CFG - Practical Session 1

IMPORTANT NOTES:
IMPORTANT_1: Submit your answers with your ESCI-UPF gmail account. You can only submit your answers once.
IMPORTANT_2: Wednesday 17th of April 2024, is the last day to submit your answers.
IMPORTANT_3: When you submit your answers, a confirmation message will appear on your screen “Your response has been recorded” but not in your email

You are working as a consultant in the Medical Genetics unit at Vall d’Hebron. A patient has been referred to because of a possible genetic disease and your team has performed a screening on a set of possible gene candidates. You are sent the following FASTA sequence as the hit candidate for the mutation of interest in your patient (*gene.fna*). Unfortunately for you, there has been an issue with the email provider and you are locked out of your account, so you can’t check the email and the information it contained. The file got automatically downloaded prior to this, but the header contains no information, so you will have to do some research on your own before you can send your report to the doctors.

Adreça electrònica *

El teu correu electrònic

Section 1

You decide to start by using the Basic Local Alignment Sequences Tool (BLAST) https://blast.ncbi.nlm.nih.gov/Blast.cgi to identify the sequence.

1.1 By looking at it, what type of sequence is it?

- (●) DNA
- ( ) RNA
- ( ) Protein

Esborra la selecció

1.2 Therefore, what alignment algorithm(s) can you use? *We advise you to click the option "Show results in a new window" before running the alignment. Also, beware that it may take 3-4 minutes. Be patient!*

- [ ] blastp
- [x] blastn
- [ ] tblastn
- [x] blastx
- [ ] tblastx

1.3 From the BLAST results, can you guess the name of the protein for which the gene encodes? (More than 1 answer may be correct)

- [x] COL1A2
- [x] Collagen Type I Alpha 2 Chain
- [x] BAC clone
- [ ] Uncharacterized protein

1.4 The top hit includes other organisms apart from *Homo sapiens*. Which is the accession number for the first hit to *Macaca mulatta*?

- ○ AC186880.2
- ○ AF004877.1
- ○ OX621291.1
- ◉ AC171642.3

Esborra la selecció

1.5 What do the first 5 organisms with the best hits have in common? (More than 1 answer may be correct)

- ☐ They are all primates
- ☑ They are all mammals
- ☑ They are all chordates
- ☐ They are all part of acoelomorpha

Section 2

Having identified your protein, you decide to look it up on UniProt (https://www.uniprot.org/).

2.1 Query the sequence name of your protein in UniProt (be precise! Follow the HGNC ID conventions. Look up "HUGO Gene Nomenclature Committee" if you don't know what that is). How many entries are there?

- (●) 922
- ( ) 288,154
- ( ) 11,187
- ( ) 46

Esborra la selecció

2.2 How many of those are manually reviewed (Swiss-Prot)? *

- ( ) 867
- (●) 55
- ( ) 721
- ( ) 28

2.3 For this protein, which organism has the most entries listed on UniProt?

- ( ) Mouse
- ( ) Zebrafish
- (●) Human
- ( ) Rat

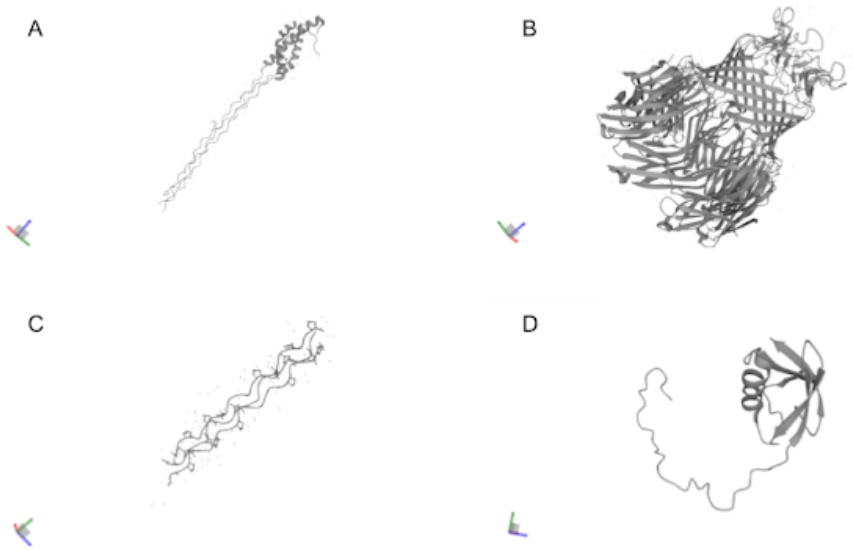
Esborra la selecció

2.4 Enter in the entry for *Homo sapiens* (XXXX_HUMAN, where XXXX is the name of your protein). Retrieve the FASTA sequence of the protein product. How many positions does it have?

- ○ 129,314
- ◉ 1,366
- ○ 1,360
- ○ 2,010

Esborra la selecció

2.5 Retrieve an image of the protein structure. Which of these options is your protein?



- ◉ A
- ○ B
- ○ C
- ○ D

Esborra la selecció

2.6 UniProt lists some disease-causing variants, which may cause a disease called Osteogenesis Imperfecta. There are several types of this illness: your patient, in particular, has no dentinogenesis imperfecta and is of normal height. Which is it?

- (●) Type I OI
- ( ) Type II OI
- ( ) Type III OI
- ( ) Type IV OI

Esborra la selecció

2.7 Is there any variant associated only to this type and not the others? *

- (●) Yes
- ( ) No

2.8 What kind of mutations are they? *

- (●) Missense
- ( ) Nonsense
- ( ) Silent
- ( ) There are no variants exclusive to this type of OI

2.9 Retrieve the amino acid sequence and run InterProScan (https://www.ebi.ac.uk/interpro/search/sequence/). What domains does the protein have according to InterProScan?

- [x] Fib_Collagen_C
- [ ] Collagen
- [x] COLFI
- [x] COLFI_2
- [x] NC1_FIB

2.10 Look up the GO Terms associated to the protein **in UNIPROT**. Which ones are associated with it?

- [x] endoplasmic reticulum lumen
- [ ] bone mineralization
- [ ] protein heterotrimerization
- [ ] odontogenesis

2.11 Look up the KEGG pathways associated to the protein. What pathway(s) is it NOT involved in?

- [ ] PI3K-Akt signaling pathway
- [ ] Relaxin signaling pathway
- [ ] Focal adhesion
- [x] mTOR signaling pathway

2.12 Look up the protein on Online Mendelian Inheritance in Man (OMIM). Which one(s) of the following is/are true?

- [x] It is located on chromosome 7 in humans
- [x] Most phenotypes have Autosomal Dominant inheritance
- [x] Mutations in this gene can cause diseases other than Osteogenesis Imperfecta
- [ ] Mutations in this gene only cause Type I OI

2.13 Check up the binary interactions listed on UniProt. Now go to STRING (linked on UniProt) and retrieve the network of protein-protein interactions described for your protein of interest. Which one(s) are NOT true?

- [x] The interaction between COL11A1 and COL1A2 is supported by evidence of gene fusion events
- [ ] The interaction between COL1A1 and CD44 is only supported through co-mentioning in PubMed abstracts
- [ ] The protein-protein interaction network has 11 nodes and 50 edges (out of the predicted 15) , and therefore there are significantly more connections than what could be expected at random
- [x] The co-occurrence of the genes is particularly strong in Opisthokonta

2.14 What is the function of these other proteins? Look them up on UniProt. Which of them have osteogenesis imperfecta-causing mutations listed?

- [ ] COL5A2
- [ ] LUM
- [ ] COL3A1
- [ ] ITGB1
- [x] COL1A1

Section 3
You additionally decide to look up the protein on Ensembl

3.1 Where is this gene located in the human genome (coordinates)?

- ○ Chromosome 8: 94,394,895-94,431,227 reverse strand.
- ○ Chromosome 7: 94,394,895-94,431,227 reverse strand.
- ◉ Chromosome 7: 94,394,895-94,431,227 forward strand.
- ○ Chromosome 8: 94,394,895-94,431,227 forward strand.

Esborra la selecció

3.2 What is the latest version of the Ensembl entry?

- ◉ 19
- ○ 11
- ○ 6
- ○ 3

Esborra la selecció

3.3 How many transcripts does this gene have?

- ◉ 12, only one is protein-coding
- ○ 12, all of them protein-coding that give rise to different isoforms
- ○ 1, the one that gives rise to COL1A2 protein
- ○ 37

Esborra la selecció

3.4 How many phenotypes has it been associated with? *

- (●) 18, including different forms of Osteogenesis Imperfecta, Ehlers-Danlos and osteoporosis
- ( ) 18, describing different forms of Osteogenesis Imperfecta
- ( ) 48
- ( ) There are no phenotypes associated

3.5 For the gold standard transcript, which one of this sentences is NOT true?

- ( ) It has 52 exons
- ( ) It is 5072 bp in length
- (●) Exon 1 is composed completely of translated regions
- ( ) Exon 2 is 11 bp long

Esborra la selecció

3.6 Which are true for GOLD colored transcripts? *

- [x] Only mouse, human and zebrafish can have gold transcripts
- [x] They are identical both in the automated pipeline and the manually curated one
- [ ] They identify the most clinically relevant transcript, i.e, the one that has the most disease-causing variants listed
- [ ] They identify the only protein-coding transcript of the gene

3.7 For the alpha-I chain (COL1A1)...

- [x] There are two protein-coding transcripts listed
- [ ] It is also located in Chromosome 7, very close to COL1A2
- [x] It is involved in many more phenotypes than COL1A2
- [ ] The gold transcript has 50 exons

3.8 Now go back to the COL1A2 **protein-coding transcript** on Ensembl. How many variant alleles are listed? *

- (•) 17856
- ( ) 208
- ( ) 1570
- ( ) None

3.9 Ensembl has listed 204 orthologs (see "about this gene"). Check them out. Which one of these are true?

- [x] There are orthologs in 214 species
- [ ] All primates have orthologs
- [x] Most of the orthologs are one-to-one
- [x] There are more one-to-one orthologs in the sauropsida dataset than in rodents

3.10 Which dataset has one-to-many orthologues? More than one may apply *

- [ ] Primates
- [ ] Placental mammals
- [ ] Rodents
- [x] Fish

You can see from this practical that UniProt is a very powerful aggregator of information from different databases, containing a lot of cross-references to different tools and platforms that allow you to get the full picture of a protein from a single page. However, be careful! Not all proteins are as well-researched as this one. In your project, you may have to go to the source databases directly, or run prediction tools.

Envia

Pàgina 1 de 1

Esborra el formulari