Wstęp do bioinformatyki

Nr ćwiczenia: 2

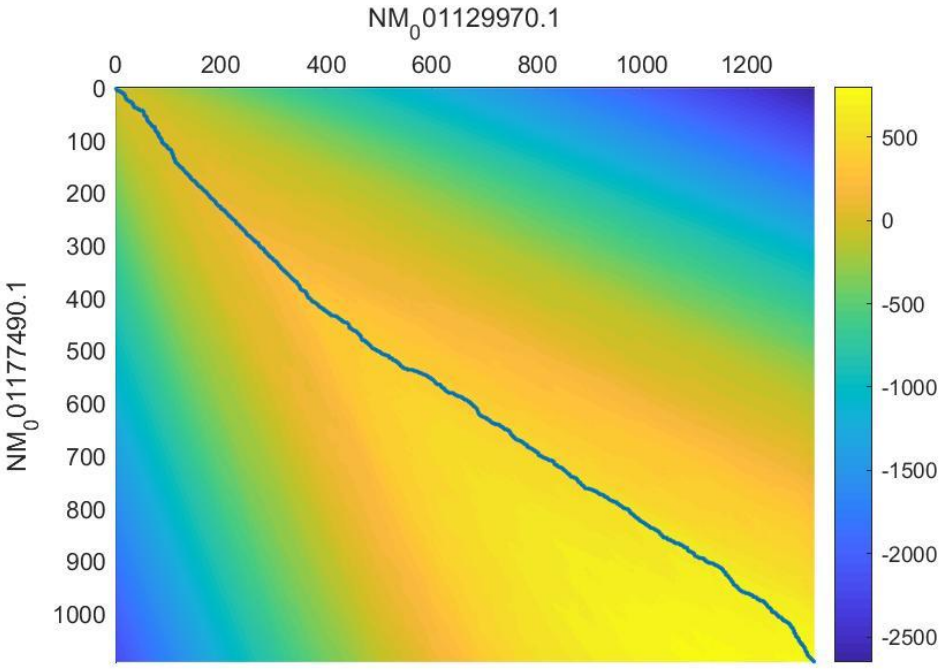
Temat ćwiczenia: Dopasowanie globalne par sekwencji

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**

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

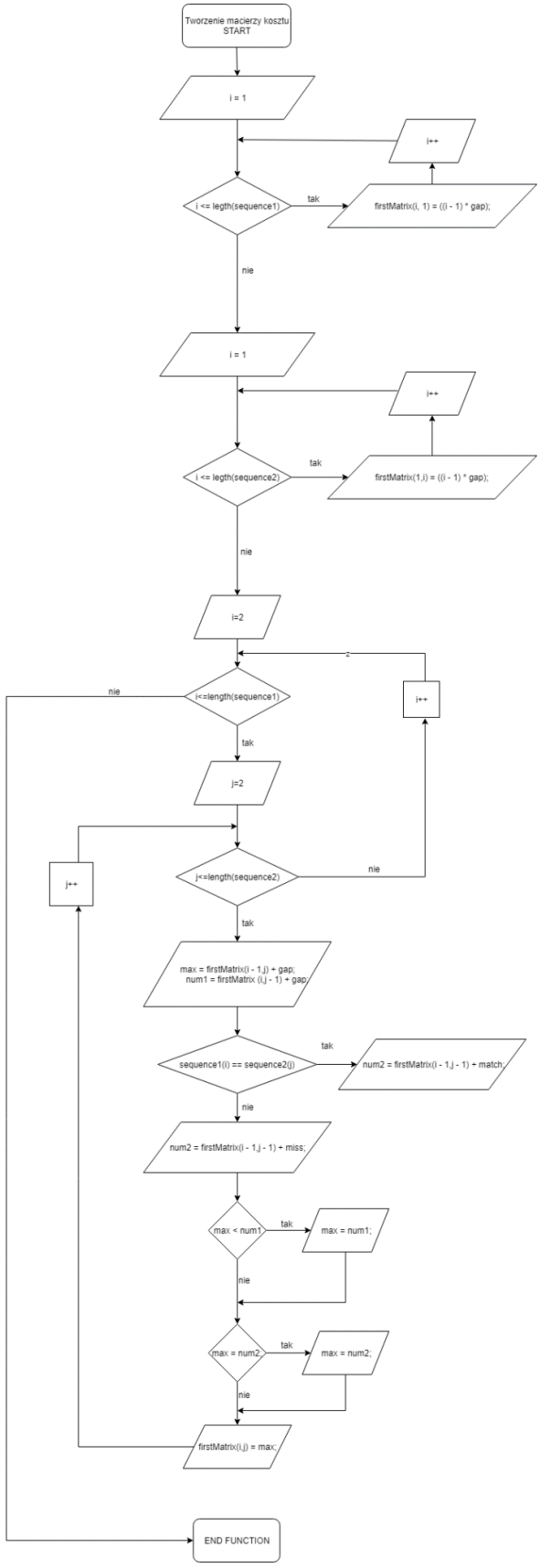
1. **Prezentacja działania programu**:



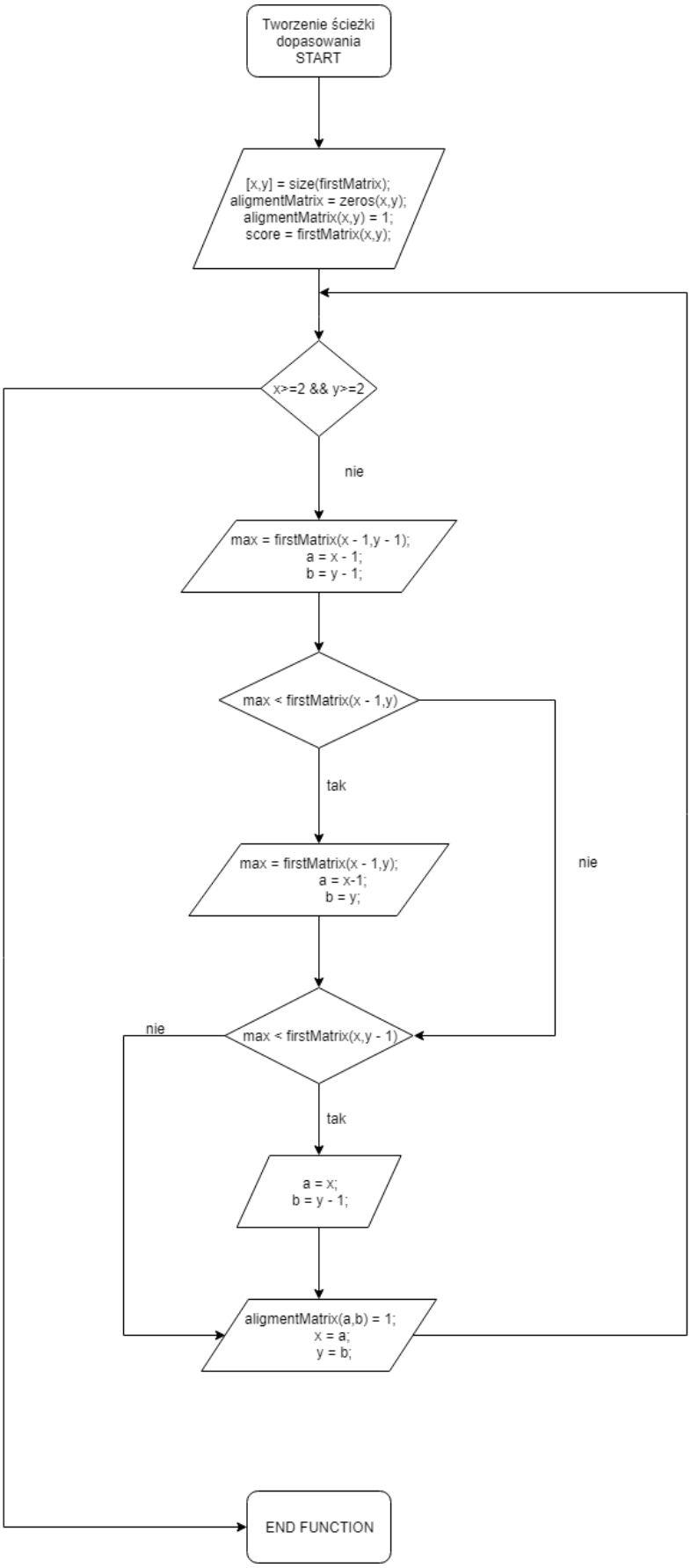
Rysunek 1 Przykład działania programu, dla genów świni i pszczoły.

```
#1: NM_001177490.1
#2: NM_001129970.1
#Mode: distance
#Match: 3
#Mismatch: -5
#Gap: -2
#Score: 803
#Length: 1538
#Indentity: 551/1538(36%)
#Gaps: 659/1538(43%)
TCAG--T-G-GTT--ACTCATCGATCGA-AG---CAT-CTATCCACG--G-T-CA-G----GCTATAAACGTCTGTAAAATTTTTCTCGCGC---GGGTGT
  ||      |    |          || ||          ||  ||       |      ||       |||||      |   |||  ||
AGAGTACGCGGGGAGA-C--C-C--AGAAAGCGGG-ACG--TC--CGGCCTGCGAGTGGTGGC-T----TGTCTG----T-TGAGCTC-GGCGAAA-----
```

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
     m+n
     createMatrix:
     m+n+mn
     readFasta:
     m+n+m+n
     searchingWay:
     mn
     writeSequence:
     mn+m

     Razem:
     m+n+m+n+mn+m+n+m+n+mn+mn+m=3mn+5m+4n≅3mn

     Złożoność obliczeniowa czasowa co najwyżej rzędu mn.
     O(mn)

   - **Pojemnościowe**

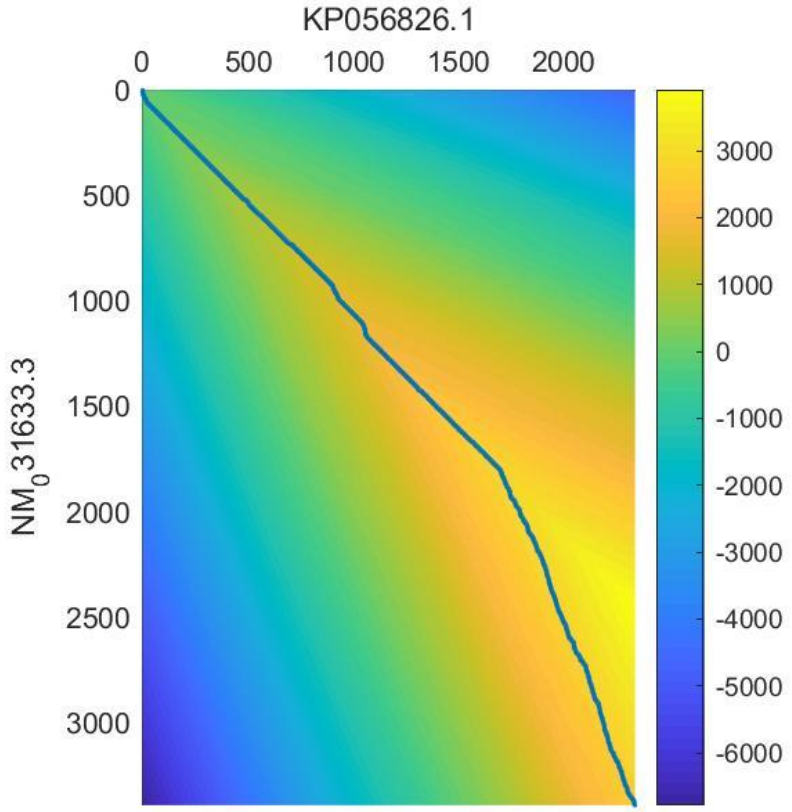
     searchingWay:
     mn+mn
     chooseFilter:
     mn+mn+m+n
     createMatrix:
     m+n+mn
     readFasta:
     m+n
     checkFile:
     m+n
     toTextFile:
     m+n+n
     writeSeq:
     m+n+mn+n

     Razem:
     mn+mn+mn+mn+m+n+m+n+mn+m+n+m+n+m+n+n+m+n+mn+n=6mn+6m+8n
     ≅6mn
     Złożoność obliczeniowa pamięciowa co najwyżej rzędu mn.
     O(mn)

4. Porównanie przykładowych par sekwencji

4.1. Ewolucyjnie powiązanych

- **Porównanie sekwencji szczura z kozą**

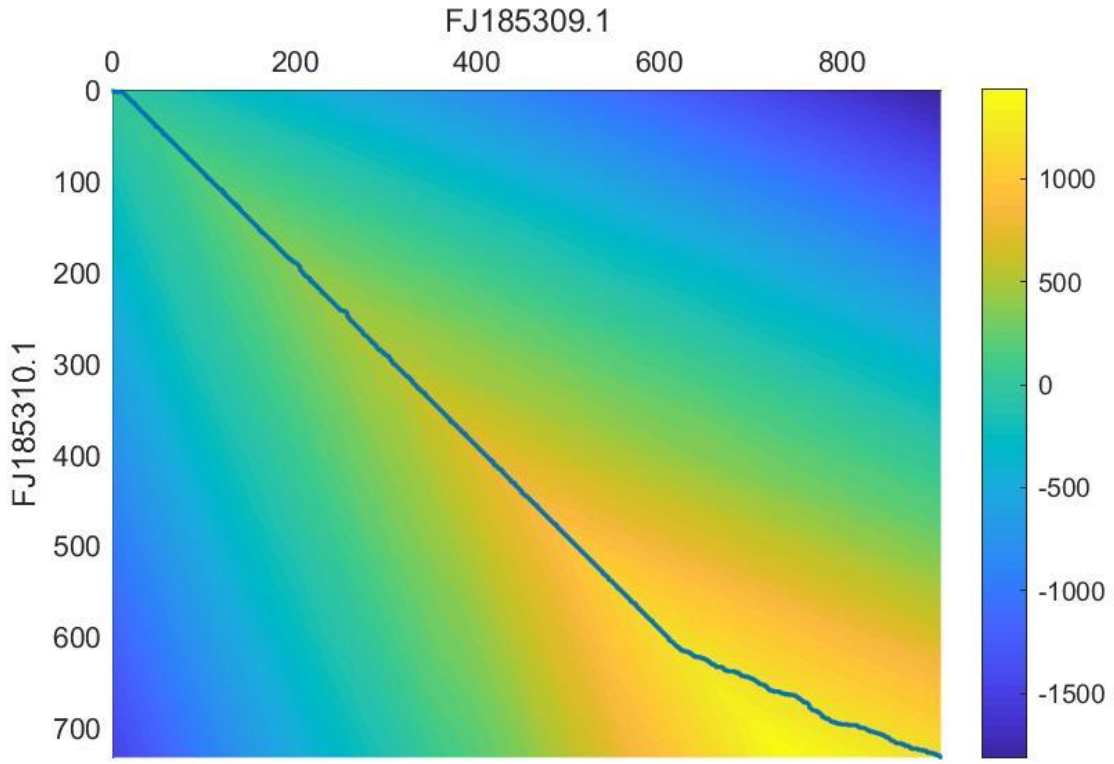


Rysunek 4 Macierz kosztu z naniesioną ścieżką optymalnego dopasowania porównująca szczura (oś y) i kozę (oś x)

```
#1: NM_031633.3
#2: KP056826.1
#Mode: distance
#Match: 3
#Mismatch: -5
#Gap: -2
#Score: 2612
#Length: 3602
#Indentity: 1596/3602(44%)
#Gaps: 1475/3602(41%)
CTGGCTCGGCCCCGCGTGGAGCAGCGGTGGCCTGTG-AGGGTCAAAGCTTGTGATTCTCG-ATGGAGAGTGAAAGCACAGCTTCATGATGAGA-ACCAGCCCCCGGCGGCCAC-TGATTCTCAAG-AGACGGA
|                 | | |      | | | | |     | |      ||   ||||||||||||||||||||| |||||||||  | |||||||||| |||||   |||||||||   ||||||||
C-----------------TA-GA--GA----TT-GGAA-GAT-----C-T-----GCT-CAATGGAGAGTGAAAGCACAGATTCATGATG-AAAACCAGCCCCCGTCGGCC-ATTGATTCTCA-AAAGACGGA
```

Rysunek 5 Konsolowy wynik działania programu dla Szczura i Kozy

- **Porównanie kota z panterą**



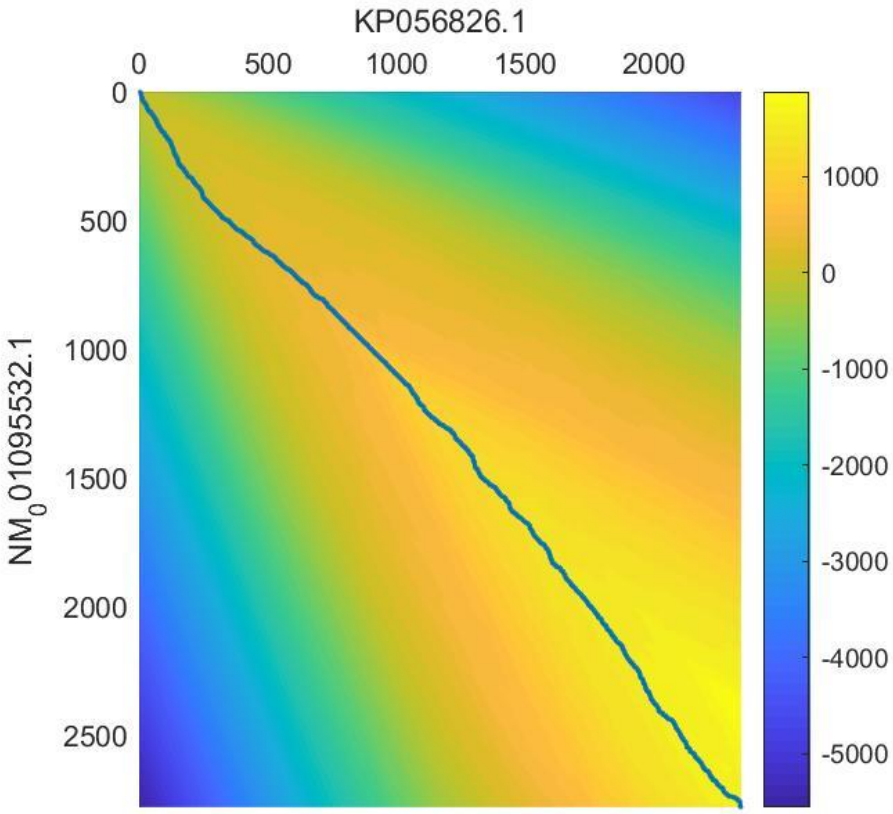
Rysunek 6 Macierz kosztu z naniesioną ścieżką optymalnego dopasowania porównująca kota (oś y) z panterą (oś x)

```
#1: FJ185310.1
#2: FJ185309.1
#Mode: distance
#Match: 3
#Mismatch: -5
#Gap: -2
#Score: 1188
#Length: 962
#Indentity: 546/962(57%)
#Gaps: 283/962(29%)
A---------TGTTCATAAACC-GTTGACTATTTTCAACTAATCACAAA-GATATTGGT-ACTCTTTACCTTTTATTTGGTGCCTGA-GCTGGCATGGTGGGGACT
|         ||||||||||||   |||||||||||||| |||||||| | ||||||||   |||||||||||||||||||||| ||||   ||||| ||||||||||||
ATTTTACCTATGTTCATAAACCGC-TGACTATTTTCAACCAATCACA-AGGATATTG-GAACTCTTTACCTTTTATTTGGCGCCT-GGGCTGGTATGGTGGGGACT
```

Rysunek 7 Konsolowy wynik działania programu dla Kota i Pantery

4.2. Ewolucyjnie niepowiązanych

- **Porównanie żaby szponiastej z kozą**



Rysunek 8 Macierz kosztu z naniesioną ścieżką optymalnego dopasowania wygenerowana dla żaby szponiastej (oś y) oraz kozy (oś x)

```
#1: NM_001095532.1
#2: KP056826.1
#Mode: distance
#Match: 3
#Mismatch: -5
#Gap: -2
#Score: 1556
#Length: 3272
#Indentity: 1188/3272(36%)
#Gaps: 1428/3272(44%)
GGCACCGGCTTATAAACAGTGATCGGATAAAAAGAAGTCTGATGC-ATTTT-GAGCTCTCAATGGCTGCCAGACGAG-C-C-G-TCACACCATG-AGAACAAGCCCACGCAGACCC-CTAATACTTAAGAGAC
          |       |     |      ||   ||||   | |   | ||          |        ||       ||     |||  ||  |  | |  |    |||
A-G--A----GA------TT---G-G-A----AG--ATCTGC-TCAA---TGGA-----G--A--GT--G-A--AAGCACAGATTC-----ATGATGA-AA-AC-C-AG---CCCCCG----T----C-G-G-
``
```

Rysunek 9 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 naturalnie, 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 dopasowani wynosi 57%. 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. Również w przypadku tej pary widzimy, że sparowane nukleotydy tworzą grupy dopasowani w rzędzie bez przerw, stąd możemy wnioskować, że podobieństwo to nie jest przypadkowe. Mówimy tutaj więc o ewolucyjnym powiązaniu. 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 44% mimo tego iż łańcuchy różnią się od siebie długością aż o około tysiąc próbek. Droga najlepszego dopasowania nie jest liniowa, jednak obserwujemy grupy sparowanych nukleotydów, które mogą świadczyć o tym, że organizmy 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 żaba szponiasta – koza. W tym przypadku udział dopasowań w optymalnej ścieżce dopasowania tych sekwencji wynosi 36%. Na tej podstawie można by wysunąć tezę, że organizmy te nie różnią się od siebie genetycznie tak, jak jest to widoczne, jednak po 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. Mimo wystarczającej liczby porównywanych nukleotydów, ze względu na losowość dopasowań, w przypadku tej pary nie możemy mówić o ewolucyjnym powiązaniu.