

Wprowadzenie do Bioinformatyki

zajęcia laboratoryjne 4

Algorytmy dopasowania dwóch sekwencji: Algorytm Needlemana-Wunscha vs. Smitha-Watermana

Bioinformatyka I rok

1 Wprowadzenie

Algorytm Needlemana-Wunscha wykorzystywany do globalnego dopasowania dwóch sekwencji. Algorytm polega na uzupełnieniu macierzy M . Warunki wstępne: pierwszy wiersz i pierwsza kolumna uzupełniane są wartościami ujemnymi (traktowane jak przerwa). Natomiast kolejne pola macierzy uzupełniane są zgodnie ze wzorem:

$$M_{i,j} = \max \begin{cases} M_{i-1,j-1} + S_{x_i,y_j} \\ M_{i-1,j} + G \\ M_{i,j-1} + G \end{cases}$$

gdzie S_{x_i,y_j} oznacza scoring - punktację macierzy za dopasowanie lub niedopasowanie, natomiast G jest to kara za przerwę.

Algorytm Smitha-Watermana wykorzystywany do lokalnego dopasowania dwóch sekwencji. Algorytm polega na uzupełnieniu macierzy M . Warunki wstępne: pierwszy wiersz i pierwsza kolumna uzupełniane są zerami. Natomiast kolejne pola macierzy uzupełniane są zgodnie ze wzorem:

$$M_{i,j} = \max \begin{cases} M_{i-1,j-1} + S_{x_i,y_j} \\ M_{i-1,j} + G \\ M_{i,j-1} + G \\ 0 \end{cases}$$

gdzie S_{x_i,y_j} oznacza scoring - punktację macierzy za dopasowanie lub niedopasowanie, natomiast G jest to kara za przerwę. W tym przypadku w macierzy nie można umieszczać wartości ujemnych, zamiast nich wprowadza się zera.

2 Przykład

Dla sekwencji *AATCG* i sekwencji *AACG* wykonano dopasowanie algorytmem Needlemana-Wunscha i Smitha-Watermana. Przyjęty system punktacji jest następujący: dopasowanie = 1, niedopasowanie = -1, przerwa = -2. Na podstawie wynikowej tabeli znaleziono rozwiązanie dopasowania tych dwóch sekwencji, w wyniku przedstawiono wartość *score* oraz dopasowanie dwóch sekwencji, gdzie "*" oznacza dopasowanie, "|" oznacza niedopasowanie, a "-" oznacza przerwę. Dopasowanie globalne znajduje się na Rysunku 1, podczas gdy dopasowanie lokalne znajduje się na Rysunku 2.

Needleman-Wunsch

		A	A	C	G
	0	-2	-4	-6	-8
A	-2	1	-1	-3	-5
A	-4	-1	2	0	-2
T	-6	-3	0	1	-1
C	-8	-5	-2	1	0
G	-10	-7	-4	-1	2

match = 1, mismatch = -1, gap = -2

```

A  A  T  C  G
*  *  |  *  *
A  A  -  C  G
    
```

score = 2

Rysunek 1: Przykładowe globalne dopasowanie dwóch sekwencji – algorytm Needlemana-Wunscha.

Smith-Waterman

		A	A	C	G
	0	0	0	0	0
A	0	1	1	0	0
A	0	1	2	0	0
T	0	0	0	1	0
C	0	0	0	1	0
G	0	0	0	0	2

match = 1, mismatch = -1, gap = -2

wynik 1:

```

A  A
*  *
A  A  score = 2
    
```

wynik 2:

```

C  G
*  *
C  G  score = 2
    
```

Rysunek 2: Przykładowe lokalne dopasowanie dwóch sekwencji – algorytm Smitha-Watermana.

3 Zadania do wykonania

- Znajdź narzędzia (zapisz linki źródłowe), które pozwalają wykonać dopasowanie sekwencji za pomocą algorytmu Needlemana-Wunscha i Smitha-Watermana. Zwróć uwagę czy w wyniku działania tych narzędzi możesz otrzymać więcej niż jedno możliwe rozwiązanie.
- Wykorzystaj algorytm Needlemana-Wunscha i Smitha-Watermana aby samodzielnie wykonać dopasowanie wybranych poniżej sekwencji. Rozwiązania do poniższych zadań można umieścić na kartce papieru lub w arkuszu kalkulacyjnym. Wynikiem dopasowania dla poszczególnych porównań jest uzupełniona macierz, znalezione dopasowanie oraz score.
 - Dopasuj poniższe sekwencje wykorzystując algorytm Needlemana-Wunscha z następującymi założeniami: dopasowanie = 1, niedopasowanie = -1, przerwa = -1:
 - AGGTC
 - AGCTTG
 - Dopasuj poniższe sekwencje wykorzystując algorytm Smitha-Watermana z następującymi założeniami: dopasowanie = 1, niedopasowanie = -1, przerwa = -2
 - AAGACC
 - ATCTACC