

# Metody bioinformatyki (MBI)

Wykład 15 - Badanie istotności wyników. Powtórzenie.

Robert Nowak

2025L

# Egzaminy

Egzaminy w terminach pokazanych w planie studiów.

- ▶ punktacja 0 – 20 pkt,
- ▶ 6 zadań,
- ▶ czas 'netto' 60 min.

# Badanie podobieństw dwu sekwencji

Algorytmy podstawowe:

- ▶ podobieństwo globalne, algorytm Needlemana-Wunscha
- ▶ podobieństwo lokalne, algorytm Smitha-Watermana

Inne:

- ▶ algorytmy o liniowym koszcie pamięciowym
- ▶ algorytmy z afiniczną funkcją kary za przerwę
- ▶ bez uwzględniania przerw, alg. Knutha-Morrisa-Prata
- ▶ algorytmy oparte o heurystyki (BLAST, FASTA)
- ▶ podobieństwo globalne, bez uwzględniania kar na końcach sekwencji

# Zagadnienia związane z podobieństwem sekwencji

- ▶ obliczanie macierzy podobieństwa BLOSUM
- ▶ obliczanie macierzy podobieństwa PAM

## Podobieństwa wielu sekwencji

- ▶ badanie podobieństwa sekwencji do profilu
- ▶ n-wymiarowy algorytm programowania dynamicznego
- ▶ metoda progresywna wykorzystująca drzewo naprowadzające

# Zagadnienia związane z sekwencjonowaniem

## Odczytywanie sekwencji genetycznych

- ▶ mapy restrykcyjne - problem częściowego strawienia
- ▶ składanie sekwencji z losowych fragmentów
  - ▶ grafy pokrycia,
  - ▶ de Bruijna
- ▶ basecalling, sekwencjonowanie urządzeniem nanopore (W. Kuśmirek)
- ▶ analiza genomu człowieka (T. Gambin)

# Analiza danych biologicznych

## Algorytmy związane z markerami genetycznymi

- ▶ badanie pokrewieństw w oparciu o obserwacje wariantów
- ▶ badanie mieszanin DNA

## Analiza danych wielowymiarowych

- ▶ grupowanie hierarchiczne
- ▶ algorytm K-średnich
- ▶ analiza składowych głównych (PCA)

## Tworzenie drzew filogenetycznych:

- ▶ metodą średnich połączeń (UPGMA),
- ▶ metodą przyłączania sąsiadów (NJ).

# Łańcuchy Markowa i ukryte modele Markowa (HMM)

- ▶ porównywanie hipotez, obliczanie prawdopodobieństw (dany HMM, sekwencja obserwacji, sekwencja stanów)
- ▶ obliczanie prawd. obserwacji, dla danego HMM:
  - ▶ algorytm prefiksowy
  - ▶ algorytm sufiksowy
- ▶ problem dekodowania (dany HMM, sekwencja obserwacji)
  - ▶ algorytm Viterbiego (najbardziej prawdopodobna sekwencja stanów)
  - ▶ algorytm dekodowania prefiksowo-sufiksowy (najbardziej prawdopodobny stan dla danego symbolu)
- ▶ estymacja parametrów ukrytego modelu Markowa
  - ▶ gdy dana sekwencja stanów
  - ▶ gdy dana sekwencja obserwacji: algorytm uczenia Viterbiego
  - ▶ j.w. : algorytm Bauma-Welcha

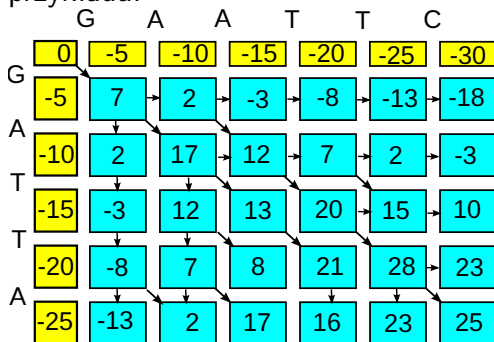
# Inne zagadnienia

- ▶ Algorytm Nussinov do badania struktur drugorzędowych RNA
- ▶ Popularne formaty danych w bazach sekwencji sekwencji biologicznych
- ▶ Analiza haplotypów, algorytm maksymalizacji oczekiwań (algorytm EM)
- ▶ Opis jakości testów binarnych: macierz pomyłek, krzywe ROC, krzywe PR, krzywe kosztu



# Przykładowe zadania

Podaj najlepsze dopasowanie globalne dla przedstawionego przykładu:



## Przykładowe zadania (2)

Zbadano profil genetyczny dla loci A i B. Warianty są oznaczane nazwą loci i numerem wariantu. Profil genetyczny dziecka to  $A1A2B1$ . Dla podanych 2 kobiet i 3 mężczyzn proszę podać, które pary mogą być rodzicami dziecka.

kobieta 1	A1 B1
kobieta 2	A1 A2 B1
mężczyzna 1	A2 B2
mężczyzna 2	A1 B1
mężczyzna 3	A2 B1

## Przykładowe zadania (2)

Zbadano profil genetyczny dla loci A i B. Warianty są oznaczane nazwą loci i numerem wariantu. Profil genetyczny dziecka to  $A1A2B1$ . Dla podanych 2 kobiet i 3 mężczyzn proszę podać, które pary mogą być rodzicami dziecka.

kobieta 1	A1 B1
kobieta 2	A1 A2 B1
mężczyzna 1	A2 B2
mężczyzna 2	A1 B1
mężczyzna 3	A2 B1

K1 i M3, K2 i M2, K2 i M3

## Przykładowe zadania (3)

Badano profil genetyczny dla loci  $A$ ,  $B$  i  $C$ , warianty są oznaczane nazwą locus i numerem wariantu. Mieszanina 2 osób (nazywanych podejrzanym i ofiarą) ma profil  $A_1A_2A_3B_1B_2C_1C_2C_3C_4$ . Dla podanych poniżej profili podejrzanych proszę podać te, są obecne w mieszaninie, zakładając, że profil ofiary to  $A_1A_2B_1B_2C_1C_3$ .

podejrzaany W	$A_1A_3B_1C_2C_4$
podejrzaany X	$A_1A_2B_1B_2C_2C_4$
podejrzaany Y	$A_3B_3C_2C_4$
podejrzaany Z	$A_3B_1B_2C_2C_4$

## Przykładowe zadania (3)

Badano profil genetyczny dla loci  $A$ ,  $B$  i  $C$ , warianty są oznaczane nazwą locus i numerem wariantu. Mieszanina 2 osób (nazywanych podejrzanym i ofiarą) ma profil  $A_1A_2A_3B_1B_2C_1C_2C_3C_4$ . Dla podanych poniżej profili podejrzanych proszę podać te, są obecne w mieszaninie, zakładając, że profil ofiary to  $A_1A_2B_1B_2C_1C_3$ .

podejrzaany W	$A_1A_3B_1C_2C_4$
podejrzaany X	$A_1A_2B_1B_2C_2C_4$
podejrzaany Y	$A_3B_3C_2C_4$
podejrzaany Z	$A_3B_1B_2C_2C_4$

W, Z

# Przykładowe zadania (4)

Posługujemy się monetami A, B, C, D (tylko moneta A jest uczciwa) i obserwujemy sekwencję rzutów (orły i reszki) **ORR**. Zakładając, że przedstawione doświadczenie jest opisywane przedstawionym ukrytym modelem Markowa, podaj najbardziej prawdopodobną sekwencję monet.

- ▶  $Q = \{A, B, C, D\}$
- ▶  $V = \{O, R\}$
- ▶  $P_A = \frac{1}{2}, P_B = \frac{1}{2}, P_C = 0, P_D = 0$

	A	B	C	D
A	$\frac{4}{5}$	$\frac{1}{10}$	$\frac{1}{20}$	$\frac{1}{20}$
B	$\frac{1}{10}$	$\frac{4}{5}$	$\frac{1}{20}$	$\frac{1}{20}$
C	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{4}{5}$	$\frac{1}{10}$
D	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{10}$	$\frac{4}{5}$

	O	R
A	$\frac{1}{2}$	$\frac{1}{2}$
B	$\frac{1}{4}$	$\frac{3}{4}$
C	0	1
D	1	0

## Przykładowe zadania (4)

Posługujemy się monetami A, B, C, D (tylko moneta A jest uczciwa) i obserwujemy sekwencję rzutów (orły i reszki) **ORR**. Zakładając, że przedstawione doświadczenie jest opisywane przedstawionym ukrytym modelem Markowa, podaj najbardziej prawdopodobną sekwencję monet.

- ▶  $Q = \{A, B, C, D\}$
- ▶  $V = \{O, R\}$
- ▶  $P_A = \frac{1}{2}, P_B = \frac{1}{2}, P_C = 0, P_D = 0$

	A	B	C	D
A	$\frac{4}{5}$	$\frac{1}{10}$	$\frac{1}{20}$	$\frac{1}{20}$
B	$\frac{1}{10}$	$\frac{4}{5}$	$\frac{1}{20}$	$\frac{1}{20}$
C	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{4}{5}$	$\frac{1}{10}$
D	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{10}$	$\frac{4}{5}$

	O	R
A	$\frac{1}{2}$	$\frac{1}{2}$
B	$\frac{1}{4}$	$\frac{3}{4}$
C	0	1
D	1	0

BBB

## Przykładowe zadania (5)

Dla przykładów A, B, C, D stosuje się algorytm grupowania hierarchicznego, który ma utworzyć dwie grupy. Macierz odległości:

	A	B	C	D
A	0	2	4	1
B		0	3	4
C			0	5
D				0

- ▶ Podaj te grupy, jeżeli odległość między grupami to pojedyncze wiązanie (minimalna odległość pomiędzy elementami)?
- ▶ Podaj te grupy, jeżeli odległość między grupami to pełne wiązanie (maksymalna odległość pomiędzy elementami)?



## Przykładowe zadania (5)

Dla przykładów A, B, C, D stosuje się algorytm grupowania hierarchicznego, który ma utworzyć dwie grupy. Macierz odległości:

	A	B	C	D
A	0	2	4	1
B		0	3	4
C			0	5
D				0

- ▶ Podaj te grupy, jeżeli odległość między grupami to pojedyncze wiązanie (minimalna odległość pomiędzy elementami)?

$\{A,B,D\}\{C\}$

- ▶ Podaj te grupy, jeżeli odległość między grupami to pełne wiązanie (maksymalna odległość pomiędzy elementami)?

$\{A,D\}\{B,C\}$

# Przykładowe zadania (6)

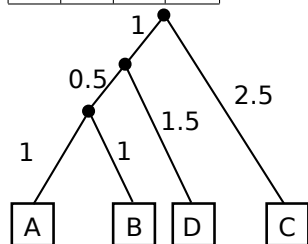
Podaj drzewo filogenetyczne dla 4 taksonów (A, B, C, D), macierz odległości pokazano poniżej. Wykorzystaj metodę grupowania parami ze średnią arytmetyczną (UPGMA).

	B	C	D
A	2	4	3
B		4	3
C			6

# Przykładowe zadania (6)

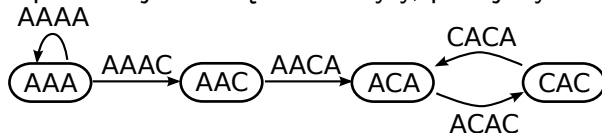
Podaj drzewo filogenetyczne dla 4 taksonów (A, B, C, D), macierz odległości pokazano poniżej. Wykorzystaj metodę grupowania parami ze średnią arytmetyczną (UPGMA).

	B	C	D
A	2	4	3
B		4	3
C			6



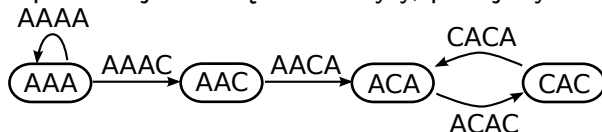
# Przykładowe zadania (7)

Dla przedstawionego poniżej grafu (graf de Bruijina 4 rzędu), który reprezentuje bezbłędne odczyty, podaj wynikową sekwencję.



# Przykładowe zadania (7)

Dla przedstawionego poniżej grafu (graf de Bruijna 4 rzędu), który reprezentuje bezbłędne odczyty, podaj wynikową sekwencję.



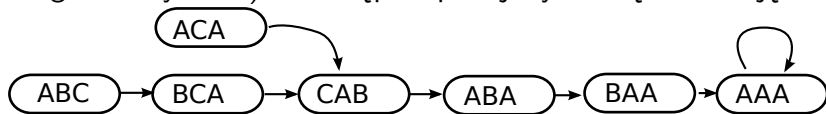
AAAACACA

## Przykładowe zadania (8)

Dla zestawu bezbłędnych odczytów: BAAAA, ABAA, ACAB, ABCABA zbuduj graf de Bruijna 4 rzędu (w wierzchołkach są sekwencje o długości 3 symboli), a następnie podaj wynikową sekwencję.

## Przykładowe zadania (8)

Dla zestawu bezbłędnych odczytów: BAAAA, ABAA, ACAB, ABCABA zbuduj graf de Bruijna 4 rzędu (w wierzchołkach są sekwencje o długości 3 symboli), a następnie podaj wynikową sekwencję.



# Przykładowe zadania (9)

Obok podano fragment pliku SAM dla sekwencji referencyjnej ACTGGGACCTA (indeks od 1). Wyznacz potencjalne warianty, podaj pokrycie i określ, czy jest on homo- czy heterozygotyczny.

POS	CIGAR	SEQ
2	5M1D1M	CTGGGC
4	3M1D3M	GTGCCT
5	2M1D4M	TGCCTA
5	2M1D4M	GGCCTA



# Przykładowe zadania (9)

Obok podano fragment pliku SAM dla sekwencji referencyjnej ACTGGGACCTA (indeks od 1). Wyznacz potencjalne warianty, podaj pokrycie i określ, czy jest on homo- czy heterozygotyczny.

POS	CIGAR	SEQ
2	5M1D1M	CTGGGC
4	3M1D3M	GTGCCT
5	2M1D4M	TGCCTA
5	2M1D4M	GGCCTA

A	C	T	G	G	G	A	C	C	T	A	
	C	T	G	G	G	*	C				5M1D1M
			G	T	G	*	C	C	T		3M1D3M
				T	G	*	C	C	T	A	2M1D4M
				G	G	*	C	C	T	A	2M1D4M

# Przykładowe zadania (9)

Obok podano fragment pliku SAM dla sekwencji referencyjnej ACTGGGACCTA (indeks od 1). Wyznacz potencjalne warianty, podaj pokrycie i określ, czy jest on homo- czy heterozygotyczny.

POS	CIGAR	SEQ
2	5M1D1M	CTGGGC
4	3M1D3M	GTGCCT
5	2M1D4M	TGCCTA
5	2M1D4M	GGCCTA

A	C	T	G	G	G	A	C	C	T	A	
	C	T	G	G	G	*	C				5M1D1M
			G	T	G	*	C	C	T		3M1D3M
				T	G	*	C	C	T	A	2M1D4M
				G	G	*	C	C	T	A	2M1D4M

Są dwa warianty:

- ▶ wariant heterozygotyczny na pozycji 5, G>T
- ▶ delecja homozygotyczny na pozycji 7, A>\_\_

# Przykładowe zadania (10)

Podaj macierz pomyłek, dla wyników testu przedstawionych w tabeli, zakładając, że traktujemy wynik testu powyżej 0.35 jako pozytywny (osoba chora), następnie dla progu 0.5 oraz progu 0.65. Narysuj dodatkowe 3 punkty na krzywej ROC (dla progu 0.35, 0.5 i 0.65). Narysuj krzywą ROC.

osoba	stan	wynik testu
A	zdrowa	0.1
B	zdrowa	0.1
C	zdrowa	0.2
D	zdrowa	0.3
E	zdrowa	0.6
F	chora	0.6
G	chora	0.7
H	chora	0.8
I	chora	0.9
J	chora	0.9

# Przykładowe zadania (10)

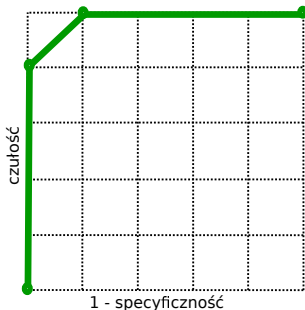
Podaj macierz pomyłek, dla wyników testu przedstawionych w tabeli, zakładając, że traktujemy wynik testu powyżej 0.35 jako pozytywny (osoba chora), następnie dla progu 0.5 oraz progu 0.65. Narysuj dodatkowe 3 punkty na krzywej ROC (dla progu 0.35, 0.5 i 0.65). Narysuj krzywą ROC.

osoba	stan	wynik testu
A	zdrowa	0.1
B	zdrowa	0.1
C	zdrowa	0.2
D	zdrowa	0.3
E	zdrowa	0.6
F	chora	0.6
G	chora	0.7
H	chora	0.8
I	chora	0.9
J	chora	0.9

Próg 0.35:	TP = 5	FP = 1
	FN = 0	TN = 4

Próg 0.5:	TP = 5	FP = 1
	FN = 0	TN = 4

Próg 0.65:	TP = 4	FP = 0
	FN = 1	TN = 5



*Dziękuję*