Bioinformatics for Biologists - Spring 2021

Study guide for Applied test 1.

1. What is this protein? What species is it from?

MAQNDSQEFAELWEKNLIIQPPGGGSCWDIINDEEYLPGSFDPNFFENVLEEQPQPSTLP PTSTVPETSDYPGDHGFRLRFPQSGTAKSVTCTYSPDLNKLFCQLAKTCPVQMVVDVAPP QGSVVRATAIYKKSEHVAEVVRRCPHHERTPDGDNLAPAGHLIRVEGNQRANYREDNITL RHSVFVPYEAPQLGAEWTTVLLNYMCNSSCMGGMNRRPILTIITLETQEGQLLGRRSFEV RVCACPGRDRKTEESNFKKDQETKTMAKTTTGTKRSLVKESSSATLRPEGSKKAKGSSSD EEIFTLQVRGRERYEILKKLNDSLELSDVVPASDAEKYRQKFMTKNKKENRESSEPKQGK KLMVKDEGRSDSD

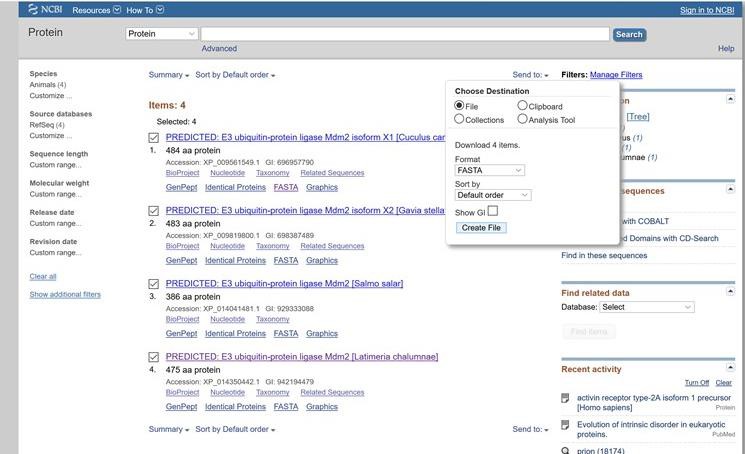
1. What species is the protein with UNIPROT accession P79734 from?
2. Build a common (species) tree for the following species (tax ids: 7757, 8030, 81824, 256737).
   1. Which tax id is the outgroup?
   2. What are the species names for these tax ids?
   3. For these species, what is the species name of the outgroup?
3. **For a given sequence, e.g. >NP\_001079353.1 bone morphogenetic protein 2-A precursor [Xenopus laevis]**
4. What is the length of the protein?
5. What is the GeneID?
6. What species is it from?
7. Are there any isoforms for the protein above? (Yes/No)
8. Are there any paralogs of this protein in the same species? (Yes/No)
9. Is NP\_001095263 an isoform or a paralog of the protein above?
10. Is XP\_018119713 an isoform or a paralog of the protein above?
11. Are there any homologs for this protein in *Gorilla gorilla*?
12. If H=yes, what is the query cover, E-value, and Ident (sequence identity) for the protein shown above and the best hit (for *Gorilla gorilla*, ranked by e-value)?
13. Are there any homologs for this ***full-length*** protein in species *Arabidopsis thaliana* in the

***refseq*** protein database?

1. Retrieve the following sequences from NCBI and align the sequences with Muscle default in Jalview or using Muscle at EBI (https:/[/www.ebi.ac.uk/Tools/msa/muscle/):](http://www.ebi.ac.uk/Tools/msa/muscle/)) XP\_009561549, XP\_009819800, XP\_014041481, XP\_014350442 (see fig below for how to extract the sequences quickly)

What is the length of the multiple sequence alignment?

A. 494 B. 504 C. 514 D. 532



1. Use NCBI BLAST to identify a template for homology modeling for a certain sequence (example below). What is the pdb id for the best template (sorted by e-value)? What is the e-value, query cover, and percent identity for the best (sorted by e-value) target-template alignment?

>XP\_009561549.1 PREDICTED: E3 ubiquitin-protein ligase Mdm2 isoform X1 [Cuculus canorus] MCNTEMSSLTDASAVTASEREALVRPKPLLLRLLKLAGAEKDTFTMKEVIFYLGQYIMSKQLYDEKQQHI VHCANDLLGDLFGVTSFSVKEHRRLYSMISRNLIAVSQQDSRHADTPEDDARFQLEEENSLKESIQELEE KQTSSNVTSRPTTSSRRRPHSESEENSSDDLHSDRRKRHKSDSISLTFDESLSWCVVSGLCRERSNSSDS TDSLSIPDLDASSLSENSDWFDHSSVSDQFSVEFEVESIYSEDYNHNEEGQELTDEDDEVYQLTIYQDED SDADSFDEDPEISLADYWKCPECSEMNPPLPRHCHRCWALREDWLPDEKSDKSVKSKLESSLPLESEEGF DVPDCKKVKMIEDKEPALEENEDKAAQISESQESEDYSQPSTSSSMFCSSQEDYKEPEKRETQDKEESIE SSLPVTSIEPCVICQSRPKNGCIVHGKTGHLMSCFTCARKLKKRNKPCPVCRQPIQMIVLTYFS

1. For a specified PDB id, e.g. 1phz,
2. What is the resolution?
3. What ligands are present in the structure?
4. Which of the following PDB structures have the highest resolution?

A) 1XFU B) 2YGG C) 3EWT D) 4HEX

1. How many homologs (with an e-value below 1e-30 [nr database]) of a given protein can you find in a given species (perhaps in a given database – although the database may also not be given)? E.g. how many homologs of protein AAA39883.1 can you find in *Plasmodium falciparum* in the nr database?

9.5 For a given sequence: MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGS LQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN

1. what is the protein?
2. what is the predominant secondary structure in this sequence based on PSIPRED prediction?
3. While your PSIPRED prediction is running: Are there any experimental protein structures of this protein (in the PDB)? Hint: BLAST
4. Based on your PSIPRED results: Which residue has the lowest confidence (e.g. P9 or F25)?
5. Use PyMOL to determine if a given protein is a monomer, dimer, trimer, etc. or what the secondary structure is of a given residue. Combine with PSIPRED to determine if it is a true positive, false positive, true negative or false negative.

**To be covered for Day 10-11**

1. PFAM
2. What are the PFAM domains of a certain protein
3. How many domain architectures are given domain found in?
4. Which species have a certain domain?
5. How many sequences are in the full alignment? How many sequences are in the seed alignment?
6. What are characteristics of the three most common amino acid found at a specific site in the HMM Logo for a given PFAM domain? For instance, site 203 in PFAM domain PF02171
7. What is the most common amino acids at site 217 in PFAM domain PF02171?
8. KEGG

Based on a given Pathway, what are specific interactions? For instance, for [KEGG PATHWAY: Pathways in](https://www.genome.jp/kegg-bin/show_pathway?hsa05200) [cancer - Homo sapiens (human) (genome.jp)](https://www.genome.jp/kegg-bin/show_pathway?hsa05200) which proteins are phosphorylated by ERK? How does MEK activate ERK? How does PIP3 contribute to the regulation of GSK-3β?

* + Remember the Help button