Algorytmy kombinatoryczne w bioinformatyce − Zadanie III

Parametry wejściowe:
akwb-zadanie-3.exe nazwa_testu długość_podsekwencji minimalna_wiarygodność
np.
akwb-zadanie-3.exe tests/test_01 6 10

Opis algorytmu:

1. **Wczytanie sekwencji z plików wejściowych**
   Otwieramy pliki fasta z sekwencjami nukleotydów i qual z wiarygodnością, nazwa wczytywana jest z pierwszego parametru programu. Czytamy po kolei bloki kolejnych sekwencji. Z nagłówka wyciągamy ID sekwencji, z zawartości nukleotydy i wiarygodność dla każdego z nich. Dla każdej wczytanej sekwencji tworzymy odpowiedni obiekt przechowujący te 3 informacje.

2. **Tworzenie grafu**
   Wszystkie wczytane wcześniej informacje wykorzystywane są w tworzeniu grafu.

   Pierwszym etapem jest utworzenie wierzchołków. Iterujemy po wszystkich sekwencjach tworząc wierzchołki zawierające n-elementowe podciągi, na podstawie drugiego parametru określana jest ich długość. Nukleotydy, które nie spełniają kryterium minimalnej wiarygodności podanej w trzecim parametrze są pomijane. Każdy wierzchołek przechowuje informacje o sekwencji z której pochodzi. Zapisywana jest także pozycję pierwszego nukleotydu podsekwencji w sekwencji pierwotnej.

   Następnie wśród wszystkich wierzchołków wyszukiwane są identyczne podciągi pochodzące z różnych sekwencji wejściowych, tworzymy pomiędzy nimi nieskierowane krawędzie.

3. **Wyszukiwanie kliki**
   Wyszukujemy w grafie wierzchołek z największą ilością krawędzi, od niego zaczynamy poszukiwanie, zapisujemy go listy odwiedzonych wierzchołków.

   Dla właśnie odwiedzonego wierzchołka sprawdzamy wszystkie krawędzie. Jeżeli wierzchołek do którego prowadzi został nie został jeszcze odwiedzony i nie ma jeszcze wierzchołka z takiej samej pierwotnej sekwencji, dodajemy go do listy i wykonujemy tą samą operację.

   Kończymy gdy wyczerpaliśmy wyniki, tzn. nie ma krawędzi prowadzącej do nowego, nieodwiedzonego wierzchołka, zwracamy listę z odwiedzonymi wierzchołkami. Wypisujemy rezultaty, tzn. nazwę sekwencji pierwotnej, pozycję podsekwencji w sekwencji pierwotnej i samą podsekwencję.

Oszacowanie złożoności algorytmu:

- W optymistycznym przypadku, gdy wszystkie nukleotydy nie będą spełniały minimalnego progu wiarygodności złożoność jest liniowa (**O(n)**)
- W pesymistycznym przypadku złożoność wielomianowa (**O(n^x)**)

Wnioski:
Im wyższy minimalny próg wiarygodności tym trudniej znaleźć powtarzające się fragmenty. Przy krótkich fragmentach algorytm dopasowuje więcej podobnych podsekwencji, zdarza się że wybrany wynik nie należy do motywu i znajduje się poza nim.

TEST 1:

>DOJHLOP01BA24W length=99 xy=0419_0782 region=1 run=R_2005_09_08_15_35_38_
TGGCTCTGGTGGCGCATTGGGGATAGGCGTCGCAGACAGGTTACTTATGTTTGAACATAG
TGTTTACACAGTTGCAAGCCCTGAAGTCTTGTGCTTCGA
>DOJHLOP01EKI27 length=119 xy=1756_3857 region=1 run=R_2005_09_08_15_35_38_
ATCAATCTGATTCTCTAATTCAACAAGAGGTTTTTCAAATTCGAGGAGAAAACGTCTAGC
CAATAAAAAAAAAGTAACCAAATGTTCAGAACTTATGTTTGAACATAGATGATTTCAAA
>DOJHLOP02F5RQK length=116 xy=2408_3866 region=2 run=R_2005_09_08_15_35_38_
TGGCTTTACATCAATAATTTTTGAGAATTTTAATTTGCTCCAAAATCAACAAGTTAGAAA
TACGACTTATGTTTGAACATACTTTACCACTGGTAATACTCATACAGCTTGGGGTG
>DOJHLOP01CMF1Y length=94 xy=0958_2612 region=1 run=R_2005_09_08_15_35_38_
CATAGTTCAAATGCAAGTATTCATCTTACTTATGTTTGACCATAGTTCTTTTGGCGAATG
GATCAGCAACACGTATATAAACGTGATATGCGGA
>DOJHLOP01DGTNK length=93 xy=1304_2702 region=1 run=R_2005_09_08_15_35_38_
CTTGAAGGAGCATCACCGTGCGCATGGTTGACAGAAGATTTGAGCTCATCTGGAGTTTAT
TGGCGAACTTATGTCCGAACATAGTGAATTGGG

>DOJHLOP01BA24W length=99 xy=0419_0782 region=1 run=R_2005_09_08_15_35_38_
32 30 25 27 32 30 27 30 26 32 29 24 31 31 32 32 30 27 25 25 20 06 31 31 31 29
24 31 31 31 31 31 29 32 31 28 31 28 28 22 29 24 32 32 28 22 31 29 29 24 23 07
28 30 25 25 31 31 29 31 26 25 24 23 07 28 28 32 31 20 26 30 25 31 32 25 18 31
27 26 13 30 27 24 16 27 12 31 30 27 28 27 32 28 31 28 30 30 32
>DOJHLOP01EKI27 length=119 xy=1756_3857 region=1 run=R_2005_09_08_15_35_38_
31 30 30 31 28 30 32 31 31 26 31 26 32 32 31 30 28 23 31 26 31 31 26 26 31 27
30 31 28 22 22 22 20 16 08 32 23 22 04 29 23 26 31 26 29 23 29 32 24 24 19 03
31 32 32 29 29 29 28 29 24 30 25 32 09 09 09 09 08 07 05 03 01 31 30 31 28 26
20 29 28 17 29 31 26 20 24 31 31 31 30 28 22 30 27 28 23 29 31 32 28 31 31 28
22 16 32 28 27 19 31 29 27 27 19 19 28 28 19
>DOJHLOP02F5RQK length=116 xy=2408_3866 region=2 run=R_2005_09_08_15_35_38_
31 31 27 32 29 28 18 28 32 31 30 30 31 27 30 27 24 21 21 20 16 09 31 30 32 31
27 24 24 22 13 28 25 27 26 18 31 32 32 31 27 24 24 22 13 31 31 28 26 31 29 27
32 31 27 29 32 29 28 16 31 32 32 32 25 32 31 32 30 27 30 27 31 27 30 27 30 28
27 18 31 27 29 28 17 30 31 27 31 30 31 31 27 29 31 27 32 29 32 31 30 30 29 30
31 27 32 32 27 25 24 24 19 04 30 32
>DOJHLOP01CMF1Y length=94 xy=0958_2612 region=1 run=R_2005_09_08_15_35_38_
30 31 31 32 32 31 28 31 29 28 16 32 32 31 27 21 32 31 28 31 27 31 31 32 32 31
27 31 28 30 31 31 28 23 32 27 27 19 31 29 31 30 32 31 26 23 22 31 23 23 21 13
29 27 32 28 31 28 29 25 23 24 31 29 28 31 31 31 28 26 30 31 30 32 31 30 18 32
28 28 19 30 27 32 31 30 31 31 29 25 31 31 28 32
>DOJHLOP01DGTNK length=93 xy=1304_2702 region=1 run=R_2005_09_08_15_35_38_

29 29 26 27 31 27 27 25 32 28 31 28 29 21 20 31 28 23 32 32 31 30 31 29 24 29
26 28 22 30 30 31 31 23 30 27 32 30 27 27 19 21 30 32 27 24 31 32 27 29 29 25
24 32 31 23 22 02 32 31 28 25 23 29 31 30 24 29 31 28 23 32 24 23 22 30 27 31
27 26 13 24 23 28 30 30 31 26 27 25 25 25 18

```
# test_01 długość = 4 wiarygodność = 10
OJHLOP01BA24W 49 TTGA
OJHLOP01EKI27 98 TTGA
OJHLOP02F5RQK 19 TTGA
OJHLOP01CMF1Y 35 TTGA
OJHLOP01DGTNK 1  TTGA

# test_01 długość = 6 wiarygodność = 20
OJHLOP01BA24W 42 ACTTAT
OJHLOP01EKI27 90 ACTTAT
OJHLOP02F5RQK 64 ACTTAT
OJHLOP01CMF1Y 27 ACTTAT
OJHLOP01DGTNK 66 ACTTAT

# test_01 długość = 8 wiarygodność = 20
OJHLOP01BA24W 42 ACTTATGT
OJHLOP01EKI27 90 ACTTATGT
OJHLOP02F5RQK 64 ACTTATGT
OJHLOP01CMF1Y 27 ACTTATGT
OJHLOP01DGTNK 66 ACTTATGT
```

TEST 2:

>DOJHLOP02GAWHC length=99 xy=2467_1630 region=2 run=R_2005_09_08_15_35_38_
CTATTCGAGCTAACCCTCCTATTAGAGCTTGATCATCAGTACCGTTCCTGTCTCCATGTA
ATTCAACCCAATCATCAACAAAACATCTGTATGTAATCG
>DOJHLOP02JHWE0 length=120 xy=3777_1226 region=2 run=R_2005_09_08_15_35_38_
TATTGTTATTTGGGTTGCAAGCATGCCTAAAACCTTAGAAATTGACCCATCTAAAATTAA
AAGACCTTGTTTAATGATCCGTTCCTGTATCCCTAAAAACTTGGGTGAGAAGTTTTCAGG
>DOJHLOP01AZPR0 length=120 xy=0289_2734 region=1 run=R_2005_09_08_15_35_38_
TATTTATCCAAGTTTGAAGATTAGACCACCACGGTCACAAAAAAAAAAGAATTCCTCAAGA
ACTTTAAGTTCTCGGTTCCTGTCTCAAGATGCAAAATATTAAATCCTTATATCACCGCTG
>DOJHLOP01BA0GQ length=63 xy=0418_1416 region=1 run=R_2005_09_08_15_35_38_
CTTTGTAAGATTCTGAGGTGAAATCTGGTCCGTTCCTGTCTCTTAGCTCTTCTTTTTCTA
CGG
>DOJHLOP02GU4JI length=107 xy=2697_3116 region=2 run=R_2005_09_08_15_35_38_
AATAGATTAGTTTTTTGCGTTCCTGTCTATGGGCAAATTGGTCTGCTCCAGGCGGTTTCA
AGGCAGTTGAGAATTCAATGTCCTTCAATAGCTGTATTGCCTTGTTT

>DOJHLOP02GAWHC length=99 xy=2467_1630 region=2 run=R_2005_09_08_15_35_38_
32 32 31 31 27 32 32 27 32 32 32 31 27 29 28 17 32 31 27 32 30 31 27 32 31 27
32 31 30 27 32 31 31 32 30 31 31 26 31 30 27 29 27 28 31 27 31 27 30 29 32 25
25 30 26 27 31 31 26 28 22 30 25 17 30 27 29 28 16 30 25 30 32 32 28 28 20 10
31 25 25 22 12 31 26 30 32 32 31 32 32 32 32 32 31 26 28 31 25
>DOJHLOP02JHWE0 length=120 xy=3777_1226 region=2 run=R_2005_09_08_15_35_38_
32 31 31 26 32 31 27 32 28 28 16 28 28 16 31 27 31 32 28 22 32 31 32 30 32 31
27 32 26 25 22 11 31 27 31 27 25 31 25 24 09 31 27 31 29 28 27 14 31 32 31 26
24 24 22 13 31 27 24 24 21 13 31 31 31 27 24 23 32 25 25 18 24 22 32 32 31 30

26 32 28 31 29 26 31 30 26 30 32 18 29 28 18 30 22 22 21 16 08 29 31 27 28 28
18 28 32 28 29 28 21 31 24 24 22 13 31 31 30 27
>DOJHLOP01AZPR0 length=120 xy=0289_2734 region=1 run=R_2005_09_08_15_35_38_
27 32 28 27 19 19 30 31 27 27 21 31 28 28 16 29 26 20 30 31 30 27 32 31 30 31
28 29 31 28 32 31 31 27 30 32 18 31 09 09 09 08 08 07 05 03 01 30 30 25 30 28
31 27 25 29 25 24 32 27 25 29 25 25 18 31 28 28 31 28 30 31 27 24 29 30 25 28
22 27 23 22 31 23 26 25 18 28 29 28 31 30 26 25 22 10 29 25 26 20 29 28 17 31
31 26 31 28 24 29 31 31 19 32 27 25 31 30 32 29
>DOJHLOP01BA0GQ length=63 xy=0418_1416 region=1 run=R_2005_09_08_15_35_38_
32 29 28 18 31 30 31 28 31 31 31 28 31 31 30 27 31 28 32 30 27 27 13 31 32 31
31 27 27 21 31 32 29 29 31 30 25 25 31 28 23 29 31 28 30 28 32 28 29 23 22 30
22 22 21 16 07 32 32 30 31 31 27
>DOJHLOP02GU4JI length=107 xy=2697_3116 region=2 run=R_2005_09_08_15_35_38_
31 27 31 32 31 31 24 22 28 31 12 11 11 11 10 08 05 02 31 27 31 31 31 31 29 32
32 26 24 32 29 28 18 30 29 28 18 29 26 31 27 32 30 31 31 29 32 31 27 28 29 27
32 31 27 28 27 18 32 31 26 31 27 32 22 30 31 27 32 31 31 26 19 30 25 30 30 25
31 30 25 31 27 30 27 20 31 27 32 23 32 31 31 32 31 32 31 26 30 30 25 24 22 32
28 28 18

```
# test_02 długość = 4 wiarygodność = 10
OJHLOP02JHWE0 82 TCCT
OJHLOP02GAWHC 16 TCCT
OJHLOP01AZPR0 51 TCCT
OJHLOP01BA0GQ 33 TCCT
OJHLOP02GU4JI 20 TCCT

# test_02 długość = 6 wiarygodność = 20
OJHLOP02GAWHC 43 GTTCCT
OJHLOP02JHWE0 80 GTTCCT
OJHLOP01AZPR0 74 GTTCCT
OJHLOP01BA0GQ 31 GTTCCT
OJHLOP02GU4JI 18 GTTCCT

# test_02 długość = 8 wiarygodność = 20
OJHLOP02GAWHC 43 GTTCCTGT
OJHLOP02JHWE0 80 GTTCCTGT
OJHLOP01AZPR0 74 GTTCCTGT
OJHLOP01BA0GQ 31 GTTCCTGT
OJHLOP02GU4JI 18 GTTCCTGT
```

TEST 3:

>DOJHLOP01C9Z1H length=99 xy=1226_3875 region=1 run=R_2005_09_08_15_35_38_
TCATCACAAAACATCTGTATGTAATCGAGTGTGCTAGGCCTTTTGGGGGGTGCCTAGCTA
CTTGAATCTTCTGAGCTGGAGTAAGTGCATTAAATATTT
>DOJHLOP02I97ND length=115 xy=3689_2983 region=2 run=R_2005_09_08_15_35_38_
TCAAGACTAAGAATTCTTCCTCCTCCAAGTCCTGGAGTAAGAGCAAGTAAAGGCTTTTGC
TGTCTGGCTACTAAAAGGAGTTCAGAAACACCAGTTCGATTGGAAAGCCAACCTA
>DOJHLOP01CJEA0 length=98 xy=0923_3738 region=1 run=R_2005_09_08_15_35_38_
AAGATGCAAAATATTAAATCCTTATATCACCGCTGCAAGAGCGCTGGAGTAAGTGTCTCT
ATTATCAAGACCTATTTCAAGAAGGGAATCTTGATAAG
>DOJHLOP02F99O5 length=106 xy=2460_0775 region=2 run=R_2005_09_08_15_35_38_
ATTAAATTCTGTAAGATCACTCAGATTGACAAATTTCACTGGAGTAAGCGCAAGCTTTTG
CATCCAACAATTTAACCGTAAGAAAAAGAAGAGCTAAGTTCTAACT
>DOJHLOP01C3AP3 length=112 xy=1150_2425 region=1 run=R_2005_09_08_15_35_38_
AGACGTTAAAAAAACTAATCTATTAAATCCCATAGAGGCTATTGCTGGAGTAAGTACAGT

AACCATTAGAAATCAAAACATCAGCTGATAGGGAACAAAGTTCCTGGGAAAG

>DOJHLOP01C9Z1H length=99 xy=1226_3875 region=1 run=R_2005_09_08_15_35_38_
32 31 31 31 32 30 32 25 25 22 11 31 32 32 32 31 32 32 31 32 31 31 30 26 31 32
29 29 32 31 31 32 31 32 30 27 28 26 28 26 24 24 19 03 19 19 18 16 10 03 29 24
30 28 31 30 32 29 29 31 30 31 26 31 26 20 26 28 28 26 30 24 27 21 30 31 31 27
25 31 29 27 30 26 31 30 30 31 17 27 25 29 28 18 31 21 25 25 18
>DOJHLOP02I97ND length=115 xy=3689_2983 region=2 run=R_2005_09_08_15_35_38_
29 30 29 23 30 30 28 31 31 27 32 29 23 30 25 32 30 27 30 24 31 30 24 32 31 26
31 26 28 31 26 31 27 27 24 27 25 28 30 25 31 31 27 31 31 27 23 31 28 28 15 31
26 29 25 24 20 07 31 31 30 31 22 32 27 31 27 29 20 17 31 28 24 24 21 13 30 27
13 27 31 27 28 21 31 25 24 08 20 32 29 23 27 32 29 24 31 24 30 27 21 28 26 23
22 03 24 26 24 20 10 31 27 32 17
>DOJHLOP01CJEA0 length=98 xy=0923_3738 region=1 run=R_2005_09_08_15_35_38_
30 25 31 31 31 30 30 25 25 20 07 31 27 30 25 27 27 13 32 29 24 31 27 31 31 32
29 32 32 31 28 32 32 32 32 31 30 25 29 30 28 30 26 32 29 27 31 31 28 32 29 29
24 29 28 17 32 31 28 29 29 31 28 32 28 31 31 28 30 32 31 28 19 26 28 28 16 31
29 23 29 23 16 25 24 08 19 09 31 30 29 24 30 31 30 30 25 32
>DOJHLOP02F99O5 length=106 xy=2460_0775 region=2 run=R_2005_09_08_15_35_38_
31 30 26 29 28 18 31 27 32 32 32 31 31 27 31 30 31 31 32 31 30 32 27 31 30 31
26 32 31 32 29 28 18 29 28 17 32 32 28 26 28 27 30 31 32 29 31 25 31 32 30 31
27 32 32 25 25 22 13 32 32 32 31 31 27 25 23 32 25 23 28 27 18 27 25 31 27 26
25 20 10 32 21 21 20 16 08 30 30 24 31 32 30 28 24 30 27 29 31 27 26 30 30 27
31 32
>DOJHLOP01C3AP3 length=112 xy=1150_2425 region=1 run=R_2005_09_08_15_35_38_
26 26 22 31 31 28 23 15 15 15 13 11 08 03 31 22 31 28 30 28 32 22 28 22 29 28
16 30 25 24 07 32 32 31 28 25 24 17 30 19 21 26 25 10 29 28 27 28 22 28 22 31
27 31 30 29 32 22 24 31 31 27 22 14 31 20 10 09 29 23 22 03 30 31 25 25 21 08
26 24 29 32 30 17 29 28 31 32 21 31 24 23 07 23 15 31 29 28 16 32 31 28 30 26
26 29 28 18 24 23 07 30

```
# test_03 długość = 4 wiarygodność = 10
OJHLOP01C9Z1H 81 TAAG
OJHLOP02I97ND 7  TAAG
OJHLOP01CJEA0 49 TAAG
OJHLOP02F99O5 11 TAAG
OJHLOP01C3AP3 50 TAAG

# test_03 długość = 6 wiarygodność = 20
OJHLOP01C9Z1H 79 AGTAAG
OJHLOP02I97ND 35 AGTAAG
OJHLOP01CJEA0 47 AGTAAG
OJHLOP02F99O5 42 AGTAAG
OJHLOP01C3AP3 48 AGTAAG

# test_03 długość = 8 wiarygodność = 20
OJHLOP01C9Z1H 75 CTGGAGTA
OJHLOP02I97ND 31 CTGGAGTA
OJHLOP01CJEA0 43 CTGGAGTA
OJHLOP02F99O5 38 CTGGAGTA
OJHLOP01C3AP3 44 CTGGAGTA
```