Exercise 1

New DB name: /home/ahmed/bioinformatics/test_data/uniprot_Atha.fasta New DB title: uniprot_Atha.fasta Sequence type: Protein Deleted existing Protein BLAST database named /home/ahmed/bioinformatics/test_data/uniprot_A Keep MBits: T Maximum file size: 3000000000B Adding sequences from FASTA; added 15719 sequences in 0.222235 seconds.

Therefore we have 15719 formatted sequences. When the number of sequences increases, the database size increases This increases BLAST E-values because the probability of obtaining that score by chance increases as its a bigger database. This means hits appear less statistically significant in larger databases than in smaller ones. source: https://sequenceserver.com/blog/blast-e-value-meaning/ https://www.metagenomics.wiki/tools/blast/evalue

Exercise 2

446 blastp -db uniprot_Atha.fasta -query test.faa -outfmt 6 > test.faa.blast 448 blastx -db uniprot_Atha.fasta -query test.fna -outfmt 6 > test.fna.blast

Exercise 3

According to the BLAST Help user manual, the default output format is pairwise (-outfmt 0) source: https://blast.ncbi.nlm.nih.gov/doc/blast-topics/resultformatoptions.html

Example from the blastx output: >sp|Q9ZTX8|ARFF_ARATH Auxin response factor 6 OS=Arabidopsis thaliana OX=3702 GN=ARF6 PE=1 SV=2 Length=935

Score = 1706 bits (4419), Expect = 0.0, Method: Compositional matrix adjust. Identities = 935/935 (100%), Positives = 935/935 (100%), Gaps = 0/935 (0%) Frame = +1

Query 1 MRLSSAGFNPQPHEVTGEKRVLNSELWHACAGPLVSLPPVGSRVVYFPQGHSEQVAASTN 180 MRLSSAGFNPQPHEVTGEKRVLNSELWHACAGPLVSLPPVGSRVVYFPQGHSEQVAASTN Sbjct 1 MRLSSAGFNPQPHEVTGEKRVLNSELWHACAGPLVSLPPVGSRVVYFPQGHSEQVAASTN 60

Query 181 KEVDAHIPNYPSLHPQLICQLHNVTMHADVETDEVYAQMTLQPLNAQEQKDPYL-PAELGV 360 KEVDAHIPNYPSLHPQLICQLHNVTMHADVETDEVYAQMTLQPLNAQEQKDPYL-PAELGV Sbjct 61 KEVDAHIPNYPSLHPQLICQLHNVTMHADVETDEVYAQMTLQPLNAQEQKD-PYLPAELGV 120

Query 361 PSRQPTNYFCKTLTASDTSTHGGFSVPRRAAEKVFPPLDYSQQPPAQELMARDLHD-NEWK 540 PSRQPTNYFCKTLTASDTSTHGGFSVPRRAAEKVFPPLDYSQQPPAQELMARDLHD-NEWK Sbjct 121 PSRQPTNYFCKTLTASDTSTHGGFSVPRRAAEKVFPPLDYSQQPPAQELMARDL-HDNEWK 180

Query 541 FRHIFRGQPKRHLLTTGWSVFVSAKRLVAGDSVLFIWNDKNQLLLGIRRANRPQTVMPSS 720 FRHIFRGQPKRHLLTTGWSVFVSAKRLVAGDSVLFIWNDKNQLLLGIRRANRPQTVMPSS Sbjct 181 FRHIFRGQPKRHLLTTGWSVFVSAKRLVAGDSVLFIWNDKNQLLLGIRRANRPQTVMPSS 240

Query 721 VLSSDSMHLGLLaaaahaaaTNSRFTIFYNPRASPSEFVIPLAKYVKAVYHTRVSVGMRF 900 VLSSDSMHLGLLAAAAHAAATNSRFTIFYNPRASPSEFVIPLAKYVKAVYHTRVSVGMRF Sbjct 241 VLSSDSMHLGLLAAAAHAAATNSRFTIFYNPRASPSEFVIPLAKYVKAVYHTRVSVGMRF 300

Exercise 4

The difference betweent the blastx and the blastp search results is that one shows the protein sequence hits themselves and therefore is a direct comparison, while the blastx searches using the transcript infromation therefore translation is needed and it can be slower. Both blasts had the Q9ZTX8 as their top

hit with 100% valid identity. blastp had a bitscore of 1915 and the blastx had a bitscore of 1706, slightly lower. Both had strong E-scores Source: https://www.nlm.nih.gov/ncbi/workshops/2023-08_BLAST_evol/e_value.html https://card.mcmaster.ca/ontology/40725

Exercise 5

the profile.out file is a position-specific scoring matrix (PSSM), which shows scores for specific amino acid positions and their likely substitutions in our query sequence (ARF6). They are generated from many sequence alignments in this case 3. They are used to detect more protein alignments beyond BLAST as instead of making the assumption that each position in the sequence is as likely as the other to change, it scores based on the positions themselves, better mirroring biological reality. source: https://www.nlm.nih.gov/ncbi/workshops/2023-08_BLAST_evol/blast_score.html

Exercise 6

Ran

```
awk '$12 > 200 {print $2}' test.faa.blast > high_score_ids.txt
while read id; do grep -A 1 "^>.*$id" uniprot_Atha.fasta; done < high_score_ids.txt > high_score_seqs.fa
clustalo -i high_score_seqs.faa -o high_score_alignment.aln --outfmt=clustal
hmmbuild ARF6.hmm high_score_alignment.aln
hmmpress ARF6.hmm
hmmscan --tblout ARF6_hmmscan.tbl ARF6.hmm uniprot_Atha.fasta > ARF6_hmmscan.out
```

We were searching for accessions from BLAST hits with high bit scores, then taking the complete protein sequences of those matches from the uniprot_Atha.fasta file. Clustal Omega performed a multiple sequence alignment of these sequences, the results from the hmmscan showed hits from homologous accessions in the Auxin Response Factor (ARF) gene family of Arabidopsis Thaliana. This is confirmed by the low E-scores and similar number of hits for both the BlastP hits and the HMM hits. The proteins of this family are very similar as many have overlapping functions. This HMM validate our blastp and blastx query results.

source: https://www.uniprot.org/citations/15659631

The table of the HMMscan hits:

```
(blastenv) ahmed@AHMEDPC:~/bioinformatics/test_data$ head -20 ARF6_hmmscan.tbl
#                                                                            --- full sequence ---- --- best 1 domain ---- --- domain number estimation ----
# target name        accession  query name           accession    E-value  score  bias   E-value  score  bias   exp reg clu  ov env dom rep inc description
 of target
#------------------- ----------  -------------------- ---------- --------- ------ ----- --------- ------ -----   --- --- --- --- --- --- --- --- -----------
----------
high_score_alignment -          sp|Q9LQE8|ARFN_ARATH -            4.7e-34  102.7   0.0   9.9e-34  101.6   0.0   1.5   1   0   0   1   1   1   1 -
high_score_alignment -          sp|Q94JM3|ARFB_ARATH -            1.8e-27   81.5   0.0   4.5e-27   80.3   0.0   1.7   1   0   0   1   1   1   1 -
high_score_alignment -          sp|Q9ZTX9|ARFD_ARATH -              1e-23   69.5   0.0   2.4e-23   68.3   0.0   1.7   1   0   0   1   1   1   1 -
high_score_alignment -          sp|P93022|ARFG_ARATH -            5.8e-27   79.9   0.1   5.8e-27   79.9   0.1   2.5   2   0   0   2   2   2   1 -
high_score_alignment -          sp|Q8L7G0|ARFA_ARATH -            2.2e-27   81.3   0.1   4.9e-27   80.1   0.1   1.7   1   0   0   1   1   1   1 -
high_score_alignment -          sp|Q9SKN5|ARFJ_ARATH -            8.4e-29   85.8   0.1   1.7e-28   84.8   0.1   1.5   1   0   0   1   1   1   1 -
high_score_alignment -          sp|Q84WU6|ARFQ_ARATH -            9.4e-25   72.8   0.4   2.3e-24   71.6   0.1   1.8   2   0   0   2   2   2   1 -
high_score_alignment -          sp|Q8RYC8|ARFS_ARATH -            5.2e-27   80.1   0.1   5.2e-27   80.1   0.1   2.6   2   0   0   2   2   2   1 -
high_score_alignment -          sp|O23661|ARFC_ARATH -            2.9e-23   68.0   0.0   5.5e-23   67.1   0.0   1.5   1   0   0   1   1   1   1 -
high_score_alignment -          sp|P93024|ARFE_ARATH -            6.9e-24   70.0   2.2   1.2e-23   69.2   0.0   2.4   2   0   0   2   2   2   1 -
high_score_alignment -          sp|Q9FGV1|ARFH_ARATH -            1.5e-26   78.5   1.2   2.2e-26   78.0   0.1   1.9   2   0   0   2   2   2   1 -
high_score_alignment -          sp|Q9XID4|ARFL_ARATH -            2.3e-33  100.4   0.0   5.8e-33   99.2   0.0   1.7   1   0   0   1   1   1   1 -
high_score_alignment -          sp|Q9LQE3|ARFO_ARATH -            5.6e-33   99.2   0.0   1.5e-32   97.9   0.0   1.7   1   0   0   1   1   1   1 -
high_score_alignment -          sp|Q93YR9|ARFP_ARATH -              4e-30   90.0   0.0   7.9e-30   89.1   0.0   1.5   1   0   0   1   1   1   1 -
high_score_alignment -          sp|Q9C8N9|ARFU_ARATH -            9.4e-33   98.5   0.0   2.1e-32   97.4   0.0   1.6   1   0   0   1   1   1   1 -
high_score_alignment -          sp|Q9LP07|ARFW_ARATH -            4.5e-34  102.7   0.1     8e-34  101.9   0.1   1.4   1   0   0   1   1   1   1 -
high_score_alignment -          sp|Q9XED8|ARFI_ARATH -            7.7e-30   89.1   0.0   2.1e-29   87.7   0.0   1.8   1   0   0   1   1   1   1 -
(blastenv) ahmed@AHMEDPC:~/bioinformatics/test_data$ # Count BLAST hits with bit score > 200
awk '$12 > 200' test.faa.blast | wc -l

22
(blastenv) ahmed@AHMEDPC:~/bioinformatics/test_data$

# Count HMM hits
grep -v "^#" ARF6_hmmscan.tbl | wc -l
23
```

Exercise 7

Run

```
grep -A 1 "ARF6" uniprot_Atha.fasta > AT1G30330.faa
```

get

> sp|Q9ZTX8|ARFF_ARATH Auxin response factor 6 OS=Arabidopsis thaliana OX=3702 GN=ARF6 PE=1 SV=2 MRLSSAGFNPQPHEVTGEKRVLNSELWHACAGPLVSLP-PVGSRVVYFPQGHSEQVAASTN

	Hit	Name	Query Boundaries	Template Boundaries	Aligned cols	Probability	E-value	Score
1	4LDU_A	Auxin response factor 5; transcription factor, DNA binding protein, nucleus; 2.15A (Arabidopsis thaliana)	17-59	46-88	43	99.26%	1.2e-11	75.22
2	4LDV_A	Auxin response factor 1; Transcription Factor, DNA binding, Nucleus, TRANSCRIPTION; HET: FMT; 1.45A (Arabidopsis thaliana)	17-58	14-55	42	99.06%	5.9e-10	66.66

	Hit	Name	Query Boundaries	Template Boundaries	Aligned cols	Probability	E-value	Score
3	8OJ2_A	Auxin response factor; Molecular caliper, Auxin Response Factor, Transcription factor, DNA binding, Nucleus, Hormone Response	17-58	19-60	42	99.01%	1.4e-9	65.13

Align	DB:ID	Source	Length	Score(Bits)	Identities(%)	Positives(%)	E-value
1	AFDB:A0A397XXN9	TIR-X1 domain-containing protein UA=A0A397XXN9 UI=A0A397XXN9_BRACM OS=Brassica campestris OX=3711 GN=BRARA_I02823	294	126.3	98.3	100.0	4.9e-34
2	AFDB:D7KET2	Auxin response factor UA=D7KET2 UI=D7KET2_ARALL OS=Arabidopsis lyrata subsp. lyrata OX=81972 GN=ARALYDRAFT_473256	891	127.5	100.0	100.0	1.5e-32
3	AFDB:A0A1J3J0G9	Auxin response factor UA=A0A1J3J0G9 UI=A0A1J3J0G9_NOCCA OS=Noccaea caerulescens OX=107243 GN=MP_TR5073_c0_g1_i1_g.13971	898	127.5	100.0	100.0	1.5e-32

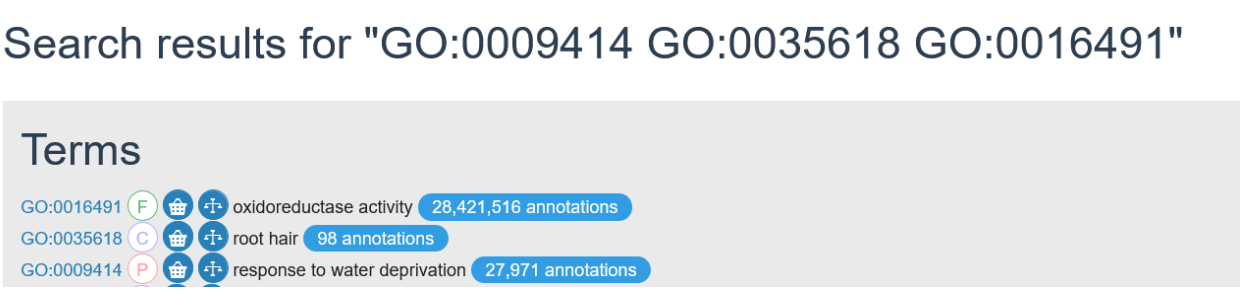
Align	DB:ID	Source	Length	Score(Bits)	Identities(%)	Positives(%)	E-value
4	AFDB:A0A1J3D2P2	Abscisic acid 2 response factor UA=A0A1J3D2P2 UI=A0A1J3D2P2_NOCCA OS=Noccaea caerulescens OX=107243 GN=GA_TR2097_c0_g1_i1_g.6851	899	127.5	100.0	100.0	1.5e-32

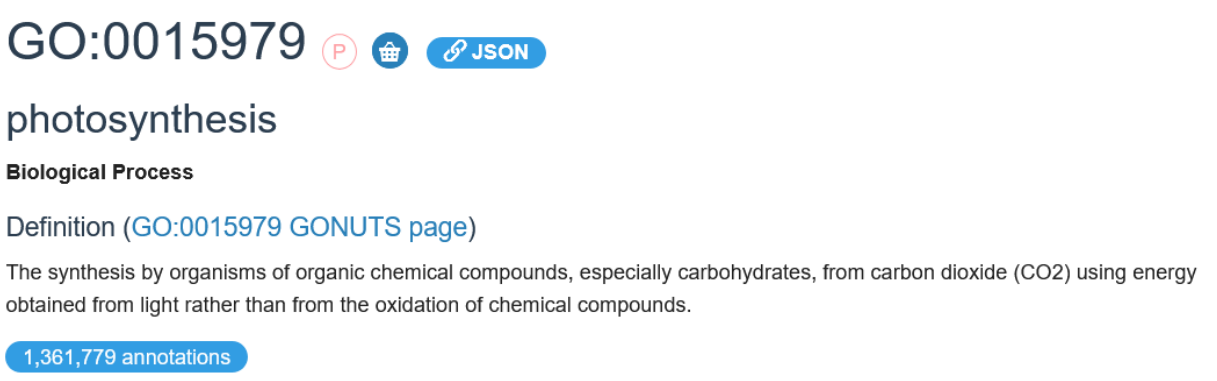
Exercise 8

Unfortunately the website http://eggnog-mapper.embl.de/ was broken. Will try again another day.

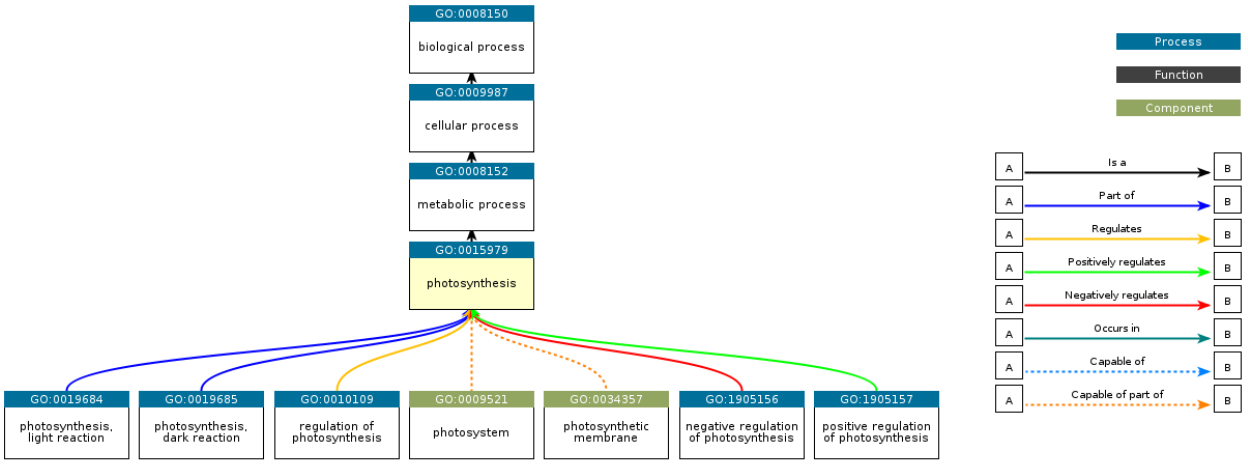
Exercise 9

1)

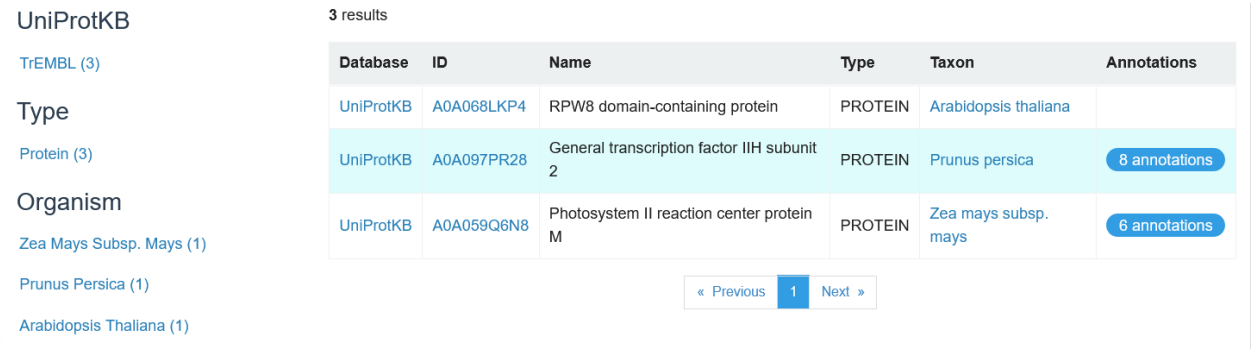




2)



3)



4)

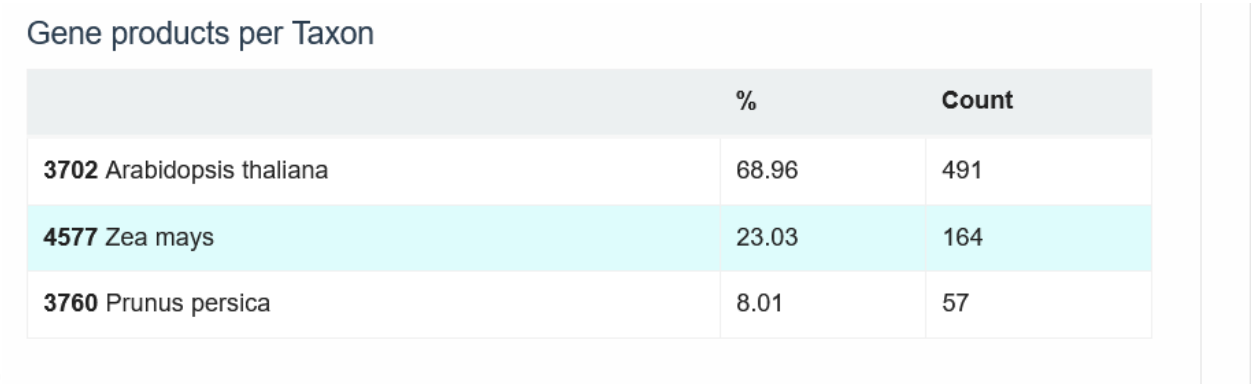
A0A068LKP4 has no GO annotation

A0A097PR28 has 8 GO annotations, GO terms for its molecular functions (MF) involved in ion binding. Biological functions (BF) involved in to DNA repair

A0A059Q6N8 has 6 annotations, BF pertaining to photosynthesis. Other annotations locate it in different cellular components (cytoplasm; chloroplast thylakoid membrane; photosystem II)

5)

GO:0048366 leaf development 21,805 annotations with 21,009 distinct gene products.



6)

491 proteins for *Arabidopsis Thaliana,* 164 for *Zea mays,* 57 for *Prunus persica* are assigned to GO:0048366

7)

After selecting the right criteria in the search for *Arabidopsis Thaliana* 3702 & *Prunus persica* 3760, the total number of Biological Process annotations and proteins supported by the experimental evidence codes (ECO:0000269) was **423** annotations.