Wstęp do bioinformatyki

Laboratorium 3

Dopasowanie lokalne par sekwencji

Magdalena Trędak

2367121

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 16.05.19)

2. Analiza złożoności programu

a) czasowa

scoringMatrix – 2pętle 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(m+n), 9 porównań, 9 przypisań . Poza pętlą while 11 przypisań. O(m+n) − złożoność czasowa co najwyżej rzędu m+n

b) pamięciowa

Macierz punktów jest macierzą o wymiarach (n * 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.

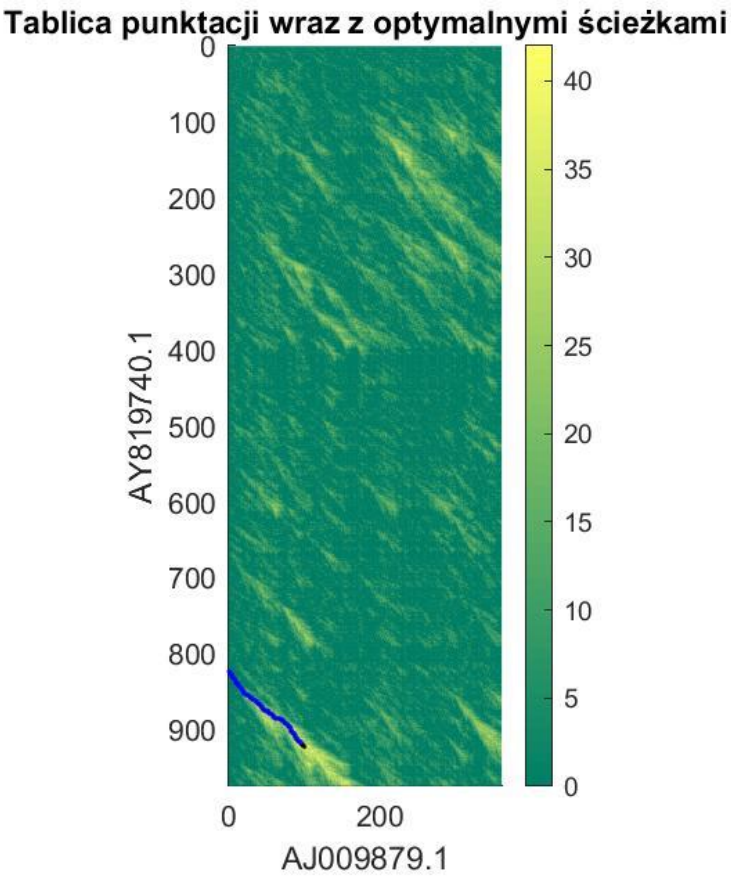
3. Porównanie przykładowych par sekwencji mitochondrialnego cytochromu b dla parametrów : match , mismatch wczytanych z macierzy substytucji, gap = -2

a) niepowiązanych ewolucyjnie

AJ009879.1- Nubian ibex
AY819740.1- African clawed frog

```
Ilosc znalezionych sciezek: 2
>seq1 822-918
>seq2 0-96
#Gap -2
#Score: 42
#Length: 121
#Gaps: 50/121 (41.3223 %)
#Identity: 71/121 (58.6777 %)
#seq1 CGCCCAATCAGCCAATGCATATTCTGAAT-CT--TGG-CAG--CA-GAT-CTACTAACA-
#seq2 C-CCC--TCAG--AATG-ATATT-TG--TCCTCATGGT-AGAACAT-ATCCTAC-AA-AT
      * ***  ****  **** ***** **  * **  ***  **  **  ** **** ** *


#seq1 -C--TCA-C-ATGAA-T-----T-GGAGGA-CAGC-CAGTCGAA-CATCCTTACAT-TAT
#seq2 GCTGT-AGCTAT-AACTAAAAATAGGAGGATCA-CAC---C-AAT-AT--TT-CATGT-T
       *  * * * ** ** *     * ****** ** * *   * **  **  ** *** * *


#seq1 T
#seq2 T
      *
```

```
>seq1 919-921
>seq2 97-99
#Gap -2
#Score: 42
#Length: 3
#Gaps: 0/3 (0 %)
#Identity: 2/3 (66.6667 %)
#seq1 ATT
#seq2 CTT
       **
```

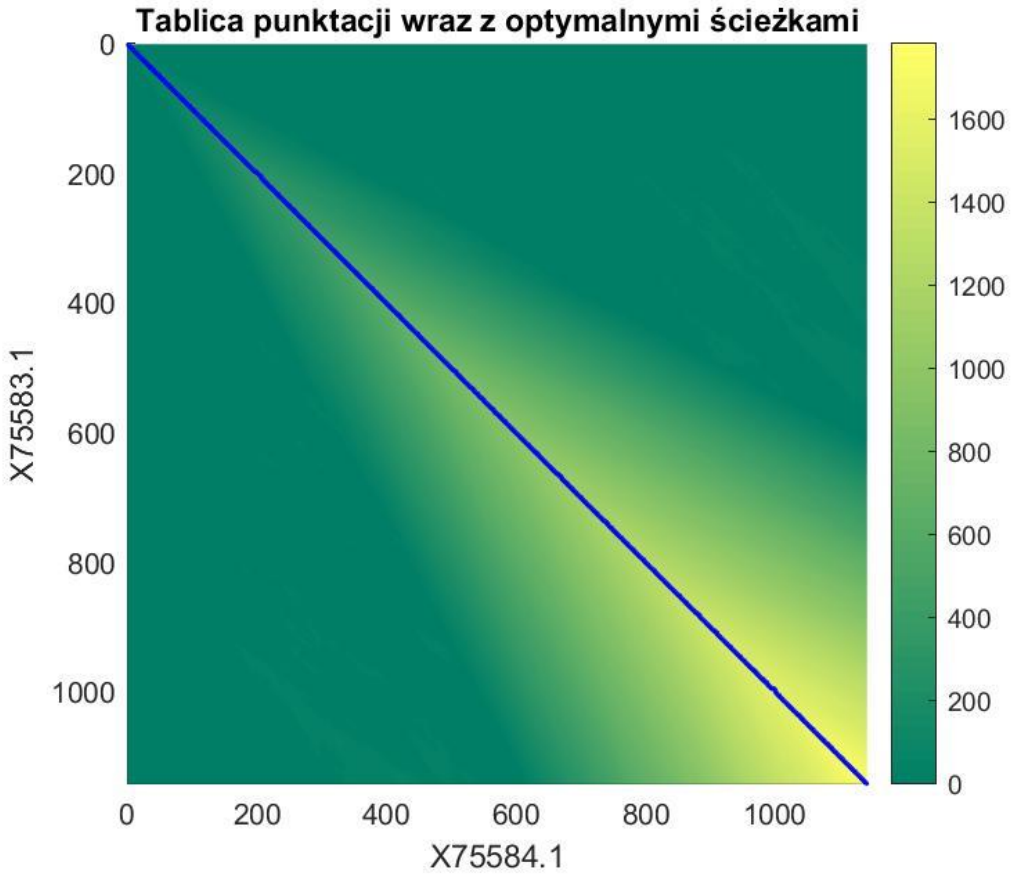


b) powiązanych ewolucyjnie

X75584.1 - humpback whale
X75583.1 - pygmy Bryde's whale

```
Ilosc znalezionych sciezek: 1
>seq1 1-1140
>seq2 1-1140
#Gap -2
#Score: 1782
#Length: 1223
#Gaps: 166/1223 (13.5732 %)
#Identity: 1057/1223 (86.4268 %)
#seq1 ATGACCAACATCCGAAAAACACACCCACTAATAAA-AATT-ATCAACGA--CACATTC-A
#seq2 ATGACCAACATCCGAAAAACACACCCACTAATAAAG-ATTG-TCAACGATG--CATTCG-
      ***********************************  ***  *******    *****


#seq1 TTGATCT-ACCCACCCCATCAAATATCTCCTCATGATGAAA-CTTCGG-TTCCCTACTCG
#seq2 TTGATCTC-CCCACCCCATCAAATATCTCCTCATGATGAAAT-TTCGGC-TCCCTACTCG
      *******  *********************************  *****  **********


#seq1 GCCT-TTGCTTAATTA-TACAAATCCTAACAGGCCTATTCCTAGCAATACACTACACACC
#seq2 GCCTC-TGCTTAATTAC-ACAAATCCTAACAGGCCTATTCCTAGCAATACACTACACACC
      ****  **********  ****************************************


#seq1 AGACACAACAACCGCCTTCTCATCAG-T-CACACA-CATCTG-TCGAGACGTAAA-TTA-
#seq2 AGACACAACAACCGCCTTCTCATCAGTTGCACACAT--T-TGC-CGAGACGTAAAC-TAC
      *************************** * ******   * **  ***********  **


#seq1 TGGCTGA-ATTATCCGATACCTACA-TGCAAA-TGG-GGCCTCCATATTCTTCATCTG-C
#seq2 -GGCTGAG-TTATCCGATACCTACAC-GCAAAC-GGA-GCCTCCATATTCTTCATCTGT-
       ******  ****************  *****  **  ********************


#seq1 CTCTACGCTCACATAGGACGAGGCCTATACTACGGCTCCTA-CGCCTTTCGAGAAACATG
#seq2 CTCTACGCTCACATAGGACGAGGCCTATACTACGGCTCCTAT-GCCTTTCGAGAAACATG
      ****************************************  *****************


#seq1 AAACATCGGAGTTAT-TCTACTATTCACAGTTATAGCCA-CTGCATTC-GTAGGCTACGT
#seq2 AAACATCGGAGTTATC-CTACTATTCACAGTTATAGCCACC-GCATTCA-TAGGCTACGT
      ***************  ********************** * ******  **********
```

```
#seq1 GCAATCCCATACATTGGTACTACCCTAGTCGAATGAATCTGGGGCGGTTT-T-TCCGTAG
#seq2 GCAATCCCATACATTGGTACTACCCTAGTCGAATGAATCTGGGGCGGTTTCTCT--GTAG
      ************************************************* * *  ****


#seq1 A-CAAAGCAACACTAACACG-TTTCTTTG-CTTTCCAC-TTCATCCTCCCCTTCA-TCAT
#seq2 AT-AAAGCAACACTAACACGCTTT-TTTGCC-TTCCACTTT-ATCCTCCCCTTCATT-AT
      *  ***************** *** **** * ****** ** ************ * **


#seq1 TAC-AGCA-TTAGCAAT-CGTCCACCTCATTTTCCTCCACGAAACAGGATCCAA-CAACC
#seq2 T-CTAGCAC-TAGCAATG-GTCCACCTCATTTTCCTCCACGAAACAGGATCCAAT-AACC
      * * ****  *******  ***********************************  ****


#seq1 CCACAGG-CA-TCCCATCCAACATAGACAAAATCCCATTCCA-CCCT-TACTACACAA-T
#seq2 CCACAGGT-ATT-CCATCCAACATAGACAAAATCCCATTCCACCCCTAT--TACACAACT
      *******  * * ***************************** **** *  ******* *


#seq1 CAAAGACA-CTCTAGGCGCCCTA-TTACTAATCCTAACCCTACTAATG-TTAACCCTATT
#seq2 -AAAGACAT-TCTAGGCGCCCTAC-TACTAATCCTAACCCTACTAATGC-TAACCCTATT
       *******  *************  ************************  **********


#seq1 AG-CTT-TCATCCCAATACTCCACACATC-CAAACAACGAAGCAT-GATGTT-TCGACCC
#seq2 AGCCTTA--ATCCCAATACTCCACACATCT-AAACAACGAAGCATA-ATGTTC-CGACCC
      ** ***   ********************  **************  *****  ******


#seq1 TTTAGCC-AGTT-TCT-GTTTTGA-AT-ACT-A-GTAGCAGAC---CTATTAGCCCT-AA
#seq2 TTTAGCCAA-TTC-CTA-TTTTGAG-TC-CTAAT-T-GCAGACTTACTA--A-CCCTG-A
      ******* * **  **  ******  *  ** *  * ******   ***  * ****  *


#seq1 CATGAATCGGCGGCCAACCCGTAGAACA-CCCATAC-ATAATCGTAGGCCAA-CTCGCAT
#seq2 CATGAATCGGCGGCCAACCCGTAGAACACCCC-TACG-TAATCGTAGGCCAAT-TCGCAT
      **************************** *** ***  **************  ******


#seq1 CCATCCTCTA-CTTCCT-CTTAA-TCCTAGTA-TTAATACCA-ATAACTAGTCTTATCGA
#seq2 CCATCCTCTAT-TTCCTCC-TAATT-CTAGTAC-TAATACCAG-TAACTAGTCTTATCGA
      **********  ***** * *** * ******  ********  ****************


#seq1 GAA-CAAACTTATAAAATGAAGA
#seq2 GAAT-AAACTTATAAAATGAAGA
      ***  ******************
```



Dla organizmów niepowiązanych ewolucyjnie porównanie genu kodującego cytochrom b generuje krótsze fragmenty sekwencji o mniejszym podobieństwie niż w przypadku porównania genu organizmów powiązanych ewolucyjnie. W tym przypadku dopasowanie lokalne staje się dopasowaniem globalnym, podobieństwo sekwencji występuje na całej długości genu i jest ono wysokie (86 %).