AKWB 3

Co i jak?

Rozdział 1: Wczytywanie instancji:

W skrócie: Mamy 2 pliki, jeden to sekwencje (fasta), drugi to oceny jakości dla każdego nukleotydu (qual). Musimy je wczytać do programu. O modyfikowaniu tych plików przed wczytaniem mówimy w punkcie: Epilog

Rozdział 2: Trimming (Przycinanie sekwencji):

Tak właściwie po chuj nam ta jakość? Ano właśnie dlatego, że musimy zaprogramować możliwość pozbycia się nukleotydów, które jakościowo nie odpowiadają userowi. Po wczytaniu plików **MUSI** być możliwość podania progu (dolnej granicy jakości). Wszystkie nukleotydy poniżej tej jakości **mają zostać wycięte z wszystkich sekwencji**. Dodatkowo jest haczyk! Program ma zapamiętać pozycje wszystkich nukleotydów, które pozostały, w nawiązaniu do sekwencji oryginalnej, licząc od 1. Co to oznacza? Przykładowo:

Sekwencja	A	C	G	T	A	A	C	T	A
Jakość	31	24	35	30	26	29	31	32	21
„Pozycja”	0	1	2	3	4	5	6	7	8

Użytkownik podał dolny limit: 30

Sekwencja	A	C	G	T	A	A	C	T	A
Jakość	31	24	35	30	26	29	31	32	21
„Pozycja”	0	1	2	3	4	5	6	7	8

Pozbywamy się z sekwencji i jakości wszystko co jest poniżej progu

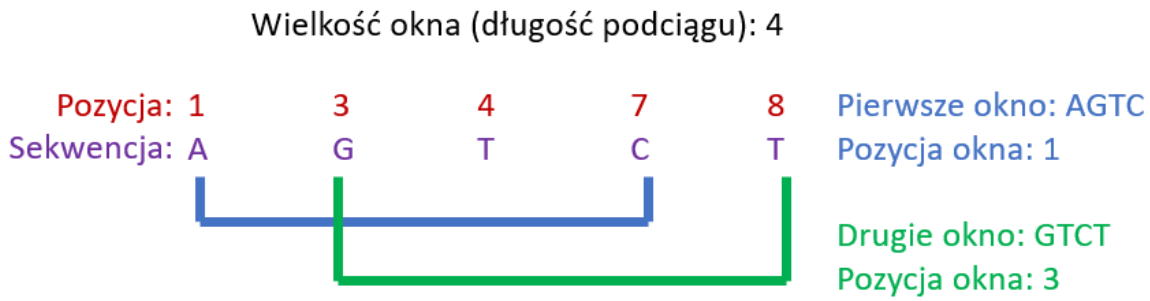
Seq	A	G	T	C	T
Qual	31	35	30	31	32
Pos	1	3	4	7	8

Numerujemy od 1 więc: 0+1, 2+1, 3+1 etc

Te pozycje możemy zrobić na różne sposoby: można stworzyć jakiś kontener przed wycięciem, po wycięciu, po wczytaniu od razu zrobić wszystkie pozycje i potem usuwać. Możliwości jest wiele. **„W nawiązaniu do sekwencji oryginalnej” oznacza, gdzie te nukleotydy znajdowały się PRZED wycięciem.** No i ofc, numerowanie od 1 bo tak każą.

Rozdział 3: Wierzchołki:

Zaczynamy jazdę. Jako że będziemy tworzyć graf, to musimy stworzyć listę wierzchołków. Czym jest wierzchołek w naszym programie? Otóż jest to kilkuliterowy podciąg (od 4 do 9 znaków), gdzie musimy przechować informacje o numerze sekwencji i pozycji, która rozpoczyna podciąg. Co to oznacza? Przykład:

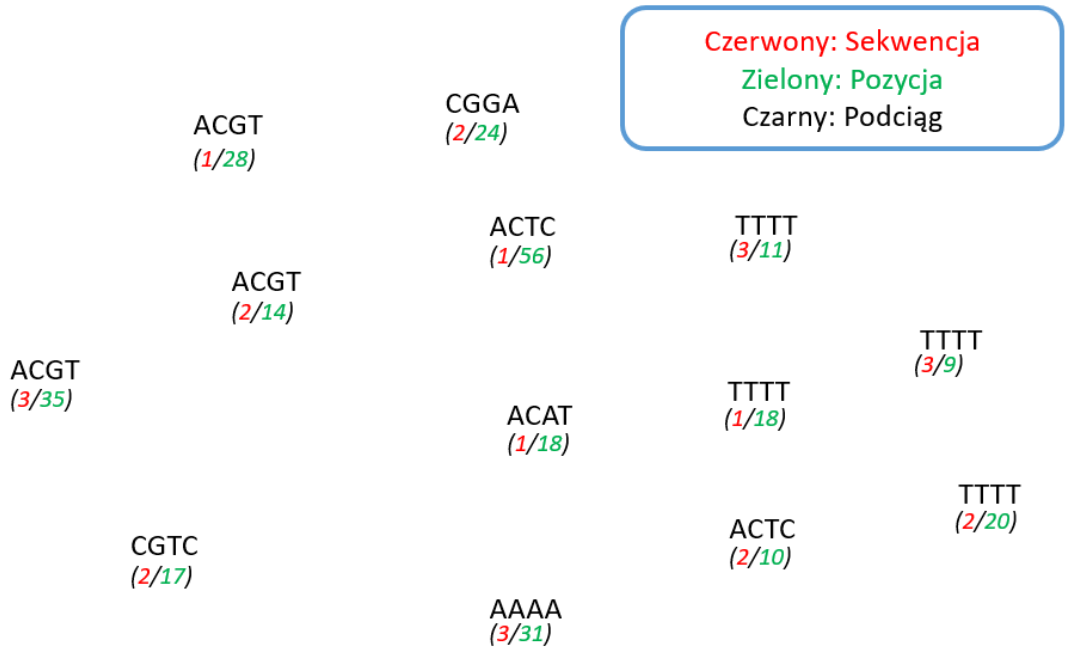


Pojedyncze okna tworzymy poprzez wzięcie ok 3 literek od znaku, na jakim jesteśmy obecnie (ofc + znak, na którym jesteśmy). Następnie przechodzimy do kolejnego znaku i znowu dokładamy 3 literki. **PRZESUWAMY SIĘ JEDEN PO JEDNYM**. Pozycja, od której rozpoczyna się podciąg to kolejne cyfry w naszym wektorze zachowanych pozycji. **Dodatkowo, musimy pamiętać o zapamiętaniu numeru sekwencji dla danego okna!.** Okna mogą się powtarzać, ponieważ ich pozycje będą różne.

Także podsumowując wierzchołek musi przechowywać 3 informacje:

- Podciąg
- Numer sekwencji
- Pozycja rozpoczynająca

Tutaj wyobrażenie sobie tego rysunkowo, jakbyśmy mieli 3 sekwencje:

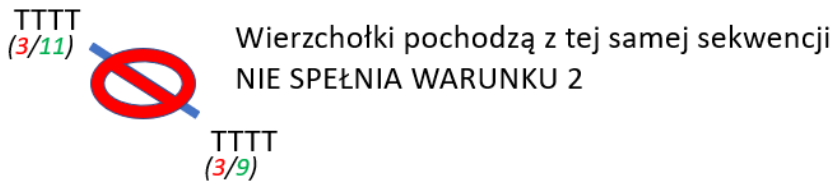


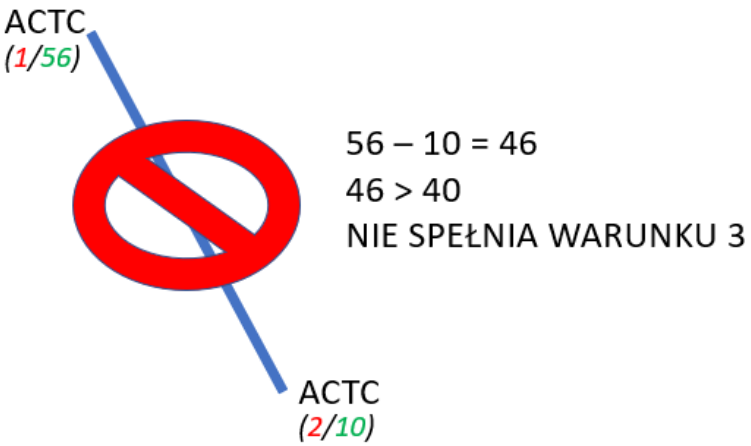
Rozdział 4: Graf

Mając graf, musimy teraz wykombinować, jak połączyć takowe wierzchołki ze sobą. W ramach projektu mamy w tej kwestii parę zasad:

1. Podciągi muszą być takie same
2. Podciągi muszą być w różnych sekwencjach (nie ma wypadku gdzie dwa podciągi z sekwencji 1 połączą się ze sobą)
3. Różnica w pozycjach podciągów nie jest większa niż dziesięciokrotność długości podciągu

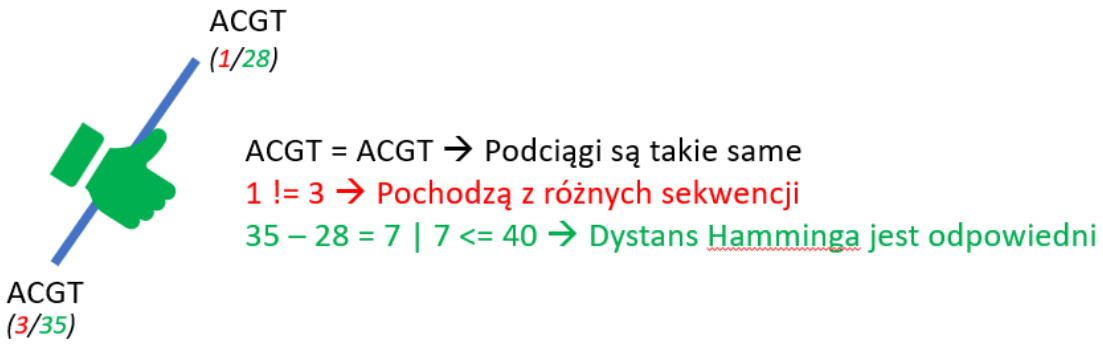
Spójrzmy na parę testowych przypadków:



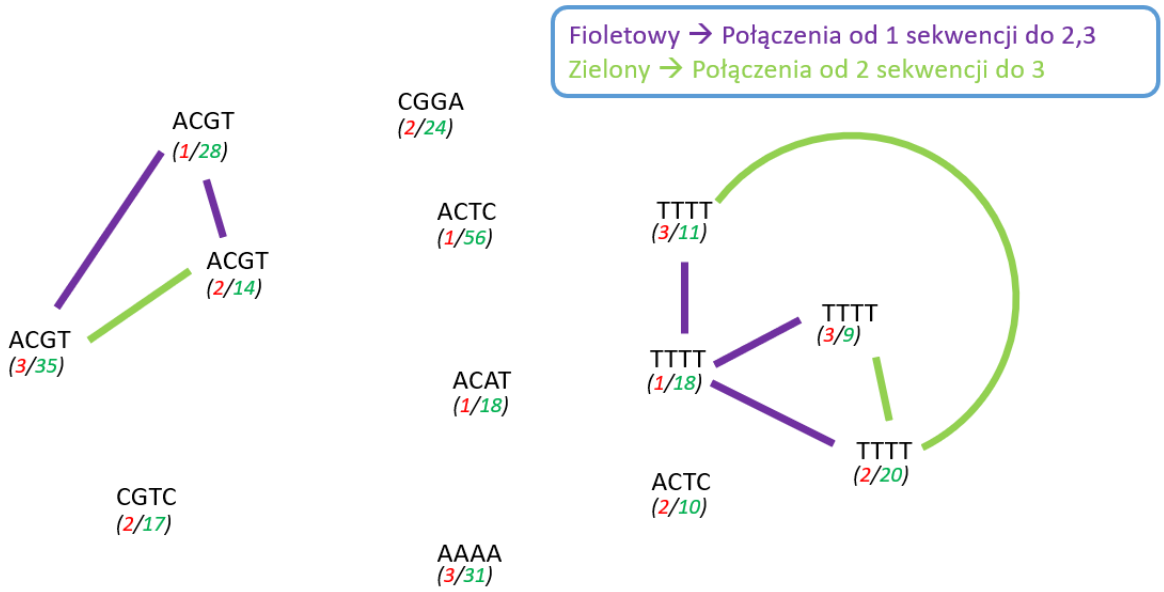


Wielkim ułatwieniem dla tego zadania jest to, że **krawędzie są nieskierowane**, więc możemy stopniowo ograniczać nasze pole przeszukiwania np.: wierzchołki z sekwencji 0 szukają połączeń z 1 2 3 4, ale już wierzchołki z sekwencji 1 szukają połączeń tylko z 2 3 4 (ponieważ już potencjalne połączenie z zerem znaleźliśmy iteracje wcześniej).

Dodatkowo należy wspomnieć, że jak badamy różnice między pozycjami, to zawsze odejmujemy większą pozycje od mniejszej.



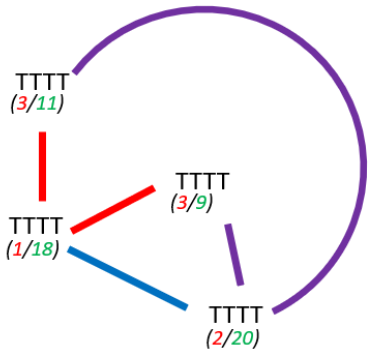
Cały nasz graf wygląda tak:



Rozdział 5: Klika/Gwiazda (właściwe wyszukiwanie motywu)

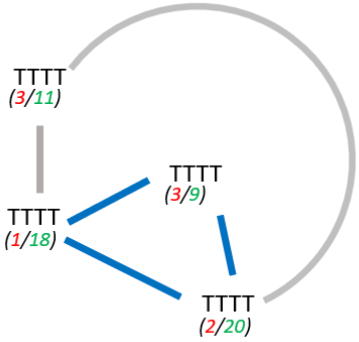
Wyszukiwanie motywu odbywa się poprzez wyszukanie kliki w naszym grafie. Czym jest klika? → Ano jest to **CZĘŚĆ** grafu, w której każdy wierzchołek jest połączony z każdym, **RÓŻNYM** od siebie. Co oznacza ta „różność" w naszym wypadku? To, że każda sekwencja jest reprezentowana **TYLKO** jednym wierzchołkiem. Dodatkowo, w naszym programie po wyszukaniu pierwszej kliki, możemy już przerwać działanie, ponieważ chcemy wyszukać tylko jeden dowolny motyw, a nie wszystkie.

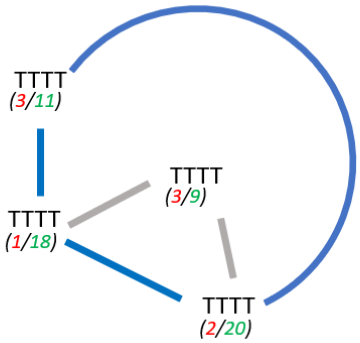
Od razu warto zaznaczyć, że szukanie struktur podobnych do kliki jest *KUREWSKO* męczące. Dlatego my skupimy się na takiej strukturze jak „gwiazda". Gwiazda jest to taka „zamknięta klika", gdzie nie ma żadnych odstępstw, innych krawędzi niż te tworzące klikę. Więcej sensu w przykładach, które zresztą są poniżej.



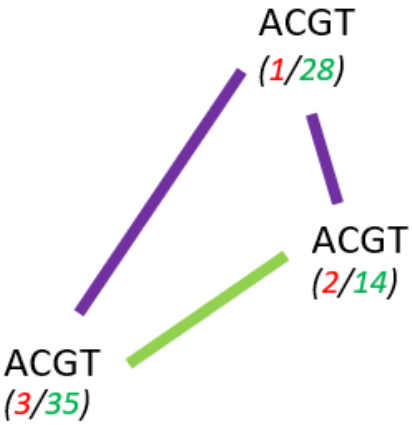
Powyższa struktura NIE JEST gwiazdą ani KLIKĄ, ponieważ, od jednej sekwencji wychodzą dwie krawędzie do tej innej, tej samej sekwencji (od 1 dwa razy do 3 ORAZ od 2 dwa razy do 3)

Więc sekwencje nie są reprezentowane pojedynczym wierzchołkiem

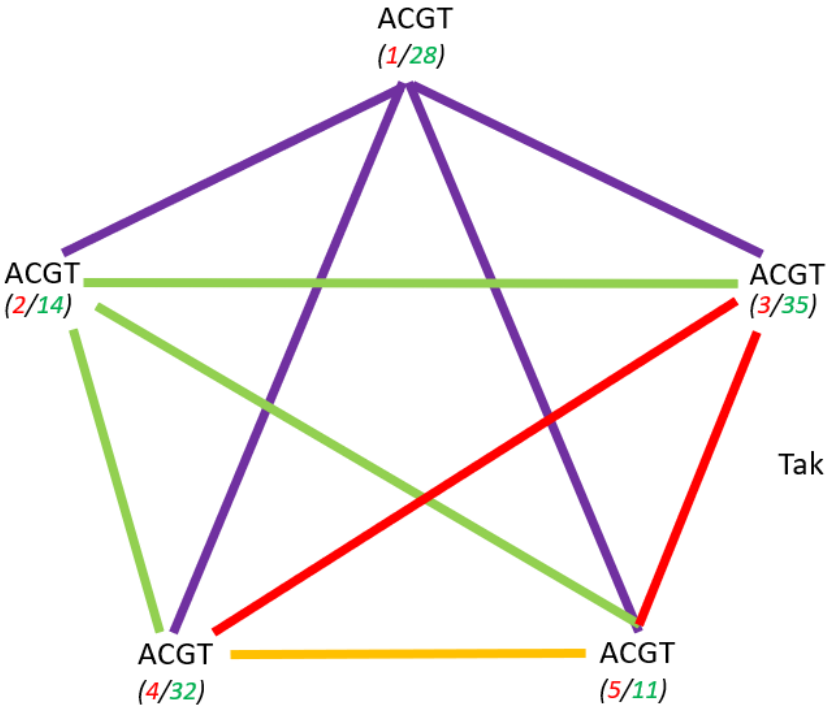




Te struktury z kolei SĄ KLIKĄ ale NIE SĄ GWIAZDĄ. Klika jest tu po prostu CZĘŚCIĄ struktury a nie CAŁĄ STRUKTURĄ



Ta struktura jest z kolei gwiazdą. CAŁA STRUKTURA jest kliką i nie ma żadnych pobocznych krawędzi



Tak by to wyglądało dla 5 sekwencji ☺

Część 6: Wynik końcowy

Gratulacje! Właśnie znalazłeś motyw. Przypominam → Wystarczy nam pierwsza znaleziona taka struktura. Nie musimy szukać wszystkich. A co do wyświetlania wyniku, projekt wymaga od nas tego byśmy uwzględnili:

- Podciąg będący motywem
- Nr sekwencji
- Pozycję

Przykładowo wyglądać to może tak:

MOTYW: ACGT
W sekwencji: 1 znaleziony na pozycji: 28
W sekwencji: 2 znaleziony na pozycji: 14
W sekwencji: 3 znaleziony na pozycji: 35
W sekwencji: 4 znaleziony na pozycji: 32
W sekwencji: 5 znaleziony na pozycji: 11

Epilog: Modyfikowanie pliku

Prawda jest taka, że znajdowanie motywów na czuja, to mozolna i *zjebana* robota. Szczególnie jeśli chcemy znajdować gwiazdy. Dlatego też pliki przed wczytaniem i wgl poza program należy sobie zmodyfikować. Generalnie chodzi o to, by samemu wybrać motyw, włożyć go do instancji (czyt do każdego pliku) (ofc za tym idzie dodanie sobie jakości). Motyw nie powinien być na początku, by nie

było za łatwo (takie zalecenie z góry) + nie chcemy, by sekwencje miały długie, powtarzające się fragmenty.