

Wiązka zadań Geny

Informację genetyczną (genotyp) każdego osobnika z galaktyki Madgen opisuje słowo (napis), w którym występują litery ze zbioru {A, B, C, D, E}. Obowiązują przy tym następujące zasady:

1. Organizmy żyjące na Madgen tworzą gatunki g_1, g_2, g_3, \dots , gdzie g_i to zbiór osobników o długości genotypu równej i .
2. W skład genotypu mogą wchodzić geny. Pierwszy gen rozpoczyna się pierwszą występującą w genotypie sekwencją AA, a kończy się najbliższą napotkaną po niej sekwencją BB. Każdy kolejny gen rozpoczyna się pierwszą sekwencją AA, występującą za końcem poprzedniego genu, i analogicznie kończy się najbliższą napotkaną sekwencją BB.
3. Geny nazywamy częścią kodującą genotypu, pozostałe fragmenty tworzą część niekodującą.

Przykład 1.

Genotyp AACDBABBBCDAAABCBBAAE

zawiera geny AACDBABB oraz AABCBB. Zwróćmy uwagę, że:

- ciąg AA występujący za genem AABCBB nie jest początkiem genu, ponieważ nie występuje za nim ciąg BB kończący gen;
- część kodująca genotypu AACDBABBBCDAAABCBBAAE jest równa AACDBABBAABCBB.

Przykład 2.

Genotyp AADBAADDDDEEEEBBEE

zawiera gen AADBAADDDDEEEEBB. Zwróćmy uwagę, że:

- pierwsze pojawienie się ciągu AA determinuje początek genu, dlatego w powyższym genotypie występuje gen AADBAADDDDEEEEBB, a nie gen AADDDDEEEEBB.

Plik dane_gen.txt zawiera genotypy 1000 osobników z galaktyki Madgen. Każdy wiersz pliku zawiera genotyp jednego osobnika o długości nie większej niż 500 znaków.

Przykład

```
ABAEACBAADAACAABBABCD  
ABAEACBADEADACACABBABCD
```

Napisz program(-y), który poda odpowiedzi na pytania postawione w poniższych zadaniach. Odpowiedzi zapisz w pliku wyniki_gen.txt. Odpowiedź do każdego zadania rozpocznij w nowym wierszu, poprzedzając ją numerem zadania.

69.1.

Podaj liczbę wszystkich gatunków, których genotypy zapisane są w pliku dane_gen.txt. Podaj największą liczbę osobników reprezentujących ten sam gatunek.

69.2.

Występowanie w jakimkolwiek **genie** ciągu BCDDC oznacza mutację powodującą małą odporność osobnika na zmęczenie. Podaj, ile osobników spośród tych, których genotypy zapisane są w pliku, ma tę mutację.

Przykład

Osobnik o genie AACBCDDCBBACDE ma mutację BCDDC (ciąg BCDDC występuje w obrębie podkreślonego genu), natomiast osobnik o genie **CB**DDCBBAACDEBB nie ma tej mutacji, gdyż występujący ciąg BCDDC nie jest ulokowany w żadnym genie.

69.3.

Wyznacz i podaj największą liczbę genów występujących u jednego osobnika. Podaj też największą długość genu zapisanego w całym pliku.

Przykład

Rozważmy plik składający się z genotypów:

```
EAABCDBBDCBAAE
EAABCDBBDCBAAEBCDBBEE
EAABCDBBECAAB
```

Pierwszy osobnik ma jeden gen (AABCDBB), drugi ma dwa geny (AABCDBB i AAEBDCBB), a trzeci osobnik ma jeden gen (AABCDBB). Zatem największa liczba genów u jednego osobnika wynosi 2, a największa długość genu to 8 (gen AAEBDCBB ma tę długość).

69.4.

Genotyp odczytywany z materiału biologicznego może być odkodowany w kierunku od strony lewej do prawej lub odwrotnie: od strony prawej do lewej. Genotyp nazywać będziemy *odpornym*, jeśli czytany od strony lewej do prawej oraz od strony prawej do lewej ma dokładnie taką samą część kodującą. Natomiast genotyp *silnie odporny* to taki, który czytany od strony lewej do prawej oraz od strony prawej do lewej daje dokładnie ten sam napis. (Inaczej mówiąc, genotyp jest silnie odporny, gdy jest palindromem).

Przykład

Rozważmy genotypy:

```
EAABCDBBDCBAAE
EAABCDBBDCBAAEBCDEE
EAABCDBBECAAB
```

Genotyp EAABCDBBDCBAAE jest silnie odporny (jest palindromem). Genotyp EAABCDBBDCBAAEBCDEE nie jest silnie odporny (nie jest palindromem), ale jest odporny, gdyż czytany od strony lewej do prawej, jak i od strony prawej do lewej ma taką samą część kodującą: AABCDBB. Natomiast genotyp EAABCDBBECAAB nie jest silnie odporny (nie jest palindromem), nie jest też odporny, gdyż czytany od strony lewej do prawej daje część kodującą AABCDBB, a czytany od strony prawej do lewej ma część kodującą równą ACEBB.

Wyznacz liczbę genotypów odpornych oraz liczbę genotypów silnie odpornych.