UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Obligatory assignment: MBV-INF4410 and MBV-INF9410
Deadline: November 30th 2015, at 09:00
Permitted materials: All written material, including all Internet resources

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

Your completed assignment must be returned, at the latest, at 9 am, Monday November 30. It should be sent by e-mail to the course coordinator Jon K. Lærdahl (e-mail address: jonkl@medisin.uio.no). Please put the course code and your name in the subject field (*e.g.* Oblig MBV-INF4410 Dolly Duck").

The oblig must 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

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 oblig can be answered by Jon K. Lærdahl (e-mail address: jonkl@medisin.uio.no).

EXTRA Obligatory assignment: MBV-INF9410 (only)

Write an essay of at least 2500 words on one of these two topics

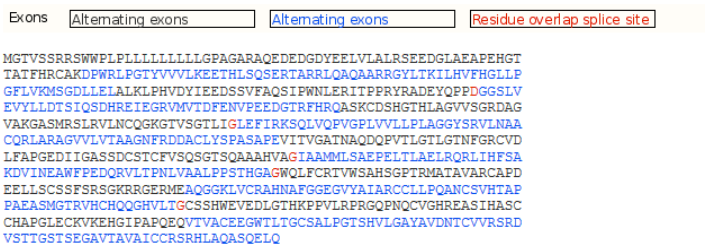
- How can some of the methods described in this course be used in your own research?
- A course relevant topic of your own choice. In this case you must get this approved by Jon K. Lærdahl *before* you write the essay

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, Entrez Gene, 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 and RefSeq. What is the transcript identifier for PCSK9 in Ensembl? What is the identifier (Accession) for the *protein* sequence corresponding to the nucleotide sequence in GenBank entry AX207686? Get the corresponding 4 protein sequences from the 4 databases and align them in Jalview. All 4 sequences are 692 residues long, but are they identical? If not, what are the differences and what might be the reason?

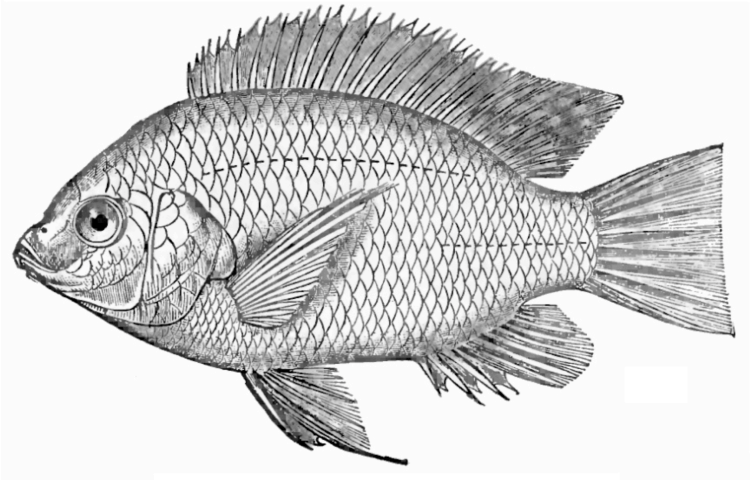


c) How many exons are there in human *PCSK9*? What is the length of the 5' intron in this gene? What is the shortest intron, and how long is it? Are there any introns that are not of the standard GU-AG type? What is the sequence of the human *PCSK9* stop codon?

d) Find the PCSK9 orthologues in Ensembl. From this resource, get the protein sequences for the PCSK9 orthologues from chimpanzee, orangutan, gibbon, pig, dolphin, Chinese softshell turtle (longest sequence variant), and tilapia (*Oreochromis niloticus*). Make a multiple sequence alignment (MSA), using Jalview and the Muscle MSA program, of these 7 sequences, together with the human PCSK9 sequence from UniProtKB. One 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 this sequence? Do you have any suggestions how you might correct this sequence? 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? What can the reason be? Correct the sequence if possible.

e) Show the full MSA (or a part of it) for the 8 sequences in a figure, and answer the following questions: What is the longest segment of the human PCSK9 protein that is 100% conserved in all 8 species? What is the sequence identity between the PCSK9 orthologues from human and chimpanzee? What about human vs. gibbon, human vs. dolphin, and human vs. tilapia? Make the MSA figure as nice as possible, and include it in your answers. Use appropriate colouring and, if you want to, add/include/alter relevant annotations and/or numbering.



f) Go to the NCBI BLAST page (blastp program) and use human PCSK9 protein as a query sequence in a search in the RefSeq proteins database. Adjust relevant settings, if needed, and explain why you did it. You get two good hits from Xenopus frog sequences. 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. 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? 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 (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? Investigate 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? Find 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 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* tilapia 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.

j) What is the easiest way to get access to and view PDB structures on a smartphone? Answer in one or two sentences.

k) The NCBI Entrez is an integrated database retrieval system that provides access to a diverse set of 40 databases that together contain at least 1.3 billion records. Some of the databases contains data submitted by the research community ("direct submission" - D) while other databases contains data generated by the NCBI and/or NLM only (internal NCBI/NLM curation – N). Give three examples of each of the two types of NCBI databases, D and N.