Assignment 12 – PSI Blast - Turkey edition

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

What do the abbreviations PSI (as in PSI blast) and PSSM stand for?

PSSM: PSI:

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

Assuming you did 10 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 10th iteration even though that match has a reported E-value of .000000001?

Assignments: (Answer questions in green)

1. Using PRSS determine if significant similarity exists between proteins with the following accession numbers AAA23671.1 (D-Ala D-Ala ligase) and P04425.1 (Glutathione synthetase). Copy the fasta formated sequences into the PRSS window. Select 1000 shuffles, then click on "Shuffle Sequence" to start the comparison.

What is the E-Value (10000) of the comparison?

Collaborate with others in the class. Some should do exercise #2, the other exercise #3. Share the results!

2. (10 minutes) Do a PSI-BLAST search (this is the same submission page as for blastp, but you check off the PSI blast radio button) with the

- Glutathione synthetase (P04425.1) as a query,
- use **UniProtKBSwiss-Prot** ,
- restrict the taxon to only search Bacteria (bacteria (taxid:2)),
- select the PAM250 substitutions matrix,
- select 20000 Max Target Sequences).
- Under "Filters and Masking", turn on the filter for low complexity (tick/check the box),
- set the E-value cut-off for inclusion in the next round ("PSI-BLAST Threshold") to 0.001 (or 1e-3) and
- change the maximum target sequences to 20000.

Note: Depending on the settings, PSI-Blast switches back and forth between the format and the result window. Place a check in the box "Show results in a new window" (at the bottom of the page).

Note that the iteration number is listed at the top of the result page. One "launches" the next PSI-Blast iteration by looking for the "Run PSI-Blast iteration ? with max" at the top (or the bottom) of the list of significant matches, and clicking "Go".

After how many iterations (do not more than 5 iterations!) do you start to pick up D-Ala D-Ala ligase, biotin carboxylase (aka Acetyl-CoA carboxylase A1) and Carbamoyl-phosphate synthase ?

Which other types of enzyme are included among the hits?

Note: Collaborate with your neighbor. One of you should do task #2, the other #3.

3. (10 minutes) Do a PSI-BLAST search (this is the same submission page as for blastp, but you check off the PSI blast radio button) with the

- D-Ala D-Ala ligase AAA23671.1 as a query,
- use **UniProtKBSwiss-Prot** as database,
- restrict the taxon to only search Bacteria (bacteria (taxid:2)),
- select the PAM250 substitutions matrix),
- select 20000 Max Target Sequences.
- Under "Filters and Masking", turn on the filter for low complexity (tick/check the box),
- set the E-value cut-off for inclusion in the next round ("PSI-BLAST Threshold") to 0.001 (or 1e-3), and
- change the maximum target sequences to 20000.

Note : Depending on the settings, PSI-Blast switches back and forth between the format and the result window. Place a check in the box Show results in a new window (at the bottom of the page).

Note that the iteration number is listed at the top of the result page. One "launches" the next PSI-Blast iteration by looking for the "Run PSI-Blast iteration ? with max" at the top (or the bottom) of the list of significant matches, and clicking "Go".

After how many iterations (do not more than 5 iterations!) do you start to pick up carbamoyl phosphate synthetases, glutathione synthetase, and biotin carboxylase (aka Acetyl-CoA carboxylase A1)?

Which other types of enzyme are included among the hits?

What might be the reason for the different results obtained in tasks 2 and 3?

4. Using PSSMs

SKIP THE COMMANDs IN BLUE!!!!!

Execute the ones in red-brown!

lab12_2023.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 three searches:

1. a blastp search of proteins annotated in as belonging to Turkey *Meleagris gallopavo* We have two databanks, one from the 5 iteration of a draft genome **Turkey.faa**, the other from the taxonomy browser downloading all proteins associated with Turkey **Turkey2.faa**/li>
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 enconded 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/usr30/lab12/Turkey.fna
2. /home/FCAM/mcb3421/usr30/lab12/Turkey.faa
3. /home/FCAM/mcb3421/usr30/lab12/Turkey2.faa

These searchable databanks were created using the commands

```
module load blast/2.13.0
makeblastdb -in Turkey_protein.faa -dbtype prot -parse_seqids
makeblastdb -in Turkey_genomic.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 lab12 directory in your account using filezilla. Transfer rt_gallus.faa and rt_gallus.pssm in the lab12 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_2023/labs/lab12_2023.zip
```
(get the files)
```
unzip lab12_2023.zip
```
(unpack the archive)
```
cd lab12
```
(change into the lab12 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/usr39/TurkeyDBs/Turkey_protein.faa -out
blastp.out -outfmt 6 -evalue 1e-8 -seg no
```

To do a PSI-blast search of the encoded proteins:
```
psiblast -db /home/FCAM/mcb3421/usr39/TurkeyDBs/Turkey_protein.faa -in_pssm rt_gallusPSSM.asn -
out PSIblastP.out -inclusion_ethresh 1e-8 -evalue 1e-8 -outfmt 6 -seg no -num_threads 2
```

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/usr39/TurkeyDBs/Turkey_genomic.fna -query rt_gallus.faa -out
tblastn.out -evalue 1e-8 -outfmt 6 -num_threads 2 -seg no
```

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

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/usr39/TurkeyDBs/Turkey_genomic.fna -in_pssm rt_gallusPSSM.asn -out
psitblastn.out -evalue 1e-8 -outfmt 6 -seg no -num_threads 2
```

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^{-3}, and selected tabular output) in a file:

```
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/usr39/TurkeyDBs/Turkey_protein.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.

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?

5. Identifying inteins.

(30 Minutes) Do a PSI-BLAST search (this is the same submission page as for blastp, but you check off the PSI blast radio button) with the following intein sequence listed below. Use **RefSeq Select Proteins** as database and restrict the taxon to only search archaea (archaea (taxid:2157))) search for 5 iterations, a max target sequences of 5000, an E-value cut-off for inclusion in the next round ("PSI-BLAST Threshold") of 0.0001, and a wordsize of 3 with the following sequence (if this takes too long, use this PSSM linked below):

DO NOT use the gi number as query. The gi number refers to the whole protein, we want to use the intein sequence only!

```
>Pab_VMA intein from gi|7436316|pir||D75028
CVDGDTLVLTKEFGLIKIKDLYKILDGKGKKTVNGNEEWTELERPITLYGYKDGKIVEIKATHVYKGFS
AGMIEIRTRTGRKIKVTPIHKLFTGRVTKNGLEIREVMAKDLKKGDRIIVAKKIDGGERVKLNIRVEQKR
GKKIRIPDVLDEKLAEFLGYLIADGTLKPRTVAIYNNDESLLRRANELANELFNIEGKIVKGRTVKALLI
HSKALVEFFSKLGVPRNKKARTWKVPKELLISEPEVVKAFIKAYIMCDGYYDENKGEIEIVTASEEAAYG
```

FSYLLAKLGIYAIIREKIIGDKVYYRVVISGESNLEKLGIERVGRGYTSYDIVPVEVEELYNALGRPYAE
LKRAGIEIHNYLSGENMSYEMFRKFAKFVGMEEIAENHLTHVLFDEIVEIRYISEGQEVYDVTTETHNFIGG
NMPTLLHNT

DO NOT use the gi number as query. The gi number refers to the whole protein, we want to use the intein sequence only!

What types of enzymes do you get as hits?
How could you verify that the target proteins contain inteins?
Which E-value cut-off for inclusion in the next round did you choose?
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)?

Save the PSSM (Position Specific Scoring Matrix, or profile) from this search. To do that choose PSSM from pull-down menu (click on the arrow) inside the download link close to the top of the result page. Save the PSSM as an ASN file on your computer.
(**If the iterations take too long**, the PSSM after 4 iterations is here

These PSSMs should work in detecting a wide variety of intein containing proteins. We will use them to search the genomes of organisms of interest. Go to PSI-BLAST. Select BlastP (on top of the page). [You can skip this with the new blast webpage: Paste intein sequence into query sequence box.] Select the non-redundant database.
Select the organisms to which the search should be restricted. You can select individual organisms or whole groups. (if you start typing, options will appear form which to select taxa)

Possible are

- Pyrobaculum aerophilum
- Aeropyrum pernix
- Sulfolobus tokodaii
- Archaeoglobus fulgidus
- Methanothermobacter thermautotrophicus
- Thermoplasma volcanium
- Methanocaldococcus
- Methanococcales
- **Halobacteria**
- Halorubrum
- Haloferax

After selecting the genomes to search, go to Algorithm parameters and under PSI-blast options select and upload your PSSM (the one you saved yourself, or the ones that are linked above; if you have time, use the different PSSMs to see if this makes any difference).

As you start your search with a PSSM matrix already, **you do not need to (and should not) do iterations**!

Which PSSM did you use?

Which genome did you search?

What are the results of your search? Did you get any significant matches? How many?

What are they?

If you have significant matches, does the match occur over the full lengths of both query and subject sequences?

What is the percent identity?

Use a hit in the databank as a query in a fast blast search to investigate if the hit indeed harbors an inteins.

What is your conclusion?

Finished?

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