Assignment 13 - Turkey edition

Your name:
Your email address: **Quiz:**

Why is the E-value not a good measure for false positives in a PSI-blast search?

Assuming you did 6 iterations in a PSI blast search with an e-value cut-off of 0.005. Why should you be skeptical of a match reported in the 6th iteration even though that match has a reported E-value of .0000001?

Assignments: (Answer questions in red)

1. Do a PSI-BLAST (use **UniProtKBSwiss-Prot** as database and restrict the taxon to only search Archaea (taxid:2157), search for at least 3 iterations, an E-value cut-off for inclusion in the next round ("PSI-BLAST Threshold") to 0.005 with the following sequence (this is the intein in the archaeal ATPsynthase catalytic subunit from *Pyrococcus abysii*). Select 5000 aximum target sequencesSelect "show results in a new window".

```
>Intein from archaeal ATPsynthase catalytic subunit from Pyrococcus abysii
CVDGDTLVLTKEFGLIKIKDLYKILDGKGKKTVNGNEEWTELERPITLYGYKDGKIVEIKATHVYKGFS
AGMIEIRTRTGRKIKVTPIHKLFTGRVTKNGLEIREVMAKDLKKGDRIIVAKKIDGGERVKLNIRVEQKR
GKKIRIPDVLDEKLAEFLGYLIADGTLKPRTVAIYNNDESLLRRANELANELFNIEGKIVKGRTVKALLI
HSKALVEFFSKLGVPRNKKARTWKVPKELLISEPEVVKAFIKAYIMCDGYYDENKGEIEIVTASEEAAYG
FSYLLAKLGIYAIIREKIIGDKVYYRVVISGESNLEKLGIERVGRGYTSYDIVPVEVEELYNALGRPYAE
LKRAGIEIHNYLSGENMSYEMFRKFAKFVGMEEIAENHLTHVLFDEIVEIRYISEGQEVYDVTTETHNFI
GGNMPTLLHN
```

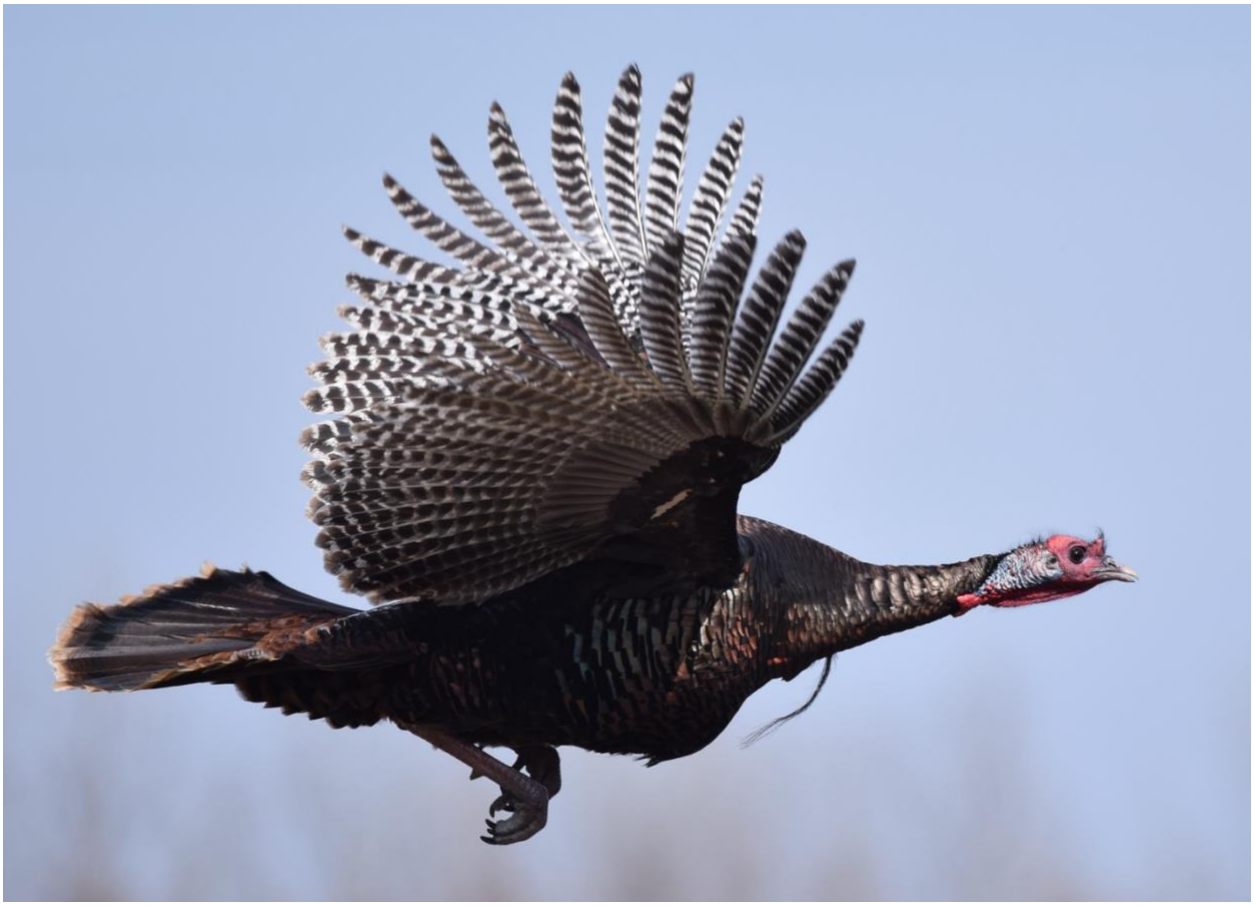
What types of enzymes do you get as hits?
1st iteration:
2nd iteration:
3rd iteration:
4th iteration:

How could you verify that the target proteins contain inteins?

Do a blast based approach to verify presence of an intein in one of the target sequences.
Which target sequence did you choose?

Does it contain an intein?

What is the percent identity of the least significant hit added in the last iteration (clicking on the score in the table will jump to the alignment)?



Wild turkey in flight from
https://upload.wikimedia.org/wikipedia/commons/c/c1/Wild_Turkey_%2825473218982%29.jpg

4. Using PSSMs

SKIP THE COMMANDs IN BLUE!!!!!

Execute the ones in green!

lab13_2024.zip contains all the files you need for this part of the exercise.
rt_gallus.faa is an amino acid sequence of a reverse transcriptase domain from an endogenous retrovirus from chicken.
The sequence is from a polyprotein that also contains capsid proteins (accession number AGL81188.1).
I used rt_gallus.faa to build a PSSM matrix: rt_gallus.pssm. Using the web interface for blastp. I used the refseq database and restricted the search to Avis. After 3 iterations, most of the entries targeted reverse transcriptases.

Alternatively, one could use

```
psiblast -db /isg/shared/databases/uniprot/uniref50/uniref50 -seg no -out
```

```
out.rt -query rt_gallus.faa -out_pssm rt_gallus.pssm -out_ascii_pssm
HE_query.asci.pssm -inclusion_ethresh 0.005 -outfmt 6 -num_iterations 5 -
num_threads 2 -save_pssm_after_last_round -max_target_seqs 5000000
```

We will use the PSSM and the original protein query for four searches using rt_gallus.faa as query:

1. a blastp search of proteins annotated in as belonging to Turkey *Meleagris gallopavo* I downloaded a draft genome from https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=9102, (Turkey_5.1) called **Turkey.faa**
2. a tblastn of the contigs from the 5th draft genomes **Turkey.fna** for significant matches to the sequence in the FASTA file rt_gallus.faa
3. a PSI-blastp search of the proteins described as encoded in the turkey genome (Turkey.faa database) (using the pre-calculated PSSM rt_gallus.pssm, and
4. a PSI tblastn search of the contigs from the turkey genome for significant matches to this PSSM

The databanks are at

1. /home/FCAM/mcb3421/usr40/lab13/Turkey.fna
2. /home/FCAM/mcb3421/usr40/lab13/Turkey.faa

These searchable databanks were created using the commands

```
module load blast/2.13.0
makeblastdb -in Turkey.faa -dbtype prot -parse_seqids
makeblastdb -in Turkey.fna -dbtype nucl -parse_seqids
```

To execute the 4 searches, we will use the xanadu cluster. Establish a terminal ssh connection to the cluster at via **xanadu-submit-ext.cam.uchc.edu** (same username mcb3421usr## and password as previously).

Skip this: Establish as filezilla sftp session to **transfer.cam.uchc.edu** (same username and password as above).
Make a lab13 directory in your account using filezilla. Transfer rt_gallus.faa and rt_gallus.pssm in the lab13 directory.

```
srun --pty --partition=mcbstudent --qos=mcbstudent --mem=2G bash
```
(we're about to do some serious computation, so onto a compute node we go...)
```
curl -O https://j.p.gogarten.uconn.edu/mcb3421_2024/labs/lab13_2024.zip
```
(get the files)
```
unzip lab13_2024.zip
```
(unpack the archive, this also creates the lab13_2024 directory)
```
cd lab13_2024
```
(change into the lab13 directory)
```
ls
```

(make sure the file is there)

```
cat rt_gallus.faa
```

(...and that the contents look OK)

```
module load blast/2.13.0
```

(load the blast module)

```
ls
```

(once more to make sure the files are present in this folder)

The creation of the PSSM takes a longer time, thus the file is provided. The script to execute the command is also in the zip archive. You could submit it to the queue using

```
sbatch shell_for_makingPSSMs_for_TurkeyRT.sh
```

Options for psiblast can be seen using

```
psiblast -help
```

or look at psiblast-help.txt

To do a normal blastp search:

```
blastp -query rt_gallus.faa -db //home/FCAM/mcb3421/usr40/lab13/Turkey.faa -out blastp.out -
outfmt 6 -evalue 1e-5 -seg no -max_target_seqs 2000
```

To do a PSI-blast search of the encoded proteins:

```
psiblast -db /home/FCAM/mcb3421/usr40/lab13/Turkey.faa -in_pssm rt_gallus.pssm -out PSIblastP.out
-inclusion_ethresh 1e-8 -evalue 1e-5 -outfmt 6 -seg no -num_threads 2 -max_target_seqs 2000
```

You will receive a fancy warning message, but it seems to works just fine. More on the warning message is here.

To do a normal tblastn search:

```
tblastn -db /home/FCAM/mcb3421/usr40/lab13/Turkey.fna -query rt_gallus.faa -out tblastn.out -
evalue 1e-5 -outfmt 6 -num_threads 2 -seg no -max_target_seqs 2000
```

To do a PSI-blast search of the 6 reading frames of the genome:

To do a PSI-blast search of the 6 reading frames of the genome:

```
tblastn -db /home/FCAM/mcb3421/usr40/lab13/Turkey.fna -in_pssm rt_gallus.pssm -out psitblastn.out
-evalue 1e-5 -outfmt 6 -seg no -num_threads 2 -max_target_seqs 2000
```

Check the contents of the out-files from the different blast searches. Counting the number of lines (corresponding to the number of significant matches; note, we had set the e-value to 10^{-5}, and selected tabular output) in a file we can use the unix word count (wc) command :

```
wc -l blastp.out
wc -l tblastn.out
wc -l PSIblastP.out
wc -l psitblastn.out
```

or

```
wc -l psitblastn.out PSIblastP.out tblastn.out blastp.out
```

To figure out how matching sequences were annotated, you can use the blastdbcmd. For example:

```
blastdbcmd -db /home/FCAM/mcb3421/usr40/lab13/Turkey.faa -entry
XP_031410439.1,XP_019468902.1,XP_010706347.1 -outfmt "%l %a   %t"
```

Note: the IDs in the search string are only separated by "," no space
The outfmt string specifies length of the sequence, accession number, and title.
To get more help on the blastdbcmd use

```
blastdbcmd -help
```

To get more information on the annotated proteins check entrez. The first match is a very long protein with many domains, and the rt is not clearly annotated, but the 2nd and 3rd match clearly are parts of a transposon. To get more information, you could paste the sequences into HHPred

How does the number of blastp matches compare to the number of tblastn, PSI-blast, and PSI-tblastn matches?

If there is a significant difference in the number of matches, can you think of a reason why this could happen?

Do the annotated proteins have an annotation that suggests the presence of the reverse transcriptase domain?

If you want to judge the quality of the matches, rerun the tblastn searches **without** the tabular option, you then get a long list of pairwise alignments of the HSPs (this follows the listing of HSPs). This also is a good way to easily get protein sequences that match the query sequence in a tblastn search.

```
tblastn -db /home/FCAM/mcb3421/usr40/lab13/Turkey.fna -query rt_gallus.faa -out tblastn.out -
evalue 1e-5 -num_threads 2 -seg no -max_target_seqs 2000
```

and

```
tblastn -db /home/FCAM/mcb3421/usr40/lab13/Turkey.fna -in_pssm rt_gallus.pssm -out psitblastn.out
-evalue 1e-5 -seg no -num_threads 2 -max_target_seqs 2000
```

Note that the listing of alignments is not in overall order of increasing e-values. The matches to the same chromosome or contig are listed together (same as in the -outfmt -6 option)

Finished?

Do not forget to email your competed worksheet to gogarten@uconn.edu and daniel.s.phillips@uconn.edu