Laboratorium 2 - AiSD 2

Piotr Duperas

Kwas deoksyrybonukleinowy (DNA) składa się z sekwencji czterech związków chemicznych: adeniny, cytozyny, guaniny, tyminy, które są tradycyjnie oznaczane za pomocą pierwszych liter ich nazw (A, C, G, T). Częstym zagadnieniem w genetyce jest sprawdzanie zgodności dwóch sekwencji DNA, aby dopasować odpowiadające sobie geny w różnych organizmach. Istotnym problemem są jednak zaburzenia w tych sekwencjach, które mogą wynikać z mutacji, bądź uszkodzenia próbek DNA. Dlatego nie jest możliwe bezpośrednie porównywanie sekwencji element po elemencie, lecz próbuje się dopasowywać do siebie poszczególne fragmenty rozciętej sekwencji tak, aby globalne dopasowanie było jak najlepsze.

Dopasowaniem dwóch sekwencji DNA nazywamy 2 ciągi tej samej długości złożone z symboli związków chemicznych: `A`, `C`, `G`, `T` oraz symbolu przerwy: `-`. W ciągach tych kolejność symboli związków chemicznych musi być identyczna jak w oryginalnych sekwencjach, z tym, że w dowolnych miejscach mogą pojawiać się symbole przerwy. Dla przykładu weźmy dwie sekwencje DNA:

```
GATAC
GTCCAG
```

Jednym z możliwych dopasowań może być:

```
GAT--AC
G-TCCAG
```

W dopasowaniu mogą wystąpić trzy różne sytuacje:

- **Poprawne dopasowanie symboli** - występuje gdy w obu ciągach na tych samych miejscach znajduje się ten sam symbol, w powyższym przykładzie poprawnie dopasowane są symbole G na początku ciągów,
- **Niepoprawne dopasowanie symboli** - występuje gdy w obu ciągach na tych samych miejscach znajdują się różne symbole, w przykładzie niepoprawnie dopasowane są symbole C i G na końcu ciągów,
- **Przerwa** - występuje, gdy symbol związku chemicznego z któregoś z ciągów nie ma odpowiadającego mu symbolu związku w tym samym miejscu w drugim ciągu. W przykładzie symbol A z pierwszego ciągu nie ma odpowiadającego mu symbolu w drugim (brak związku oznaczamy symbolem przerwy `-`).

Celem zdania jest implementacja algorytmu znajdowania najlepszego dopasowania dwóch sekwencji DNA w dwóch wariantach z różnymi kryteriami oceny dopasowań.

1 Wariant I

Dla danego dopasowania możemy określić prostą ocenę jego jakości jako sumę punktów:

- +1 za każde poprawne dopasowanie symboli,
- -3 za każde niepoprawne dopasowanie symboli,
- -2 za każdą przerwę.

W powyższym przykładzie dopasowania mamy 3 poprawnie dopasowane symbole, 1 niepoprawnie dopasowany symbol oraz 3 przerwy, co daje ocenę $3 * 1 + 1 * (-3) + 3 * (-2) = -6$. Interesuje nas znalezienie dopasowania o **najwyższej** ocenie.

2 Wariant II

W zastosowaniach algorytmu dopasowywania sekwencji DNA przyjmuje się, że lepsze jest dopasowanie, które ma mało długich przerw, niż takie, które ma wiele przerw krótkich. Dlatego też jeśli mamy 2 dopasowania tych samych sekwencji:

```
CAAAAAATTTTTG
C-AA--AT-T-TG
```

oraz

```
CAAAAAATTTTTG
C---AAA--TTTG,
```

to za lepsze uznaje się to drugie, gdyż ma tylko dwie długie przerwy zamiast wielu krótkich.
W tym wariancie należy zmodyfikować sposób oceny dopasowania, tak aby:

- pierwsza przerwa w serii przerw była oceniana jako -5,
- każda kolejna przerwa w serii przerw była oceniana jako -2.

Dopasowanie:

```
---TTGA
CCCT-AA
```

posiada dwie serie przerw o długościach 3 oraz 1, zatem przerwy oceniane są jako $(-5+2*(-2))+(-5+0*(-2))=-14$, natomiast całe dopasowanie ma ocenę $-14+2+(-3)=-15$.

3 Uwagi

- Oczekiwana złożoność czasowa w obu wariantach to $O(mn)$, gdzie m i n to długości sekwencji wejściowych - skorzystaj z programowania dynamicznego,
- Na wejściu algorytmu zawsze będą niepuste ciągi złożone z wielkich liter A, C, G, T,
- Na wyjściu algorytmu oczekiwane są dwa ciągi znaków równej długości złożone z wielkich liter A, C, G, T oraz znaków - (myślnik),
- Jeśli wiele różnych dopasowań ma najlepszą ocenę, można zwrócić dowolne z nich.

4 Etapy punktacji

- Etap 1 - Wariant I - zwrócenie jedynie oceny najlepszego dopasowania: 1pkt,
- Etap 2 - Wariant I - zwrócenie najlepszego dopasowania wraz z oceną: 0,5pkt,
- Etap 3 - Wariant II - zwrócenie jedynie oceny najlepszego dopasowania: 0,5pkt,
- Etap 4 - Wariant II - zwrócenie najlepszego dopasowania wraz z oceną: 0,5pkt.

5 Wskazówki

- Wariant I: rozwiąż wszystkie podproblemy polegające na dopasowaniu i pierwszych wyrazów pierwszej sekwencji do j pierwszych wyrazów drugiej sekwencji, dla wszystkich par (i, j),
- Wariant II: rozważ dodatkowe podproblemy, w których dopasowania odpowiednich fragmentów kończą się przerwą w pierwszej lub w drugiej sekwencji,
- Pomocne może być zapamiętywanie z jakiego stanu dociera się do danego stanu dopasowania,
- Pomysłem rozwiązania wariantu II może być utworzenie dodatkowych tablic, które będą śledzić stan wstawiania przerw.