Advanced Programming for Biology

2021/22

T08 - (Mini) programming project*

* Based on an exercise by Lourens-Jan Ugen and Bioinformatics Algorithms by M. Rocha and P. Ferreira

Ribosome

- Ribosomes are macromolecular machines, found within all living cells, that perform biological protein synthesis (mRNA translation). Ribosomes link amino acids together in the order specified by the codons of messenger RNA (mRNA) molecules to form polypeptide chains. Ribosomes consist of two major components: the small and large ribosomal subunits. Each subunit consists of one or more ribosomal RNA (rRNA) molecules and many ribosomal proteins (RPs or r-proteins). The ribosomes and associated molecules are also known as the translational apparatus.

Goal of the project

- The program that accepts a DNA sequence and returns the amino-acid sequence of the proteins that it encodes.

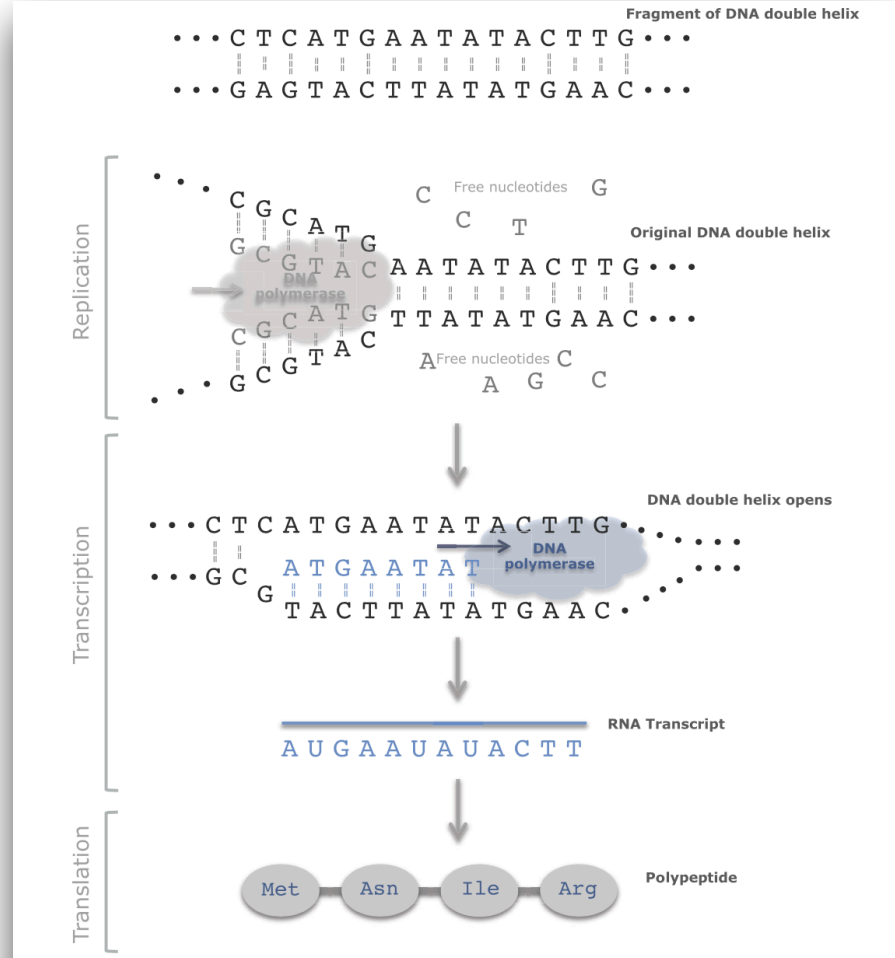

Figure 3.1: Genetic information flow. Example of a fragment of a DNA double helix is shown at the top. Replication of DNA forms two identical DNA double helices. Transcription allows the synthesis of an mRNA transcript. Translation synthetizes a polypeptide from the mRNA transcript.

Functions

1. Transcribing DNA

   Write a function that can read DNA and return a sequence of RNA. Both DNA and RNA strands are a sequence of nucleotides. The four nucleotides found in DNA are adenine (A), cytosine (C), guanine (G) and thymine (T). The four nucleotides found in RNA are adenine (A), cytosine (C), guanine (G) and uracil (U). Given a DNA strand, its transcribed RNA strand is formed by replacing each nucleotide with its complement:

   G -> C

   C -> G

   T -> A

   A -> U

   Make sure you can store the RNA made by the function so you can use it later.

2. Reading encodings

   Write a function that can read all the encodings (e.g. mapping between codons and amino acids) from the file `genetic_code.txt`. This file contains the information needed to insert the right amino-acids and stop at the right moment. Pick an adequate data structure for storing the encodings.

Functions

3. Translating RNA

   Write a function that will read a sequence of RNA and stores relevant parts in a string. Considerer what would happen if the ribosome starts reading one or two base–pairs after the sequence has started?

4. Creating a protein sequence

   Write a function that takes as input the encodings read previously and a string containing RNA. It should return a string containing the amino-acid codes for a protein.

5. Create all protein sequences

   Write a function that reads all the possible sequences from the RNA sequence. Note that our python ribosome cannot distinguish between reading the triplets of bases and switching one or two bases over and start there. This means that the amount of possible proteins is larger than in a natural setting.

Program

A. For the first version of the program use the small subsequence of the genome of E-coli given in file `e_coli_subseq.txt`.

B. For the second version your program should be able to analyse the complete genome of E-coli given in file `e_coli.txt` (in FASTA format).

C. For the extended version do the following tasks:

   i) Adapt your function 3. to consider the three possible offsets (called reading frames). Each offset will generate a different sequence of proteins.

Table 3.2: An example of an mRNA sequence and the three reading frames.

AAUGCUCGUAAUUUAG
AAU-GCU-CGU-AAU-UUA → Asn-Ala-Arg-Asn-Leu
AUG-CUC-GUA-AUU-UAG → Met-Leu-Val-Ile-Stop
UGC-UCG-UAA-UUU → Cys-Ser-Stop-Stop

Program

C. For the extended version do the following tasks:

   ii) Do some basic statistics analysis for the sequence of proteins generated for the different offsets, including number of proteins; average size of a protein sequence; frequency of each protein.