Algorytmy w bioinformatyce sekwencji

biologia syntetnczna, algorytm Nussinov, powtórzenie

Robert Nowak

2024

Powtórzenie

Badanie podobieństw sekwencji

Badanie podobieństw dwu sekwencji:

- bez uwzględniania przerw, alg. Knutha-Morrisa-Prata,
- podobieństwo globalne, algorytm Needlemana-Wunscha,
- podobieństwo lokalne, algorytm Smitha-Watermana,
- algorytmy o liniowym koszcie pamięciowym,
- algorytmy z afiniczną funkcją kary za przerwę.

Podobieństwa wielu sekwencji:

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

Zagadnienia związane z badanim podobieństw

- Obliczanie macierzy podobieństwa BLOSUM,
- obliczanie macierzy podobieństwa PAM,
- istotność przy algorytmach podobieństwa sekwencji.

Bazy danych sekwencji:

- popularne formaty danych w bazach sekwencji sekwencji biologicznych,
- algorytmy oparte o heurystyki (BLAST, FASTA),
- istotność wyników wyszukiwania w bazie danych,
- badanie jakości testów binarnych: macierz pomyłek, krzywe ROC, krzywe PR, krzywe kosztu.

Sekwencjonowanie

- Generacje sekwenatorów,
- resekwencjonowanie, pliki SAM,
- asemblacja de-novo, nadmiarowość sekwencjonowania,
- algorytmy asemblacji oparte o:
  - grafy pokrycia,
  - podgrafy de Bruijna.
- mapy restrykcyjne, algorytm rozwiązujący problem częściowego strawienia.

Analiza danych wielowymiarowych

Grupowanie:

- grupowanie hierarchiczne,
- algorytm K-średnich.

Tworzenie drzew filogenetycznych w oparciu o odległości:

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

Redukcja wymiarów algorytmem analiza składowych głównych (PCA).

Algorytmy związane z markerami genetycznymi:

- badanie pokrewieństw w oparciu o obserwacje wariantów,
- badanie mieszanin DNA,
- odtwarzanie haplotypów, algorytm EM.

Inne zagadnienia

Ukryte modele Markowa (HMM):

- porównywanie hipotez, obliczanie prawdopodobieństw (dany HMM, sekwencja obserwacji, sekwencja stanów)
- problem dekodowania (dany HMM, sekwencja obserwacji)
  - algorytm Viterbiego (najbardziej prawdopodobna sekwencja stanów)
  - algorytm dekodowania prefiksowo-sufiksowy (najbardziej prawdopodobny stan dla danego symbolu)
- obliczanie prawd. obserwacji, dla danego HMM:
  - algorytm prefiksowy
  - algorytm sufiksowy

Algorytm Nussinov do znajdowania struktur drugorzędowych.

Egzaminy

Zaliczenie przedmiotu:

- egzamin: 0 – 50pkt
- ćwiczenia: 0 – 50pkt

91 – 100 pkt.	ocena 5
81 – 90 pkt.	ocena $4\frac{1}{2}$
71 – 80 pkt.	ocena 4
61 – 70 pkt.	ocena $3\frac{1}{2}$
51 – 60 pkt.	ocena 3
0 – 50 pkt.	ocena 2

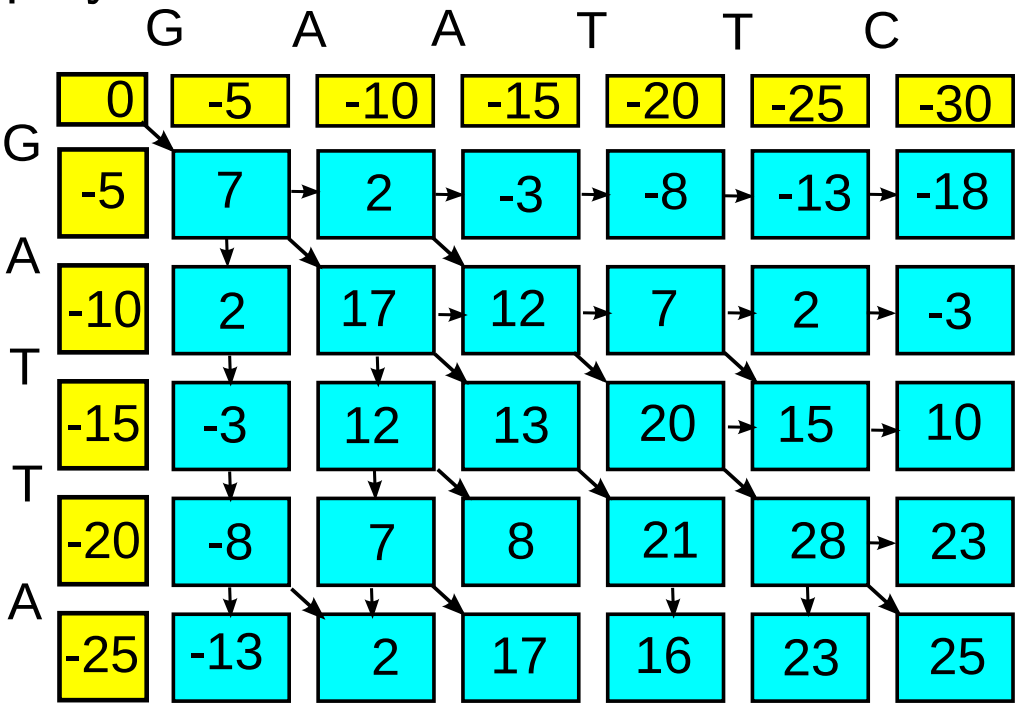
Egzamin online na platformie MS Teams, zakładka Zadania.

- 5 zadań, każde oceniane w skali 0-10 pkt, 60 minut,
- można mieć notatki,
- w trakcie egzaminu trzeba być na kanale wykładowym na MS Teams i mieć włączoną kamerę.

Proszę wykonać zadanie próbne - to jedynie test komunikacji.

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 $A_1A_2B_1B_2C_1C_3$.

podejrzany W	$A_1A_3B_1C_2C_4$
podejrzany X	$A_1A_2B_1B_2C_2C_4$
podejrzany Y	$A_3B_3C_2C_4$
podejrzany 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 $A_1A_2B_1B_2C_1C_3$.

podejrzany W	$A_1A_3B_1C_2C_4$
podejrzany X	$A_1A_2B_1B_2C_2C_4$
podejrzany Y	$A_3B_3C_2C_4$
podejrzany 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)

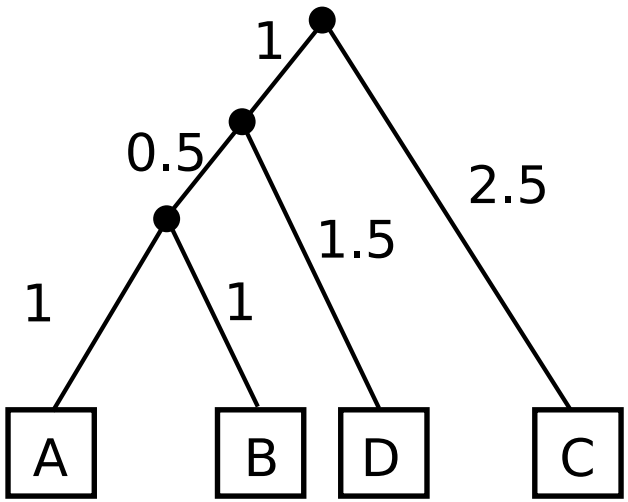
Podaj drzewo filogenetyczne dla 4 taksonów (A, B, C, D), macierz odległości pokazano poniżej. Wykorzystaj metodę grupowania parami ze średnią arytmetyczną (UPGMA).

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

Przykładowe zadania (6)

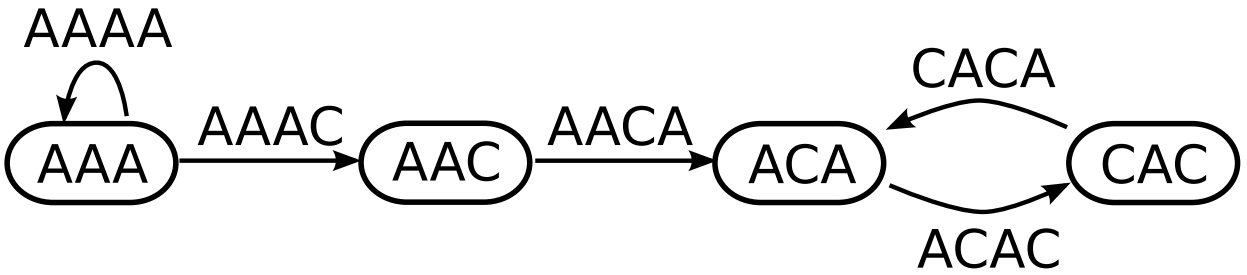
Podaj drzewo filogenetyczne dla 4 taksonów (A, B, C, D), macierz odległości pokazano poniżej. Wykorzystaj metodę grupowania parami ze średnią arytmetyczną (UPGMA).

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



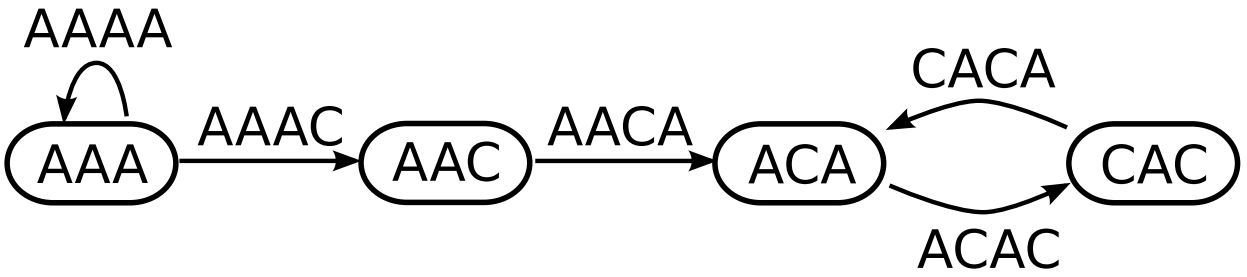
Przykładowe zadania (7)

Dla przedstawionego poniżej grafu (graf de Bruijna 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.



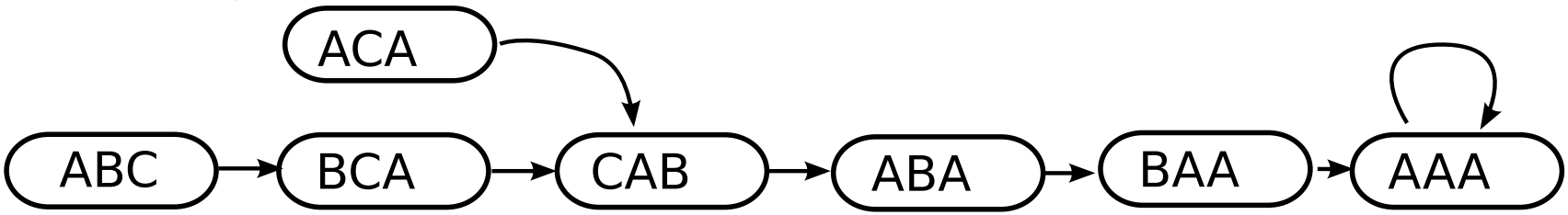
AAAACACA

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

Sa 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:

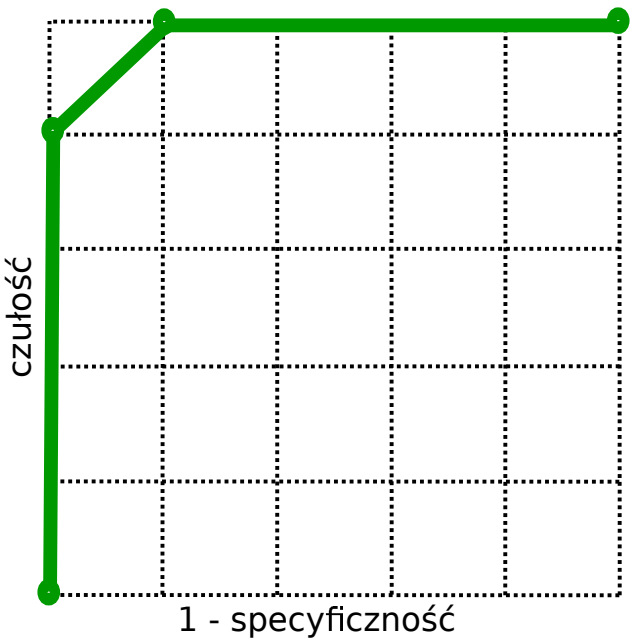
TP = 5	FP = 1
FN = 0	TN = 4

Próg 0.5:

TP = 5	FP = 1
FN = 0	TN = 4

Próg 0.65:

TP = 4	FP = 0
FN = 1	TN = 5



Dziękuję