Algorytmy biologii molekularnej
Zadania domowe, seria I
Termin oddania: 24.11.99

Uwaga: rozwiązanie każdego zadania proszę napisać na osobnej kartce.

Zadanie 1. Znajdź dwa krótkie słowa S_1, S_2, tej samej długości, takie że optymalne globalne uliniowienie nie jest jednoznaczne. Jako funkcję podobieństwa dla tego zadania przyjmijmy

$$s(x,y) = \begin{cases} 1 & \text{gdy } x = y, \\ -1 & \text{gdy } x \neq y. \end{cases}$$

oraz $s(x,-) = -2$, dla $x, y \in \Sigma$.

Zadanie 2. Znajdź algorytm, który w czasie $O(mn)$ wyznacza liczbę wszystkich optymalnych globalnych uliniowień dla dowolnej pary słów. Algorytm powinien działać dla dowolnej funkcji podobieństwa.

Zadanie 3. W rozwiązaniu tego zadania należy skorzystać z przeglądarki WWW i bazy danych dostępnej pod adresem `http://www.ncbi.nlm.nih.gov/`.

Wyniki w postaci pliku `.tar.gz` lub `.zip` zawierającego wszystkie sekwencje, z których się korzystało oraz opisu wykonywanych czynności proszę przesłać na adres `luke@mimuw.edu.pl`.

(a) Wyszukaj w bazie sekwencję kodującą ludzki hormon wzrostu (GH1).

- Jakich słów i w jakich polach użyłeś aby wyszukać rekordy?
- Ile rekordów zostało znalezionych?
- Jakie sekwencje zostały znalezione oprócz szukanego genu?

(b) Pobierz rekordy dla mRNA, cDNA i DNA.

- Jakie są ich numery? (accession numbers)
- Znajdź w DNA kasetę TATA (TATA-box), który zazwyczaj znajduje się przed genem. Podaj jego pozycje.

(c) Wymyśl sekwencje DNA i z pomocą algorytmu BLAST porównaj ją z bazą danych. Opisz wyniki.

Zadanie 4. W problemie *lokalnego uliniowienia* dwóch różnych sekwencji A i B należy znaleźć parę podsłów, jedno zawarte w A a drugie w B, będących do siebie maksymalnie podobne.

Rozważmy problem znajdowania podobnych podsłów w tej samej sekwencji A. Nazwiemy ten problem *optymalnym niedokładnym powtórzeniem*. Skonstruuj algorytmy rozwiązujące problem optymalnego niedokładnego powtórzenia w następujących szczególnych przypadkach, różniących się umiejscowieniem powtarzających się podsłów w słowie A.

(a) Interesują nas tylko przypadki, w których słowa się *nie* nakładają.

(b) Dwa podsłowa muszą być sąsiednie w sekwencji A, to znaczy $A = uw_lw_rv$ oraz w_l i w_r są do siebie najbardziej podobne spośród wszystkich sąsiadujących ze sobą podsłów sekwencji A. Tego typu zjawisko (nazywane ang. *tandem repeat*) spotyka się bardzo często w sekwencjach biologicznych.

(c) Podobnie jak w punkcie (b) z dodatkowym wymogiem, że poszukujemy największego podobieństwa pomiędzy słowem w_l a lustrzanym odbiciem słowa komplementarnego do w_l, tzn. $w_r = ((w_r)^c)^R$, gdzie dla dowolnego słowa w, w^R jest lustrzanym odbiciem w, a w^c oznacza sekwencję DNA komplementarną do w.

Dla każdego algorytmu zanalizuj jego złożoność czasową i pamięciową (wyżej punktowane będą rozwiązania o mniejszej złożoności).