Algorytmy Kombinatoryczne W Bioinformatyce projekt 3

Natalia Michałkiewicz 147902

Cel projektu:

Program, którego funkcją będzie znalezienie struktury gwiazdy(struktury podobnej do kliki), w podanych na wejściu pięciu sekwencjach. Jako parametry mamy do dyspozycji długość podciągu i minimalny próg wiarygodności, które wprowadza użytkownik.

Opis funkcji użytych w programie:

Odczyt:

Wczytywanie podanych przez użytkownika dwóch plików: .fasta z sekwencjami i .qual z jakością. Zawiera pętle znajdujące znak „>" czyli początek, omija go i wczytuje sekwencje i jakość do dwóch osobnych wektorów.

Usuwanie:

Na początku używając pętli for zakropkowuję ('.') miejsca w sekwencji poniżej progu, to samo robię dla jakości, lecz tam wpisuję '0' w miejsca poniżej progu. Następnie tworzę wektor wektorów na nowe pozycje, czyli przepisuję wszystkie z pominięciem '.' i '0'. Dzięki temu pozwala mi to na zachowanie oryginalnych numerów pozycji sekwencji.

Tworzenie grafu:

Stworzyłam w tym celu dwie funkcje pomocnicze, które opierają się na warunku z zadania („połączenie wierzchołków nieskierowanymi krawędziami, jeśli odpowiadają one takim samym podciągom występującym w różnych sekwencjach, a różnica w pozycjach podciągów wewnątrz sekwencji nie jest większa niż dziesięciokrotność długości podciągu").

Samo tworzenie grafu opiera się na czterech pętlach for, gdzie dwie pierwsze przechodzą po pierwszym zbiorze, a dwie następne dlatego że mamy połączenia z innych sekwencji- w której warunek if sprawdza możliwość stworzenia krawędzi: muszą być takie same podciągi, pochodzić z innych sekwencji i 10krotność podciągu. Na wyjściu otrzymujemy graf w postaci wierzchołków-podciągów, z informacją o numerze sekwencji w jakiej się znajduje i na której pozycji i jego następnikami.

Opis algorytmu:

Wyszukanie motywu :

Na początku sprawdzam krawędź po krawędzi w funkcji *validateqlique* czy następniki łączą się z kolejnymi sekwencjami i przyznaję punkty za połączenie. Jeśli wierzchołek posiada wszystkie punkty (>0) może być potencjalną kliką. Następnie tworzę mapę zawierającą parametry: nr i pozycję w pierwotnej sekwencji oraz wszystkie krawędzie. Później w funkcji pomocniczej *uzupelnijmape* przechodzę po wszystkich następnikach pierwszej sekwencji i dla każdego wierzchołka tworzę numer, którego nie mogę użyć (ponieważ w następnikach może się pojawić np. sytuacja: 1,2, 2, 3, 4 – nie może być powtórzeń). Później do wektora tymczasowego przypisuję dozwolone krawędzie z pominięciem użytych wcześniej. Następnie w głównej funkcji *motywy* przechodzę po każdej krawędzi za pomocą pętli for, stamtąd do każdej pojedynczej sekwencji i jeśli pozycje się zgadzają z elementem mapy to porównuję wektory. Do tego służy ostatnia już funkcja pomocnicza *porownajwektory*. Sprawdza czy każdy podciąg pokrywa się z każdym w kolejnych czterech sekwencjach. Zwracana jest pierwsza napotkana struktura gwiazdy.

Złożoność obliczeniowa:

O(n^4) – obliczone na podstawie ilości zgnieżdżonych w sobie pętli

Instancje:

1 INSTANCJA

.fasta:

```
>DOJHLOP01BC4SV length=99 xy=0442_2045 region=1 run=R_2005_09_08_15_35_38_
CAGGCGTCGCAGACAGGTTACTTATGTTTGAACATAGTGTTTACACAGTTGCAAGCCCTGTAGCTTGTGCTTCGAATCTATGGAGGG
ATGCTGGCAAGG

>DOJHLOP02GF4GK length=95 xy=2526_3586 region=2 run=R_2005_09_08_15_35_38_
CTGGATGCCCTGAAGCTTGTGCTTCGAGTCTATGGAGGGATGCTGGCAAGGCTCCGGAAGCAGCATCAGCAATTTAAAAAATTACTG
GACCTGAT

>DOJHLOP01CC1S3 length=91 xy=0851_2517 region=1 run=R_2005_09_08_15_35_38_
AGTTGCAAGCCCTGAAGCTTGTGCTTCGACTCTATGGAGGGATGCTGGCAAGGCTCCGGAAGCAGCATCAGCAATTTAAAAAATTAC
TGGA

>DOJHLOP01C1GFF length=99 xy=1129_2521 region=1 run=R_2005_09_08_15_35_38_
TTTACACAGTTGCAAGCCCTGAAGCTTGTGCTTCGATTCTATGGAGGGATGCTGGCAAGGCTCCGGAAGCAGCATCAGCAATTTAAA
AAATTACTGGGA

>DOJHLOP02I6MO4 length=95 xy=3648_3798 region=2 run=R_2005_09_08_15_35_38_
TACTTATGTTTGAACATAGTGTTTACACAGTTGCAAGCCCTGAAGCTTGTGCTTCGAATGTATGGAGGGATGCTGGCAAGGCTCCGG
AAGCAGCA
```

.qual:

```
>DOJHLOP01BC4SV length=99 xy=0442_2045 region=1 run=R_2005_09_08_15_35_38_
31 31 31 26 31 31 31 32 32 31 29 32 32 31 31 31 26 31 27 30 31 30 25 28 31 32 26 25 9
32 26 20 31 32 31 31 32 31 31 29 28 17 24 31 30 31 28 32 31 28 31 31 28 23 32 28 28 16
32 31 26 20 31 32 31 28 29 27 32 31 31 28 31 32 32 30 27 31 31 31 31 24 23 30 29 28 17
20 32 25 32 32 29 27 31 31 28 27 25

>DOJHLOP02GF4GK length=95 xy=2526_3586 region=2 run=R_2005_09_08_15_35_38
31 27 31 32 29 24 31 29 28 17 31 32 26 19 31 31 31 27 30 32 32 32 30 27 32 32 30 29 26
30 30 30 31 23 22 25 28 27 18 28 32 29 31 32 31 27 32 28 22 29 26 26 31 31 26 26 24 24
17 29 31 29 29 32 32 31 30 30 30 31 31 26 23 22 3 19 19 18 16 11 3 27 24 23 21 31 23 22
25 31 27 28 30 32 17

>DOJHLOP01CC1S3 length=91 xy=0851_2517 region=1 run=R_2005_09_08_15_35_38
32 31 28 23 32 31 29 23 32 29 28 18 31 30 28 22 31 32 31 28 28 32 31 29 31 28 31 32 31
28 26 31 30 26 32 27 25 28 29 28 18 24 30 28 29 32 31 28 30 31 26 30 28 24 31 31 27 27
25 29 23 29 31 30 30 31 31 29 26 31 31 32 27 21 23 22 5 19 19 18 16 11 3 30 28 27 29 29
28 26 30

>DOJHLOP01C1GFF length=99 xy=1129_2521 region=1 run=R_2005_09_08_15_35_38
28 28 16 26 31 32 31 28 32 31 27 30 31 29 23 32 29 28 16 31 31 29 24 32 31 28 26 31 31
31 32 30 27 32 31 31 29 27 31 32 30 32 25 23 31 29 28 18 26 31 27 32 32 31 28 31 27 21
31 28 24 32 31 27 28 26 25 19 31 32 29 31 31 32 30 31 30 32 32 29 24 23 22 4 18 18 17
16 11 5 28 26 25 26 32 23 22 2 31

>DOJHLOP02I6MO4 length=95 xy=3648_3798 region=2 run=R_2005_09_08_15_35_38_
32 30 32 31 26 30 29 32 29 28 17 32 25 18 31 31 32 31 32 27 32 29 28 16 30 32 32 32 30
29 31 27 31 32 29 23 32 28 28 18 31 28 27 21 32 32 29 27 30 30 31 29 29 26 30 32 28 27
25 31 30 31 32 23 22 32 26 26 18 27 28 21 31 29 30 27 31 30 25 25 23 17 31 31 27 25 23
26 19 28 28 32 29 32 32
```

test 1:

próg: 26

długość podciągu: 8

motyw: GCTTGTGC

sekwencja: 1, nr pozycji: 63

sekwencja: 2, nr pozycji: 15

sekwencja: 3, nr pozycji: 17

sekwencja: 4, nr pozycji: 24

sekwencja: 5, nr pozycji: 45

test 2:

próg: 21

długość podciągu: 5

motyw: GCCTG

sekwencja: 1, nr pozycji: 55

sekwencja: 2, nr pozycji: 7

sekwencja: 3, nr pozycji: 9

sekwencja: 4, nr pozycji: 16

sekwencja: 5, nr pozycji: 37

test 3:

próg: 30

długość podciągu: 4

Nie znaleziono motywu

test 4:

próg: 28

długość podciągu: 6

motyw: AGCCTG

sekwencja: 1, nr pozycji: 53

sekwencja: 2, nr pozycji: 5

sekwencja: 3, nr pozycji: 7

sekwencja: 4, nr pozycji: 14

sekwencja: 5, nr pozycji: 35

2 INSTANCJA

.fasta:

```
>DOJHLOP02GOMKX length=105 xy=2623_3007 region=2 run=R_2005_09_08_15_35_38_
CAAGAGTCGTATTTCTAACTTGTTGATTTTGGAGCAATGACGCTGATGGCAAAATTATTGATGTAAAGCCACCTAGAGCAGTGATAC
TGATCTCTTTCTTTTGCG

>DOJHLOP02HH1SG length=109 xy=2958_3582 region=2 run=R_2005_09_08_15_35_38_
TGCAAGAAGAAAGGTTATGACGCTGATGGCGTTGGCGCAAAAGAAAGAGATCAGTATCACTGCTCTAGGTGGCTTTACATCAATTAA
TTTTTGAGTAATTTTTAATTTG

>DOJHLOP01DBEEU length=106 xy=1242_3620 region=1 run=R_2005_09_08_15_35_38_
CGCTTCAAAACTGCAAGAAGAAAGGTTTTGAATGCAATATGACGCTGATGGCAAGAAAGAGATCAGTATCACTGCTCTAGGTGGCTT
TACATCAAATTAATTTTTG

>DOJHLOP02GU0TZ length=98 xy=2696_2405 region=2 run=R_2005_09_08_15_35_38_
CCACCTAGAGCAGTGATACTGATCTCTTTCTTTTGCGCCAACTCCATTGCATTCAAAACCTTTCTATGACGCTGATGGCAGCGACTT
AGCATCTCTGG
>DOJHLOP02I4IZJ length=98 xy=3624_3981 region=2 run=R_2005_09_08_15_35_38_
AGCAGTGATACTGATCATGACGCTGATGGCAACTCCATTGCATTACAAACCTTTCTTCTTGCTAGTTTGAAGCGACTTAGCATCTCT
AGGAAACAAAG
```

.qual

```
>DOJHLOP02GOMKX length=105 xy=2623_3007 region=2 run=R_2005_09_08_15_35_38_
32 31 26 32 29 32 32 31 32 30 31 29 28 17 32 31 31 27 31 31 27 31 30 27 31 25 25 25 22
13 28 26 27 31 31 28 27 18 27 25 24 24 22 13 31 27 32 31 30 22 22 21 16 7 30 27 27 31
27 31 23 32 31 32 28 27 15 31 30 27 32 31 26 32 31 32 29 31 31 31 29 32 31 31 32 32 32
32 31 27 32 29 24 31 25 24 18 30 25 25 20 7 32 30 25

>DOJHLOP02HH1SG length=109 xy=2958_3582 region=2 run=R_2005_09_08_15_35_38_
32 32 32 30 26 28 31 27 30 29 28 17 31 27 24 24 20 6 30 27 20 31 31 32 31 26 31 30 27
21 23 31 27 29 26 32 32 32 24 24 21 13 30 29 28 16 25 32 32 28 30 32 32 32 32 27 32 28
31 32 32 32 27 30 32 31 29 28 25 30 31 26 26 27 26 18 28 31 28 30 32 29 26 21 11 30 27
22 22 21 16 7 26 32 31 10 31 26 23 23 21 13 1 31 27 28 28 16 31

>DOJHLOP01DBEEU length=106 xy=1242_3620 region=1 run=R_2005_09_08_15_35_38_
29 32 26 31 28 32 24 24 19 4 32 30 31 32 30 27 31 31 28 32 25 24 9 30 25 24 24 19 5 31
26 20 26 29 32 30 26 31 31 28 28 29 31 27 29 24 25 31 31 23 23 20 13 1 24 28 28 16 28
31 24 32 32 31 29 29 27 32 31 31 32 31 29 31 25 29 31 25 31 31 28 28 23 15 23 29 28 18
22 30 16 32 27 23 22 3 22 13 31 27 22 22 20 14 3 25

>DOJHLOP02GU0TZ length=98 xy=2696_2405 region=2 run=R_2005_09_08_15_35_38_
31 27 31 29 27 32 32 32 32 32 31 31 32 31 31 30 32 32 32 31 32 27 31 32 29 32 26 26 18
32 26 25 22 11 31 31 29 30 27 31 26 26 30 31 27 31 31 27 32 32 30 29 26 29 25 25 22 10
31 27 25 25 18 32 30 27 32 31 27 31 32 20 30 25 24 22 13 31 29 23 32 29 32 32 30 30 27
30 31 30 32 30 25 32 32 26 28 25

>DOJHLOP02I4IZJ length=98 xy=3624_3981 region=2 run=R_2005_09_08_15_35_38_
32 26 31 29 28 31 30 32 30 26 30 31 29 32 32 31 18 28 29 28 18 27 24 24 18 27 30 28 31
26 25 18 29 30 23 15 25 27 21 26 28 31 30 24 13 30 25 25 18 28 21 28 28 16 32 29 26 31
29 23 31 24 20 25 32 24 24 18 30 21 12 30 31 23 31 29 28 22 28 22 30 28 30 20 24 18 31
22 30 26 23 22 3 28 27 27 18 31
```

test 1:

próg: 15

długość podciągu: 5

motyw: GACGC

sekwencja: 1, nr pozycji: 39

sekwencja: 2, nr pozycji: 19

sekwencja: 3, nr pozycji: 41

sekwencja: 4, nr pozycji: 68

sekwencja: 5, nr pozycji: 19

test 2:

próg: 20

długość podciągu: 6

nie znaleziono motywu

test 3:

próg: 10

długość podciągu: 6

motyw: GACGCT

sekwencja: 1, nr pozycji: 39

sekwencja: 2, nr pozycji: 19

sekwencja: 3, nr pozycji: 41

sekwencja: 4, nr pozycji: 68

sekwencja: 5, nr pozycji: 19

test 4:

próg: 24

długość podciągu: 4

motyw: TGAT

sekwencja: 1, nr pozycji: 24

sekwencja: 2, nr pozycji: 24

sekwencja: 3, nr pozycji: 27

sekwencja: 4, nr pozycji: 14

sekwencja: 5, nr pozycji: 6

3 INSTANCJA

.fasta:

```
>DOJHLOP02JGLS0 length=105 xy=3762_2258 region=2 run=R_2005_09_08_15_35_38_
GATGGCATCATCCATATCGGCATATTTTCTGGCAACCTATTAAGAGTCTTGATTAAAATATTTTCAGCAAAAAAAGTATCAGTAATC
AAAGGTAAATGTGGTCCA

>DOJHLOP01ATT3O length=103 xy=0222_2834 region=1 run=R_2005_09_08_15_35_38_
CATGGACCACATTTACCTTTGATTACTGAGTATCAGTATTGCTGAAAATATTTTAATCAAGACTCTTAATAGGTTGCCAGAAAATAT
GCCGATATGGATGATG

>DOJHLOP01E4H5T length=107 xy=1984_1887 region=1 run=R_2005_09_08_15_35_38_
CATGGACCACATTTACCTTTGATTACTGATTACTTTGTATCAGTAGAAAATATTTTAATCAAGACTCTTAATAGGTTGCCAAGAAAA
ATATGCCGATATGGATGATG

>DOJHLOP01E3GGD length=94 xy=1972_2171 region=1 run=R_2005_09_08_15_35_38_
CATCCATATCGGTATCAGTATCTGGCAACCTATTAAGAGTCTTGAATTAAAATTATTTTCAGAACAAAAAAAGTATCAAGTAATACA
AAGGTAA

>DOJHLOP01CY96H length=106 xy=1104_3511 region=1 run=R_2005_09_08_15_35_38_
ATCGGCATATTTTCTGGCAACCTATTAAGAGTCTTGATTAAAATATTTTCAAGCAAAAAAAGTATCAGTAACGGGAAGTAAATGTGG
TCCATGTTGTTCACATGCA
```

.qual:

```
>DOJHLOP02JGLS0 length=105 xy=3762_2258 region=2 run=R_2005_09_08_15_35_38_
27 30 29 31 27 31 30 32 31 31 28 31 27 32 30 32 28 30 31 26 30 30 31 32 25 25 20 7 27
31 26 19 30 27 20 31 26 31 30 30 24 31 26 31 25 28 32 29 31 27 28 32 31 26 25 25 22 10
28 32 25 25 21 9 28 28 30 27 15 15 15 14 12 8 3 31 31 32 28 31 31 28 32 30 25 26 29 29
28 18 31 26 31 24 23 7 32 30 32 23 15 15 22 13 22
>DOJHLOP01ATT3O length=103 xy=0222_2834 region=1 run=R_2005_09_08_15_35_38_
32 32 32 31 28 28 31 27 29 31 31 29 28 18 29 30 26 28 28 18 32 32 31 28 26 31 25 31 29
22 30 32 18 17 16 10 19 18 5 2 32 31 30 32 24 24 19 4 27 19 25 25 22 12 31 28 32 31 31
28 32 30 32 26 32 31 28 30 27 31 31 31 28 30 25 32 30 25 20 32 25 25 21 9 31 28 26 30
29 24 30 26 30 30 30 30 27 27 31 28 30 31 30
>DOJHLOP01E4H5T length=107 xy=1984_1887 region=1 run=R_2005_09_08_15_35_38_
32 30 31 31 27 29 31 27 30 31 32 28 27 19 29 29 23 29 28 18 31 32 31 28 29 31 31 32 29
25 18 31 32 13 13 13 12 10 8 15 18 32 31 32 10 31 25 24 20 5 17 20 25 25 22 12 31 28 31
32 29 26 32 28 26 16 31 26 24 23 22 19 30 26 24 31 28 31 29 26 31 27 30 22 22 21 15 5
21 32 25 24 26 19 28 28 30 31 31 31 27 25 32 29 30 28 31
>DOJHLOP01E3GGD length=94 xy=1972_2171 region=1 run=R_2005_09_08_15_35_38_
32 26 32 29 23 22 27 28 21 32 29 24 28 26 19 19 32 24 24 19 5 32 32 28 22 18 30 26 28
22 21 17 24 17 31 27 31 29 32 16 29 30 25 31 24 17 30 27 25 25 22 11 25 18 32 24 24 18
3 30 20 30 31 28 31 17 17 16 15 12 6 1 31 31 13 30 28 18 8 31 25 26 20 31 32 30 27 26
18 22 14 18 29 26
>DOJHLOP01CY96H length=106 xy=1104_3511 region=1 run=R_2005_09_08_15_35_38_
29 27 29 29 24 28 31 18 29 25 25 22 11 29 28 23 15 31 24 17 31 28 26 30 25 18 31 28 31
29 29 29 29 26 24 29 28 31 27 24 24 22 12 27 29 25 25 21 9 32 17 6 31 32 14 14 14 13 11
8 4 24 29 32 32 30 24 31 31 27 20 26 29 29 28 17 31 28 29 26 25 11 18 23 32 18 8 22 22
13 31 29 29 30 26 28 29 24 31 28 17 21 29 30 31 24
```

test 1:

próg: 25

długość podciągu: 4

motyw: TAAT

sekwencja: 1, nr pozycji: 54

sekwencja: 2, nr pozycji: 52

sekwencja: 3, nr pozycji: 54

sekwencja: 4, nr pozycji: 48

sekwencja: 5, nr pozycji: 79

test 2:

próg: 6

długość podciągu: 8

motyw: GTATCAGT

sekwencja: 1, nr pozycji: 76

sekwencja: 2, nr pozycji: 30

sekwencja: 3, nr pozycji: 37

sekwencja: 4, nr pozycji: 12

sekwencja: 5, nr pozycji: 62

test 3:

próg: 18

długość podciągu: 5

motyw: TTGAT

sekwencja: 1, nr pozycji: 49

sekwencja: 2, nr pozycji: 19

sekwencja: 3, nr pozycji: 19

sekwencja: 4, nr pozycji: 42

sekwencja: 5, nr pozycji: 34

Wnioski:

Im dłuższy podciąg tym znalezienie struktury gwiazdy jest znacznie trudniejsze.

Zmiany progu wiarygodności często uniemożliwiają przypisanie nukleotydów do podciągu.

Zmniejszenie wymogów ułatwia znalezienie struktury.