Bioinformatyka Sprawozdanie 2

Dopasowanie globalne par sekwencji

Michał Marciniak 244811

1 Kod do repozytorium

`https://gitlab.com/MMarciniak103/bioinformatics`

2 Analiza złożoności obliczeniowej czasowej i pamięciowej:

Algorithm 1 Dopasowanie globalne

procedure INICJALIZACJA
    $m \leftarrow$ długość sekwencji x
    $n \leftarrow$ długość sekwencji y
    $ins \leftarrow$ koszt insercji
    $del \leftarrow$ koszt delecji
    $sub \leftarrow$ koszt substytucji
    $R \leftarrow$ macierz o wymiarach $n \times m$
    for $i \leftarrow 1$ *to* n **do**
        $R[i, 0] = R[i-1, 0] + ins$
    for $j \leftarrow 1$ *to* m **do**
        $R[0, j] = R[0, j-1] + del$
    for $i \leftarrow 1$ *to* n **do**
        for $j \leftarrow 1$ *to* m **do**

$$R[i,j] = \min \begin{cases} R[i-1,j] + del \\ R[i,j-1] + ins \\ R[i-1,j-1] + sub \times (x[i] \neq y[j]) \end{cases}$$

procedure ODTWARZANIE ŚCIEŻKI
    $i \leftarrow n$
    $j \leftarrow m$
    aln1 $\leftarrow$ ""
    aln2 $\leftarrow$ ""
    while True **do**
        if i == 0 and j ==0 **then**
            Break
        else
            if i > 0 and j > 0 and $R[i, j] == R[i-1, j-1] + sub \times (x[i] \neq y[j])$ **then**
                aln1 = $x[i]$ + aln1, aln2 = $y[j]$ + aln2,$i \leftarrow i-1$, $j \leftarrow j-1$
            else
                if i > 0 and $R[i, j] == R[i-1, j] + ins$ **then**
                    aln1 = "-" + aln1, aln2 = $y[j]$ + aln2, $j \leftarrow j-1$
                else
                    aln1 = $x[i]$ + aln1, aln2 = "-" + aln2, $i \leftarrow i-1$

Złożoność obliczeniowa:
$O(n^2)$ ← Iteracja po wszystkich komórkach macierzy R i zapełnienie jest wartościami kosztu
Złożoność pamięciowa:
$O(n^2)$ ←macierz o rozmiarach $n \times m$

3 Opis programu

Po wczytaniu sekwencji (dostępne 3 sposoby wczytywania) użytkownik może kliknąć przycisk global alignment (przycisk local pełni funkcję placeholder'a), co w przypadku poprawnego wczytania sekwencji otworzy drugie okno Fig. 1. Tam użytkownik ma możliwość podać koszt każdej z możliwych dróg do przejścia z jednej komórki macierzy do sąsiedniej. Aby zostały wykonane obliczenia muszą być podane wszystkie 4 koszty. W innym wypadku zostanie wyświetlony odpowiedni komunikat Fig. 2.
Po wykonaniu obliczeń podstawowe informacje o uzyskanym rezultacie zostają wyświetlone w zielonym polu. Użytkownik ma również możliwość zapisania pełnej informacji do pliku tekstowego Fig. 4. Dodatkowo zostaje wyświetlona mapa cieplna przedstawiająca punktacje macierzy dopasowania, wraz z zaznaczoną optymalną ścieżką Fig. 3.

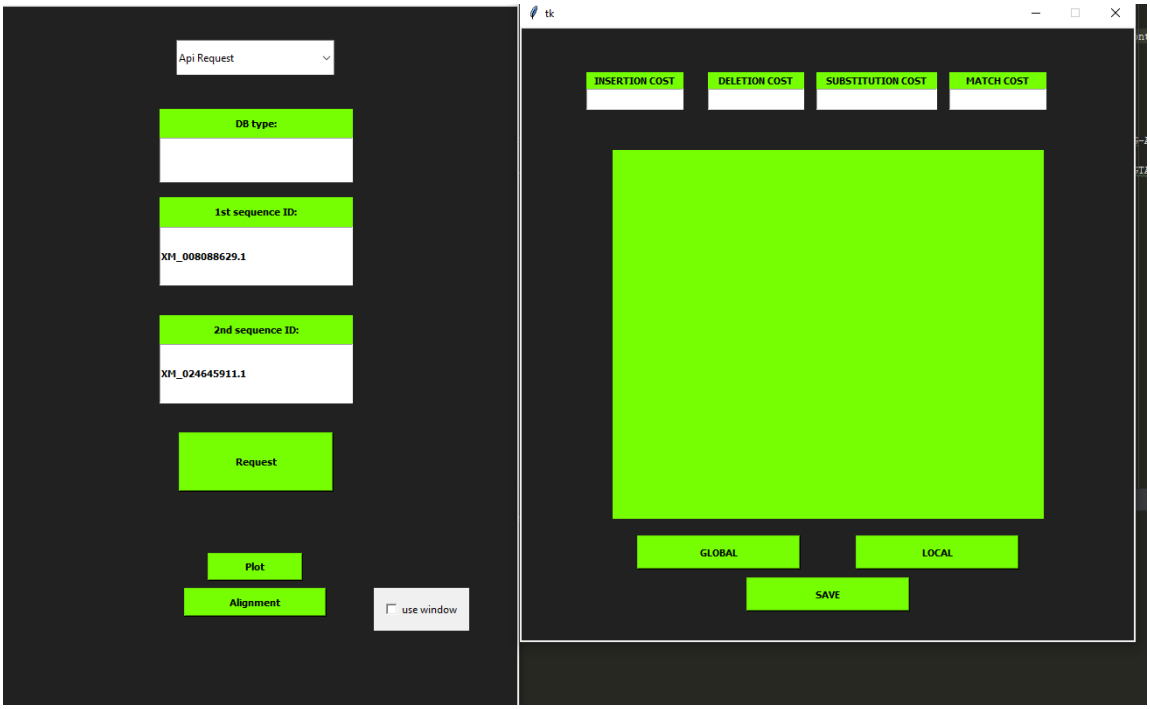
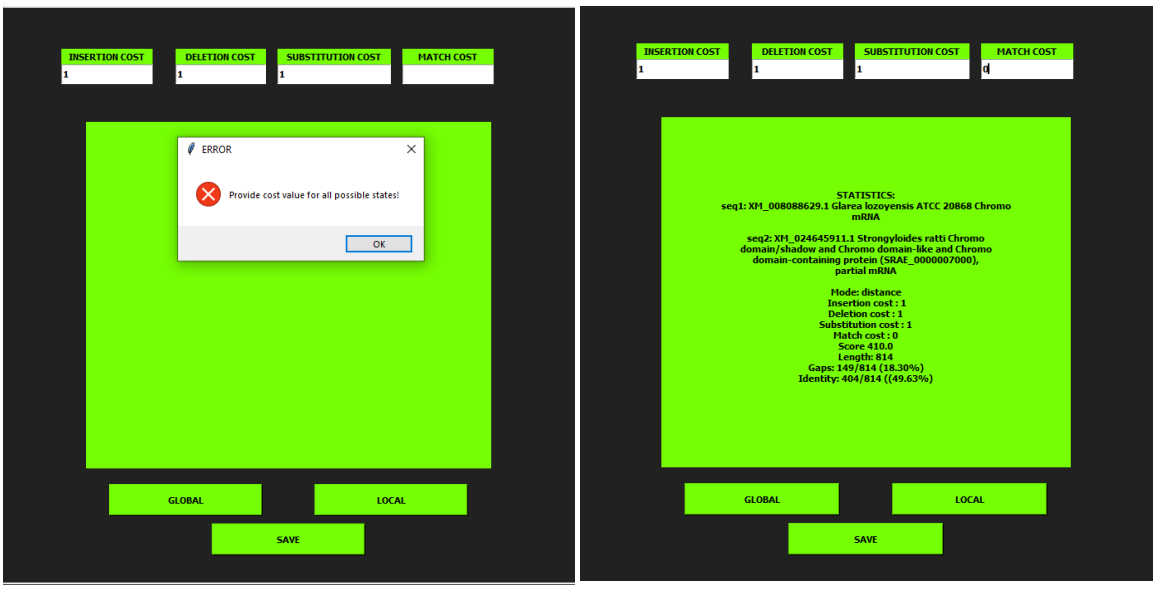


Figure 1: Okno dopasowania globalnego



(a) Komunikat błędu - nie podano wszystkich kosztow

(b) Wynik obliczeń

Figure 2: Prezentacja programu

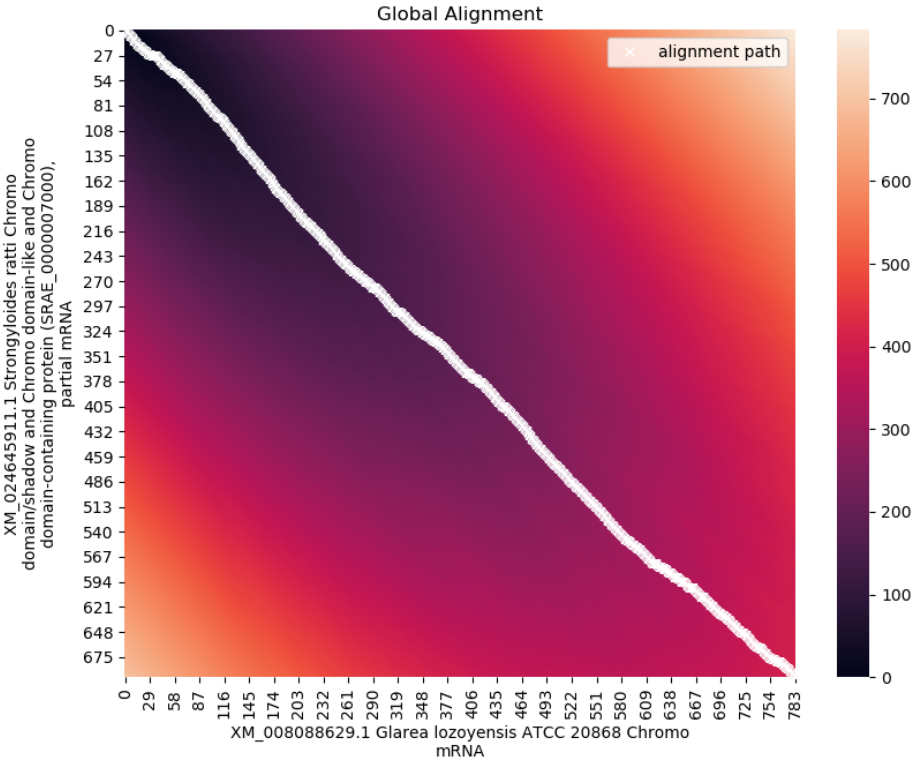


Figure 3: Mapa cieplna kosztów z zaznaczona optymalną ścieżką

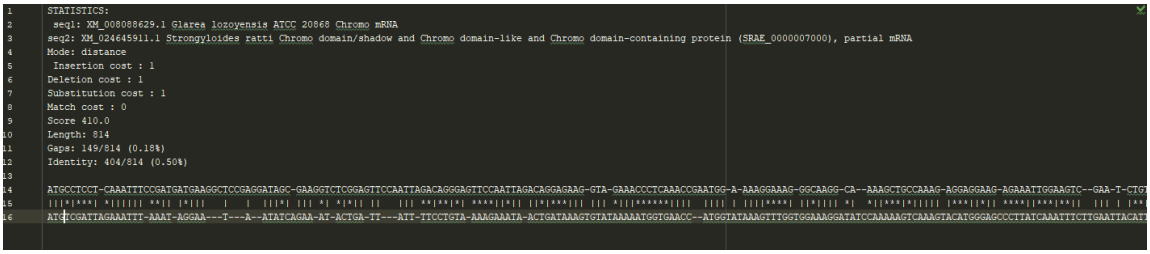


Figure 4: Format zapisu wyników do pliku tekstowego

4 Porównanie sekwencji

4.1 Ewolucyjnie powiązanych

seq1: pdb|1ASY|S Chain S, Class Ii Aminoacyl Transfer Rna Synthetases: Crystal Structure Of Yeast Aspartyl-Trna Synthetase Complexed With Trna Asp

ser2: pdb|1ASY|R Chain R, Class Ii Aminoacyl Transfer Rna Synthetases: Crystal Structure Of Yeast Aspartyl-Trna Synthetase Complexed With Trna Asp

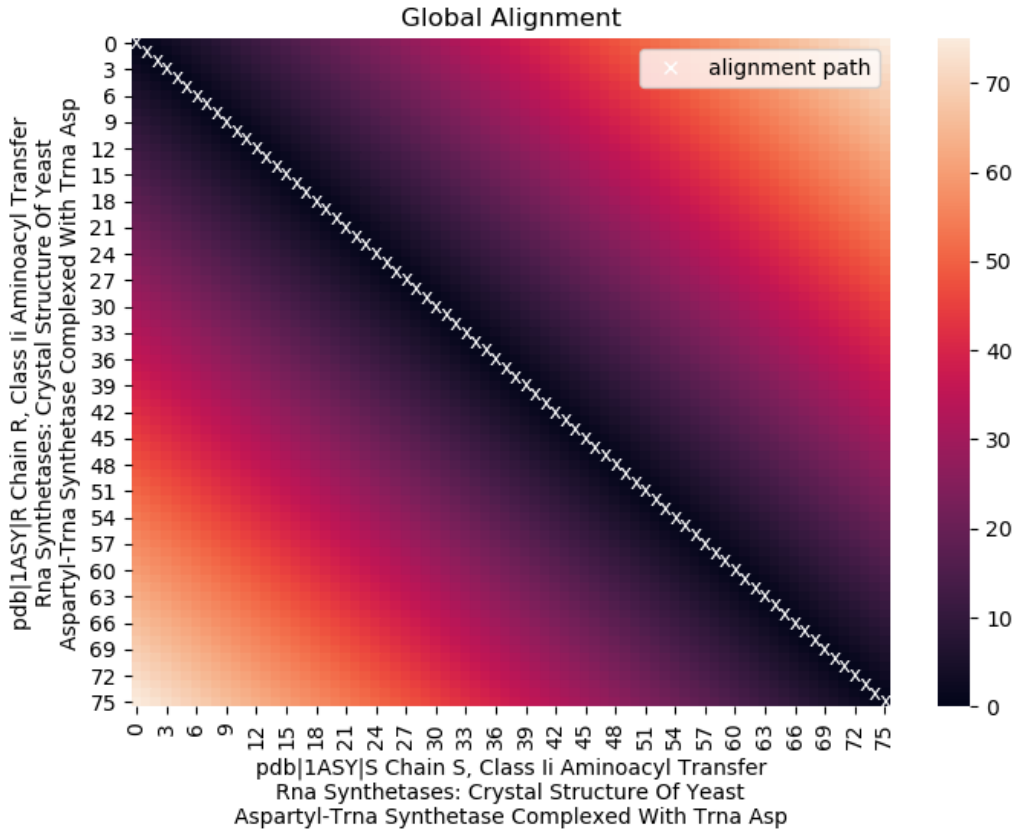


Figure 5: Mapa cieplna kosztów z zaznaczona optymalną ścieżką

Figure 6: Okno dopasowania globalnego

Jak widać jest to ta sama sekwencja. Wskazuje na to score = 0 oraz identity 100%. Optymalna ścieżka przebiega dokładnie po przekątnej. (Ta sama para była porównywana na 1 liście). Przyglądając się mapie cieplnej można zaobserwować symetrię kosztu względem przekątnej macierzy dla tego przypadku.

seq1: Human alphoid repetitive DNA sequence (alpha-R1-110, 2'monomer).
seq2: Chimpanzee alphoid repetitive DNA sequence (C-alpha-RI(680),104,2'monomer I)

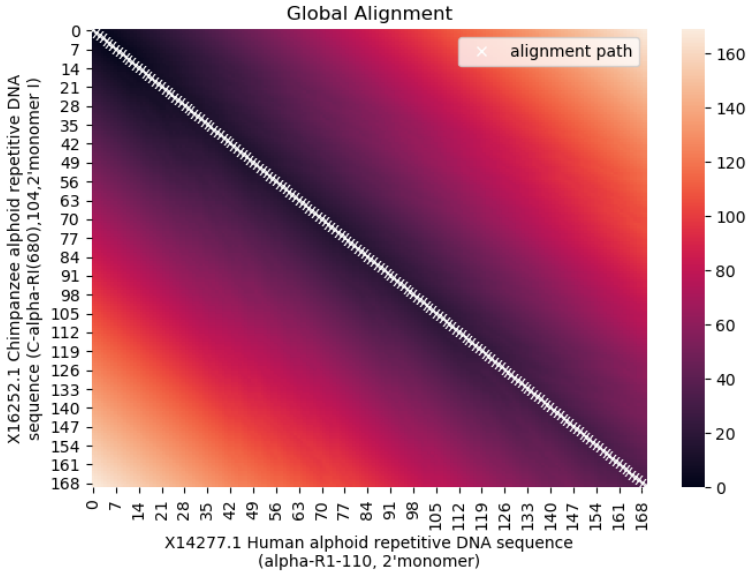


Figure 7: Mapa cieplna kosztów z zaznaczona optymalną ścieżką

Figure 8: Okno dopasowania globalnego

Tutaj również widać duże podobieństwo. W wyniku dopasowania uzyskano identity ok. 81% z 0 przerwami. Dla długości 169 uzyskano wynik 32. Przy braku przerw oznacza to, że tylko 32 pary były różne.

Dla porównania w następującym doświadczeniu ustawiono koszt delecji i insercji na 1 a koszt substytucji 10 krotnie większy. W wyniku tego algorytm preferował wstawienie przerwy, co można zaobserwować na poniższych wykresach. Również w wyniku tego widać ze ścieżka nie przebiega już równo po przekątnej. Wystąpiło aż 54 przerw.

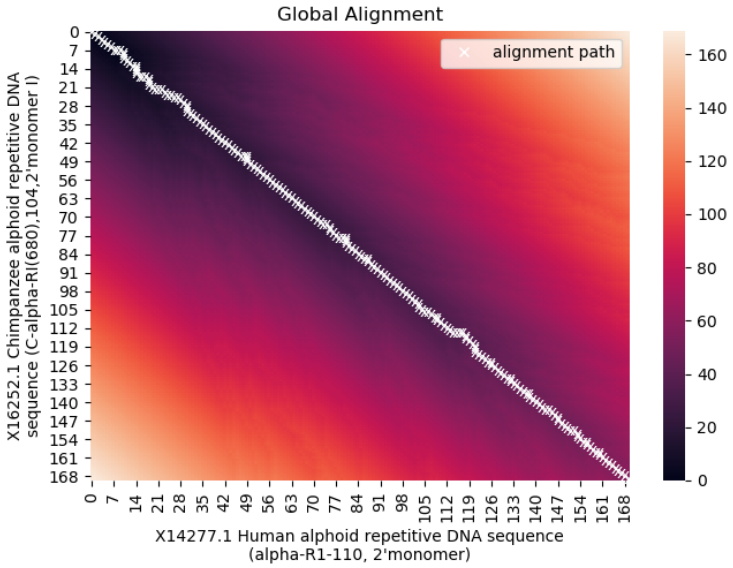


Figure 9: Mapa cieplna kosztów z zaznaczona optymalną ścieżką

Figure 10: Okno dopasowania globalnego

seq1: Human chromosome 4 sequence-tagged site STS4-85, sequence tagged
seq2: Mus musculus chromosome 4 map between D4Mit178 and D4Mit7, sequence tagged site

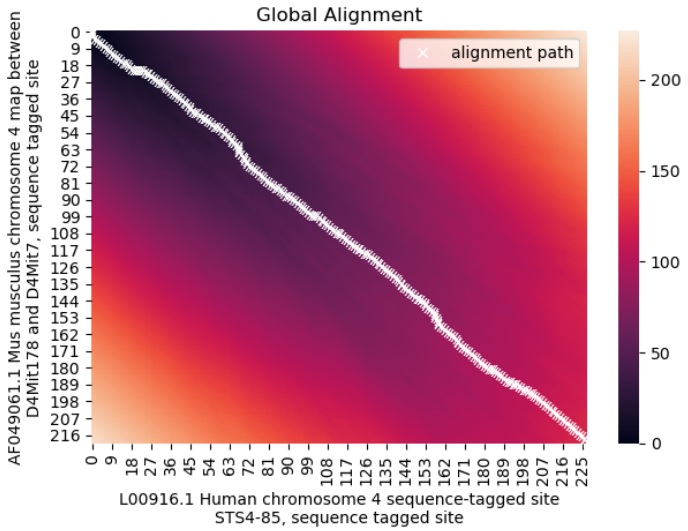


Figure 11: Mapa cieplna kosztów z zaznaczona optymalną ścieżką

Figure 12: Okno dopasowania globalnego

W tym przypadku są analizowane małe fragmenty chromosomu 4 człowieka i myszy. Wyniki wskazują tozsamość na poziomie 50% i ilość przerw wynoszącą ok. 17%. Biorąc pod uwagę, że są to małe fragmenty danego genu , nie należy oczekiwać idealnego dopasowania.

Natomiast jeżeli uwzględnimy, że substytucja ma większe prawdopodobieństwo zajścia niż insercja/delecja to należy ustawić inne wartości kosztu. Przy zaproponowanych poniżej wartościach uzyskujemy inny wynik. Zwiększenie wartości kosztu mniej prawdopodobnych przypadkow o 1 prowadzi do spadku ilości przerw z ok. 17% do ok. 5,5%. Jednocześnie spada również wartość identity z około 50% do 43%. Również można zauważyć, że w tym przypadku ścieżka dopasowania bardziej pokrywa się z przekątną macierzy.

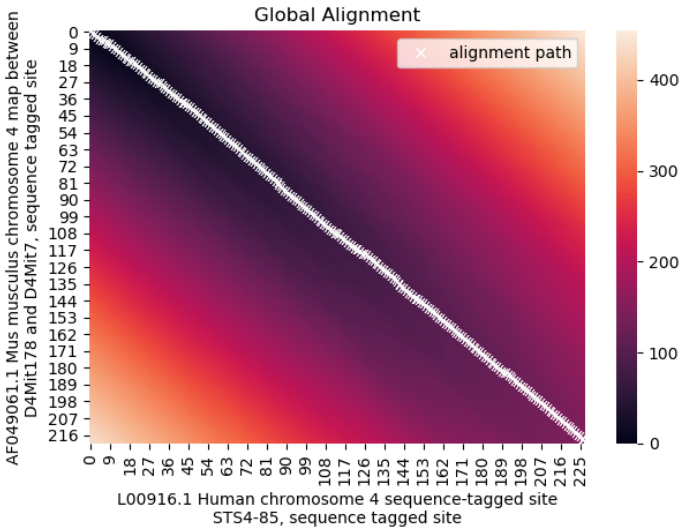


Figure 13: Mapa cieplna kosztów z zaznaczona optymalną ścieżką

Figure 14: Okno dopasowania globalnego

Jeżeli jeszcze bardziej obciążymy insercje oraz delecje wyższym kosztem (co jest bliższe sytuacji w przyrodzie) to uzyskamy następujące wyniki. Zwiększył się score z wartości 122 dla równych kosztów oraz z 143 dla poprzedniego przypadku. W tym eksperymencie wyniósł on 172. Jednocześnie znowu zmalała wartość przerw (ok. 3%) oraz wartość identity (ok.36,5%). Wynika z tego, że te sekwencje nie są zgodnie dopasowane, ale należy pamiętać że są to małe fragmenty, które odpowiadają innym miejscom w kodzie tego genu.

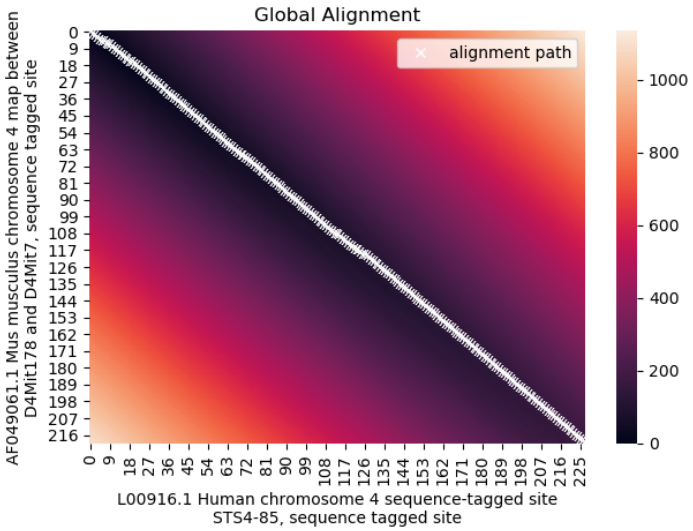


Figure 15: Mapa cieplna kosztów z zaznaczona optymalną ścieżką

Figure 16: Okno dopasowania globalnego

4.2 Ewolucyjnie niepowiązanych

seq1: Human peripheral and cord blood Homo sapiens STS genomic, sequence tagged site.
seq2: Pan troglodytes chromosome 8 genomic scaffold.

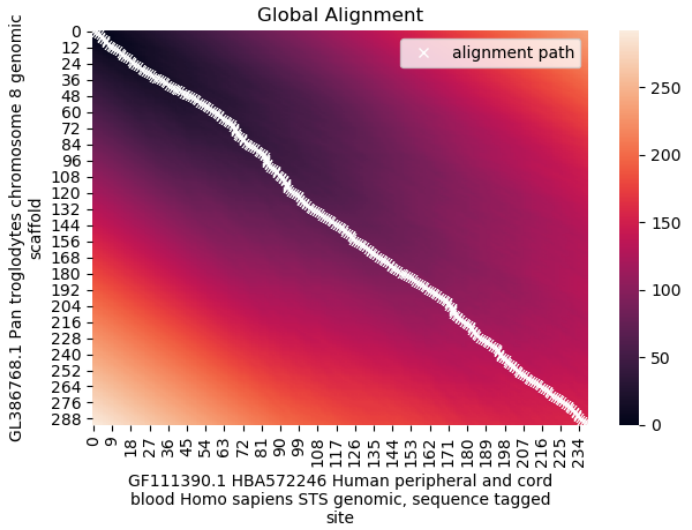


Figure 17: Mapa cieplna kosztów z zaznaczona optymalną ścieżką

Figure 18: Okno dopasowania globalnego

Sekwencje dotyczą 2 różnych genów. Algorytm próbuje je jak najbardziej dopasować do siebie poprzez minimalizację odległości, uzyskując wynik 154 dla długośći 296. W następnym przykładzie algorytm spróbuje dopasować te sekwencje do siebie ale z kosztem insercji/delecji równej 10.

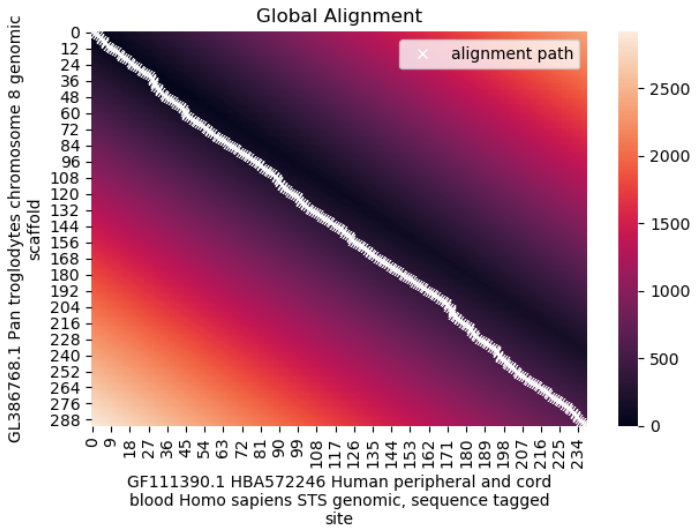


Figure 19: Mapa cieplna kosztów z zaznaczona optymalną ścieżką

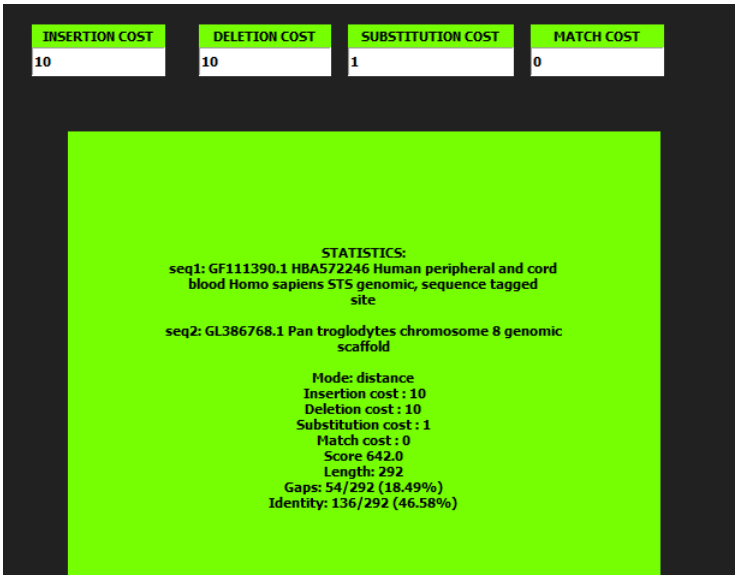


Figure 20: Okno dopasowania globalnego

Pomimo drastycznego zwiększenia kosztu insercji/delecji uzyskano podobną ilość przerw (62 oraz 54) i wynik identity(ok. 48% oraz 45.6%). Taka ilość przerw i substytucji oznacza, że te sekwencje nie są podobne.

seq1: Anti CDH6 antibodies and anti CDH6 antibody drug conjugates.
seq2: Pan troglodytes chromosome 8 genomic scaffold.

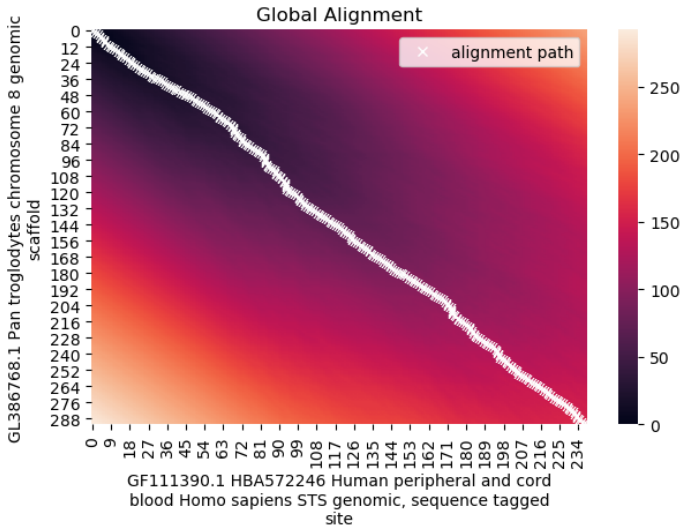


Figure 21: Mapa cieplna kosztów z zaznaczona optymalną ścieżką

Figure 22: Okno dopasowania globalnego

Jedna sekwencja dotyczy 8 chromosomu szympansa, a druga to sztucznie wytworzona w laboratorium dotycząca pewnych antyciał. Spora ilość przerw (ok. 21%) oraz wartość identity wynosząca ok. 47% wskazuje na brak podobieństwa pomiędzy tymi sekwencjami.