

Bioinformatyka Sprawozdanie 4

Progresywne dopasowanie wielosekwencyjne

Michał Marciniak 244811

1 Kod do repozytorium

<https://gitlab.com/MMarciniak103/bioinformatics>

2 Analiza złożoności obliczeniowej czasowej i pamięciowej:

Algorithm 1 Dopasowanie wielosekwencyjne - algorytm gwiazdowy

```
1: procedure INICJALIZACJA
2:    $X \leftarrow$  tablica sekwencji  $x$ 
3:    $gap \leftarrow$  koszt przerwy
4:    $sub \leftarrow$  koszt substytucji
5:    $match \leftarrow$  koszt dopasowania
for each:  $x_k, x_l \in X$ 
6:    $P^{kl} \leftarrow$  dopasowanie globalne pary  $x_k$  i  $x_l$ 
7:    $s(P^{kl}) \leftarrow$  koszt dopasowania
8: procedure DOPASOWANIE WIELOSEKWENCYJNE
9:    $M \leftarrow$  tablica zawierająca dopasowanie wielosekwencyjne
10:   $c = \operatorname{argmin}_k \sum_{1 \leq l \leq N, l \neq k} s(P^{kl})$ 
11:   $x^c \leftarrow$  sekwencja stanowiąca centrum dopasowania
12:   $M+ = x^c$ 
for each:  $x^i \in X - x^c$ 
13:    wybierz najlepsze dopasowanie pomiędzy  $x^c$  and  $x^i$ 
14:    dodaj  $x^i$  do  $M$ , uwzględniając propagację przerw w sekwencjach, które już znajdują się w tablicy  $M$ 
```

Złożoność obliczeniowa wynosi : $O(k^2n^2) \leftarrow k$ - ilość sekwencji, n - długość sekwencji (Obliczenie dopasowania globalnego dla każdej pary)

Złożoność pamięciowa: $O(k^2n^2) \leftarrow$ macierz zawierająca dopasowania globalne dla wszystkich par

3 Opis programu

Po wybraniu przez użytkownika opcji 'multiple sequences' zostaje wyświetlone okno pozwalające na wczytanie sekwencji zarówno w postaci pliku fasta jak i wysłania zapytania do API NCBI. Na ten moment użytkownik ma do wyboru tylko algorytm gwiazdy. W przypadku niepoprawnego wczytania sekwencji zostanie wyświetlony komunikat, co zostało zaprezentowane na Fig. 2.



Figure 1: Okno dopasowania wielosekwencyjnego

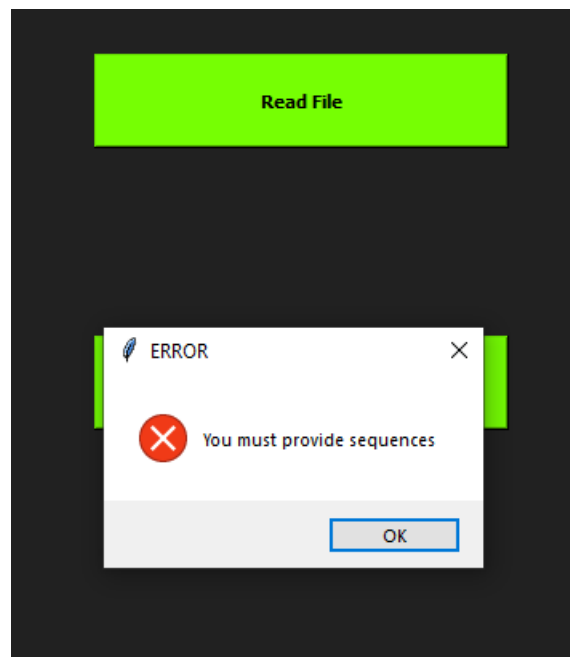


Figure 2: Komunikat błędu

Po poprawnym wczytaniu sekwencji i wciśnięciu przycisku STAR zostaje wyświetlone okno, w którym użytkownik musi podać wartości kosztów za przerwy,substytucje oraz dopsaowania.

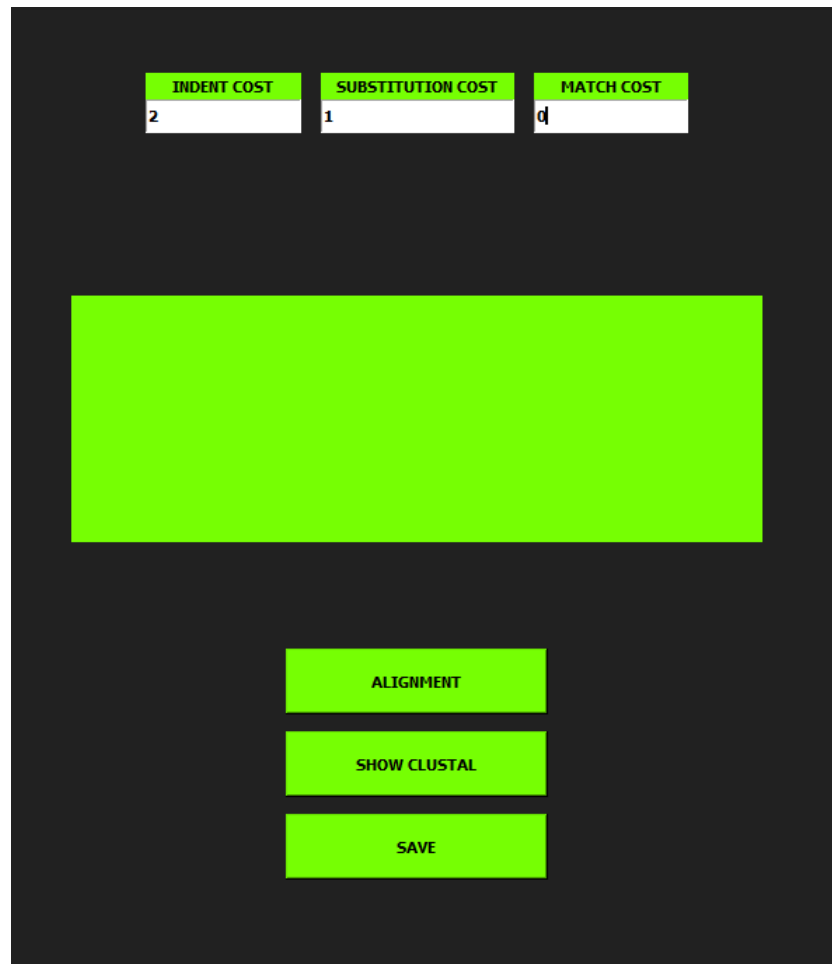
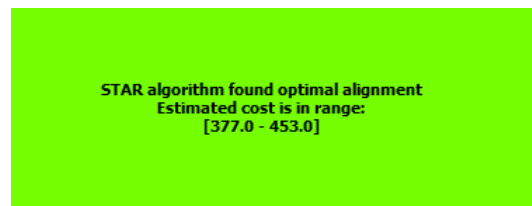


Figure 3: okno STAR

Po zakończeniu obliczeń w zielonym polu wyświetlone zostaną informacje o znalezionym dopasowaniu, tj. oszacowany przedział w jakim znajduje się wynik dopasowania. Zaprezentowane jest to na Fig. 4.



STAR algorithm found optimal alignment
Estimated cost is in range:
[377.0 - 453.0]

Figure 4: okno STAR

Klikając SHOW CLUSTAL użytkownik ma możliwość wyświetlić dopasowanie w formacie CLUSTAL. Zaprezentowano to na zdjęciu poniżej.

```

x1: -----CAGG----TAACAAACC-----AA-CCAA-
x2: ATTAAAGGTTTATACCTTCCCAGG----TAACA-----AACCAACC
x3: ATAT-----TAGGTTTTACCTACCCAGGAAAAGCCAAC
x4: -----

x1: C--TTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAAAATCTGT
x2: AACTTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAAAATCTGT
x3: CAACCTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAAAATCTGT
x4: -----AGATCTGTTCTCTAAACGAACTTTAAAAATCTGT
      *****

x1: GTGGCTGTCA
x2: GTGGCTGTCA
x3: GTAGCTGTCG
x4: GTGGCTGTCA
      ** *****

```

Figure 5: okno CLUSTAL

Możliwe jest również zapisanie wyników w formacie FASTA.

```

1 >Multiple Sequences Alignment ,seq1: x1 ,seq2: x2 ,seq3: x3 ,seq4: x4
2 -----CAGG----TAACAAACC-----AA-CCAA-C--TTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAAAATCTGTGIGGCTGTCA
3 ATTAAAGGTTTATACCTTCCCAGG----TAACA-----AACCAACCAACTTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAAAATCTGTGIGGCTGTCA
4 ATAT-----TAGGTTTTACCTACCCAGGAAAAGCCAACCAACCTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAAAATCTGTGTAGCTGTCG
5 -----AGATCTGTTCTCTAAACGAACTTTAAAAATCTGTGIGGCTGTCA
6

```

Figure 6: zapis w formacie FASTA

4 Porównanie sekwencji powiązanych ewolucyjnie

seq1: Pongo pygmaeus partial mRNA for transcription factor A (tfam gene)

seq2: Pan paniscus partial mRNA for putative mitochondrial transcription factor A (tfam gene), isoform A.

seq3: Pan troglodytes partial mRNA for putative mitochondrial transcription factor A (tfam gene), isoform A.

seq4: Papio anubis partial mRNA for transcription factor A (tfam gene).

Wszystkie 4 sekwencje dotyczą rodziny małp. Pierwsza pochodzi od orangutana berneńskiego, druga od szympansa karłowatego, trzecia od szympansa zwyczajnego, a czwarta od pawiana oliwkowego.

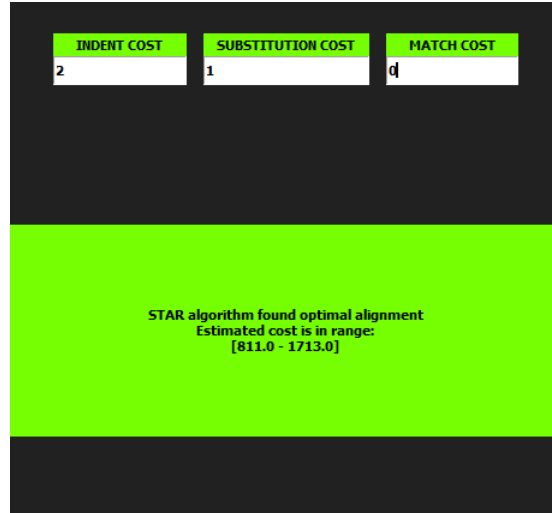


Figure 7: Wynik dopasowania multisekwencyjnego

Obserwując wynik dopasowania w formacie CLUSTAL na Fig. 8 oraz Fig. 9 można zauważyć dobrze zachowany początkowy fragment sekwencji, powtarzający się pośród wszystkimi czterema sekwencjami (Oznaczają to gwiazdki). Dopiero w około połowie sekwencji zaczynają występować różnice pomiędzy tymi 4 osobnikami.

```

AJ841768: GAACAACCTACCCAAATTTAAAGCTCAGAACCCAGATGCAAAAACCTACAGA
AJ971484: GAACAGCTACCCATATTTAAAGCTCAGAACCCAGATGCAAAAACCTACAGA
AJ971485: GAACAGCTACCCAAATTTAAAGCTCAGAACCCAGATGCAAAAACCTACAGA
AJ841770: GAACAGCTACCCAAATTTAAAGCTCAGAACCCAGATGCAAAAACCTACGGA
*****
AJ841768: ACTAATTAGAAGAATTGCCAGCGTTGGAGGGAACCTCCTGATTCAAAGA
AJ971484: ACTAATTAGAAGAATTGCCAGCGTTGGAGGGAACCTCCTGATTCAAAGA
AJ971485: ACTAATTAGAAGAATTGCCAGCGTTGGAGGGAACCTCCTGATTCAAAGA
AJ841770: ACTAATTAGAAGAATTGCCAGCGTTGGAGGGAACCTCCTGATTCAAAGA
*****
AJ841768: AAAAAATATATCAAGATGCTTATAGGGCGGAGTGGCAGGTATATAAAGAA
AJ971484: AAAAAATATATCAAGATGCTTATAGGGCGGAGTGGCAGGTATATAAAGAA
AJ971485: AAAAAATATATCAAGATGCTTATAGGGCGGAGTGGCAGGTATATAAAGAA
AJ841770: AAAAAATATATCAAGATGCTTATAGGGCGGAGTGGCAGGTATATAAAGAA
*****
AJ841768: GAGATAAGCAGATTTAAAGAACAGCTAACTCCAAGTCAGATTATGTCTTT
AJ971484: GAGATAAGCAGATTTAAAGAACAGCTAACTCCAAGTCAGATTATGTCTTT
AJ971485: GAGATAAGCAGATTTAAAGAACAGCTAACTCCAAGTCAGATTATGTCTTT
AJ841770: GAGATAAGCAGATTTAAAGAGCAGCTAACTCCAAGTCAGATTATGTCTTT
*****
AJ841768: GGAAAAAGAAATCATGGACAAACA--T-----T--T-----
AJ971484: GGAAAAAGAAATCATGGACAAACA-----
AJ971485: GGAAAAAGAAATCATGGACAAACATTTAAAGGAAAGC-T-ATGACAAA
AJ841770: GGAAAAAGAAATCACGGACAAACA-----
*****
AJ841768: -----AA-A-----A-A-----GG-----A-A-----
AJ971484: -----
AJ971485: AAA--AAAAG--GTTGACA-CTGCTTGGA---ACCAAAAAGACCTCG
AJ841770: -----
*****
AJ841768: --AG-----C---TATG--A--C-----A---AA--AA--
AJ971484: -----
AJ971485: TT---CAGCT-T-ATAACG-TTATGT-AGCT----GA-----AAG
AJ841770: -----TTTAAAAAGGAAAGCTATGG

```

Figure 8: Clustal v1

```

AJ841768: -----A-A-----A---A-----A-GAAAAGCTGAAGACTGTAA
AJ971484: TTTAAAAAGGAAAGCTATGACAAAAAAGAAAGCTGAAGACTGTAA
AJ971485: ATTCCAAGAGCTAAGGGTGATTACCCGACGAAAAGCTGAAGACTGTAA
AJ841770: CAAAAAAGAGTTAACACTGCTTGGAACCAAAAAGACCTCGTTCA
*
AJ841768: AGGAAAACCTGGAAAAATCTGTCTGACTCTGAAAAGGAATTATATATTCAG
AJ971484: AGGAAAACCTGGAAAAATCTGTCCGACTCTGAAAAGGAATTATATATTCAG
AJ971485: AGGAAAACCTGGAAAAATCTGTCTGACTCTGAAAAGGAATTATATATTCAG
AJ841770: GCTTATAACGTTTATGTAGCTGAAAGATTCGAAGAAGTTCAAGGATTC
*
AJ841768: CATGCTAAAGAGGACGAAACTCGTTATCATAATGAAATGAAATCCTGGGA
AJ971484: CATACTAAAGAGGACGAAACTCGTTATCATAATGAAATGAAATCCTGGGA
AJ971485: CATGCTAAAGAGGACGAAACTCGTTATCATAATGAAATGAAATCCTGGGA
AJ841770: ACCACAGGAAAAGCTGAAGACTCTAAAGGAAAACCTGGAAAAATCTGTCTG
*
AJ841768: AGAACAAATG-----
AJ971484: AGAACAAATG-----
AJ971485: AGAACAGATG-----
AJ841770: ACTCTGAAAAGGAATTATATATTCAGTATGCTAAAGAGGATGAAACTCGT
*
AJ841768: -----
AJ971484: -----
AJ971485: -----
AJ841770: TATCATAATGAAATGAAATCCTGGGAAGAGCAGATG

```

Figure 9: Clustal v2

seq1: Gallus lafayetii DNA, chicken repeat 1 (CR1) element, clone:CJFCR1d1.
seq2: Gallus sonneratii DNA, chicken repeat 1 (CR1) element, clone:GyJFINDR1d1..
seq3: Gallus varius DNA, chicken repeat 1 (CR1) element, clone:GJFCR1d1.
seq4: Gallus gallus DNA, chicken repeat 1 (CR1) element, clone:NLAOCR1d2

Pierwsza sekwencja pochodzi od kura cejlońskiego, druga od kura siwego, trzecia od kura zielonego, a czwarta od kura bankiwa.

W tym przypadku możemy zaobserwować ciekawą sytuację, a mianowicie bardzo niski koszt dopasowania co świadczy o prawie całkowitym zachowaniu całej sekwencji w obrębie wszystkich 4 osobników. Potwierdza to zapis w formacie CLUSTAL i ilość gwiazdek jaką można tam dostrzec. Występują tam pojedyncze substytucje.

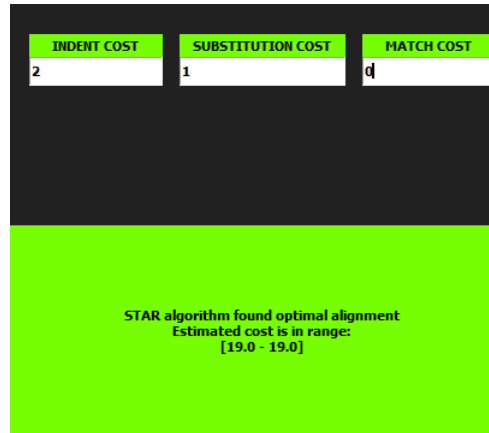


Figure 10: Wynik dopasowania multisekwencyjnego

```

AB166711: GAATCATAGAACAATTGGGTTGGAAGAGCCCCCTAGTGATTATCTAGTC
AB166714: GAATCATAGAACAATTGGGTTGGAAGAGCCCCCTAGTGATTATCTAGTC
AB166713: GAATCATAGAACAATTGGGTTGGAAGAGCCCCCTAGTGATTATCTAGTC
AB166717: GAATCATAGAACAATTGGGTTGGAAGAGCTCCCTAGTGATTATCTTGTG
*****
AB166711: CATCCCTCCACCATGGGCAGGGACACATCCCACTCCTATGTCATGTTGCC
AB166714: CATCCCTCCACCATGGGCAGGGACACATCCCACTCCTATGTCATGTTGCC
AB166713: CATCCCTCCACCATGGGCAGGGACACATCCCACTCCTATGTCATGTTGCC
AB166717: CATCCCTCCACCATGGGCAGGGACACATCCCACTCCTATGTCATGTTGCC
*****
AB166711: CAAAGGCCCATCTAGCCTGGTCTTGAATGCCTCCAGAGAGCGGGCACCCA
AB166714: CAAAGGCCCATCTAGCCTGGTCTTGAATGCCTCCAGAGAGCGGGCACCCA
AB166713: CAAAGGCCCATCCAGCCTGGTCTTGAATGCCTCCAGAGAGCGGGCACCCA
AB166717: CAAAGGCCCATCCAGCCTGGTCTTGAATGCCTCCAGAGAGCGGGCACCCA
*****
AB166711: CAGTCCTCTGGACAGCCTGTCCAGCATCTTGCCACTTTGTTGTGGAAA
AB166714: CAGTCCTCTGGACAGCCTGTCCAGCATCTTGCCACTTTGTTGTGGAAA
AB166713: CAGTCCTCTGGACAGCCTGTCCAGCATCTTGCCACTTTGTTGTGGAAA
AB166717: CAGTCCTCTGGACAGCCTGTCCAGCATCTTGCCACTTTGTTGTGGAAA
*****
AB166711: ACTTCCTCCTTATGTGCAGTTTCCTTCAGCTTAAAACCATGCCCCCTTGG
AB166714: ACTTCCTCCTTATGTGCAGTTTCCTTCAGCTTAAAACCATGCCCCCTTGG
AB166713: ACTTCCTCCTTATGTGCAGTTTCCTTCAGCTTAAAACCATGCCCCCTTGG
AB166717: ACTTCCTCCTTATGTGCAGTTTCCTTCAGCTTAAAACCATGCCCCCTTGG
*****

```

Figure 11: Clustal v1

```

AB166711: CCTGTCCCTACAGTTTGGTAAAAGGTTTACTCCAGCTGTAGTATATAC
AB166714: CCTGTCCCTACAGTTTGGTAAAAGGTTTACTCCAGCTGTAGTATATAC
AB166713: CCTGTCCCTACAGTTTGGTAAAAGGTTTACTCCAGCTGTAGTATATAC
AB166717: CCTGTCCCTACAGTTTGATAAAAGGTTTACTCCAGCTGTAGTATATAC
*****
AB166711: TGCAAAGCTGCAGCAAGGTCTCCCTGGAACCTTTCCACGCTCAGTGACC
AB166714: TGCAAAGCTGCAGCAAGGTCTCCCTGGAACCTTTCCACGCTCAGTGACC
AB166713: TGCAAAGCTGCAGCAAGGTCTCCCTGGAACCTTTCCACGCTCAGTGACC
AB166717: TGCAAAGCTGCAGCAAGGTCTCCCTGGAACCTTTCCACGCTCAGTGACC
*****
AB166711: CCAACTCTCCCTGCCTTTCCTCATAGAAGAGGTGTCCAGCTTGCTAATA
AB166714: CCAACTCTCCCTGCCTTTCCTCATAGAAGAGGTGTCCAGCTTGCTAATA
AB166713: CCAACTCTCCCTGCCTTTCCTCATAGAAGAGGTGTCCAGCTTGCTAATA
AB166717: CCAACTCTCCCTGCCTTTCCTCATAGAAGAGGTGTCCAGCTTGCTAATA
*****
AB166711: ATTCTTGTGGCCTTTCTCTGTCCACGTCCTCCTTGTGCTGGGGACCCAAT
AB166714: ATTCTTGTGGCCTTTCTCTGTCCACGTCCTCCTTGTGCTGGGGACCCAAT
AB166713: ATTCTTGTGGCCTTTCTCTGTCCACGTCCTCCTTGTGCTGGGGACCCAAT
AB166717: ATTCTTGTGGCCTTTCTCTGTCCACGTCCTCCTTGTGCTGGGGACCCAAT
*****
AB166711: A
AB166714: A
AB166713: A
AB166717: A
*
```

Figure 12: Clustal v2