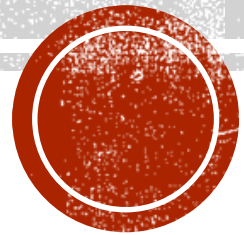
KONWERTER NOTACJI STRUKTURY DRUGORZĘDOWEJ RNA

Bartosz Bukowiec



PODTYTUŁ

- Dla zadanego na wejściu pliku ze strukturą drugorzędową zapisaną w jednym z akceptowanych formatów wygeneruj zapisy struktur we wszystkich pozostałych formatach.
- Program powinien automatycznie wykrywać typ formatu wejściowego oraz ew. błędne dane. Domyślnie generowane są wszystkie pozostałe formaty, możliwość wybrania konkretnego formatu docelowego.

OBSŁUGIWANE FORMATY

- Dot-bracket

- Jest to łańcuch składający się z kropek oraz nawiasów, dzięki którym oznacza się niesparowane nukleotydy "." oraz sparowaną podstawową parę nukleotydów "()". Za pomocą nawiasów kwadratowych "[]", ostrych "<>", wąsiastych "{}" przedstawiane są psudowęzły.

- Zapis może się różnić pomiędzy programami

```
CAGCACGACACUAGCAGUCAGUGUCAGACUGCAIACAGCACGACACUAGCAGUCAGUGUCAGACUGCAIACAGCACGACACUAGCAGUCAGUGUCAGACUGCAIA
..(((((...(((((...(((((...(((((.....)))))...)))))).....(((((...(((((.....)))))...)))))).....)))))...)))))..
```

OBSŁUGIWANE FORMATY

```
1 C 0
2 A 0
3 G 103
4 C 102
5 A 101
6 C 100
7 G 99
8 A 0
9 C 0
10 A 0
11 C 95
12 U 94
13 A 93
14 G 92
15 C 91
   .
   .
   .
```

- BPSEQ (Basepair and sequence information)

- Plik ten zawiera informacje na temat par zasad, zapisanych w trzech kolumnach:
  - 1 kolumna zawiera pozycję sekwencji (licząc od jednego)
  - 2 kolumna zawiera pojedynczą zasadę zapisaną jedną literą
  - 3 kolumna zawiera pozycję sparowanego nukleotydu lub zero gdy nukleotyd jest niesparowany

OBSŁUGIWANE FORMATY

```
1  C  0  2  0  1
2  A  1  3  0  2
3  G  2  4 103  3
4  C  3  5 102  4
5  A  4  6 101  5
6  C  5  7 100  6
7  G  6  8  99  7
8  A  7  9  0  8
9  C  8 10  0  9
10 A  9 11  0 10
11 C 10 12 95 11
12 U 11 13 94 12
13 A 12 14 93 13
14 G 13 15 92 14
15 C 14 16 91 15
       .
       .
       .
```

- CT (Connect table)

- Plik ten zawiera informacje na temat par zasad, zapisanych w sześciu kolumnach:
  - 1 kolumna zawiera pozycję sekwencji (licząc od jednego)
  - 2 kolumna zawiera pojedynczą zasadę zapisaną jedną literą
  - 3, 4, 6 kolumna zawiera pozycję sekwencji odpowiednio poprzednią następną i aktualną
  - 5 kolumna zawiera pozycję sparowanego nukleotydu lub zero gdy nukleotyd jest niesparowany

PLAN

1. Wczytanie pliku
   - Plik zostanie wczytany z wiersza poleceń jako argument
2. Ustalenie notacji
   - Poprawności pliku
   - Komunikaty o ewentualnych błędach zawartości
3. Dobór funkcji konwertujących
4. Zapis plików
   - Wybór pomiędzy zapisem we wszystkich formatach lub jednym wybranym

ZAŁOŻENIA

1. Zawartość:
   - Poprawna zawartość w przypadku **dot-bracket** to trzy linijki, w których odpowiednio znajduje się: podpis sekwencji, sekwencja, łańcuch kropkowo-nawiasowy. Dla **ct** oraz **bpseq** będzie to linijka z podpisem oraz odpowiednia, wymagana liczba kolumn dla danego formatu jak i wierszy świadczących o długości sekwencji. Wymagane odstępy pomiędzy wartościami w wierszu.
2. Nagłówek:
   - Nagłówek po znaku ">" zawiera tytuł oraz po kropce odpowiednie rozszerzenie (.bpseq .ct .dbn)
3. Oznaczenie:
   - "." – oznaczać będzie niesparowane nukleotydy
   - "()" – oznaczać będzie parę nukleotydów
   - "[], {}, <>, Aa, Bb, Cc, Dd, Ee, Ff, Gg" – oznaczać będzie pseudowęzły odpowiednio od stopnia pierwszego do stopnia dziesiątego

MOŻLIWE BŁĘDY

- Głównym problemem będzie poprawność pliku zadanego na wstępie:
  1. Brak poprawnego nagłówka lub niezgodnego z danymi zawartymi poniżej
  2. Zawartość główna
     - Dla **.ct** i **.bpseq** będą to problemy związane z liczbą kolumn lub z połączeniami, które nie mogą występować (powtórzenia)
     - Dla **.dbn** będzie to sytuacja gdzie niezgodna jest długość sekwencji do łańcucha struktury lub występują nieprawidłowe ,,otwarcia, zamknięcia,, nawiasów. Problem także z ustaleniem nawiasowania w przypadku wysokiej rzędowości pseudowęzłów