UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Obligatory assignment: MBV-INF4410 and MBV-INF9410
Deadline: *Latest* December 2nd 2018, at 23:59
Permitted materials: All written material, including all Internet resources

Only students that get this assignment approved will be permitted to take the final exam.

Your completed assignment must be returned, at the latest, at 23:59, December 2nd. It should be sent by e-mail to the course coordinator Jon Bråte (e-mail address: jon.brate@ibv.uio.no). Please put the course code and your name in the subject field (e.g. Oblig MBV-INF4410 Dolly Duck").

The assignment should be handed in as a single PDF document (Microsoft Word or an Open Office Document is also acceptable). **Please also include your name and course code in the document and in the document title.**

You are encouraged to use screenshots and other figures in order to improve your explanations.

THE WORK MUST REPRESENT YOUR OWN ANSWERS AND BE WRITTEN IN YOUR OWN WORDS.

Answers should contain only what is asked for. Some questions have multiple parts. Your answers may be given in English or in Norwegian. Technical questions about the assignment can be answered by Jon Bråte (e-mail address: jon.brate@ibv.uio.no).

EXTRA Obligatory assignment - MBV-INF9410 students only:

Write an essay of at least 2500 words on the topic:

- How can some of the methods (two or more) covered in this course be used in your own research? See more info about the essay in the email sent on 19 November.

Exercise on human PCSK9

Open your Internet browser and go to the website http://www.genenames.org. Here you find the database of the HUGO Gene Nomenclature Committee (HGNC) approved gene names. Search for the gene PCSK9 and go to the PCSK9 page. Use the links and resources pointed to on this page to answer the questions below. In particular, the UniProt, GenBank, RefSeq, Ensembl, and OMIM links will be useful and you will find all the information you need in these databases. If you wish, you may also use other resources or databases on the web, or the scientific literature, to answer the questions, but this is not necessary. Human gene PCSK9 (most likely) has only a single biologically relevant splice variant. The full-length PCSK9 protein, including signal sequence and pro-segment, has 692 residues. We will ignore any other putative splice variants here. Using some (relatively few) screenshots to show what you do to answer the questions below might be helpful.

a) What is the approved symbol and name for this human gene, and on which chromosome is it found? Are there any other names/symbols used in the literature? What are they? Briefly, in no more than 5-10 sentences, describe the biological function of the PCSK9 protein. List your sources to this information.

b) What are the identifiers (IDs, Accession, etc.) for the human PCSK9 protein sequences (not gene or transcript, for example) in UniProtKB? What is the *gene* identifier for PCSK9 in Ensembl? And what is the *transcript* identifier for the correct isoform in Ensembl? What is the identifier (Accession) for the protein sequence corresponding to the nucleotide sequence with the NCBI gene id 255738?

c) How many exons are there in human PCSK9? What is the length of the 5' intron (first intron) in this gene? What is the shortest intron, and how long is it? Are there any introns that are not of the standard GU-AG type? What is the sequence of the human PCSK9 stop codon?

d) Find the PCSK9 orthologues in Ensembl. From this resource, get the protein sequences for the PCSK9 orthologues (longest isoforms) from chimpanzee (*Pan troglodytes*), orangutan (*Pongo abelii*), gorilla (*Gorilla gorilla*), pig (*Sus scrofa*), Zebrafish (*Danio rerio*), Chinese softshell turtle (*Pelodiscus sinensis*), and tilapia (*Oreochromis niloticus*). What are the lengths of the protein sequences? Make a multiple sequence alignment (MSA), using JalView and the Muscle algorithm, of these 7 sequences, together with the human PCSK9 sequence from UniProtKB. Two of the sequences stands out as being quite different from the others. What can the reason be? Is it likely that there is something wrong with these sequences? Do you have any suggestions how you might correct them? If you do, correct it. There is one more sequence that appears to be wrong at one of the ends. Which one, and what is the problem? (Hint: start codon). Correct the sequence if possible.

e) Go to the NCBI BLAST page (blastp program) and use human PCSK9 protein as a query sequence in a search against the RefSeq protein database. Restrict the blast search to vertebrates and set "Max target sequences" to 1000. Keep the other settings at default. Select the sequence with the lowest e-value (or highest score if same e-value) from the same 7 species as the previous exercise. Check the boxes to the left of the target

sequence in the list and click "Download" and select "GenBank" format. You will get a file called *sequences.gb*. Now, use Biopython to print the lengths of the downloaded sequences. You can use the Python script provided in the mail as a starter. But you need to change a small ting (hint: what is inside *record.seq* and how do you get the length of things in Python?) *Remember, on Freebee you need to do module load python2 in addition to module load python3*. Attach a screen shot of the python script you used. Are the sequences of the same lengths as the corresponding ones from Ensembl? What about the orangutan and chimp sequences? And what about the Turtle, does it have the same error as the Ensemble one? What could be the reason for these differences? Now save the sequences to a file in fasta format. You can use Biopython for this also, and the previous script is a good starting point. (hint: there are a few commented lines where you can simply remove the comments). Show a screen shot of the script you used. Align the sequences the same way as before. Which dataset seems "better" or more reliable, or are they basically similar?

f) Go back to the Blastp hit list, download the two best hits from *Xenopus* sp. (two frog species). What are the RefSeq identifiers/accessions for these two sequences? Get the sequences and add them to the 8 sequences you already have in Jalview (the Ensembl sequences). Re-align all 10 sequences with Muscle. Show the multiple sequence alignment in a figure. What is the sequence identity between the two frog PCSK9 orthologues? What about human vs. the frog sequences? *Tip: go back to the alignment, deselect everything with the Esc button. Then select human and the two frog sequences before calculating the pairwise identity.* Several human families with very high blood serum LDL levels (high levels of "bad cholesterol") have mutations in the residue Asp374. Based on the information in the MSA you have generated, is it likely that the PCSK9 Asp374Tyr mutation will alter the function of human PCSK9? Explain why you come to this conclusion.

g) Between residues 450 and 692 in the human sequence (the C-terminus) there are 18 Cys residues in human PCSK9. Locate these residues in the MSA. Are they conserved or not? Find information about the human single-nucleotide polymorphism (SNP) rs562556, for example in Ensembl. This variant changes residue Val474 in human PCSK9 to something else. What is the mutation? What is the reference codon and what is the alternative variant codon? the population genetics for this SNP. In the 1000 Genomes project data, what is the allele frequency of the minor allele in the Iberian population in Spain. What is it in the Han Chinese in Beijing population? the affected residue in the MSA you generated above. Is it likely that this mutation will affect the function of PCSK9 severely? Why/why not?

h) The 3D structure of human PCSK9 can be found in the PDB with the identifier 2P4E. Which method was used to determine the 3D structure? Download the 2P4E PDB file and open it in PyMOL. Show the protein with "cartoon" rendering and colour by secondary structure. Locate the 18 Cys residues you investigated in (g) above. Make a selection containing only these 18 residues and show them as "sticks" with an appropriate colouring. Explain why these 18 residues are conserved/not conserved. What appears to be their function? With PyMOL, make a nice image of the PCSK9 3D structure showing the 18 Cys residues, and include it in your answers.

i) One of your colleagues is planning to use *Oreochromis niloticus* as a model organism to study the function of PCSK9. Is it possible to make a reliable 3D structure model of

tilapia PCSK9? Why/why not? Which method should be used to generate the protein 3D model? Briefly, list the steps involved in making the 3D structure model. Do not generate the model.