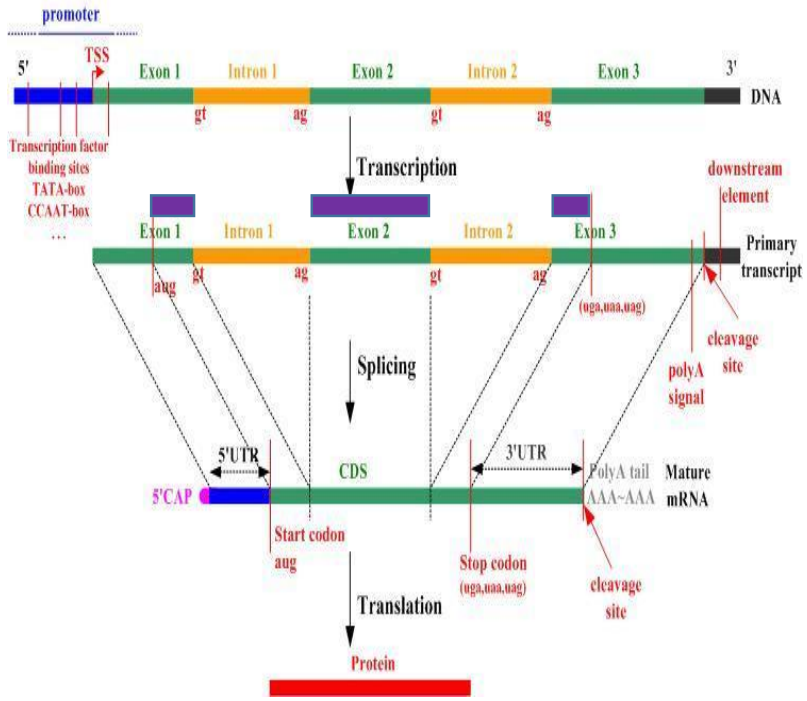
Lab Exercise 2

1)

The file **orf_exons_chr17.txt** contains a list of genes on chromosome 17 and their ORF exon sequences. ORF exons are the parts of the exons of a gene that contribute to its ORF/CDS (i.e. without the UTRs). It means that exons that are contained entirely in the UTRs (and thus have no contribution to the ORF at all) are not included in the list.

For illustration, see the purple regions in the following figure:



A)

Parse the file into the following data structure:

{

'gene_symbol1': ['orf_exon_seq1', 'orf_exon_seq2', 'orf_exon_seq3', …],

'gene_symbol2': ['orf_exon_seq1', 'orf_exon_seq2', 'orf_exon_seq3', …],

'gene_symbol3': ['orf_exon_seq1', 'orf_exon_seq2', 'orf_exon_seq3', …],

…

}

B)

How many genes are there in this data? How many ORF-contributing exons in total? What is the average number of ORF-contributing exons per gene?

C)

Count the number of genes per number of ORF-contributing exons (i.e. genes that have that number of ORF-contributing exons).

D)

Find the five genes with the highest number of ORF-contributing exons.

E)

Write the names of all the genes with only a single ORF-contributing exon into a file.

F)

What is the average length of the ORF-contributing parts of exons?

G)

Find the five shortest ORF-contributing parts of exons.