Monika Reguła
236689

Wstęp do bioinformatyki

Dopasowywanie par sekwencji – algorytm kropkowy

Link: https://github.com/monikaRegula/Bioinformatics/commits/dotPlot

1. Analiza złożoności obliczeniowej dla funkcji:

```
function dotPlot = createDotPlot(comparison,window,mistake)
%Function using comparison matrix of sequences and parameter such as window siz
%and mistake threshold. In loop of window size checks values from comparison.
%There is counting for cells that equal 1. If counter is within acceptable
%limits (equal or more than differnece between windows size and mistake)
%then new matrix's cell equal 1.
counter = 0;
[size1, size2] = size(comparison);
dotPlot = zeros(size1,size2);
m = size2 - (window-1);
n = size1 - (window-1);

for x=1:n
    for y=1:m
        for z = 1:window
            %counting cells that are equal 1
            if(comparison(x+z-1,y+z-1) == 1)
                counter = counter+1;
            end
        end
        %checking if counter exceeds acceptance limit
        %and creating new data for plot
        if (counter >= (window-mistake))
            for z = 1:window
                dotPlot(x+z-1,y+z-1) = 1;
            end
        end
        %restarting value of counter
        counter = 0;
    end
end
```

Pamięć:

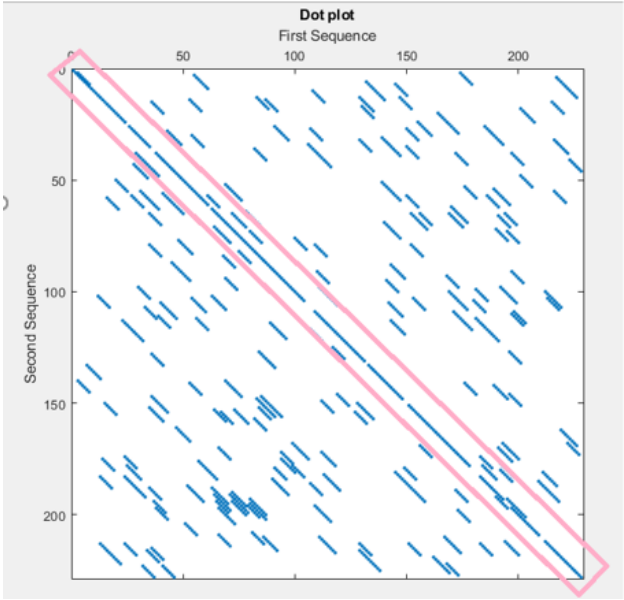
size1*size2 + counter + size1*size2 + m + n + window + mistake + x + y + z = 1+ 2nm +1 +1+1+1+1+1+1= **8+2nm**

Jest to pamięć przeznaczona dla argumentów funkcji (macierz logiczna (nm),rozmiar okna(1), próg błędu(1)) oraz zmienne: licznik counter(1), macierz dotPlot(ab), rozmiar m (1), rozmiar n (1), licznik x(1), y(1), z(1)

Złożoność czasowa:

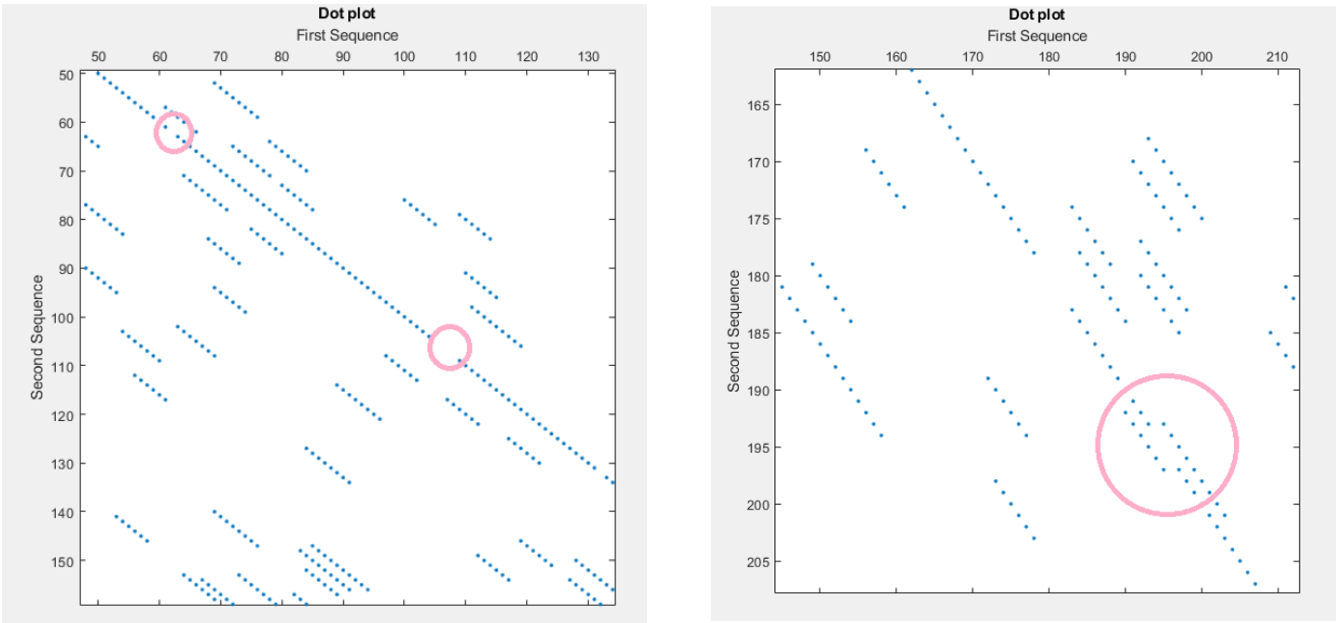
((z *2)+ z) *y*x = 3xyz , gdzie x=n y=m
O(m) = 3nmz ~ O(m^2) przy założeniu, że rozmiar okna znacznie mniejszy niż długość sekwencji

2. Porównanie par sekwencji

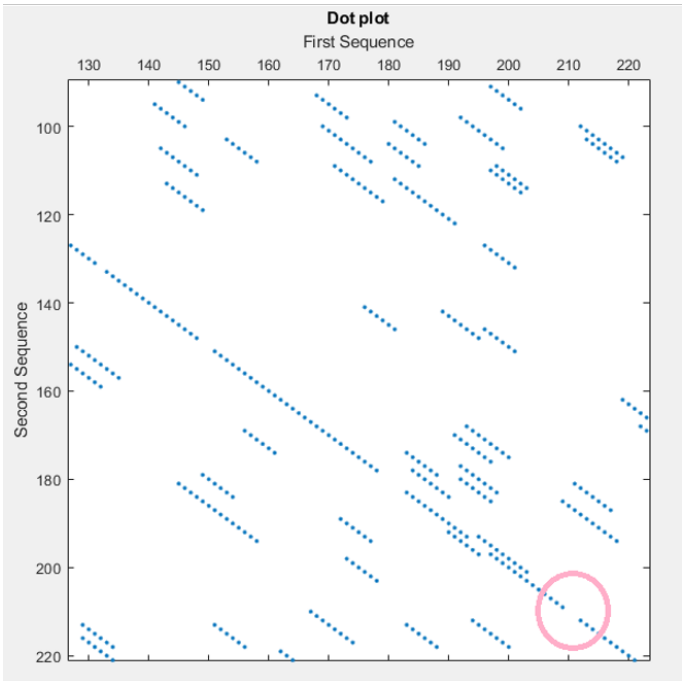


Rysunek 1 Macierz kropkowa dla sekwencji nosorożca oraz mastodonta dla parametrów: rozmiar okna = 6 próg błędu = 1

Za pomocą wygenerowanej macierzy kropkowej można dostrzec powiązanie między sekwencjami cytochromu b dla nosorożca oraz mastodonta (rys. 1). Świadczy o tym linia przechodząca wzdłuż głównej przekątnej ograniczona różowym prostokątem. Przerwy w linii sygnalizują substytucję towarzyszącej mutacji białek (rys. 2). Aczkolwiek insercja czy delecja zachodzi w przypadku, gdy dwie części są delikatnie przesunięte względem siebie (rys3). Istnieje również mutacja w postaci duplikacji(replikacji) widoczna w pobliżu lokalnych zgrubień na wykresie. Dzięki wysokiej ciągłości przekątnej można stwierdzić, że nosorożec oraz mastodont są homologami.

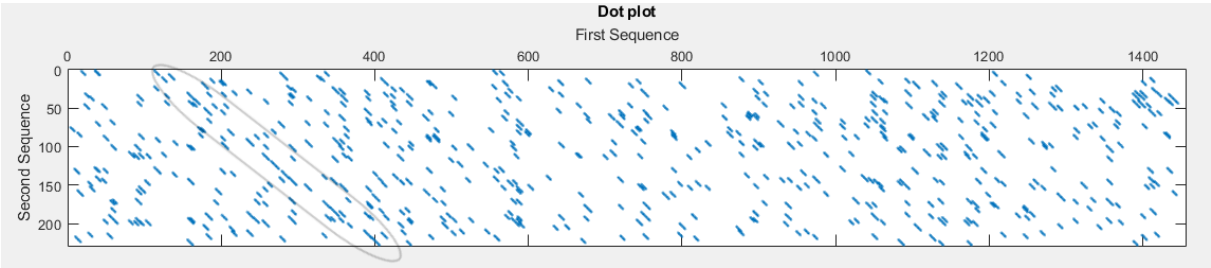


Rysunek 2 Powiększenie wykresów w miejscach substytucji oraz duplikacji

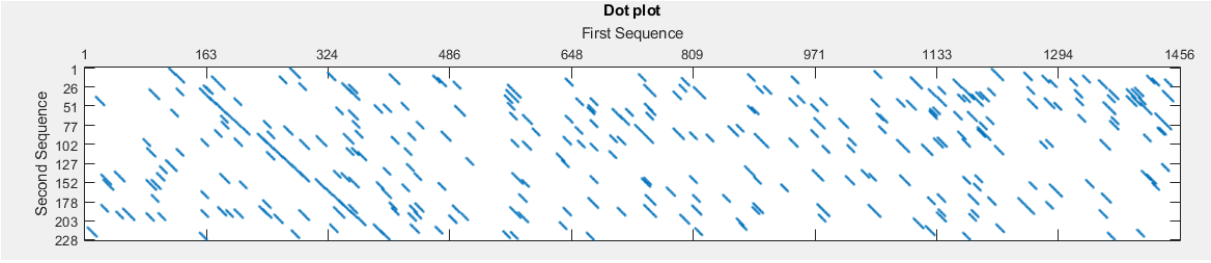


Rysunek 3 Powiększenie wykresu w miejscu prawdopodobnej delecji/insercji

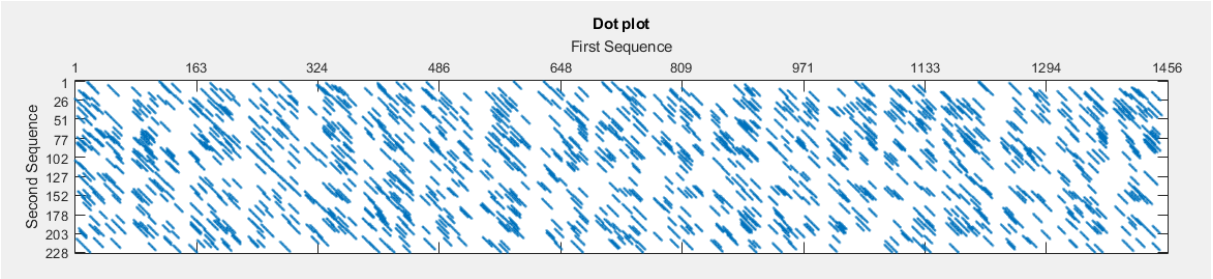
Na rys. 4 widać brak ciągłości linii wzdłuż przekątnej. Jest to skutkiem doboru parametrów, które są mniej optymalne niż te na rys. 5, na którym widać wyraźne fragmenty przekątnej. Jednakże próg błędu 3 dla okna 10



Rysunek 4 Macierz kropkowa dla sekwencji nosorożca i człowieka dla parametrów: rozmiar okna = 6, granica błędu = 1

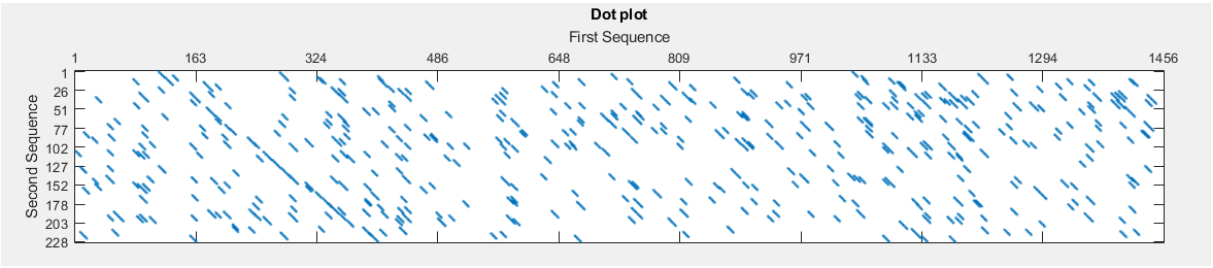


Rysunek 5 Macierz kropkowa dla sekwencji nosorożca i człowieka dla parametrów: rozmiar okna = 10, granica błędu =3

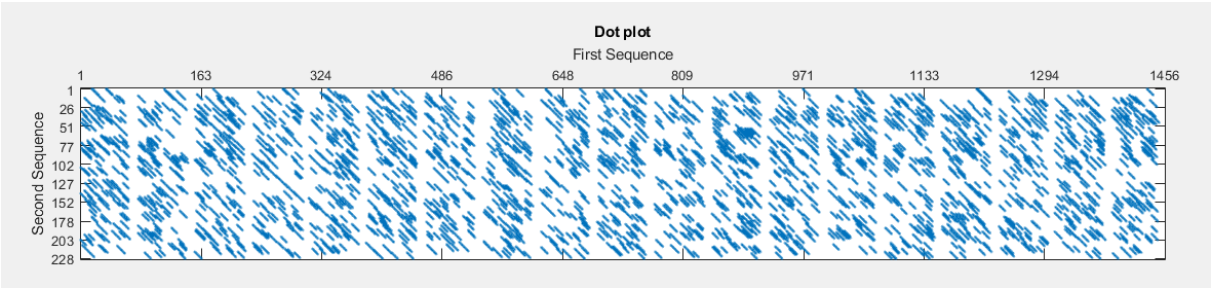


Rysunek 6 Macierz kropkowa dla sekwencji nosorożca i człowieka dla parametrów: rozmiar okna = 10, granica błędu =4

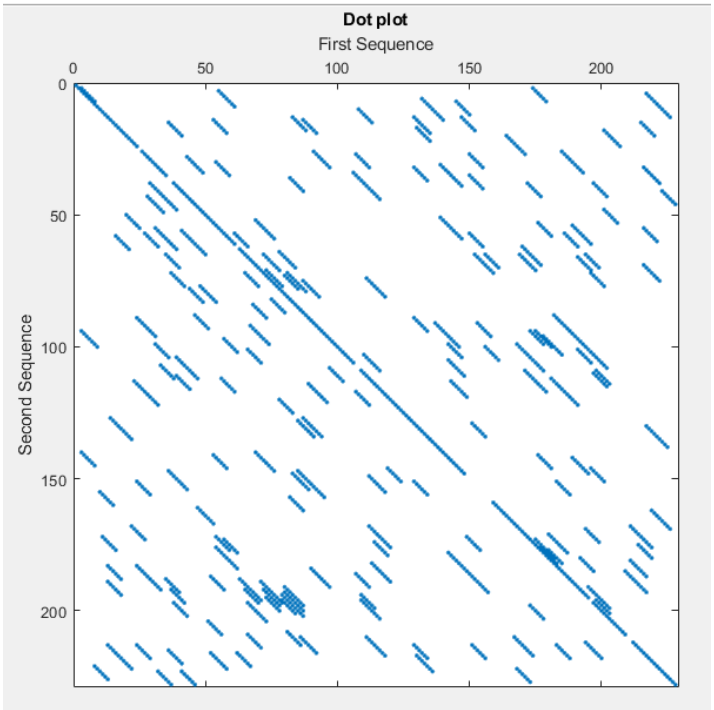
Na rys. 7 można dostrzec przekątną jednak wokół znajduje się wiele fragmentów, które zmniejszają przejrzystość. Porównując z rys. 5 można zauważyć tutaj większą ilość krótszych fragmentów wykazujących podobieństwo. Wynika to z doboru progu błędu, który tutaj wynosi 25% a dla rys. 5 wynosi 30%.
Niewielka zmiana progu błędu skutkuje wysoką nieczytelnością wykresu. Znakomicie widać dla rozmiaru okna równego 8 oraz błędu równego 2 oraz błędu równego 3. Najbardziej optymalny wygląda zakres błędu 27%-32%.



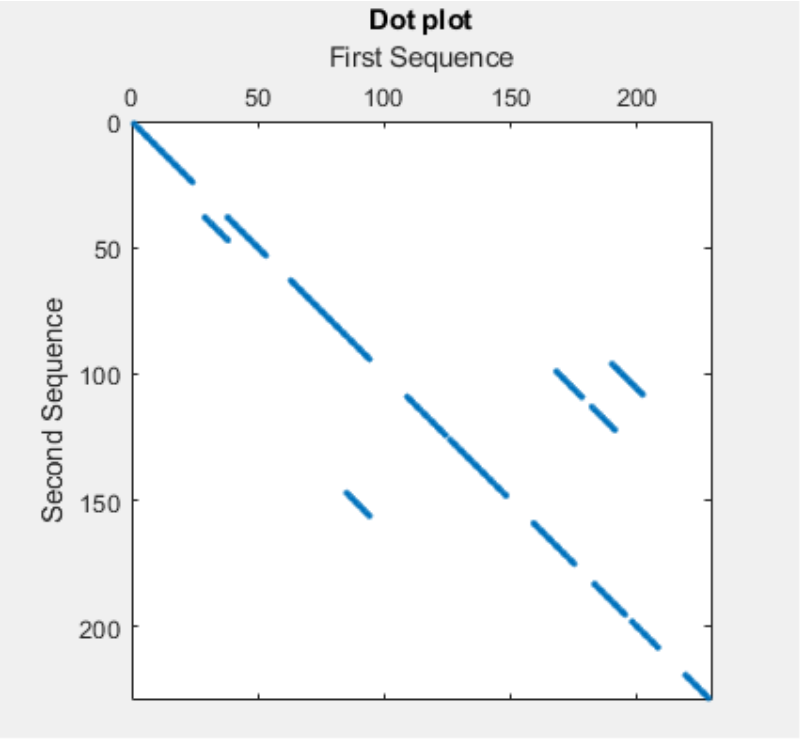
Rysunek 7 Macierz kropkowa dla sekwencji nosorożca i człowieka dla parametrów: rozmiar okna =8, granica błędu =2



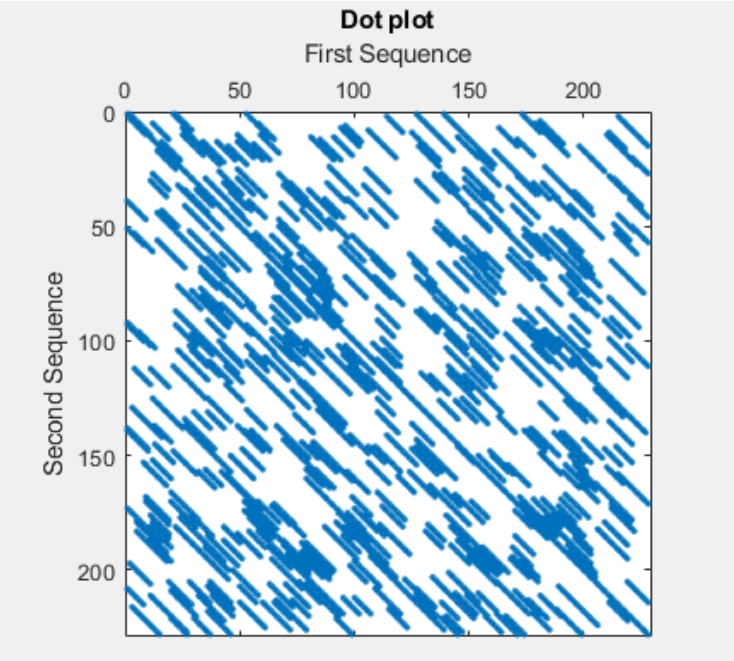
Rysunek 8 Macierz kropkowa dla sekwencji nosorożca i człowieka dla parametrów: rozmiar okna =8, granica błędu =3



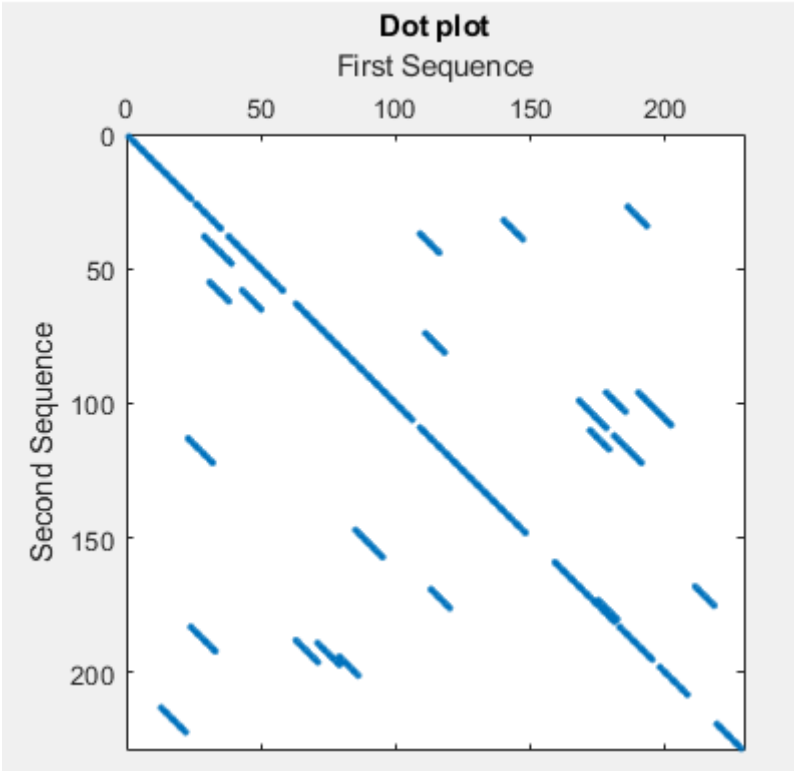
Rysunek 9 Macierz kropkowa dla sekwencji słonia azjatyckiego oraz nosorożca dla parametrów: rozmiar okna = 6, próg błędu = 1



Rysunek 10 Macierz kropkowa sekwencji słonia azjatyckiego i nosorożca dla parametrów: rozmiar okna = 10 próg błędu = 1



Rysunek 11 Macierz kropkowa sekwencji słonia azjatyckiego i nosorożca dla parametrów: rozmiar okna = 8, próg błędu = 3



Rysunek 12 Macierz kropkowa sekwencja słonia azjatyckiego i nosorożca dla parametrów: rozmiar okna = 8 próg błędu = 1

Wybranie parametrów dla algorytmu kropkowego ma wpływ na jakość odczytu macierzy. Niski próg wraz z wysokim rozmiarem okna powoduje, że macierz jest bardziej przejrzysta. Wyższy próg błędu powoduje, że macierz jest trudna do odczytu. Najbardziej optymalne parametry dla wybranych sekwencji to te z rys 8, ponieważ można dostrzec mutacje m.in. duplikacje ,których nie można odczytać z rys 6 oraz insercje/delecje, których nie można odczytać z rys 7.

WNIOSKI: Dzięki zastosowaniu algorytmu kropkowego możliwe było wykazanie pokrewieństwa między sekwencjami. Słoń azjatycki, mastodont czy nosorożec są homologami o czym świadczy główna przekątna. Inaczej jest w przypadku porównania sekwencji człowieka z nosorożcem, gdzie pomimo pewnego zarysu przekątnej organizmy nie są homologami.
Na jakość macierzy końcowej mają wpływ parametry, które powinny być dopasowane do konkretnego zestawu sekwencji. Jednakże można wywnioskować, że im większy próg błędu tym mniejsza czytelność macierzy. Może to skutkować utratą informacji o występujących mutacjach.