UNIVERSITY
OF SKÖVDE

School of Bioscience

WRITTEN EXAMINATION

Course: Bioinformatics – Concepts and Methods

Examination: Module 5

Course code: BI760A

Credits for written examination: 1.5

Date: 19 April 2024

Examination time: 2 hours

Examination responsible: Zelmina Lubovac

Teachers concerned: Björn Olsson

Aid at the exam/appendices: None

Other

Instructions

- ☐ Take a new sheet of paper for each teacher.
- ☒ Take a new sheet of paper when starting a new question.
- ☒ Write only on one side of the paper.
- ☒ Write your name and personal ID No. on all pages you hand in.
- ☒ Use page numbering.
- ☒ Don´t use a red pen.
- ☒ Mark answered questions with a cross on the cover sheet.

Grade points: 0-15 = F; 16-18 = E; 19-21 = D; 22-24 = C; 25-27 = B; 28-30 = A

Examination results should be made public within 18 working days

Good luck!

Total number of pages: 6

UNIVERSITY
OF SKÖVDE

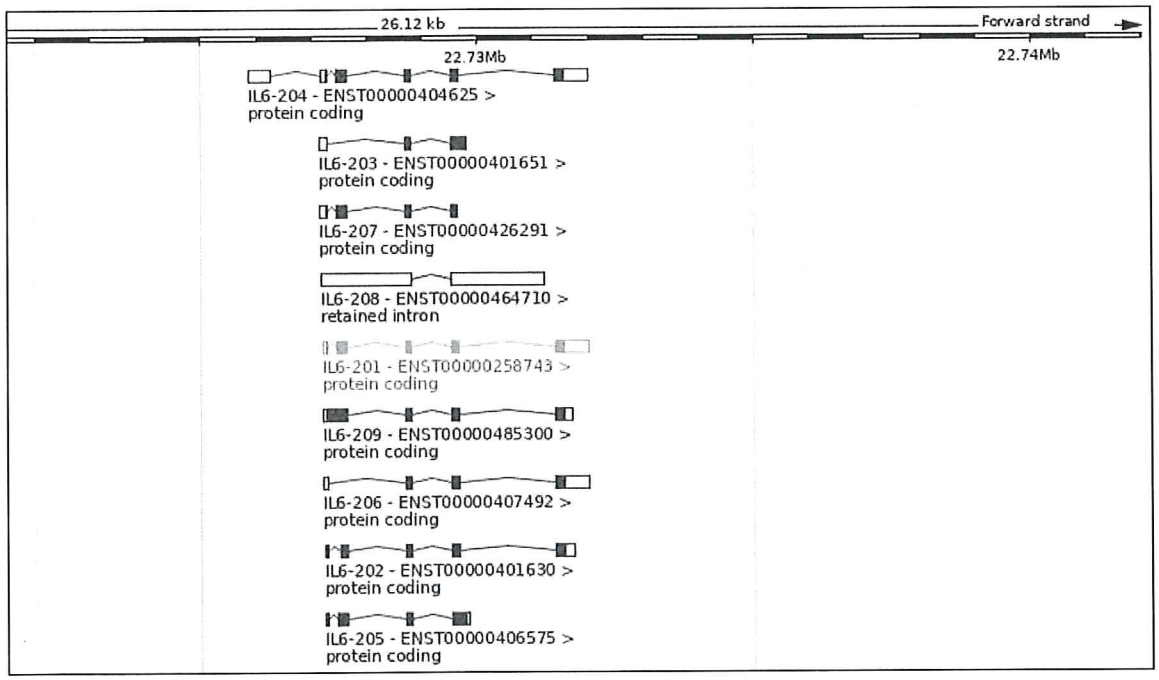
Question 1 (6p)

The figure below shows a region overview from Ensembl. It was created by searching for the IL6 gene (interleukin 6) in the human genome.

1a) How many transcripts (splice variants) does the IL6 gene have, and how many of those transcripts are protein coding? **(2p)**

1b) Describe what we can learn about the intron-exon structure of a transcript from the visualization in the region overview. You should pick one of the protein coding IL6 transcripts as an example to comment on in your description. Make sure to include the accession number of the chosen transcript in your answer. **(2p)**

1c) Now compare the transcript from question 1b with another IL6 transcript from the region overview. Explain how they differ from each other. For the two chosen transcripts you should mention, for example, which one has the largest number of exons and which one will produce the longest amino acid sequence. **(2p)**



UNIVERSITY
OF SKÖVDE

Question 2 (6p)

For each of the following claims about Ensembl, state if the claim is true or false. You do not need to give any motivations in your answer, just writing (for each claim) "true" or "false" is sufficient.

2a) The Ensembl genome database only contains data about the human genome. **(1p)**

2b) The data in Ensembl is based on user-submissions, i.e. any researcher who has sequenced a new genome can upload it to the Ensembl database. **(1p)**

2c) There is only one way to access the contents of the Ensembl database - by viewing it in the genome browser. **(1p)**

2d) You can find data about single-nucleotide polymorphisms (SNPs) in Ensembl. **(1p)**

2e) The Ensembl database was founded in the 1970s. **(1p)**

2f) Ensembl includes the genomes of model species, such as mouse, fruitfly and zebrafish. **(1p)**

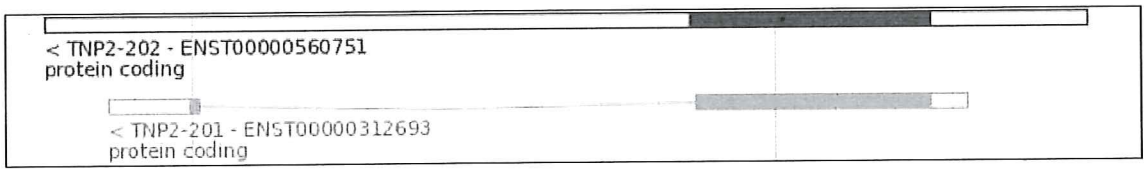
UNIVERSITY
OF SKÖVDE

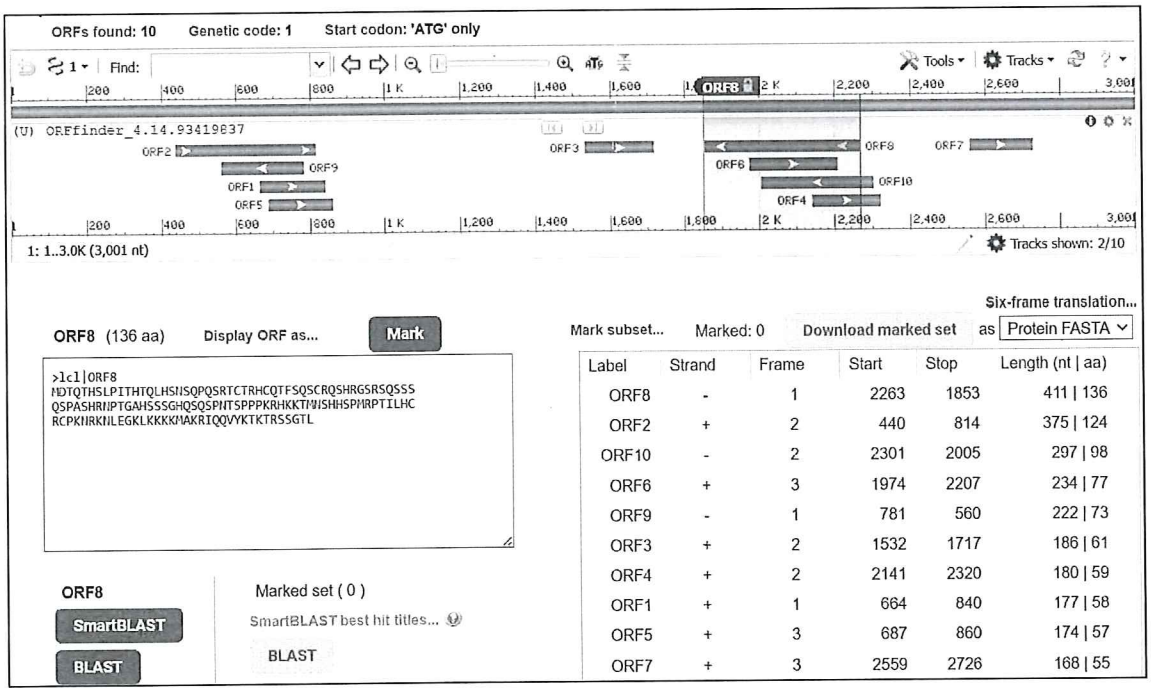
Question 3 (6p)

When predicting the locations of protein coding genes, it can be useful to start with a more basic task, namely to predict the locations of Open Reading Frames (ORFs). Explain the following:

3a) How does tool "ORF-Finder" predict open reading frames? Explain briefly in words the main ideas the ORF-Finder algorithm is based on. **(3p)**

3b) Compare the two figures below. The upper one is from Ensembl and shows two transcript variants of the human gene TNP2. The lower figure shows the ORF predictions for the same genomic region, where the TNP2 gene is located. The longest predicted ORF (ORF8) corresponds in length and location exactly to the longest exon of the TNP2 gene. Based on these two figures, discuss the possibilities and limitations of using ORF finding as a method for gene prediction. Would it be possibile to find all human genes by ORF finding without also finding many falsely predicted genes? What can be done with parameter settings to try to improve the accuracy of the predictions? Would it be possible to predict the intron-exon structure of genes in euakaryote genomes only by ORF finding? Can ORF finding give us information about different transcript variants? **(3p)**





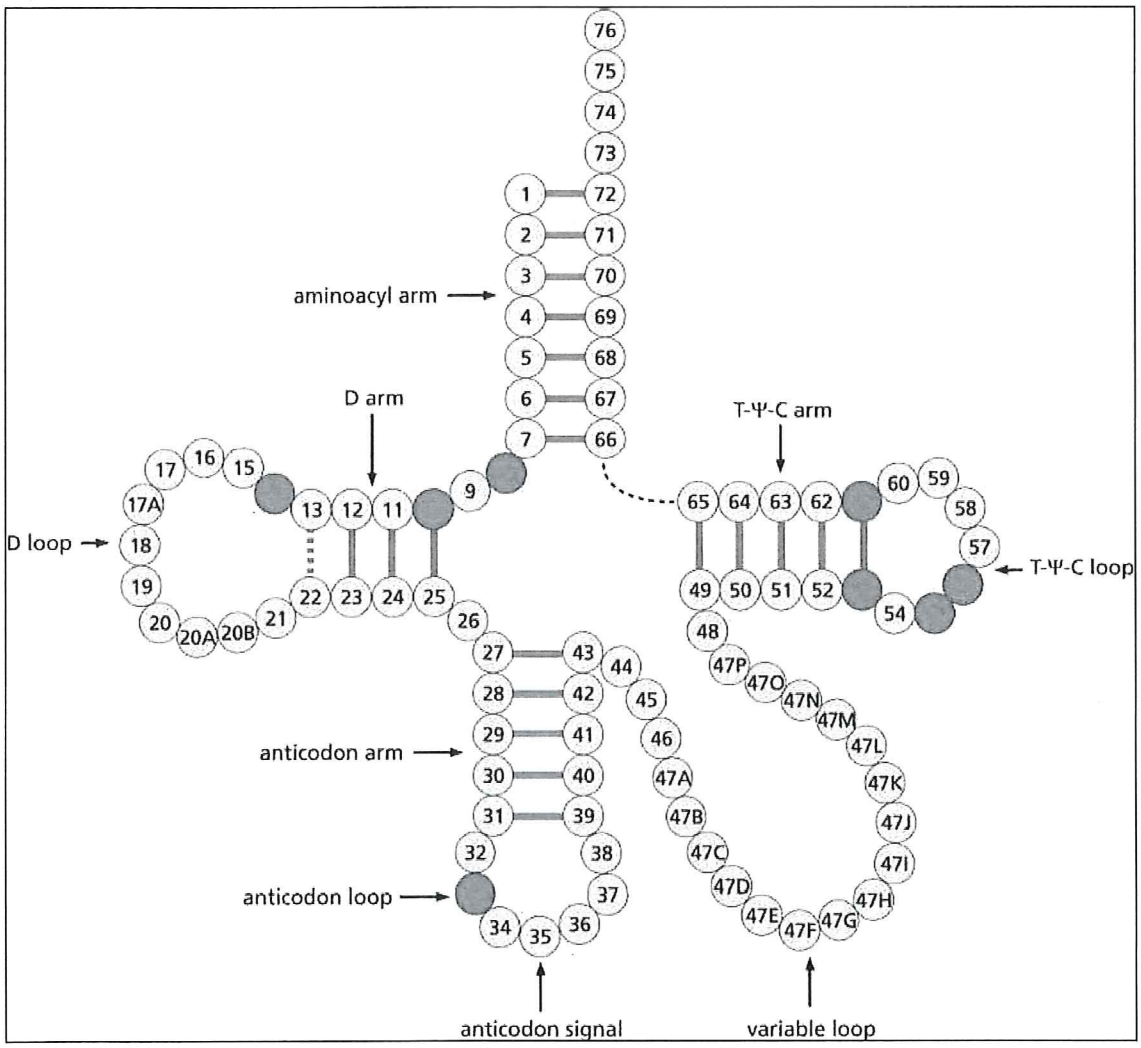
Label	Strand	Frame	Start	Stop	Length (nt | aa)
ORF8	-	1	2263	1853	411 | 136
ORF2	+	2	440	814	375 | 124
ORF10	-	2	2301	2005	297 | 98
ORF6	+	3	1974	2207	234 | 77
ORF9	-	1	781	560	222 | 73
ORF3	+	2	1532	1717	186 | 61
ORF4	+	2	2141	2320	180 | 59
ORF1	+	1	664	840	177 | 58
ORF5	+	3	687	860	174 | 57
ORF7	+	3	2559	2726	168 | 55

UNIVERSITY
OF SKÖVDE

Question 4 (6p)

The tRNAscan algorithm predicts tRNA genes using a rules-based approach. The figure below is a schematic drawing of the structure of a tRNA molecule. When answering the question below, you can refer to this figure to illustrate and exemplify your explanations.

Describe <u>two</u> of the rules that tRNAscan applies when classifying a sequence as either being a tRNA gene sequence or not. You can get up to **3p** for each described rule (depending on correctness and level of detail), for a total of maximum **6p**.



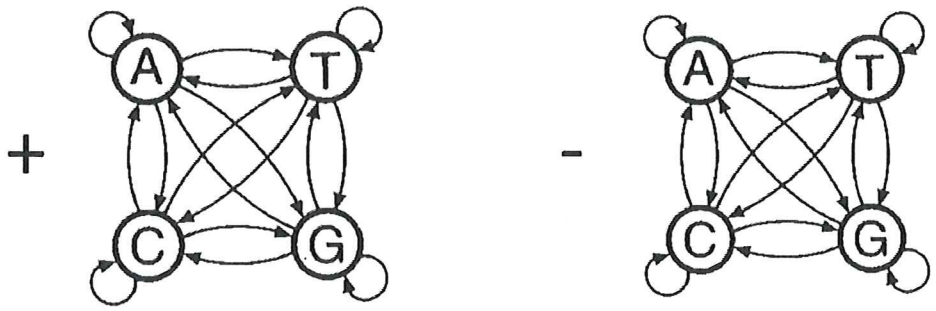
UNIVERSITY
OF SKÖVDE

Question 5 (6p)

Markov models have become the most commonly used method for predicting protein coding genes in eukaryote genomes.

5a) Explain what a Markov model is. What "components" does it consist of? You can refer to the figures below to illustrate and explain the "building blocks" of a Markov model. **(3p)**

5b) Explain the difference between a Markov model and a Hidden Markov model (HMM). Explain how the two models shown below (marked with plus and minus) could be joined into one HMM. **(3p)**



+	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

−	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292