**Elementy Bioinformatyki**

**Projekt - sprawozdanie**

Zestawienia optymalne 2 sekwencji II

Anna Cholewczyńska 143198

Kamil Gersten 143220

Dominik Stefaniak 143133

# Opis programu

Po uruchomieniu programu, z pliku *input.txt* znajdującego się w tym samym folderze co program, pobrane zostają dwa wyrazy (sekwencje nukleotydowe).

Następnie program oblicza tablicę odległości edycyjnych (D), oraz tablice A, B, C oraz S, potrzebne do wyznaczenia podobieństwa wyrazów.

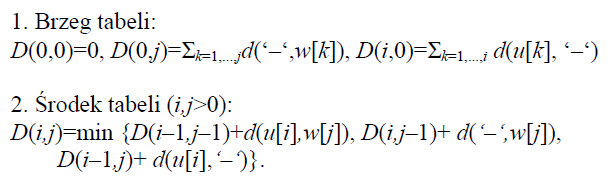
Parametry takie jak wartość metryki, wartość podobieństwa znaków czy funkcja kary za przerwy są zdefiniowane w kodzie programu.

Następnie program wypisuje na konsolę sekwencje wejściowe, wartość podobieństwa wyrazów i ich optymalne dopasowanie w algorytmie z funkcją kary.

Wywołana jest metoda wyznaczająca dopasowanie na podstawie algorytmu Hirshberga.

# Zastosowane algorytmy

Przy wypełnianiu tablicy odległości edycyjnych D skorzystano z następujących wzorów:

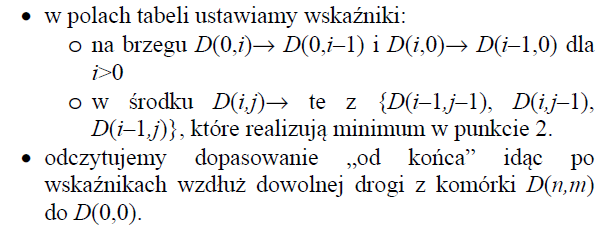


, gdzie *i* i *j* to długości słów odpowiednio pierwszego i drugiego.

Funkcja *d(a,b)* zdefiniowana jest w ten sposób, że wynosi 0, gdy *a=b*, oraz 1 w innych przypadkach. Wartość tą można zmienić w kodzie programu.

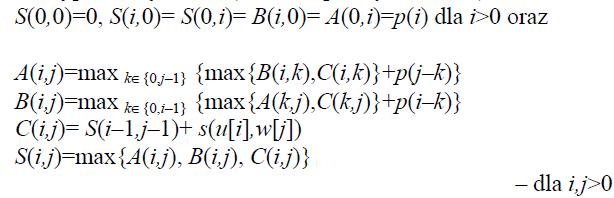
Wartość odległości edycyjnej odczytywana jest z komórki D(i,j).

Optymalne dopasowanie wyznaczono za pomocą algorytmu



## Optymalne dopasowanie z funkcją kary

Dla wypełniania tablic A, B, C, S, skorzystano z wzorów:



, gdzie *p(x)* to funkcja kary za przerwy, zdefiniowana w kodzie.

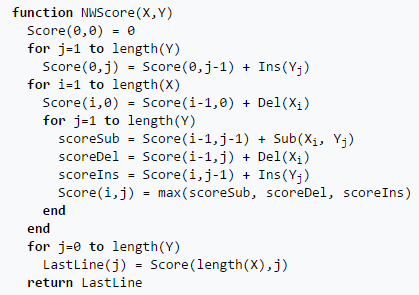
Wartość podobieństwa sekwencji odczytywana jest z pola *S(i,j)*.

Dopasowanie optymalne dla podobieństw wyrazów wyznaczane jest poprzez wyznaczanie najdłuższej ścieżki wiodącej z *S(i,j)*, do *S(0,0)I.* Wykorzystano tutaj identyczny algorytm co w przypadku odległości edycyjnych, z tą różnicą że tutaj szukano maksimum zamiast minimum.

## Algorytm Hirshberga

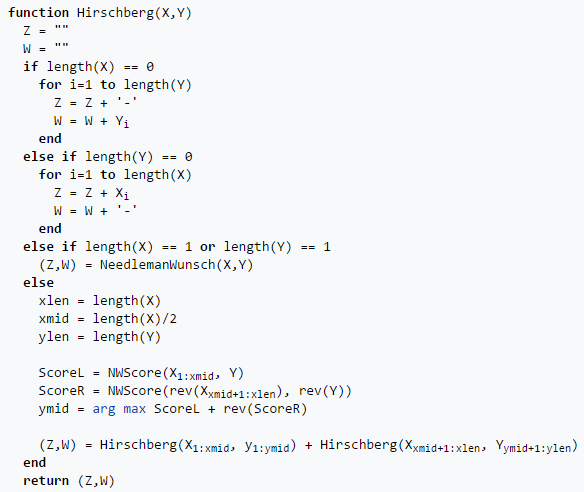
Algorytm Hirshberga jest opisywany jako wersję „dziel i rządź” algorytmu Needlemana-Wunscha (wykorzystanego i opisanego w poprzednim przypadku). Jest to algorytm rekurencyjny, wywoływany tak długo, aż sekwencje wejściowe osiągną długość 1.

Definiując funkcję *NWScore(X,Y),* jako funkcję zwracającą ostatni wiersz macierzy S algorytmu Needlemana-Wunscha:



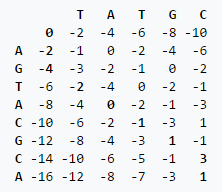
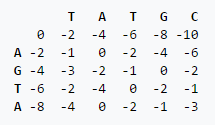
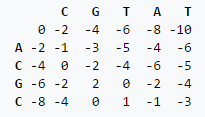
, gdzie Sub, Del i Ins to są metryki za podmienienie, usunięcie i wstawienie znaku.

Sam algorytm Hirshberga wygląda następująco:

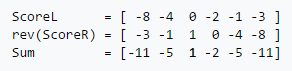


, tj.:

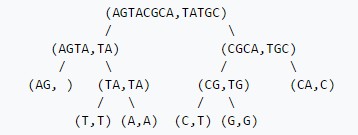
* Dla długości jednej z sekwencji wejściowej = 0, sekwencja wyjściowa odpowiadająca tej równej 0 będzie wypełniona znakami ‘-‘, a druga sekwencja będzie przepisana
* Dla długości jednej z sekwencji wejściowej = 1, wykonywany jest zwyczajny algorytm Needlemana-Wunscha
* W innych przypadkach:
  + Sekwencja wejściowa X, dzielona jest w połowie (rys 1,2,3)
  + Na podstawie sumy ostatnich rzędów obu połówek sekw. X, wyznaczane jest miejsce rozdzielenia sekwencji Y. (Rys. 4)
  + Na obu zestawach połówek wywoływany jest algorytm Hirshberga.
  + Sekwencjami wyjściowymi są sekwencje powstałe na skutek złączenia sekwencji wyjściowych funkcji Hirshberga na obu zestawach połówek

Rys. 1 Rys. 2 Rys. 3



Rys. 4



Rys. 5

# Bibliografia

* *Liniowe dopasowanie dwóch sekwencji*, Materiały pomocnicze do przedmiotu *„Elementy Bioinformatyki”*, Krzysztof Giaro
* *Computational Biology*, Slajdy wykładowe do przedmiotu „*Elementy Bioinformatyki”*, Krzysztof Giaro
* *Hirschberg's algorithm,* https://en.wikipedia.org/wiki/Hirschberg's\_algorithm