

Introduction to Bioinformatics

Course NR. 22111

Virtual Ribosome & The protein database



Stefan Olevinskiy
s246026

Polina Krasikova
s245850

Virtual Ribosome

Question 1:

1. How is a STOP codon displayed?

Stop codons are displayed as ***

2. How is a START codon displayed?

On the PLUS strand:

- >>> : strict start codon (usually ATG)
-))) : alternative start codon (like TTG or GTG, if used at the start)

On the MINUS strand:

- <<< : strict start codon
- (((: alternative start codon

3. Does a start-codon always code for Methionine (M)?

Yes — when a codon is used as the start codon, it is translated as Methionine (M).

In our example (>Yeast_ACT1), the start codon is ATG, which codes for Methionine.

Other codons like TTG or GTG could also serve as start codons and then would code for Methionine, but in this sequence TTG appears only inside the gene and is translated as Leucine.

4. What is the difference between the two types of start codons?

Strict start codon (ATG): always codes for Methionine when used as the first codon.

Alternative start codons (TTG, GTG): normally code for other amino acids inside the gene, but if used as the first codon, they are translated as Methionine.

Question 2:

1. Did the translation succeed (i.e. did it yield a long amino acid sequence unbroken by stop codons)?

For translation table: Standard SGC0, translation failed because many premature stop codons appear, so the protein is not a long continuous sequence.

- 2. Nothing is wrong with the DNA sequence. Can you come up with some good reasons for the result?**

The translation fails under the standard code because COX1 is a mitochondrial gene (biological context: one of three mitochondrially-encoded subunits, from: [COX1 | SGD](#)). In mitochondria, some codons that are stops in the standard code (e.g., TGA) actually code for amino acids. Therefore, we need to use the Yeast Mitochondrial code (SGC2) to obtain a correct, long amino acid sequence.

Question 3:

- 1. What is the difference in the use of STOP codons?**

Some codons that are stops in the standard code (like TGA) encode amino acids (e.g., Trp) in yeast mitochondria.

- 2. What is the difference in the use of START codons?**

Alternative start codons (like ATA) can be used at the beginning and encode Methionine.

- 3. Are codons coding for completely different amino acids?**

Only a few codons change meaning; most codons code for the same amino acids as in the standard code.

Question 4:

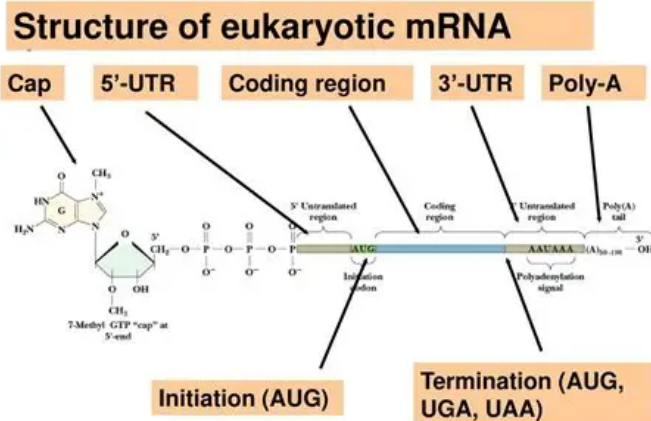
- 1. Yeast has introns in some genes, could this be a major problem in this case?**

No, it is not a major problem because the sequence is an mRNA, which has already had introns spliced out.

- 2. Can an mRNA molecule contain more sequence than the gene in question? (Can it be longer than the CDS coding for the protein).**

Yes, an mRNA molecule can be longer than the coding sequence (CDS) because it contains additional regions and modifications that are not translated into protein. These include the 5' untranslated region (5' UTR), the 5' cap, the 3' untranslated region (3' UTR), and the 3' poly-A tail.

All of these elements increase the total length of the mRNA beyond the coding sequence.



Question 5:

Why is this?

Only the positive reading frames are needed because mRNA is read 5' → 3'.

Question 6:

What reading frame is most likely the right one?

The most likely reading frame is frame 1 on the positive strand, as it produces a long continuous protein sequence with proper start and stop codons, whereas frames 2 and 3 contain many stop codons and interruptions.

Question 7:

1. How does the DNA sequence in the output look (is it identical to the one you input)?

The DNA sequence looks identical but is displayed in the 3'→5' direction for the negative strand.

2. In what direction shall it be read (left-to-right or right-to-left)?

For negative chain DNA should be read right-to-left (3'→5').

3. In what direction shall the protein-sequence be read (left-to-right or right-to-left)?

Left-to-right.

Question 8:

1. How many DNA strings are displayed?

Six reading frames are displayed: three on the positive strand and three on the negative strand.

2. Why is this?

This is because each DNA strand can be read in three different ways, depending on where it starts. Consequently, a single DNA sequence can potentially encode multiple protein sequences, depending on the reading frame used.

Question 9:

1. Does the result fit to what you found earlier?

Yes, frame 1 on the positive strand produces the longest continuous protein, as we found manually.

2. Would it make any difference to the result if we had only a partial sequence where the last part of the sequence with the STOP codon is missing?

The open reading frame would be shorter and may not be recognized as complete.

3. What would happen if the first 50 nucleotides (with the START codon) were missing?

The open reading frame would not start with ATG, so the correct protein would not be detected.

The protein database

Question 1.1:

We got 9722 hits, out of which 1764 are from Swiss-Prot. The Accession code for human insulin is P01308 and the entry name is INS_HUMAN

Question 1.2:

1915 hits left after filtering for human only, out of which 1181 are from Swiss-Prot.

Question 1.3:

After using advanced search to filter for human insulin using the organism and protein fields, we got 202 hits, of which 60 were from Swiss-Prot.

Question 1.4:

After filtering out “insulin-like” protein names, the results did not change.

Question 1.5:

After filtering out insulin receptors by using the protein name field, we got 151 hits, out of which 50 were from Swiss-Prot.

Question 2.1:

When looking at the publications for the human insulin entry, there are only 31 Swiss-Prot references. Insulin is a highly investigated protein though, obviously, as it underlies diabetes, hyperproinsulinemia and basically all of human biology and the biology of other organisms, as energy regulation is the foundation of biological function.

Question 2.2:

Insulin is secreted outside of the cell, as it regulates energy utilization of the entire organism, not of an individual cell.

Question 2.3:

Signal peptide is 24 AA long, and the propeptide is 30 AA long.

Question 2.4:

Well, beta strands are formed from positions 48-50, 56-58, 74-76, and 98-101.

Question 3.1:

Now, filtering only by having any signal peptide (ft_signal:*), we get 18656301 hits, out of which 45100 are Swiss-Prot.

Question 3.2:

Filtering by the experimental evidence (ft_signal_exp:*) got us 3891 results, all of which are Swiss-Prot and 734 are human (q3.3).

Question 3.4:

In *Bacillus subtilis*, there are 18297 hits, of which only 63 are Swiss-Prot. Search string is (organism_id:1423).

Question 3.5:

Expanding the search to include lower taxonomic ranks, we now get 35762 hits, out of which are 4280 are from Swiss-Prot.

Question 3.6 and 3.7:

Using the search string (length:[1 TO 10]) AND (existence:1) to filter for very search protein that have evidence of existence at protein level, we now get 1350 hits, out of which 1202 are Swiss-Prot. Without the existence:1 filter, we get 46068 results, of which 1225 are from Swiss-Prot. This is because many of these are actually mistakes directly from nucleotide sequences that have no evidence of them being protein coding.

Question 3.8 and 3.9:

Further narrowing the search by excluding fragments using the search string (length:[1 TO 10]) AND (fragment:false) AND (existence:1), we get 905 hits, out of which 841 are Swiss-Prot, out of which only 6 are human – Erythrocyte membrane glycopeptide, Gastric juice peptide 1,

Phagocytosis-stimulating peptide, Pneumadin, T cell receptor delta diversity 1, and Urine glycopeptide:

```
>sp|P0DPR3|TRDD1_HUMAN T cell receptor delta diversity 1 OS=Homo sapiens
OX=9606 GN=TRDD1 PE=1 SV=1
EI
>sp|P01858|TUFT_HUMAN Phagocytosis-stimulating peptide OS=Homo sapiens OX=9606
PE=1 SV=1
TKPR
>sp|P02729|GLUR_HUMAN Urine glycopeptide OS=Homo sapiens OX=9606 PE=1 SV=1
CEHSHDGA
>sp|P01358|GAJU_HUMAN Gastric juice peptide 1 OS=Homo sapiens OX=9606 PE=1 SV=1
LAAGKVEDSD
>sp|P02728|GLEM_HUMAN Erythrocyte membrane glycopeptide OS=Homo sapiens OX=9606
PE=1 SV=1
CEGHSHDHGA
>sp|P22103|PNEU_HUMAN Pneumadin OS=Homo sapiens OX=9606 PE=1 SV=1
AGEPKLDAGV
```

Question 4:

Find out how many proteins from Escherichia coli (all strains) there are in UniProt.

Using the search string (taxonomy_id:562) AND (fragment:false) AND (existence:1), we got 4303 hits, out of which 3752 are Swiss-Prot.

How many of these are from the notorious pathogenic serotype O157:H7 (including its sub-strains)?

Using the search string (taxonomy_id:83334) AND (fragment:false) AND (existence:1), we got 162 hits, of which 101 hits are from Swiss-Prot.

Find insulin from as many organisms as possible, without including entries that are not insulin.



Using the search string (gene:INS) AND (protein_name:Insulin), we got 978 hits, out of which 96 are from Swiss-Prot.

Find alpha-globin (the alpha subunit of hemoglobin) from as many ruminants as possible (see the GenBank exercise).

Using the search string (protein_name:"Alpha Globin") AND (taxonomy_id:9845), we got 35 results, out of which 6 are from Swiss-Prot.

Find alpha-A globin and alpha-D globin from *Columba livia* (Hint: You can use a "*" to perform the search with one search string).

Using the search string (protein_name:*) AND (organism_id:8932) and manually finding alpha-A and alpha-D.

<input checked="" type="checkbox"/>	O12985		HBAD_COLLI	Hemoglobin subunit alpha-D[...]	HBAD	Columba livia (Rock dove)	140 AA
<input checked="" type="checkbox"/>	P21871		HBA_COLLI	Hemoglobin subunit alpha-A[...]	HBAA	Columba livia (Rock dove)	142 AA