Bioinformatyka Sprawozdanie 1

Dopasowanie par sekwencji - algorytm kropkowy

Michał Marciniak 244811

1 Kod do repozytorium

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

2 Analiza złożoności obliczeniowej czasowej i pamięciowej:

Algorithm 1 Algorytm Kropkowy

		operations number n
1:	**procedure** INICJALIZACJA	
2:	$m \leftarrow$ długość sekwencji x	1
3:	$n \leftarrow$ długość sekwencji y	1
4:	$R \leftarrow$ macierz o wymiarach $n \times m$	1
5:	**for** $i \leftarrow 1$ *to* n **do**	n
6:	**for** $j \leftarrow 1$ *to* m **do**	m
7:	**if** $x[i] == y[i]$ **then**	1
8:	$R[i, j] = 1$	1
9:	**else**	1
10:	$R[i, j] = 0$	1

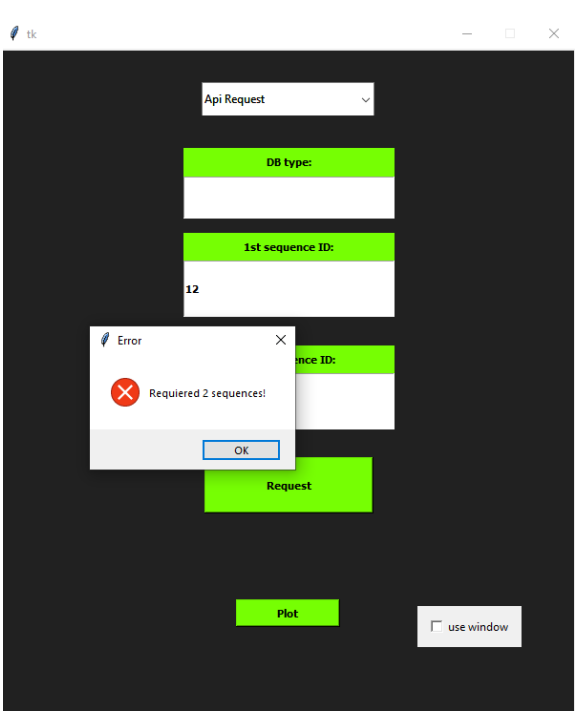
Złożoność obliczeniowa: $3 + n \times (3 + m \times (2 + 2)) = 3 + 3 \times n + 4 \times n \times m \rightarrow O(n^2)$
Złożoność pamięciowa: $O(n^2) \leftarrow$ tablica o rozmiarach $n \times m$

3 Prezentacja programu

Główne okno programu umożliwia wybór metody za pomocą której chcemy wczytać sekwencje.

Figure 1: Główne okno programu

Decydując się na wczytanie pliku FASTA, należy upewnić się, że zawiera on 2 sekwencje. Inaczej zostanie wyświetlony komunikat o niepoprawnej zawartości pliku. Analogicznie w przypadku wysyłania zapytania przez API. Jeżeli w zwróconej odpowiedzi nie są zwarte 2 sekwencje użytkownik zostanie o tym poinformowany za pomocą komunikatu.
W przypadku ręcznego wpisywania sekwencji użytkownik ograniczony jest do korzystania tylko z znaków 'ATCG'. W przypadku złamania tego ograniczenia zostaje wyświetlony komunikat informujący o błędzie.
W przypadku próby narysowania wykresu kropkowego bez podania poprawnej sekwencji również zostaje wyświetlony komunikat informujący o niepoprawnej sekwencji.
Po zaznaczeniu opcji 'use window' zostają wyświetlone kontrolki odpowiedzialne za wybór parametrów okna filtrującego.



(a) Komunikat błędu niepoprawnej sekwencji

(b) Komunikat błędu odnoszący się do użytych znaków

Figure 2: Komunikaty błędu

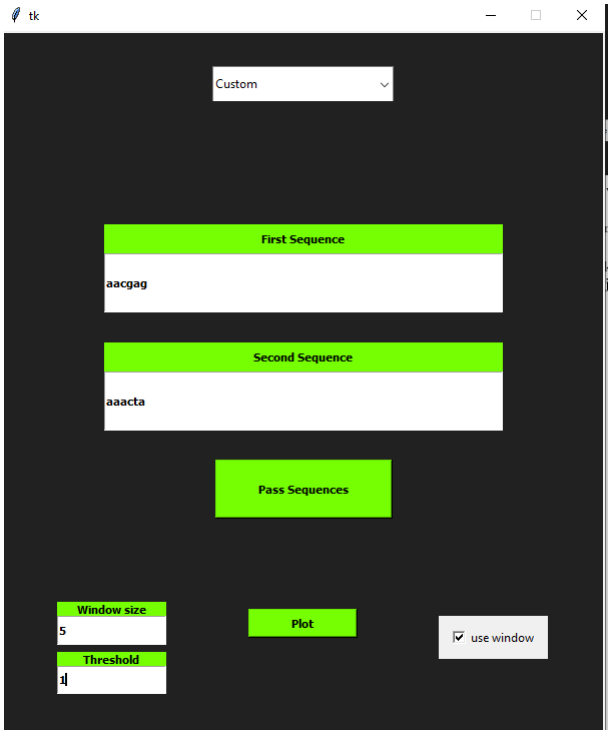


Figure 3: Wybór parametrów okna filtrującego

4 Porównanie par

4.1 Ewolucyjnie powiązanych

X: S Chain S, Class Ii Aminoacyl Transfer Rna Synthetases: Crystal Structure Of Yeast Aspartyl-Trna Synthetase Complexed With Trna Asp
Y: R Chain R, Class Ii Aminoacyl Transfer Rna Synthetases: Crystal Structure Of Yeast Aspartyl-Trna Synthetase Complexed With Trna Asp

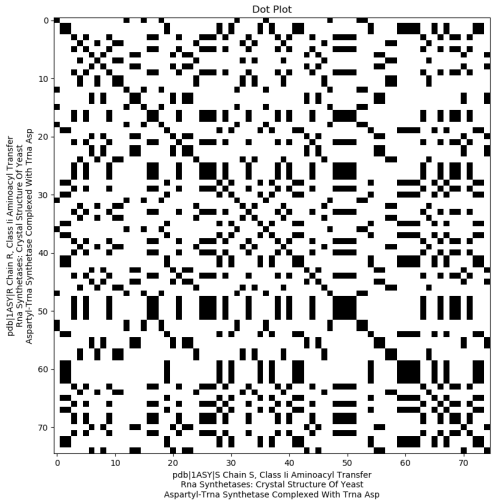


Figure 4: Macierz kropkowa podanych sekwencji

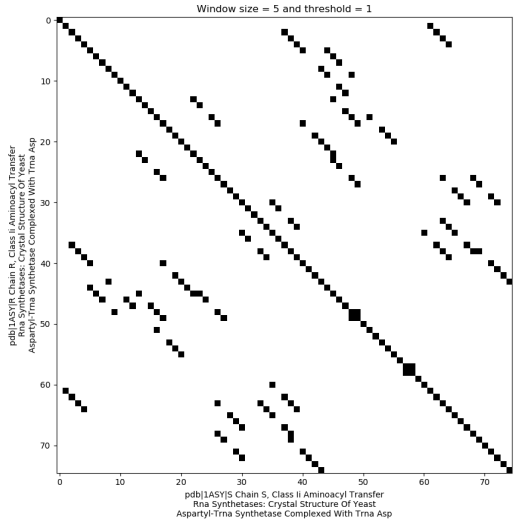


Figure 5: Macierz kropkowa po zastosowaniu okna o rozmiarze 5 i wartości progowej 1

Obie sekwencje są takie same, co widać patrząć na przekątną macierzy. Dodatkowo występują powtarzające się fragmenty w obrębie całej sekwencji, co objawia się na powyższym wykresie jako linie równoległe do przekątnej.

X: Homo sapiens proline dehydrogenase 2 (PRODH2), transcript variant 1, mRNA
Y: Mus musculus proline dehydrogenase (oxidase) 2 (Prodh2), mRNA

(a) Macierz kropkowa podanych sekwencji

(b) Macierz kropkowa po zastosowaniu okna o rozmiarze 5 i wartości progowej 1

Figure 6: Macierze kropkowe

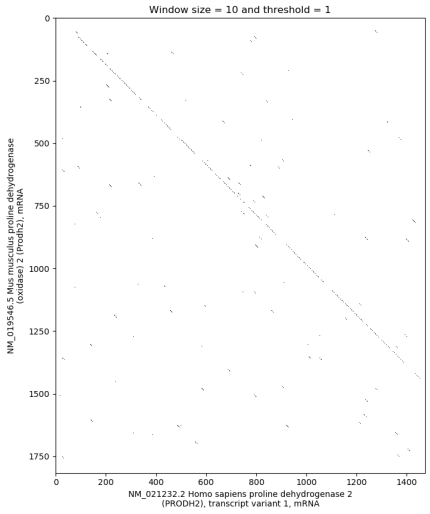


Figure 7: Macierz kropkowa po zastosowaniu okna o rozmiarze 10 i wartości progowej 1

Obie sekwencje kodują dehydrogenazę prolinową PRODH2. Jedna pochodzi od człowieka a druga od myszy. Jak widać po przefiltrowaniu oknem o rozmiarze 10×10 i progu 1, wykazują one podobieństwo (przekątna macierzy). Dodatkowo widać jak duży wpływ na prezentację wyników ma dobór rozmiaru okna filtrującego. Po przybliżeniu na przekątną macierzy kropkowej można dostrzec miejscowe mutacje objawiające się przerwą w ciągłości linii.

X: V(D)J recombination-activating protein 1 isoform X1 [Pan troglodytes]
Y: V(D)J recombination-activating protein 1 [Danio rerio]

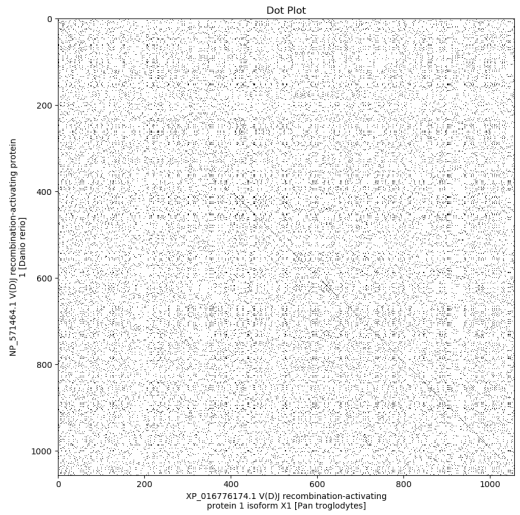


Figure 8: Macierz kropkowa podanych sekwencji

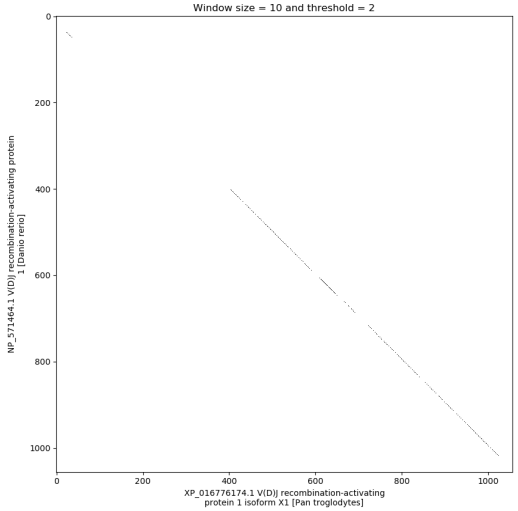


Figure 9: Macierz kropkowa po zastosowaniu okna o rozmiarze 10 i wartości progowej 2

Obie sekwencje kodują gen RAG1. Pierwsza pochodzi od szympansa a druga od ryby zwanej Danio pręgowany. Obie sekwencje wykazują duże podobieństowo. Analizując wykres kropkowy wygląda na to że początkowe fragmenty obu sekwencji są zupełnie różne, ale potem zaczyna się dłuższy fragment identyczny w obu sekwencjach.

X: GATOR complex protein MIOS isoform 1 [Homo sapiens].
Y: GATOR complex protein MIOS [Gallus gallus]

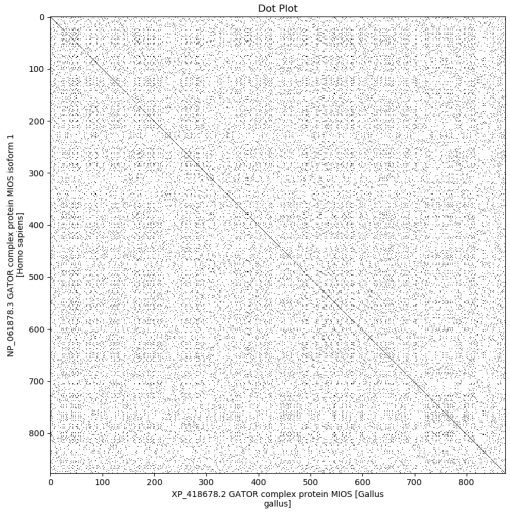


Figure 10: Macierz kropkowa podanych sekwencji

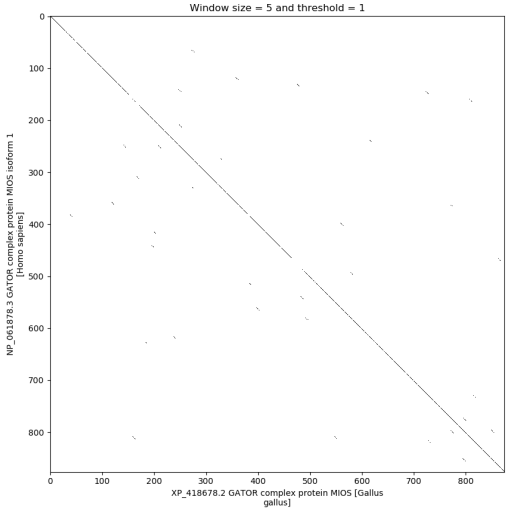


Figure 11: Macierz kropkowa po zastosowaniu okna o rozmiarze 5 i wartości progowej 1

Obie sekwencje wykazują znaczące podobieństowo. Mocno zarysowana przekątna jest widoczna nawet przed procesem filtracji przy pomocy okna. W kilku fragmentach sekwencji musiało dojść do mutacji punktowej, ponieważ na wykresie występują przerwania w ciągłąści linii.

X: Campylobacter coli strain cco117 putative NADP-dependent alcohol dehydrogenase (nadP) gene, partial cds
Y: Glarea lozoyensis ATCC 20868 Chromo mRNA

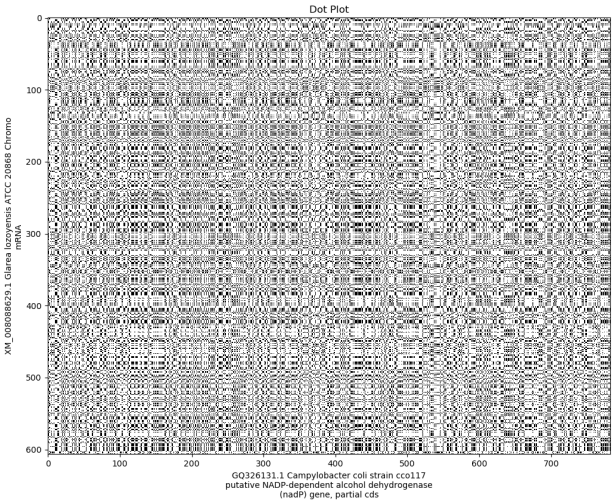


Figure 12: Macierz kropkowa podanych sekwencji

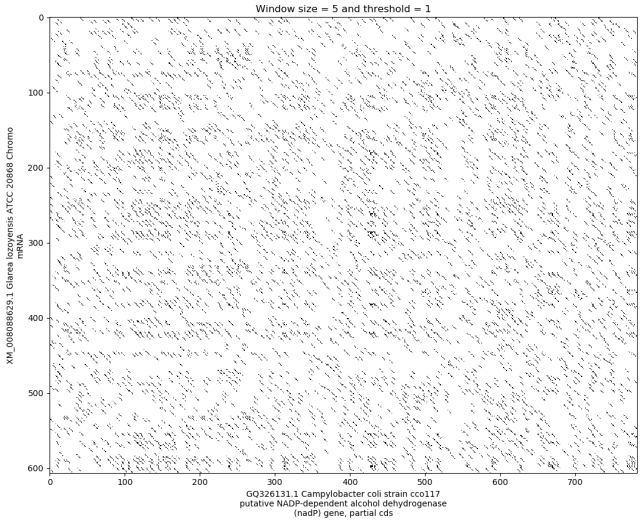


Figure 13: Macierz kropkowa po zastosowaniu okna o rozmiarze 5 i wartości progowej 1

Sekwencje nie są powiązne ewolucyjnie. Wykres kropkowy nie ukazuje żadnych znaczących wzorców. Ciemne obszary zdają się występować losowo. Przy zastosowaniu większego okna można jeszcze bardziej wyczyścić macierz z takich losowych skupisk, co zostało pokazane niżej.

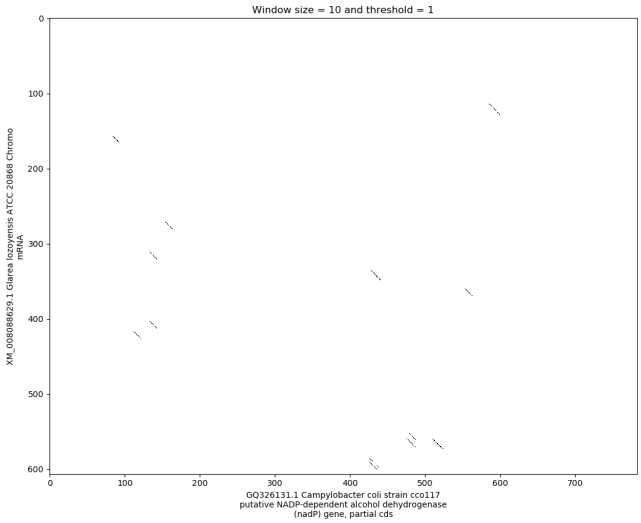


Figure 14: Macierz kropkowa po zastosowaniu okna o rozmiarze 10 i wartości progowej 1

Po zastosowaniu surowszego kryterium filtracji okazuje się, że obie sekwencje posiadają mało rejonów wskazujących na jakiekolwiek podobieństwo.
Widoczne fragmenty zdają się być losowe biorąc pod uwagę rozmiar sekwencji.

X: Escherichia coli strain RF5A NODE 647 length 1892 cov 15.2312, whole genome shotgun sequence
Y: Synthetic construct Homo sapiens clone CCSBHm 00015721 ASPA (ASPA) mRNA

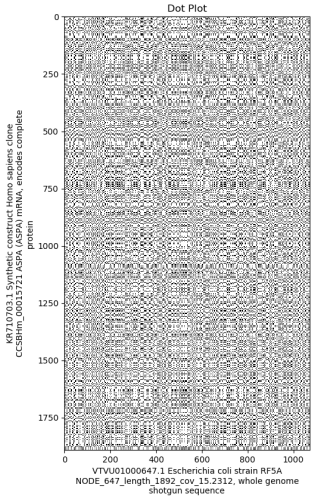


Figure 15: Macierz kropkowa podanych sekwencji

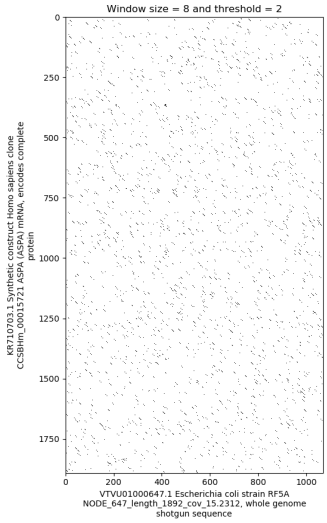


Figure 16: Macierz kropkowa po zastosowaniu okna o rozmiarze 8 i wartości progowej 2

Tutaj również obie sekwencje nie wykazują istotnego podobieństwa.

5 Wnioski

Algorytm kropkowy w prosty sposób umożliwia nam wizualne porównanie dwóch sekwencji w celu wykrycia wzorców świadczących o ich podobieństwie. W przypadku długich sekwencji wykres staję się bardziej zaszumiony i ciężko coś z niego odczytać. W takim przypadku należy analizować poszczególne fragmenty wykresu w powiększeniu, oraz stosować filtrację przy użyciu okna.
Poprzez zwiększanie rozmiaru okna i jednoczesne zmniejszanie wartości progowej sprawiamy, że filtr jest bardziej krytyczny w uznaniu podobieństwa pomiędzy sekwencjami. Umożliwia nam to na oczyszczanie wykresu z losowych wzorców.