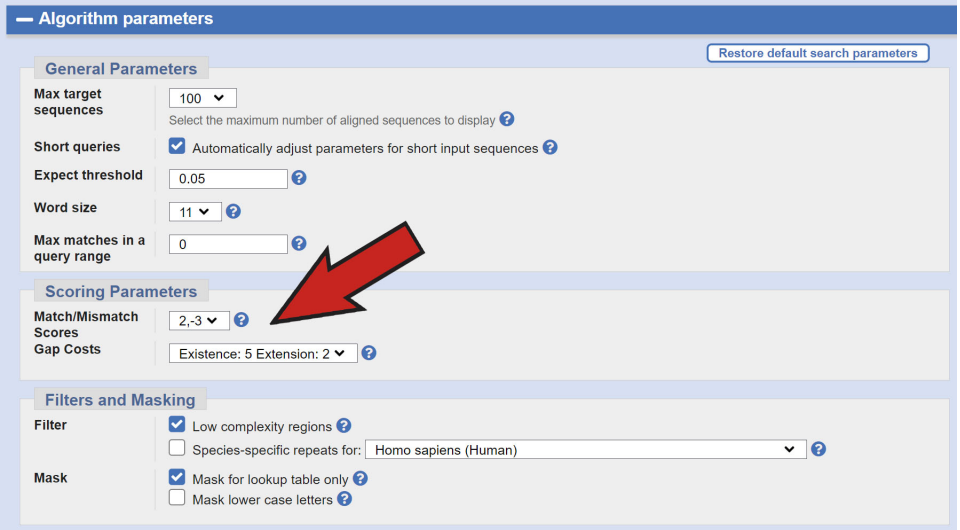
University of Pennsylvania
BIOL4536 Fall 2023

HW#6
(BLAST)

Assigned October 11th
Due October 19th, 11:59pm

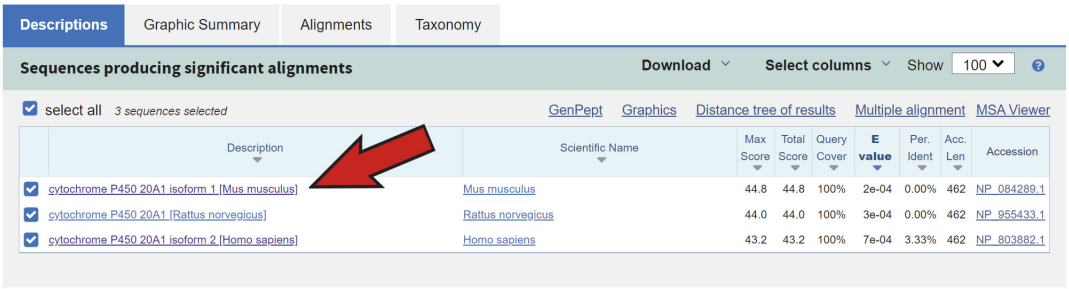
Question (1) Go to the NCBI BLAST page (`https://blast.ncbi.nlm.nih.gov/`) and select "Nucleotide BLAST" and under "Program Selection" choose "blastn" (should be the third one). We learned in class that you can align DNA with a scoring scheme that scores mismatches between pyrimadines differently from mismatches between purines. Expand the "Algorithm parameters" section and examine the "Scoring parameters". Is it possible to configure BLAST to score mismatches between pyrimadines differently from mismatches between purines? If so, explain how.

ANSWER: No, it only offers one score for mismatches regardless of type.

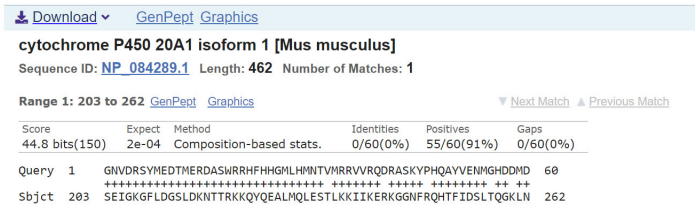


Question (2) What matrix works best? Based on the matrix that works, what do you conclude about the evolutionary distance between this sequence and its closest homologs in the database. What gene is it? What species is the closest hit? Show the best alignment. How many amino acids are unchanged? How many are "positives"? What does "positive" mean here? What's the *E*-value?

ANSWER: It takes the PAM250 matrix to achieve significance. This means the evolutionary distance is large.



The gene is “cytochrome P450 20A1 isoform 1” (Mus musculus). The alignment is:

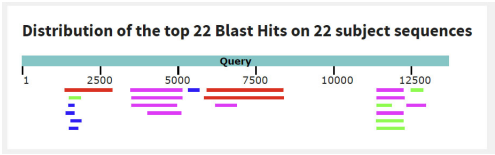


Zero amino acids are unchanged (Identities = 0/60). There are 55 of 60 “positives”. Positives are pairs which have a positive score in the (PAM250) substitution matrix. The *E*-value is 2*e*-04.

Question (3) How many results were returned? How many species are involved in the hits?

ANSWER: 22 results were returned involving from two species.

Click on “Graphic Summary”. Supply a screen shot of the graphic.

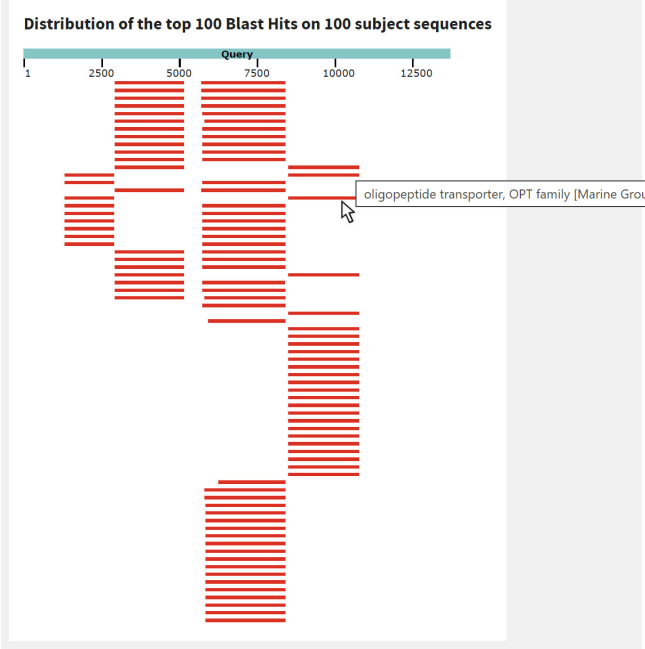


Hover over the colored bars to figure out how many genes there appear to be on this contig.

ANSWER: There are six different genes.

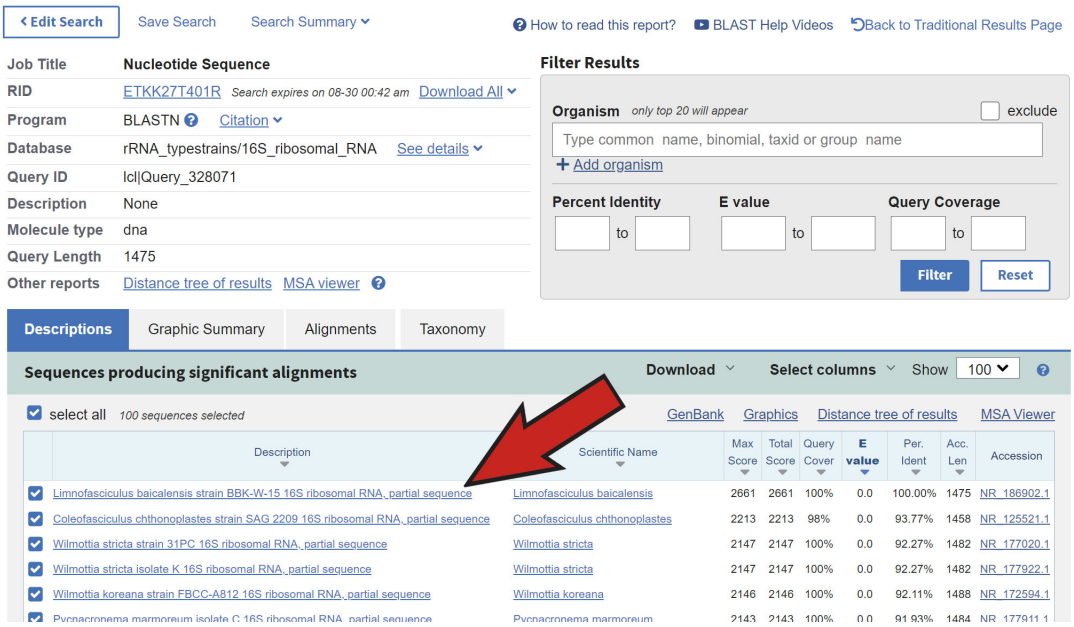
Go back and run BLAST again but this time change database to “Non-redundant protein sequences (nr)” (again restrict to Archaea). Does this reveal another possible gene on the contig? If so, what’s the gene’s name?

ANSWER: This search does reveal a 7th gene (oligopeptide transporter)



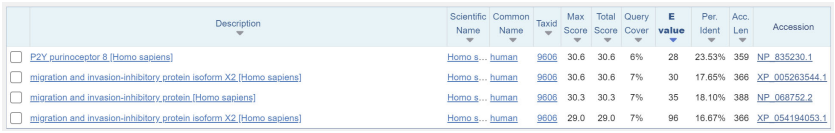
Question (4) Metagenomics. How much higher is the score of the top hit compared to the second hit? Based on the top hit, what is the species?

ANSWER: The top hit has score 2661 and the second hit has score 2213, so a difference of 448.



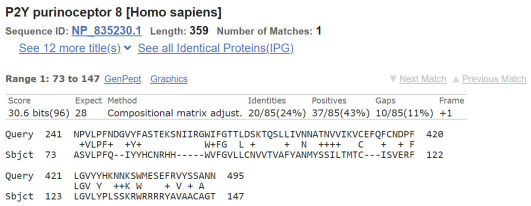
Question (5) How many genes are returned (give a screen shot of the Descriptions). What is the top hit? What's the E-value of the top hit? Intepret the E-value. Show the actual alignment.

ANSWER: Four genes are returned. The top hit is "P2Y purinoceptor 8". The descriptions look like this:



Description	Scientific Name	Common Name	Taxid	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
P2Y purinoceptor 8 [Homo sapiens]	Homo s...	human	9606	30.6	30.6	6%	28	23.53%	359	NP_835230.1
migration and invasion-inhibitory protein isoform X2 [Homo sapiens]	Homo s...	human	9606	30.6	30.6	7%	30	17.65%	366	XP_005263544.1
migration and invasion-inhibitory protein [Homo sapiens]	Homo s...	human	9606	30.3	30.3	7%	35	18.10%	388	NP_068752.2
migration and invasion-inhibitory protein isoform X2 [Homo sapiens]	Homo s...	human	9606	29.0	29.0	7%	96	16.67%	366	XP_054194053.1

The *E*-value of the top hit is 28. That means we expect 28 hits this good or better even from a random database. So that's not very convincing. The actual aignment looks like this:



P2Y purinoceptor 8 [Homo sapiens]

Sequence ID: NP_835230.1 Length: 359 Number of Matches: 1

See 12 more title(s) See all Identical Proteins(IPG)

Range 1: 73 to 147 GenPept Graphics Next Match Previous Match

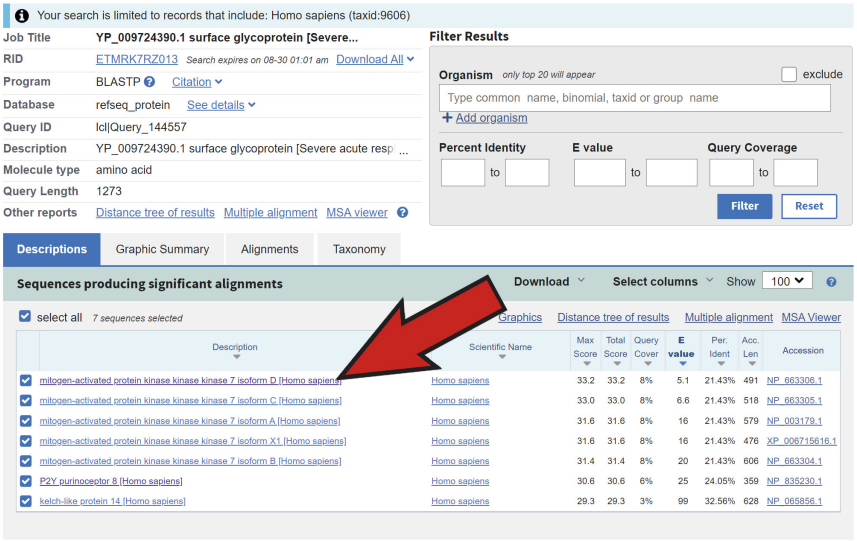
Score	Expect	Method	Identities	Positives	Gaps	Frame
30.6 bits(96)	28	Compositional matrix adjust.	20/85(24%)	37/85(43%)	10/85(11%)	+1

```
Query  241  NPVLPFNDGVVFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPF  420
            +VLPF+ +V+              W+FG  L +       + N  ++++   C   + F
Sbjct  73   ASVLPFQ--IYVHCNRHH-----WVFGVLLCNVVTVAFYANMYSSILTMTC---ISVERF  122

Query  421  LGVVYHKNNKSWMESEFRVYSSANN  495
            LGV Y  ++K W     + V + A
Sbjct  123  LGVLYPLSSKRWRRRRYAVAACAGT  147
```

Now go back and get the protein sequence of this gene and do it again with blastp. What is the top hit and *E*-value now? Draw a conclusion - in other words, should we worry when targeting the spike protein about off target side-effects due to homology to human proteins?

ANSWER: The top hit now is "mitogen-activated protein kinase kinase kinase 7 isoform D".. The *E*-value is 5.1. We conclude that there are (probably) no human proteins that are similar enough to the spike protein to worry about off-target effects.



Your search is limited to records that include: Homo sapiens (taxid:9606)

Job Title	YP_009724390.1 surface glycoprotein [Severe...
RID	ETMRK7RZ013 *Search expires on 08-30 01:01 am* Download All
Program	BLASTP Citation
Database	refseq_protein See details
Query ID	lcl|Query_144557
Description	YP_009724390.1 surface glycoprotein [Severe acute resp ...
Molecule type	amino acid
Query Length	1273
Other reports	Distance tree of results Multiple alignment MSA viewer

Filter Results

Organism *only top 20 will appear* exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity to E value to Query Coverage to

Filter Reset

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Select columns Show 100

select all *7 sequences selected* Graphics Distance tree of results Multiple alignment MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
mitogen-activated protein kinase kinase kinase 7 isoform D [Homo sapiens]	Homo sapiens	33.2	33.2	8%	5.1	21.43%	491	NP_663306.1
mitogen-activated protein kinase kinase kinase 7 isoform C [Homo sapiens]	Homo sapiens	33.0	33.0	8%	6.6	21.43%	518	NP_663305.1
mitogen-activated protein kinase kinase kinase 7 isoform A [Homo sapiens]	Homo sapiens	31.6	31.6	8%	16	21.43%	579	NP_003179.1
mitogen-activated protein kinase kinase kinase 7 isoform X1 [Homo sapiens]	Homo sapiens	31.6	31.6	8%	16	21.43%	476	XP_006715616.1
mitogen-activated protein kinase kinase kinase 7 isoform B [Homo sapiens]	Homo sapiens	31.4	31.4	8%	20	21.43%	606	NP_663304.1
P2Y purinoceptor 8 [Homo sapiens]	Homo sapiens	30.6	30.6	6%	25	24.05%	359	NP_835230.1
kelch-like protein 14 [Homo sapiens]	Homo sapiens	29.3	29.3	3%	99	32.56%	628	NP_065856.1