

## **Zadanie 7.18**

### **Translator kodów nukleotydowych na aminokwasowe**

#### **Spis treści:**

Rozdział I. Sformułowanie treści zadania.	1
Rozdział II. Dopuszczalne zestawy danych.	2
Rozdział III. Algorytm działania programu.	2

#### **Rozdział I. Sformułowanie treści zadania.**

Kwasu nukleinowe, to związki przechowujące informację genetyczną w organizmach żywych, m.in DNA czy RNA, ich podstawową jednostką budulcową są nukleotydy, których trójki tworzą kodony. Dla przykładu trójka nukleotydów o oznaczeniu AUG koduje aminokwas zwany metioniną. Format FASTA służy natomiast do zapisu sekwencji nukleotydów lub aminokwasów.

Celem stworzonego programu jest pobranie od użytkownika danych w postaci zapisu ciągu liter, takich jak: AAGGGAAAAAATTTT, czyli naszych nukleotydów. Ciąg ten zostaje zapisany w formacie FASTA, następnie program dla podanej sekwencji nukleotydowej generuje odpowiadający tej sekwencji ciąg aminokwasów (dokonuje transkrypcji trójek nukleotydowych na aminokwas). Transkrypcja rozpoczyna się od sekwencji START (AUG), a kończy się na sekwencji STOP (UGA, UAA, UAG).

Aby nasz program działał poprawnie, konieczny jest format pliku z konkretnymi danymi. W naszych organizmach najpierw zachodzi translacja, czyli przepisanie informacji genetycznej z nici DNA na nić mRNA. Dopiero na podstawie sekwencji nukleotydów zawartych na pojedynczej nici matrycowego RNA odczytywana jest sekwencja aminokwasów. My uprościliśmy trochę ten mechanizm i przedstawiliśmy go w dalszych rozdziałach.

## Rozdział II. Dopuszczalne zestawy danych.

Nasz program obsługuje dane zawarte w pliku w formacie FASTA. Dane są w postaci sekwencji nukleotydowej nici matrycowej DNA, tzn. nici niekodującej - komplementarnej do nici mRNA. Bardzo ważnym jest, aby plik zawierał prawidłowe dane, czyli występujące w nici matrycowej DNA, abyśmy otrzymali satysfakcjonujący nas wynik. W nici DNA nie występuje U (uracyl) tylko T (tymina), a kod genetyczny w programie definiowany jest na podstawie danych nici DNA. Nukleotydy z nici DNA dzielone są na trójki aminokwasowe, które transkrybujemy na nić mRNA i na podstawie tych kodonów odczytywane są aminokwasy.

Plik rozpoczyna się znakiem „>”, po którym następuje nagłówek z najważniejszymi informacjami. Od następnej linii zaczyna się sekwencja nukleotydów kodonem START - ‚TAC‘, a kończy się jednym z kodonów STOP. W dalszej części pliku nie występują już dane, które były w pliku pierwotnym - nukleotydy zostały zamienione na nazwy aminokwasów.

Wynikiem, który generuje nasz program, jest nowy plik tekstowy w formacie FASTA. Występują w nim dane w postaci przetłumaczonych kodonów - trzyliterowych skrótów nazw aminokwasów, kodowanych przez trójki aminokwasowe (kodony). W nagłówkach występuje informacja, czy sekwencja uległa transkrypcji, czy nie. Jeśli sekwencja nie rozpocznie się kodonem START - informacja zostanie przekazana w nagłówku.

## Rozdział III. Algorytm działania programu.

Na podstawie danych zawartych w pliku FASTA (szeregu nukleotydów w nici matrycowej DNA) nasz program generuje nowy plik FASTA zawierający szereg aminokwasów. Na podstawie tabeli kodu genetycznego, w której trójki aminokwasowe tłumaczone są na podstawie nici mRNA, zdefiniowaliśmy naszą tabelę kodu genetycznego (kodonów), która na podstawie sekwencji nukleotydowej w nici DNA odczytuje aminokwas kodowany przez dany kodon.

Zdefiniowaliśmy funkcję, która otwiera i odczytuje dane z pliku wejściowego i zapisuje je w pliku wyjściowym w zmienionym formacie. Program sprawdza, czy sekwencja nukleotydowa rozpoczyna się od kodonu START. Jeśli tak to przystępuje do następnego kroku, czyli tłumaczenia trójek nukleotydów na aminokwasy. Jeśli kodon znajduje się w tabeli to zostaje mu przypisana dana nazwa aminokwasu.

W programie, aby odczytać następny kodon, przesuwamy się o 3 nukleotydy. Jeśli podczas odczytywania trójek napotkamy kodon STOP (którym powinna być zakończona dana sekwencja) przerywamy nasz proces. Sprawdzamy całą sekwencję nukleotydów zawartą w pliku wejściowym, do momentu napotkania kodonu STOP.

Aby użytkownik mógł wybrać z jakiego pliku chce, aby aminokwasy zostały odczytane, prosimy go o podanie ścieżki do pliku oraz ścieżki do pliku wyjściowego, w którym ma zostać przekazany wynik pracy programu.

Po zakończeniu pracy programu, dostajemy odpowiedź o zakończonej transkrypcji. W naszym nowym pliku, rozpoczynającym się w ten sam sposób, co plik pierwotny, w

miejscu sekwencji nukleotydów pojawia się sekwencja aminokwasów. Jeśli nasz plik wejściowy nie spełnia wymagań, najprawdopodobniej sekwencja nie będzie wcale odczytana (brak kodonu start), zostanie przerwana w trakcie lub otrzymamy wynik w postaci samych znaków „?”. Ten symbol pojawia się wtedy, kiedy nie ma zdefiniowanego kodonu odczytanego przez program.