7
B		9	13	12
C			5	7
D				2

Zadanie 4 (3 pkt)

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

$$Q = \{A, B\},$$

$$V = \{O, R\},$$

$$P_A = 0, P_B = 1.$$

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

	O	R	R	R
A	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
B	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>

Podaj najbardziej prawdopodobną sekwencję użytych monet:

Pytanie (1 pkt)

Jakie zagadnienie / zagadnienia warto byłoby dodać do MBI

Pytanie (1 pkt)

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

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 odczytów.

sekwencja referencyjna:
>chrY:1-24
gcctatgatggacatgcatgtgca

plik SAM:
1~gcctatgat
5~atgatggacatgt
14 atgtatgtg
10 ggacatgta

- Określ sekwencję aminokwasów, które są kodowane przez tę 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)?

Jeżeli tak:

- Określ typ wariantu(ów) (SNV / Indel ?),
- Jaki jest ich efekt na białko,
- Ile jest odczytów wskazujących na wystąpienie wariantu?

- Dodaj trzy odczyty (podaj ich lokalizację i sekwencje), które zmieniają każdy wariant homozygotyczny zidentyfikowany w punkcie 3 na wariant heterozygotyczny.

Tablica kodonów:

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

Notatki lub uwagi do Tomasz Gambina:

MBI, 31.01.2025, czas: 60 min

Imię i nazwisko:

Zadanie 6 (3 pkt)

Zaproponuj ile przestrzeni dyskowej należy przygotować na skompresowany plik w formacie FASTQ. Parametry sekwencjonowania:

- sekwencjonowanie Whole Exome Sequencing
- okna sekwencjonowania pokrywają $\approx 5\%$ długości genomu
- długość genomu referencyjnego: 3 Gbp
- pokrycie okna sekwencjonowania odczytami: $50\times$.

W szacowaniu pomiń identyfikatory sekwencji, znaki nowych linii oraz linie oddzielające sekwencje DNA od symboli jakości. Przyjmij, że współczynnik kompresji to 4.

Notatki lub uwagi do W. Kuśmirka: