Wstęp do bioinformatyki

Laboratorium 3

Dopasowanie lokalne par sekwencji

Magdalena Trędak

236712

1. Schemat blokowy algorytmu dopasowania lokalnego

Schematy blokowe algorytmów tworzenia macierzy punktowej oraz optymalnej ścieżki dopasowania ze względów dużych rozmiarów i umożliwienia poprawy ich czytelności zamieszczono w repozytorium jako pliki graficzne o nazwach: SchematBlokowyGenerowaniaMacierzyPunktów.jpg, SchematBlokowyGenerowaniaŚcieżkiDopasowania.jpg.
Do wygenerowania schematów użyto programu online znajdującego się na stronie: http://www.algorytm.org/narzedzia/edytor-schematow-blokowych.html (data dostępu 15.04.19)

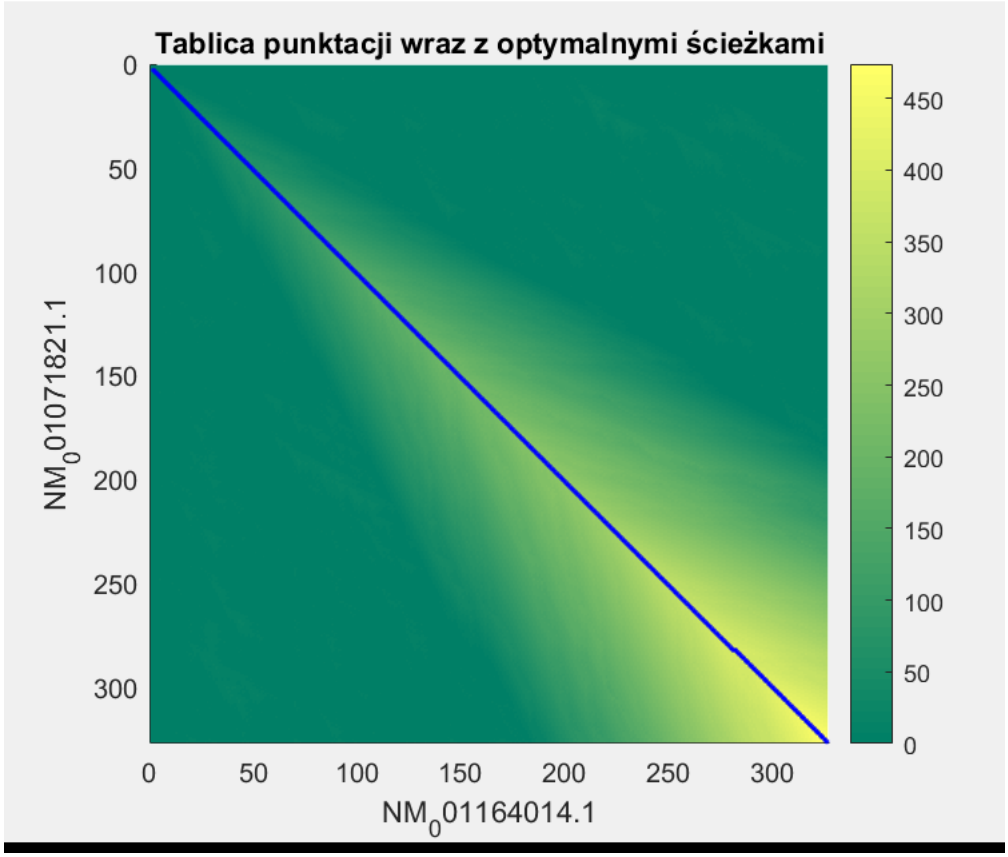
2. Analiza złożoności obliczeniowej czasowej i pamięciowej

- Oszacowanie złożoności czasowej
  scoringMatrix – 2 pętle for (m*n), 2 pętle for(g*f), gdzie g i f to rozmiary macierzy substytucji, znacznie mniejsze od długości sekwencji, 1 warunek if, 5 przypisań wartości w pętlach, 15 poza nimi $O(mn) - złożoność\ czasowa\ co\ najwyżej\ rzędu\ mn$
  tracBackMatrix – k-razy pętla for, k to ilość maksimów macierzy punktów > 0,
  pętla while – w skrajnym przypadku maksymalny element znajduje się w końcu macierzy punktów a ostatnie 0 na jej początku, wtedy wyszukanie ścieżki odbywa się po całej macierzy (n*m) – rząd co najwyżej O(n,m),6 porównań, 10 przypisań
  $$O(mn) - złożoność\ czasowa\ co\ najwyżej\ rzędu\ mn$$
  createInfo – 3 pętle for. Maksymalnie rzędu m*n, ponieważ dopasowanie lokalne inne niż po przekątnej całej macierzy punktacji pozwala skrócić porównywane sekwencje
  $$O(mn) - złożoność\ czasowa\ co\ najwyżej\ rzędu\ mn$$

- Oszacowanie złożoności pamięciowej
  Macierz punktów jest macierzą o wymiarach (n x m) i tyle też zajmuje miejsca w pamięci – rząd m*n. Pozostałe tworzone macierze są takich samych rozmiarów lub mniejsze. Przypisania poszczególnych zmiennych lub wektorów są znacznie mniejsze niż rozmiarów n*m. Pozwala to przyjąć założenie, że macierz punktów (scoringMatrix) i macierz ścieżki optymalnego dopasowania (tracBackMatrix) są największymi obiektami, dlatego złożoność pamięciowa programu $O(mn) - \ co\ najwyżej\ rzędu\ mn$

3. Porównanie przykładowych par sekwencji

Porównanie cytochromu c konia (Equus caballus) - NM_001164014.1 i szympansa zwyczajnego (Pan troglodytes) - NM_001071821.1 – porównanie nr 1



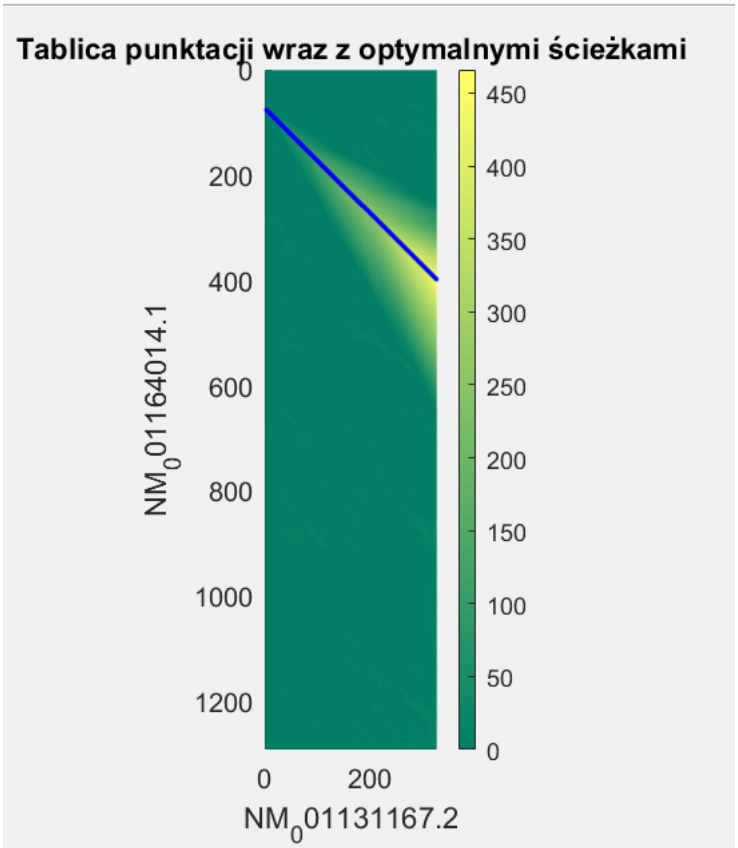

LocalAlignment — Notatnik

Plik Edycja Format Widok Pomoc

```
# Substitution matrix:
#A C G T
A 2-7-5-7
C -72-7-5
G -5-72-7
T -7-5-72
>seq1 3-325
>seq2 2-325
#Gap: -2
#Length: 318
#Gaps: 1/318 (0.31447 %)
#Identity: 127/318 (39.9371 %)
TGGGTGATGTTGAGAAGGGCAAGAAGATTTTTGTTCAGAAGTGTGCCCAGTGCCATACCGTGGAAAAGGGAGGCAAGCACAAGACTGGGCCAAACCTCCATGGTCTATTTGGGCG
  ||      |    | | | |   ||||  | |  |      ||     |    |     ||| ||  |  |     |     || | ||    |   |     || ||
ATGGGTGATGTTGAGAAAGGCAAGAAGATTTTTATTATGAAGTGTTCCCAGTGCCATACCGTTGAAAAGGGAGGCAAGCACAAGACTGGGCCAAATCTCCATGGTCTCTTCGGGC
```

Porównanie cytochromu c orangutana (Pongo abelii) – NM_001131167 i konia

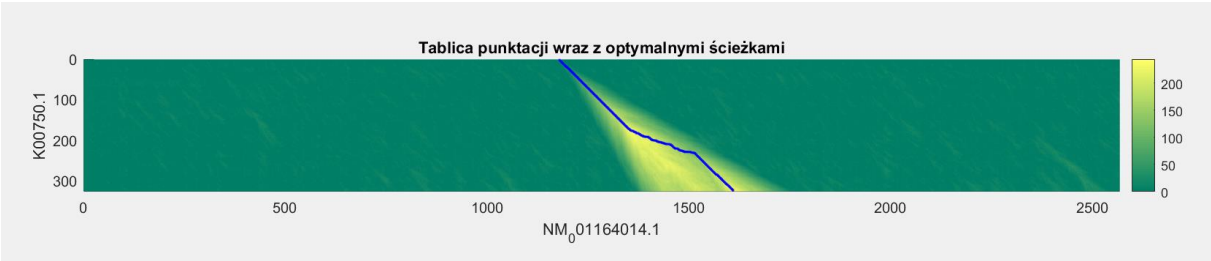
(Equus caballus) - NM_001164014.1 – porównanie nr 2

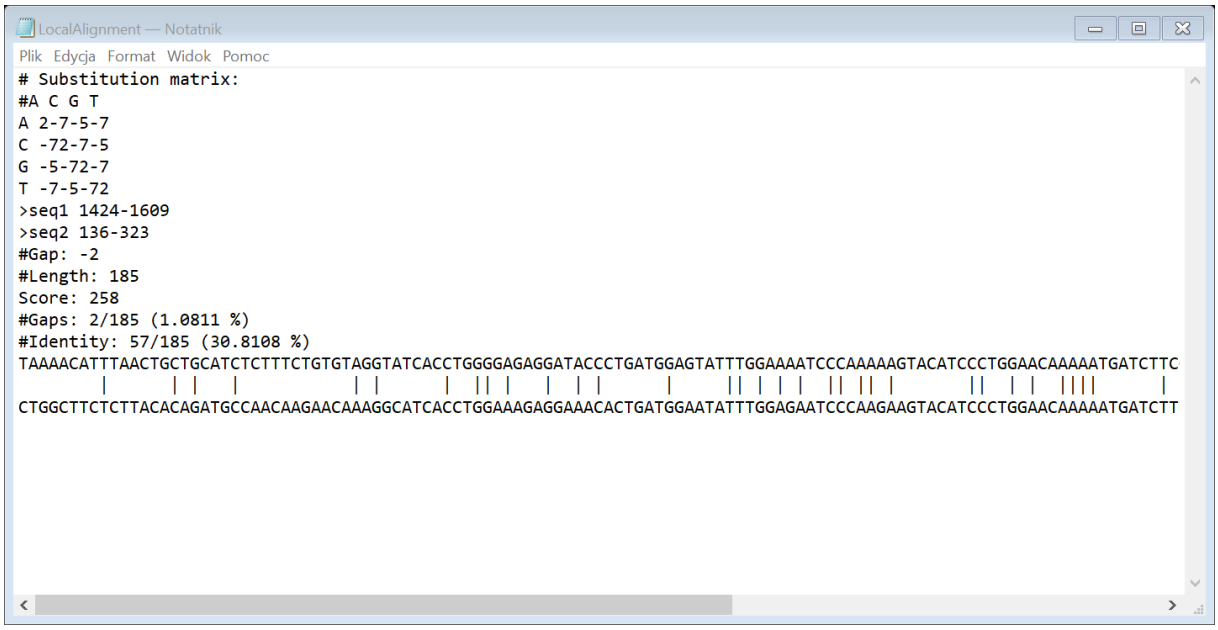



```
>seq1 76-397
>seq2 3-325
#Gap: -2
#Length: 317
Score: 466
#Gaps: 1/317 (0.31546 %)
#Identity: 100/317 (31.5457 %)
GGTGATGTTGAGAAAGGCAAGAAGATTTTTATTATGAAGTGTTCGCAGTGCCACACCGTTGAAAAGGGAGGCAAGCACAAGACTGGGCCAAATCTCCATGGTCTCTTCGGGCGGA
 | |   |   || | |    |  |  ||| |  | |   ||  ||   |   ||        ||  | |       ||  |    |    |   |       |||    | |
TGGGTGATGTTGAGAAGGGCAAGAAGATTTTTGTTCAGAAGTGTGCCCAGTGCCATACCGTGGAAAAGGGAGGCAAGCACAAGACTGGGCCAAACCTCCATGGTCTATTTGGGCG
```

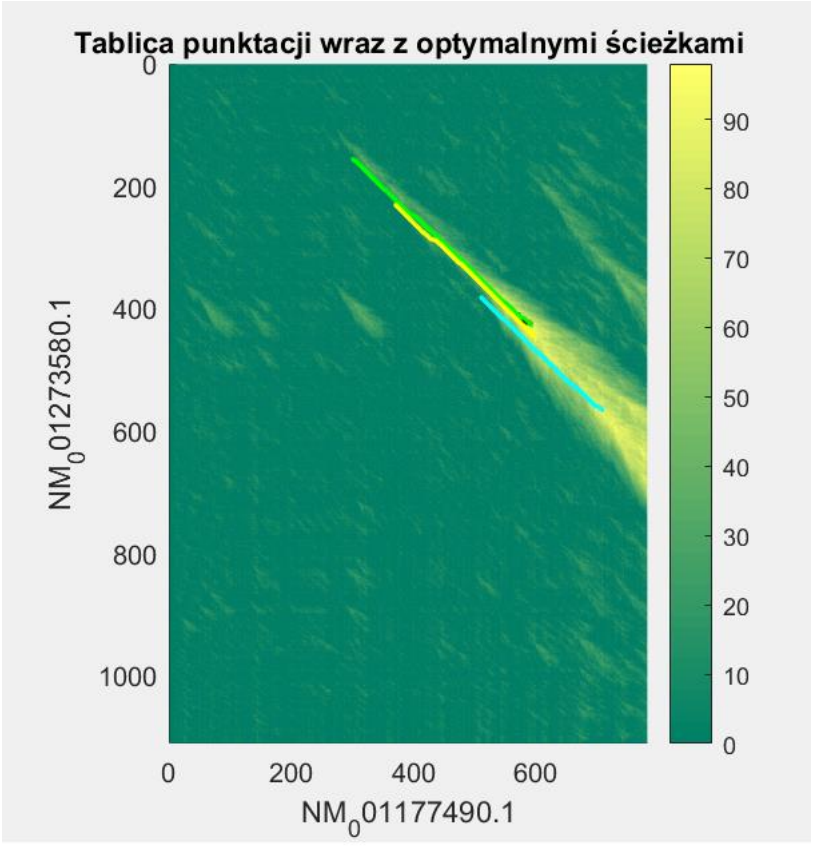
Porównanie cytochromu c szczura wędrownego (Rattus norvegicus) – K00750.1 i konia

(Equus caballus) - NM_001164014.1 – porównanie nr 3






Porównanie cytochromu c muszki owocowej (Drosophila melanogaster) variant B - NM_001273580.1 i pszczoły miodnej (Apis mellifera) - NM_001177490.1 – porównanie nr 4


Wnioski:

Porównano ten sam gen dla organizmów powiązanych i niepowiązanych ewolucyjnie. Wszystkie wyniki otrzymano poprzez wykorzystanie algorytmu dopasowania lokalnego podobieństwa o parametrach: gap = -2 oraz parametrach zawartych w macierzy substytucji dołączonej jako plik sMatrix.txt

*Tabela 1. Porównanie otrzymanych wyników dopasowań globalnych dla różnych par organizmów*

| Powiązanie ewolucyjne organizmów | Nr porównania | Score [-] | Length [-] | Gap [%] | Identity [%] |
|---|---|---|---|---|---|
| Tak | 1 | 460 | 318 | 0,31 | 39,93 |
| Tak | 2 | 466 | 317 | 0,31 | 31,54 |
| Nie | 3 | 258 | 185 | 1,08 | 31,81 |

Na podstawie wyników zawartych w Tabeli [1] można zauważyć następujące zależności:

- Dla organizmów powiązanych ewolucyjnie score jest dwukrotnie większy niż dla niepowiązanych. Dopasowane sekwencje są dłuższe a odsetek gap jest mniejszy.
- Dla wykorzystanej do analizy macierzy substytucji podobieństwa dopasowanych sekwencji są tego samego rzędu (30 %)