Algorytmy Kombinatoryczne w Bioinformatyce – Projekt III

Michał Konon  
Nr Albumu: 152499  
Semestr III, Stopień I  
Bioinformatyka  
Wydział Informatyki i Telekomunikacji

# Opis Algorytmu

Algorytm odczytuje dane z pliku w formacie:

>ID  
<SEKWENCJA JAKOŚCI ODDZIELONE SPACJĄ>  
<SEKWENCJA DNA>

Następnie tworzy nową sekwencję z usuniętymi nukleotydami, których jakość była poniżej wyznaczonego poziomu. Dla każdej sekwencji tworzony jest graf. Proces przełożenia sekwencji na graf, rozpoczyna się podzielenia nici na oligonukleotydy o długości wcześniej podanej przez użytkownika i wstawienia do struktury jego sekwencji oraz pozycji. Następnie dla każdego wierzchołka utworzona zostaje lista następników, poprzez porównanie ze sobą każdej pary sekwencji. Następnik to para zawierająca informacje o indeksie sekwencji oraz pozycji w zmodyfikowanej nici DNA. Następnie odnajdujemy najkrótszą sekwencje i sortujemy wierzchołki w grafie od największej ilości następników do najmniejszej. Następnie program zaczyna przechodzić po kolejnych wierzchołkach i sprawdza czy motyw został znaleziony w każdej sekwencji, jeżeli tak to wyświetla jego pozycje (zaczynając od 1).

# Oszacowanie Złożoności algorytmu

W najgorszym wypadku algorytm ma złożoność O(n4). Ze względu na wielokrotną pętle podczas tworzenia listy następników.

# Instancje Testowe oraz Przeprowadzone Testy

Program został przetestowany na 5 instancjach z różnymi parametrami. Oto 3 najciekawsze instancje.

## Instancja 1

Stały niski poziom akceptowalnej jakości z zmieniającą się długością oligonukleotydu.

Poziom jakości = 20  
Długość oligonukleotydu = 6-9

### Wyniki

> .\main.exe 20 5 instance\_1.txt

CAGTT

sekwencja nr 1 pozycja 38

sekwencja nr 2 pozycja 30

sekwencja nr 3 pozycja 36

sekwencja nr 4 pozycja 17

sekwencja nr 5 pozycja 12

> .\main.exe 20 6 instance\_1.txt

CAGTTG

sekwencja nr 1 pozycja 38

sekwencja nr 2 pozycja 30

sekwencja nr 3 pozycja 36

sekwencja nr 4 pozycja 17

sekwencja nr 5 pozycja 12

> .\main.exe 20 7 instance\_1.txt

CAGTTGC

sekwencja nr 1 pozycja 38

sekwencja nr 2 pozycja 30

sekwencja nr 3 pozycja 36

sekwencja nr 4 pozycja 17

sekwencja nr 5 pozycja 12

> .\main.exe 20 8 instance\_1.txt

BRAK WYNIKU

### Dane

>DOJHLOP02JAKIE

27 32 24 24 19 04 26 20 32 32 32 32 31 26 30 31 27 29 29 27 32 31 27 31 28 25 30 27 25 25 21 09 30 30 31 27 31 32 23 27 31 27 31 32 31 27 27 24 31 30 25 31 27 28 22 32 29 28 18 30 27 32 24 24 19 04 26 20 32 32 32 32 31 26 30 31 27 29 29 27 32 31 27 31 28 25 30 27 25 25 21 09 30 30 31 27 31 32 23 27 31 27 31 32 31 27 27 24 31 30 25 31 27 28 22 32 29 28 18 30

ATCCCCAATGCGCCACCAGAGCCACCTTCCCCTATTACAGTTGCAATTATTGGAACTTTAATCCCCAATGCGCCACCAGAGCCACCTTCCCCTATTACAGTTGCAATTATTGGAACTTTA

>DOJHLOP02FU6TN

32 32 31 27 32 31 27 32 30 27 31 27 32 31 29 26 30 27 32 24 31 31 27 31 31 27 31 30 31 32 31 31 31 32 27 21 32 31 32 29 32 20 18 25 23 22 03 18 07 18 29 28 18 29 31 26 32 24 22 28 32 32 31 27 32 31 27 32 30 27 31 27 32 31 29 26 30 27 32 24 31 31 27 31 31 27 31 30 31 32 31 31 31 32 27 21 32 31 32 29 32 20 18 25 23 22 03 18 07 18 29 28 18 29 31 26 32 24 22 28

AGAAGAACAAGGACAAGGAGAGGCAATCGCAGTTGCTCTACGAGAAATTGTTTAGGCTTAAGAAGAACAAGGACAAGGAGAGGCAATCGCAGTGAATCTACGAGAAATTGTTTAGGCTTA

>DOJHLOP02GKA8Z

26 32 31 28 31 31 32 31 31 27 32 30 27 29 31 26 27 21 30 28 30 27 30 24 32 29 28 27 20 29 26 20 32 28 31 25 32 31 31 22 29 24 29 28 31 17 25 28 17 25 26 25 10 19 24 24 23 06 30 21 26 32 31 28 31 31 32 31 31 27 32 30 27 29 31 26 27 21 30 28 30 27 30 24 32 29 28 27 20 29 26 20 32 28 31 25 32 31 31 22 29 24 29 28 31 17 25 28 17 25 26 25 10 19 24 24 23 06 30 21

CATAGCAGAAGAACAAGGACAAGGAGAGGCAATCGCAGTTGCTCTACGAGAAATGTTTAGCATAGCAGAAGAACAAGGACAAGGAGAGGCAATCGCAGTGAATCTACGAGAAATGTTTAG

>DOJHLOP01DBA3Z

31 27 21 30 27 25 23 25 25 22 12 27 32 30 25 29 32 21 28 31 27 31 27 23 16 26 24 32 31 27 30 25 30 25 31 29 28 18 31 28 26 31 28 28 28 27 19 31 25 29 28 18 28 29 31 32 32 31 32 25 31 27 21 30 27 25 23 25 25 22 12 27 32 30 25 29 32 21 28 31 27 31 27 23 16 26 24 32 31 27 30 25 30 25 31 29 28 18 31 28 26 31 28 28 28 27 19 31 25 29 28 18 28 29 31 32 32 31 32 25

CCACCTTCCCCTATTACAGTTGCAATTATTGGAACTTTAAGCCTAAACATTTCTCGTAGACCACCTTCCCCTATTACAGTTGCAATTATTGGAACTTTAAGCCTAAACATTTCTCGTAGA

>DOJHLOP01BGY5C

29 24 24 24 18 02 30 31 29 24 23 31 24 29 29 24 31 30 30 25 31 27 27 31 26 29 24 25 18 31 27 26 13 31 27 30 29 24 32 29 28 17 32 31 27 26 13 32 29 30 31 28 30 28 31 29 24 32 31 31 29 24 24 24 18 02 30 31 29 24 23 31 24 29 29 24 31 30 30 25 31 27 27 31 26 29 24 25 18 31 27 26 13 31 27 30 29 24 32 29 28 17 32 31 27 26 13 32 29 30 31 28 30 28 31 29 24 32 31 31

TTCCCCTATTACAGTTGCAATTATTGGAACTTTAAGCCTAAACATTTCTCGTAGATTCACTTCCCCTATTACAGTTGCAATTATTGGAACTTTAAGCCTAAACATTTCTCGTAGATTCAC

## Instancja 2

Poszukiwanie motywu o stałej długości ze zmienną jakością nukleotydów.

Poziom jakości = 20-24  
Długość oligonukleotydu = 4

### Wyniki

> .\main.exe 21 4 instance\_2.txt

CGAT

sekwencja nr 1 pozycja 88

sekwencja nr 2 pozycja 40

sekwencja nr 3 pozycja 45

sekwencja nr 4 pozycja 44

sekwencja nr 5 pozycja 99

>.\main.exe 22 4 instance\_2.txt

GATC

sekwencja nr 1 pozycja 79

sekwencja nr 2 pozycja 41

sekwencja nr 3 pozycja 46

sekwencja nr 4 pozycja 45

sekwencja nr 5 pozycja 101

>.\main.exe 23 4 instance\_2.txt

GATC

sekwencja nr 1 pozycja 79

sekwencja nr 2 pozycja 41

sekwencja nr 3 pozycja 46

sekwencja nr 4 pozycja 45

sekwencja nr 5 pozycja 101

> .\main.exe 24 4 instance\_2.txt  
BRAK MOTYWU

### Dane

>DOJHLOP02IK9M4

18 26 19 24 24 23 06 31 32 29 28 17 31 25 32 25 30 32 31 23 27 21 31 31 28 32 28 31 27 30 24 16 28 27 21 30 32 29 25 17 26 19 28 31 26 20 31 27 31 30 24 31 30 24 28 30 28 32 31 26 18 26 19 24 24 23 06 31 32 29 28 17 31 25 32 25 30 32 31 23 27 21 31 31 28 32 28 35 37 38 34 30 24 27 21 30 32 29 25 17 26 19 28 31 26 20 31 27 31 30 24 31 30 24 28 30 28 32 31 26

GCCTAAACATTTCTCGTAGATTCACTGCGATTGCCTCTCCTTGTCCTTGTTCTTCTGCTAGCCTAAACATTTCTCGTAGATTCACTGCGATTGCCTCTCCTTGTCCTTGTTCTTCTGCTA

>DOJHLOP02F21K2

30 24 22 24 25 23 29 32 21 30 29 29 28 27 18 27 25 30 29 28 16 26 12 31 27 31 27 28 23 21 28 25 23 29 26 28 32 24 31 27 30 32 31 31 30 20 28 28 18 28 21 27 21 26 24 31 27 24 23 06 30 24 22 24 25 23 29 32 21 30 29 29 28 27 18 27 25 30 29 28 16 26 12 31 27 31 27 38 33 31 38 20 24 29 26 28 32 24 31 27 30 32 31 21 30 20 28 28 18 28 21 27 21 26 24 31 27 24 23 06

TGGAGGCTACAGAAAGGCTTTGAGGCTGATGGACCATGCCGATCGCTTTAATCTTCCAAATGGAGGCTACAGAAAGGCTTTGAGGCTGATGGACCATGCCGATCGCTTTAATCTTCCAAA

>DOJHLOP01C6EES

27 21 31 31 31 29 23 28 29 24 26 30 32 32 31 31 31 31 29 32 23 22 05 30 23 22 02 30 26 30 24 17 31 26 21 13 31 20 30 25 29 28 16 32 39 31 34 31 18 31 26 30 29 31 28 22 31 29 23 31 27 21 31 31 31 29 23 28 29 24 26 30 32 32 31 31 31 31 29 32 23 22 05 30 23 22 02 30 26 30 24 17 31 26 21 13 31 20 30 25 29 28 16 32 29 31 24 31 28 31 26 30 29 31 28 22 31 29 23 31

AAGCTCCAGGAGTATCTATGAAAGAAATTATTGGAAGATTAAAGCGATCGGCATGGTCCAAAGCTCCAGGAGTATCTATGAAAGAAATTATTGGAAGATTAAAGCGATCGGCATGGTCCA

>DOJHLOP01C8Z6P

28 23 30 31 31 29 24 30 25 19 32 32 22 31 28 31 32 28 31 31 24 22 30 23 22 03 28 23 30 27 21 25 18 23 15 31 31 21 11 25 24 09 32 34 32 31 28 18 30 25 26 29 29 25 18 31 27 20 32 29 28 23 30 31 31 29 24 30 25 19 32 32 22 31 28 31 32 28 31 31 24 22 30 23 22 03 28 23 30 27 21 25 18 23 15 31 31 21 11 25 24 09 32 24 32 31 28 28 30 25 26 29 29 25 18 31 27 20 32 29

AAGCTCCAGGAGTATCTATGAAGAAATTATTGGAAGATTAAAGCGATCGGCATGGTCCATAAGCTCCAGGAGTATCTATGAAGAAATTATTGGAAGATTAAAGCGATCGGCATGGTCCAT

>DOJHLOP02FH72B

31 28 28 22 28 20 10 31 30 31 24 20 32 24 23 05 26 19 28 24 22 27 19 23 15 29 31 32 32 18 24 17 27 20 10 28 32 30 23 15 27 30 27 29 31 15 32 27 27 13 31 26 31 24 23 14 17 29 23 17 31 28 28 22 28 20 10 31 30 31 24 20 32 24 23 05 26 19 28 24 22 27 19 23 15 29 31 32 32 18 24 17 27 20 10 28 32 30 33 35 37 30 20 39 31 15 32 27 27 13 31 26 31 24 23 14 17 29 23 17

CTGGAGGCTACAGAAAGGCTTGAGGCTGATGGACCATGCCGATCGTCTTTAATCTTCAATCTGGAGGCTACAGAAAGGCTTGAGGCTGATGGACCATGCCGATCGTCTTTAATCTTCAAT

## Instancja 3

### Wyniki

PS C:\Users\Admin\Desktop\akwb\AKwB\_Project3\results> .\main.exe 20 6 instance\_3.txt

CCGTAC

sekwencja nr 1 pozycja 35

sekwencja nr 2 pozycja 28

sekwencja nr 3 pozycja 4

sekwencja nr 4 pozycja 48

sekwencja nr 5 pozycja 26

> .\main.exe 20 7 instance\_3.txt  
BRAK WYNIKU

> .\main.exe 25 6 instance\_3.txt

GTACTG

sekwencja nr 1 pozycja 37

sekwencja nr 2 pozycja 30

sekwencja nr 3 pozycja 6

sekwencja nr 4 pozycja 50

sekwencja nr 5 pozycja 28

> .\main.exe 25 7 instance\_3.txt

CGTACTG

sekwencja nr 1 pozycja 36

sekwencja nr 2 pozycja 29

sekwencja nr 3 pozycja 5

sekwencja nr 4 pozycja 49

sekwencja nr 5 pozycja 27  
Dane

>DOJHLOP01A0CMG

24 27 17 17 31 24 24 19 04 18 08 18 07 31 30 22 13 22 22 20 13 02 31 28 31 28 27 20 20 28 31 27 30 22 31 32 34 35 38 34 23 26 31 22 21 15 05 27 21 28 18 29 23 28 32 31 17 28 26 25 22 11 32 31 31 24 17 30 32 24 27 17 17 31 24 24 19 04 18 08 18 07 31 30 22 13 22 22 20 13 02 31 28 31 28 27 20 20 28 31 27 30 22 22 21 15 05 27 21 28 18 29 23 28 32 31 17 28 26 25 22 11 32 31 31 24 17 30 32

TAGTGAAAATTCCGTCCTTTTTAATTGGACAACACCGTACTTGAAAAGGTAGAGATACAAAAGAGAATGTAGTGAAAATTCCGTCCTTTTTAATTGGACAACAAAAAGGTAGAGATACAAAAGAGAATG

>DOJHLOP01ANIOR

24 09 15 24 23 06 21 12 31 31 29 28 28 31 27 21 31 24 31 28 32 31 22 22 20 13 02 31 32 34 35 38 34 23 26 31 31 31 28 31 21 29 23 26 25 22 10 27 21 31 27 27 21 31 32 25 23 30 22 22 21 16 07 29 24 29 31 31 27 24 09 15 24 23 06 21 12 31 31 29 28 28 31 27 21 31 24 31 28 32 31 22 22 20 13 02 31 31 28 31 21 29 23 26 25 22 10 27 21 31 27 27 21 31 32 25 23 30 22 22 21 16 07 29 24 29 31 31 27

GGAGGGTTAGCTCGAATAGGTGAAAAACCGTACTTGTCCGTCCTTTTAATTGGACAACAAAAAGGTAGAGGAGGGTTAGCTCGAATAGGTGAAAAATCCGTCCTTTTAATTGGACAACAAAAAGGTAGA

>DOJHLOP02FEHEE

31 32 32 31 32 34 35 38 34 23 26 31 32 31 29 29 31 26 32 31 27 29 31 27 25 25 21 09 32 30 26 31 29 31 30 26 27 27 18 31 28 27 15 32 31 26 30 31 30 26 28 30 31 26 31 30 29 25 25 22 10 30 30 31 31 32 31 31 22 31 32 32 32 31 29 29 31 26 32 31 27 29 31 27 25 25 21 09 32 30 26 31 29 31 30 26 27 27 18 31 28 27 15 32 31 26 30 31 30 26 28 30 31 26 31 30 29 25 25 22 10 30 30 31 31 32 31 31 22

TCTCCGTACTTGGTAGCCTCCAGGTTTTGCCATGCCAAAGTTTCTTGCAACATTCTCTTTTGTATCTCTTCTGTAGCCTCCAGGTTTTGCCATGCCAAAGTTTCTTGCAACATTCTCTTTTGTATCTCT

>DOJHLOP02GV797

25 24 20 07 31 30 24 29 32 31 28 22 27 27 13 32 29 28 17 32 31 27 32 32 31 27 28 31 31 27 32 24 31 26 25 22 11 27 32 27 27 31 31 29 17 32 29 31 32 34 35 38 34 23 26 31 24 22 22 21 16 07 31 31 27 31 32 31 27 25 24 20 07 31 30 24 29 32 31 28 22 27 27 13 32 29 28 17 32 31 27 32 32 31 27 28 31 31 27 32 24 31 26 25 22 11 27 32 27 27 31 31 29 17 32 29 24 22 22 21 16 07 31 31 27 31 32 31 27

TTTTGCCATGCCAAAGTTTCTTGCAACATTCTCTTTTGTATCTCTACCCGTACTTGCTTTTTGTTGTCCTTTTGCCATGCCAAAGTTTCTTGCAACATTCTCTTTTGTATCTCTACCTTTTTGTTGTCC

>DOJHLOP02JGPQ7

25 25 22 10 30 31 26 31 31 32 31 26 28 28 16 31 28 27 18 32 31 27 32 32 29 31 32 34 35 38 34 23 26 31 23 31 31 31 27 32 21 30 26 25 22 10 27 28 30 32 32 32 30 19 27 31 26 21 21 20 16 08 31 30 27 31 31 31 26 25 25 22 10 30 31 26 31 31 32 31 26 28 28 16 31 28 27 18 32 31 27 32 32 29 23 31 31 31 27 32 21 30 26 25 22 10 27 28 30 32 32 32 30 19 27 31 26 21 21 20 16 08 31 30 27 31 31 31 26

TTTTGCCATGCCAAAGTTTCTTGCACCGTACATGACATTCTCTTTTGTATCTCTACCTTTTTGTTGTCCTTTTGCCATGCCAAAGTTTCTTGCAACATTCTCTTTTGTATCTCTACCTTTTTGTTGTCC

# Wnioski

Algorytm znajduje motyw w zależności od podanych parametrów. Dla krótkich oligonukleotydów jest większa szansa że podany algorytm znajdzie sekwencję ze względu na większą ilość wyszukiwanych rozwiązań (Instacja 1). Zmiana jakości również wpływa na wynik, ze względu na usunięte nukleotydy (Instacja 2). Zmiana jakości może umożliwić znalezienie dłuższego motywu (Instancja 3).