Wstęp do bioinformatyki

Nr ćwiczenia: 2

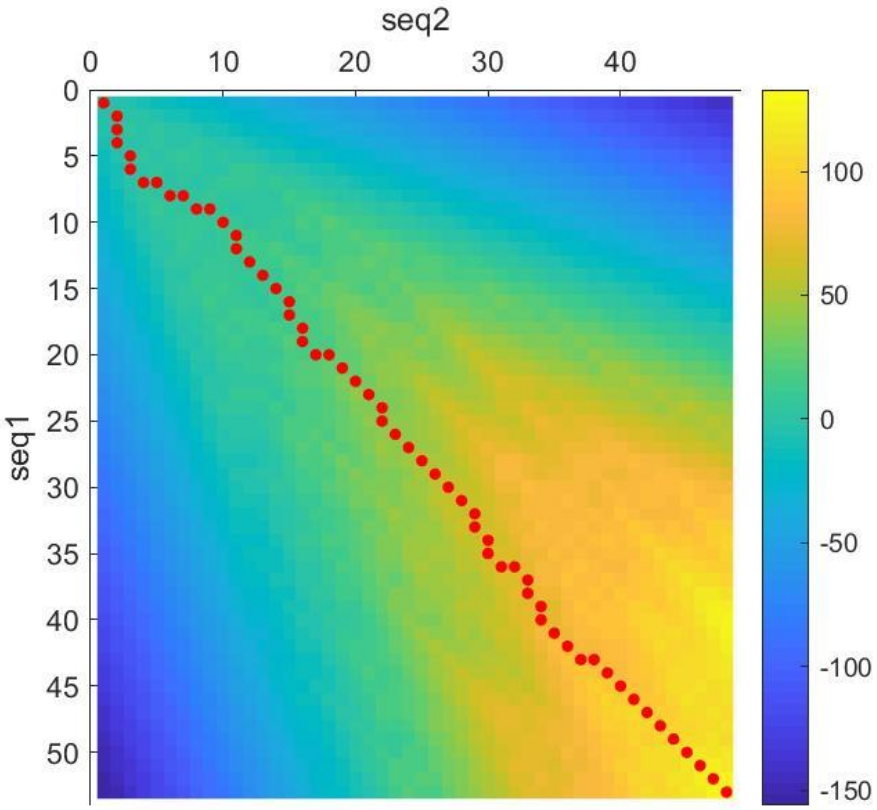
Temat ćwiczenia: Dopasowanie globalne par sekwencji

Nazwisko i Imię prowadzącego kurs: dr inż. Witold Dyrka

Wykonawcy:	
Imię i Nazwisko **Nr indeksu, wydział**	**Edyta Krukowska 217097, WPPT**
Termin zajęć: dzień tygodnia, godzina	**Piątek 11.15**
Data oddania sprawozdania	**04.02.2019 /poprawa 18.04**

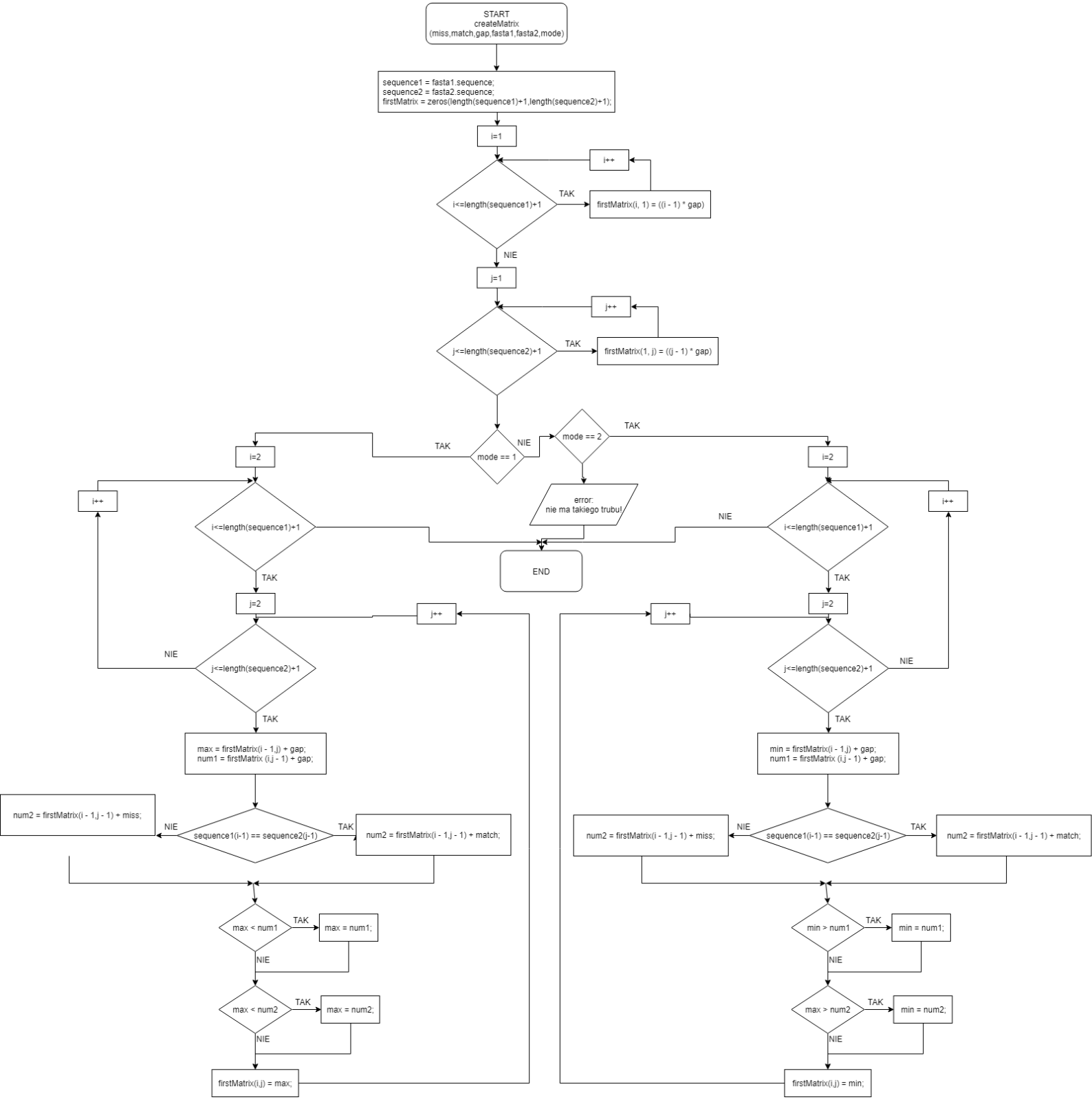
Repozytorium: https://github.com/Edie1995/Bioinformatyka/tree/zad2

1. **Prezentacja działania programu**:

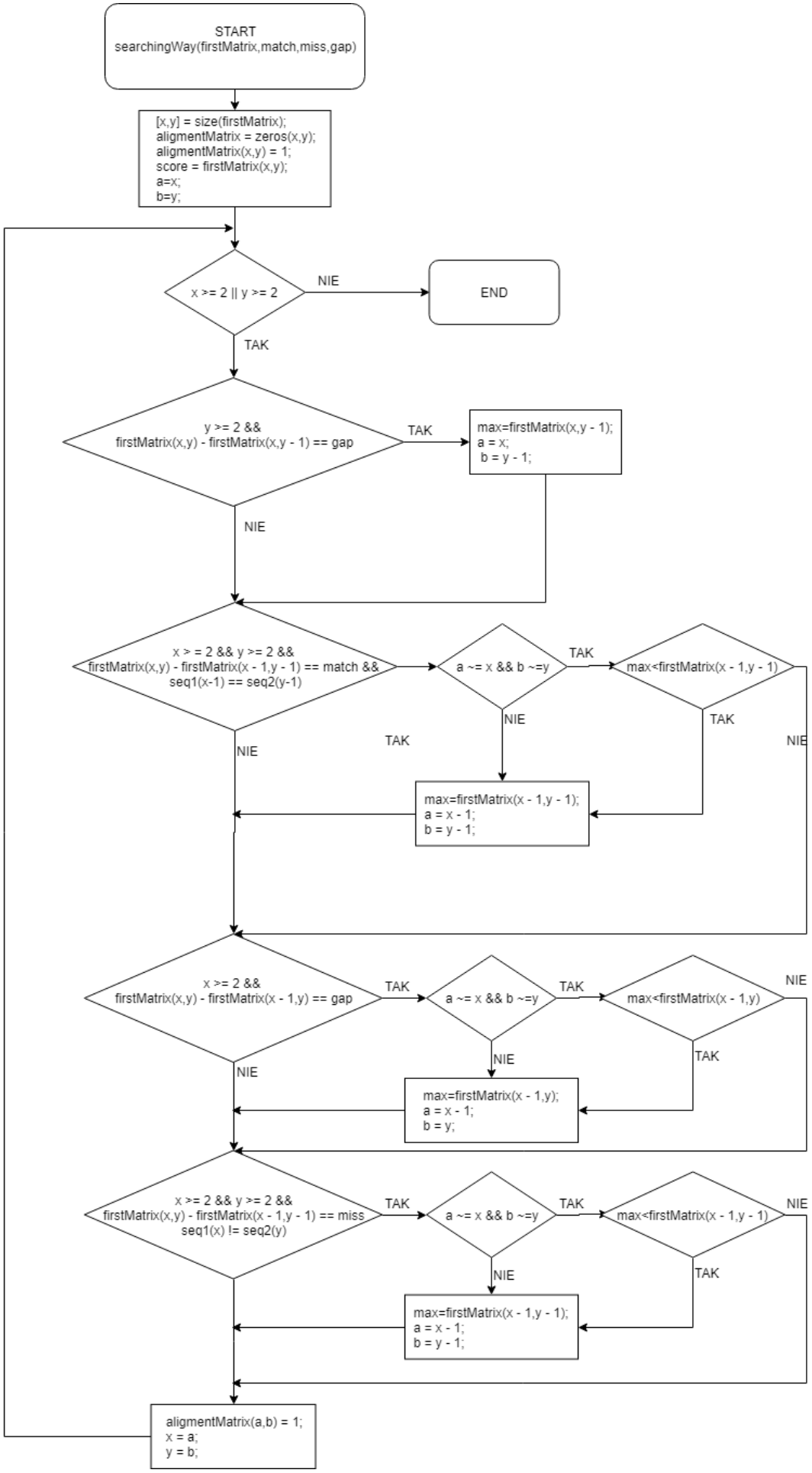


Rysunek 1 Przykład działania programu, dla przykładowych sekwencji wpisanych z klawiatury

2. Schematy blokowe algorytmu dopasowania



Rysunek 2 Schemat blokowy - worzenie macierzy kosztu



Rysunek 3 Schemat blokowy - worzenie ścieżki optymalnego dopasowania sekwencji

3. **Oszacowanie złożoności czasowej obliczeniowej i pamięciowej kodu poszczególnych funkcji i całego programu:**
   - **Czasowe**
     - checkFile:

     Funkcja ta zawiera jedną pętlę o rozmiarze m lub n, czyli jest maksymalnie rzędu O(m) lub O(n)
     - createMatrix:

     jedno przypisanie *n, drugie *m, jedno przypisanie *m*n
     jedna pętla*n, druga*m, największa pętla for :m*n; funkcja maksymalnie rzędu m*n, czyli O(mn)
     - readFasta: ponieważ brana jest pod uwagę ilość linii a nie znaków, maksymalnie rzedu o(mn)
     - searchingWay: jedno przypisanie wartości m*n, jedna pętla m*n, funkcja maksymalnie rzędu mn, O(mn)
     - writeSequence: jedno przypisanie m, jedno przypisanie rzędu n, pętla o rozmiarach m*n, funkcja maksymalnie rzędu O(mn)

   Pozostałe funkcje programu są stałe O(1) lub liniowe O(m) lub O(n), wynika z tego, że cały program jest maksymalnie rzędu c*m*n (c>0), O(mn).

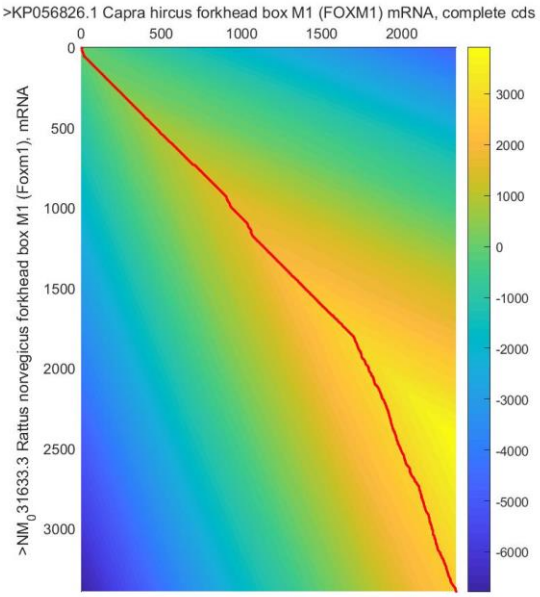
   - **Pojemnościowe**

     Zmienna o maksymalnych rozmiarach w programie wynosi m*n pozostałe zmienne są rzędu o(mn) (o(mn) < O(mn)). Ponieważ najwyższa zmienna jest rozmiarów mn, cały program to c*m*n (c>0), cały program: O(mn).

4. Porównanie przykładowych par sekwencji

4.1. Ewolucyjnie powiązanych

- **Porównanie sekwencji szczura z kozą**

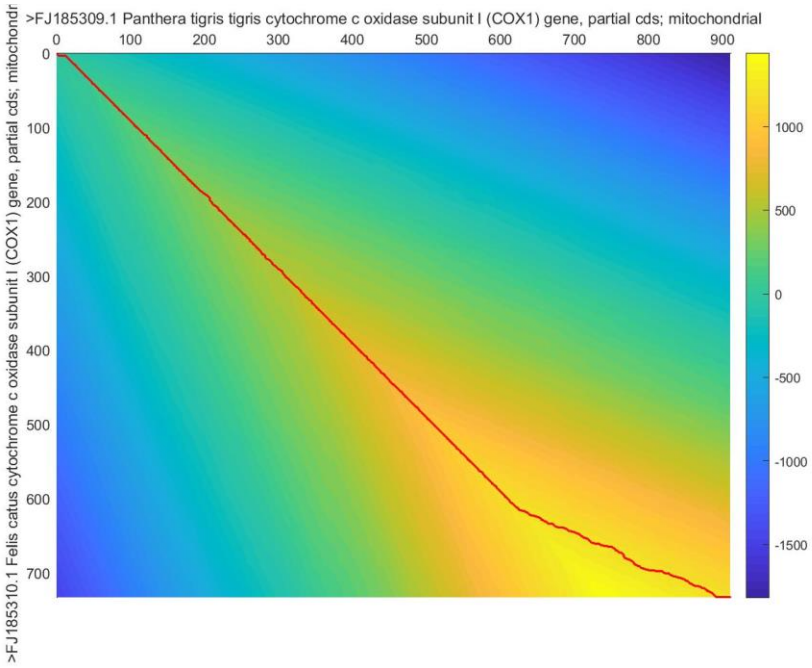


Rysunek 4 Macierz kosztu z naniesioną ścieżką optymalnego dopasowania porównująca M1 mRNA szczura (oś y) z M1 mRNA kozy (oś x)

```
#1: >NM_031633.3 Rattus norvegicus forkhead box M1 (Foxm1), mRNA
#2: >KP056826.1 Capra hircus forkhead box M1 (FOXM1) mRNA, complete cds
#Mode: similarity
#Match: 3
#Mismatch: -5
#Gap: -2
#Score: 2615
#Length: 3721
#Indentity: 1496/3721(40%)
#Gaps: 1709/3721(46%)
GCCTGGCTCGGCCCCGCGTGGAGCAGCG-GTGGCCTGTGAGGGTCAAAGCTTGTGA-T-TCTCG-ATGGAGAGTGAAAGCACAG-CTTCATGATGAGA-ACCAGC
  |                  |  |          |  |        |      ||    ||   ||||||||||||||||||   |||||||||  | ||||||
--C-----------------TA-GA---GA----TT-GG-----A--A------GATCTGCT-CAATGGAGAGTGAAAGCACA-GATTCATGATG-AAAACCAGC
```

Rysunek 5 Konsolowy wynik działania programu dla Szczura i Kozy

- **Porównanie kota z panterą**



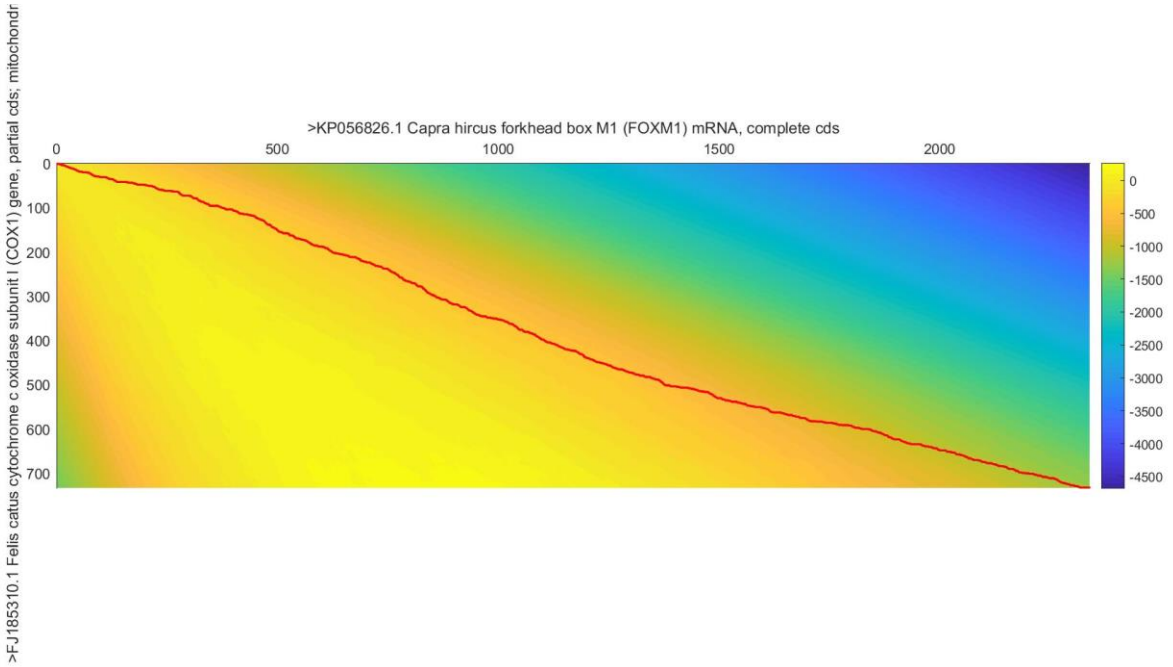
Rysunek 6 Macierz kosztu z naniesioną ścieżką optymalnego dopasowania porównująca cytochrom c kota (oś y) z cytochromem c pantery (oś x)

```
#1: >FJ185310.1 Felis catus cytochrome c oxidase subunit I (COX1) gene, partial cds; mitochondrial
#2: >FJ185309.1 Panthera tigris tigris cytochrome c oxidase subunit I (COX1) gene, partial cds; mitochc
#Mode: similarity
#Match: 3
#Mismatch: -5
#Gap: -2
#Score: 1191
#Length: 1003
#Indentity: 511/1003(51%)
#Gaps: 363/1003(36%)
A---------TGTTCATAAACC-GTTGACTATTTTCAAC-TAATCACAAA-GATATTGGT-ACTCTTTACCTTTTATTTGG-TGCCTGA-GCTGG-CATGGTG
|         ||||||||||||   ||||||||||||||   |||||||  | |||||||   ||||||||||||||||||||   ||||   ||||   ||||||
ATTTTACCTATGTTCATAAACCGC-TGACTATTTTCAA-CCAATCACA-AGGATATTG-GAACTCTTTACCTTTTATTTG-GCGCCT-GGGCTG-GTATGGTG
```

Rysunek 7 Konsolowy wynik działania programu dla Kota i Pantery

4.2. Ewolucyjnie niepowiązanych

- **Porównanie cytochromu c kota z M1 mRNA kozy**



Rysunek 9 Macierz kosztu z naniesioną ścieżką optymalnego dopasowania wygenerowana dla cytochromu c kota (oś y) oraz M1 mRNA kozy (oś x)

```
#1: >FJ185310.1 Felis catus cytochrome c oxidase subunit I (COX1) gene, partial cds; mitochondrial
#2: >KP056826.1 Capra hircus forkhead box M1 (FOXM1) mRNA, complete cds
#Mode: similarity
#Match: 3
#Mismatch: -5
#Gap: -2
#Score: -1275
#Length: 2376
#Indentity: 327/2376(14%)
#Gaps: 1680/2376(71%)
------A-T-----G-T-T--C-A--T-A---A---A-C-C-GT---TG------A---------C-----TAT--TTT--CA----A------C----------
      | |           | |                   |   ||      |               ||   ||  ||    |      |
CTAGAGATTGGAAGATCTGCTCAATGGAGAGTGAAAGCACAGATTCATGATGAAAACCAGCCCCCGTCGGCCATTGATTCTCAAAAGACGGAGGCTGCCCCTTCC
```

Rysunek 8 Konsolowy wynik działania programu dla Żaby szponiastej i Kozy

Podobieństwo sekwencji możemy interpretować na różne sposoby, trzeba przy tym pamiętać, że nie liczy się jedynie procentowa wielkość dopasowania dwóch sekwencji, bardzo ważnym czynnikiem jest tutaj również długość sekwencji, które porównujemy, oraz fakt, czy podobieństwa tworzą się grupami, czy losowo, na długości całego łańcucha. Na podstawie powyższych przykładów obserwujemy, że największy procent dopasowania występuje między kotem a panterą. Współczynnik ten byłby jeszcze większy, gdyby nie fakt, że długości wybranych łańcuchów znacznie się od siebie różnią, co powoduje konieczność początkowego przesunięcia krótszego łańcucha. Mimo różnic w długości, widzimy, że w przypadku tej pary udział dopasowania wynosi aż 51%. Należy również zauważyć, że do miejsca , w którym sekwencje zaczynają różnić się długością, uzyskujemy niemalże liniowe dopasowanie. Ponieważ długość porównywanych sekwencji w przypadku obu sekwencji wynosi ponad 800 elementów, mówimy tutaj o dopasowaniu znaczącym. Dodatkowo w przypadku tej pary widzimy, że sparowania występują grupami, stąd możemy wnioskować, że

podobieństwo to nie jest przypadkowe. Różnice między tymi sekwencjami, świadczą o mutacjach, które zaszły w genach tych organizmów po odłączeniu od wspólnego przodka.

Mniej oczywistym przykładem jeśli mówimy o wspólnym pochodzeniu, jest para szczur i koza. Obserwujemy jednak, że podobieństwo dopasowania wynosi aż 40%, mimo tego iż łańcuchy różnią się od siebie znacząco długością łańcucha próbek. Droga najlepszego dopasowania jest przybliżenie liniowa, dodatkowo obserwujemy grupy sparowanych nukleotydów, które mogą świadczyć o tym, że pary te są spokrewnione i podobieństwo poszczególnych nukleotydów w tym dopasowaniu nie jest przypadkowe. Ponieważ badamy sekwencje o długościach 2500-3500 próbek, podobieństwa te są znaczące i możemy uznać tę parę za ewolucyjnie powiązaną.

Najbardziej odległą genetycznie od siebie parą jest para dwóch różnych komórek cytochrom c kota oraz M1 mRNA kozy. W tym przypadku udział dopasowań w optymalnej ścieżce dopasowania tych sekwencji wynosi jedynie 14%. Na tej podstawie można by już wysunąć tezę, że sekwencje te różnią się od siebie genetycznie tak, jak jest to widoczne. Po dalszym przeanalizowaniu występowania tych dopasowań wzdłuż łańcucha widzimy, że nie tworzą one większych grup podobieństwa. W takim przypadku można powiedzieć, że jest do dopasowanie przypadkowe. Ponieważ analizujemy rozmieszczenie 4 nukleotydów wzdłuż całego łańcucha, prawdopodobieństwo, że te same pary znajdą się na tych samych pozycjach jest znaczne, jednak im więcej miałoby takich połączeń wystąpić obok siebie, tym to prawdopodobieństwo spada i można wysnuć wniosek, że wówczas uznanie takiego podobieństwa za nieprzypadkowe jest słuszne. Z powodu niewystarczającej liczby porównywanych nukleotydów oraz ze względu na losowość dopasowań, w przypadku tej pary nie możemy mówić o ewolucyjnym powiązaniu.