# Bioinformatyka – zadanie 1

## Opis problemu

Badanym problemem jest problem uliniowienia sekwencji. Podane zadanie zostało wykonane z użyciem dwóch algorytmów Needlemana-Wunscha oraz Watermana-Smitha. Oba algorytmy zostały zaimplementowane w języku Python. Działanie podanych algorytmów zostało sprawdzone podczas porównania dwóch sekwencji z użyciem kilku różnych funkcji scorujących (scoring function).

## Metody

Dla podanego eksperymentu zostały użyte dwa algorytmy stosowane do problemu dopasowania sekwencji :

* Algorytm Needlemana-Wunscha – algorytm stosowany do znalezienia optymalnego uliniowienia globalnego. Podany algorytm stosuje ideę programowania dynamicznego do rozwiązania problemu dopasowania sekwencji. Przy użyciu tablicy, porównywane są mniejsze pod problemy, które pomagają uzyskać najlepszy wynik dla całych sekwencji.
* Algorytm Smitha-Watermana – algorytm bardzo podobny do algorytmu Needlemana-Wunscha. Główną różnicą pomiędzy algorytmami jest fakt, że algorytm Smitha-Watermana wybiera maksimum z wykonywanych działań oraz 0, co sprawia że w macierzy na, której pracuje znajdują się wyłącznie dodatnie wartości. Podany algorytm znajduje rozwiązania lokalne.

## Opis eksperymentów

Powyższe algorytmy zostały przetestowane dla poniższych problemów :

1. Porównanie dwóch genów homologicznych, zostały użyte geny :
   1. APOE człowieka - <https://www.ncbi.nlm.nih.gov/gene/348>
   2. APOE makaka królewskiego - <https://www.ncbi.nlm.nih.gov/gene/714623>
2. Porównanie sekwencji białkowej insuliny :
   1. ludzkiej - <https://www.ncbi.nlm.nih.gov/gene/3630>
   2. chomika - <https://www.ncbi.nlm.nih.gov/gene/100760716>

Eksperymenty zostały przeprowadzone dla trzech różnych funkcji scorujących :

* Funkcja, która w znacznym stopniu nagradza za dobre dopasowanie :
  + 10 punktów za dobre dopasowanie
  + -1 punkt za złe dopasowanie
  + -1 punkt za przerwę w sekwencji
* Funkcja, która karze za złe dopasowanie
  + 2 punkty za dobre dopasowanie
  + -6 punktów za złe dopasowanie
  + -2 punkty za przerwę w sekwencji
* Funkcja, która karze za przerwy w sekwencji
  + 2 punkty za dobre dopasowanie
  + -2 punkty za złe dopasowanie
  + -5 punkty za przerwę w sekwencji

Dla każdego z eksperymentów policzony został wynik według zadanej funkcji scorującej oraz ilość pasujących liter, niepasujących oraz przerw. Przerwa oraz litera liczona było jako niepasująca litera.

## Wyniki

[10, -1, -1]

Needleman-Wunsch

APOE : Score : 32777 Matches : 3345 Mismatches : 673 Gaps : 512

INS : Score : 6027 Matches : 681 Mismatches : 783 Gaps : 699

Smith-Waterman

APOE : score : 32824 Matches : 3340 Mismatches : 760 Gaps : 676

INS : Score : 6119 Matches : 642 Mismatches : 898 Gaps : 851

[2, -6, -2]

Needleman-Wunsch

APOE : Score : 5022 Matches : 3345 Mismatches : 834 Gaps : 834

INS : Score : -372 Matches : 681 Mismatches : 867 Gaps : 867

Smith-Waterman

APOE : Score : 5156 Matches : 3343 Mismatches : 756 Gaps : 674

INS : Score : 96 Matches : 538 Mismatches : 1062 Gaps : 971

[2, -2, -5]

Needleman-Wunsch

APOE : Score : 3856 Matches : 3337 Mismatches : 671 Gaps : 492

INS : Score : -2149 Matches : 653 Mismatches : 778 Gaps : 633

Smith-Waterman

APOE : Score : 4181 Matches : 3230 Mismatches : 784 Gaps : 504

INS : Score : 3 Matches : 353 Mismatches : 1177 Gaps : 831

## Wnioski

Dla dwóch pierwszych funkcji scorujących liczba dopasowanych liter jest większa niż w przypadku trzeciej funkcji scorującej. Może być to spowodowane większą karą za przerwę w porównaniu do kary za złe dopasowanie.

Pomiędzy funkcjami scorującymi 2 i 3 widać różnicę pomiędzy ilością nieprawidłowych dopasowań. Można zauważyć, że różnice te działają w przeciwne strony dla obydwu algorytmów. Dla algorytmu Needlemana-Wunscha więcej złych dopasowań jest w przypadku, gdy zastosowana została większa kara na złe dopasowanie. Dla algorytmu Watermana-Smitha rezultat jest odwrotny.

W przypadku kary za wystąpienie przerwy, rezultat jest zgodny z oczekiwaniami. Występuje ich mniej jeśli nałożona została wyższa kara za wystąpienie przerwy.