Post Lab: Week 4

Human # 1955791§

§ Northfield Academy of Microbial Analysis and Engineering

I. Check for Understanding

1. Top results from blastp and tblastn of toy dataset.

On the ‘blastp’ side, the top hit was “MFS transporter [Sulfolobus acidocaldarius].” It had an e-value of 0.0 and 100 percent identity. This protein is part of the Major Facilitator Superfamily (MFS), a group of transporters that facilitates transport across cytoplasmic or internal membranes. It comes from the organism *Sulfolobus acidocaldarius* which is a thermoacidophilic archaeon. The thermostable restriction enzyme *SuaI* is obtained from this organism. (Could this be the protein that a past student discovered was overly represented in databases due to an error?)

On the ‘tblastn’ side, the top hit appears to be “Sulfolobus acidocaldarius strain DG1, complete genome,” with an e-value of 0.0 and an identity percent of 95%. So we meet the mighty *Sulfolobus acidocaldarius* once again. (It exists as only a single cell, but don’t let that fool you—it is large, it contains multitudes!) This is not a single protein, but rather the whole genome as sequenced by Mao, D. and Grogan, D. for their hitherto unpublished paper “Genome diversification in the archaeon Sulfolobus acidocaldarius.”

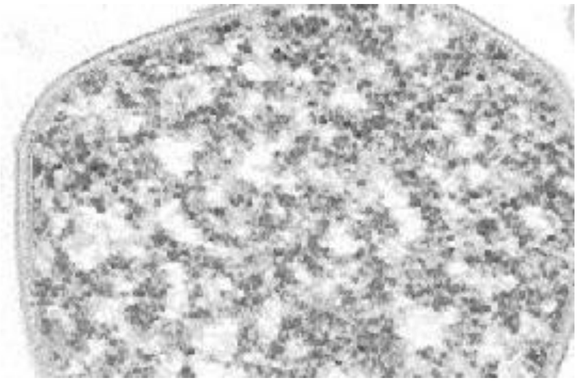


Figure 1. Behold! The mighty Sulfolobus acidocaldarius!

2. Following top results from from blastp and tblastn of toy dataset.

On the ‘blastp’ side of things, one finds a whole lot of MFS transporter genes running all down the page, with miniscule e-values. (Though these small magnitudes may be overly influenced by the fact that we are using a very short amino acid sequence.) There is some variability in the potential organism to which the transporter belongs (it could be *Sulfolobus tokodaii* or *Sulfolobus islandicus* or even *Moorella thermoacetica*!) but the results indicate very strongly that this protein is a MFS transporter protein.

Hopping back into ‘tblastn’ land, we discover many of the top hits to be for complete genomes, mostly for various strains of *Sulfolobus acidocaldarius*. The ‘blastp’ method compares the given AA sequence to a non-redundant protein database, while the ‘blastn’ method compares the given AA sequence to translations of known nucleotide sequences, including entire genomes. It makes sense, then, that the ‘blastp’ would return single proteins, while the ‘tblastn’ could return whole genomes.

3 & 4. Best results from Interproscan. The top results in my interproscan output file, when sorted by e-value using Excel, do not match the BLAST. The top result is now for "arabinose_DH_like (cd05284)", a group of arabinose dehydrogenases (AraDH) and related alcohol dehydrogenases.[1] The second-best result is for "Hydantoinase_B (PF02538)", a family which "includes N-methylhydaintoinase B which converts hydantoin to N-carbamyl-amino acids, and 5-oxoprolinase (P97608) which catalyses the formation of L-glutamate from 5-oxo-L-proline."[2]

5.

a) **Which protein among your Pfam query sequences had the best hit?**

Looking for smallest e-value and largest 'percent identity,' I find that the best hit is a protein with the query sequence name "Q4J793_SULAC/8-222." When I look this up on uniprot.org, I find it to be the organism "‘Sulfolobus acidocaldarius (strain ATCC 33909 / DSM 639)". So *Sulfolobus acidocaldarius* is back baby! The protein name is given simply as "conserved protein," and the gene name is given as "Saci_2043."

b) **What was the percent identity?**

100%!

c) **What organism does the matching Pfam protien query sequence come from?**

Sulfolobus acidocaldarius!

d) **Which of your ORFs did it match?**

I can see "subject start coordinates" is 8 and "subject end coordinates" is 222. I'm not sure how to tell which ORF this matched to.

e) **Does this ORF have hits to other sequences within your query file? What do you think this means?**

Don't know.

6. How do these BLAST results differ from your previous BLAST? Explain why.

The e-values are certainly much larger than for previous blasts, the smallest being 0.54, and the largest being 6.2 (see Appendix).

7. Describe the protein you chose and how you found the sequence for that protein.

I chose the Zur protein, hailing from *Bacillus velezensis*. I found it after perusing a Wikipedia listing of bacterial proteins, clicking "Zinc uptake regulator" and discovering that the zinc uptake regulator (Zur) gene is a bacterial gene that codes for a transcription protein involved in zinc homeostasis. Being a long-time zinc homeostasis enthusiast, I knew this was just the protein for me! I searched "Zur" at www.ncbi.nlm.nih.gov/protein/ and took the top result, which is the form of the protein belonging to the fearsome *Bacillus velezensis*.[3]

8. Show the command you executed for the blast.

```
$ nano ZUR_velezensis.faa

$ blastp -query ZUR_velezensis.faa -db
toy_assembly_ORFs.faa -outfmt 6 -out
ZUR_velezensis_vs_prodigal_ORFs_toy.bla
stp
```

[1] https://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=176187

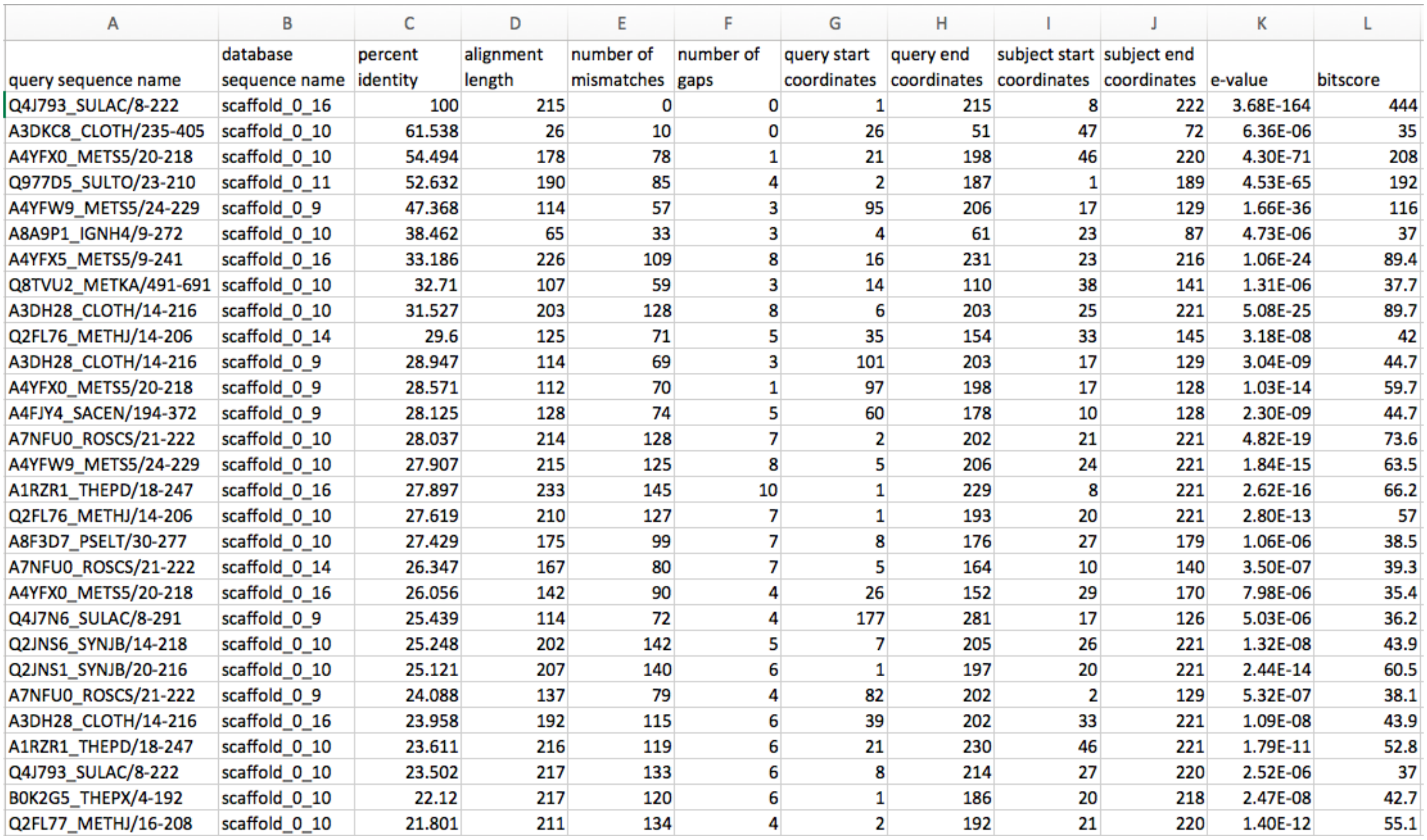
[2] http://pfam.xfam.org/family/PF02538

[3] Make no mistake, contrary to popular belief, Bacillus velezensis is *not* a later heterotypic synonym of Bacillus amyloliquefaciens (Dunlap et al., 2015).

9. Where there any matches? If so, which contig was the best match?

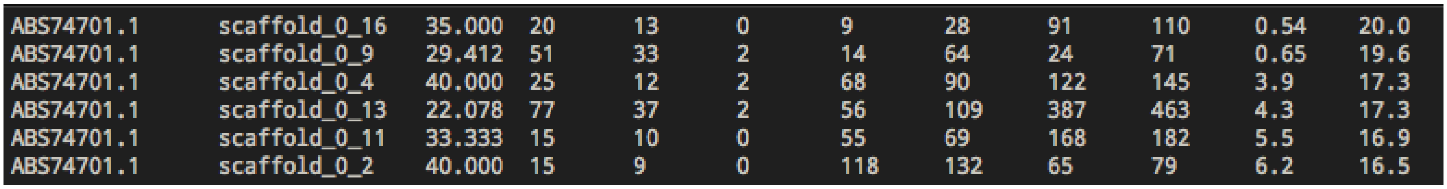
The best match seems to be with "ABS74701.1," which was not found in uniprot.

Appendix



A	B	C	D	E	F	G	H	I	J	K	L
query sequence name	database sequence name	percent identity	alignment length	number of mismatches	number of gaps	query start coordinates	query end coordinates	subject start coordinates	subject end coordinates	e-value	bitscore
Q4J793_SULAC/8-222	scaffold_0_16	100	215	0	0	1	215	8	222	3.68E-164	444
A3DKC8_CLOTH/235-405	scaffold_0_10	61.538	26	10	0	26	51	47	72	6.36E-06	35
A4YFX0_METS5/20-218	scaffold_0_10	54.494	178	78	1	21	198	46	220	4.30E-71	208
Q977D5_SULTO/23-210	scaffold_0_11	52.632	190	85	4	2	187	1	189	4.53E-65	192
A4YFW9_METS5/24-229	scaffold_0_9	47.368	114	57	3	95	206	17	129	1.66E-36	116
A8A9P1_IGNH4/9-272	scaffold_0_10	38.462	65	33	3	4	61	23	87	4.73E-06	37
A4YFX5_METS5/9-241	scaffold_0_16	33.186	226	109	8	16	231	23	216	1.06E-24	89.4
Q8TVU2_METKA/491-691	scaffold_0_10	32.71	107	59	3	14	110	38	141	1.31E-06	37.7
A3DH28_CLOTH/14-216	scaffold_0_10	31.527	203	128	8	6	203	25	221	5.08E-25	89.7
Q2FL76_METHJ/14-206	scaffold_0_14	29.6	125	71	5	35	154	33	145	3.18E-08	42
A3DH28_CLOTH/14-216	scaffold_0_9	28.947	114	69	3	101	203	17	129	3.04E-09	44.7
A4YFX0_METS5/20-218	scaffold_0_9	28.571	112	70	1	97	198	17	128	1.03E-14	59.7
A4FJY4_SACEN/194-372	scaffold_0_9	28.125	128	74	5	60	178	10	128	2.30E-09	44.7
A7NFU0_ROSCS/21-222	scaffold_0_10	28.037	214	128	7	2	202	21	221	4.82E-19	73.6
A4YFW9_METS5/24-229	scaffold_0_10	27.907	215	125	8	5	206	24	221	1.84E-15	63.5
A1RZR1_THEPD/18-247	scaffold_0_16	27.897	233	145	10	1	229	8	221	2.62E-16	66.2
Q2FL76_METHJ/14-206	scaffold_0_10	27.619	210	127	7	1	193	20	221	2.80E-13	57
A8F3D7_PSELT/30-277	scaffold_0_10	27.429	175	99	7	8	176	27	179	1.06E-06	38.5
A7NFU0_ROSCS/21-222	scaffold_0_14	26.347	167	80	7	5	164	10	140	3.50E-07	39.3
A4YFX0_METS5/20-218	scaffold_0_16	26.056	142	90	4	26	152	29	170	7.98E-06	35.4
Q4J7N6_SULAC/8-291	scaffold_0_9	25.439	114	72	4	177	281	17	126	5.03E-06	36.2
Q2JNS6_SYNJB/14-218	scaffold_0_10	25.248	202	142	5	7	205	26	221	1.32E-08	43.9
Q2JNS1_SYNJB/20-216	scaffold_0_10	25.121	207	140	6	1	197	20	221	2.44E-14	60.5
A7NFU0_ROSCS/21-222	scaffold_0_9	24.088	137	79	4	82	202	2	129	5.32E-07	38.1
A3DH28_CLOTH/14-216	scaffold_0_16	23.958	192	115	6	39	202	33	221	1.09E-08	43.9
A1RZR1_THEPD/18-247	scaffold_0_10	23.611	216	119	6	21	230	46	221	1.79E-11	52.8
Q4J793_SULAC/8-222	scaffold_0_10	23.502	217	133	6	8	214	27	220	2.52E-06	37
B0K2G5_THEPX/4-192	scaffold_0_10	22.12	217	120	6	1	186	20	218	2.47E-08	42.7
Q2FL77_METHJ/16-208	scaffold_0_10	21.801	211	134	4	2	192	21	220	1.40E-12	55.1

Figure 2. PF03787_vs_prodigal_ORFs_toy.blastp, in Xcel / with headers



```
ABS74701.1      scaffold_0_16   35.000  20      13      0       9       28      91      110     0.54    20.0
ABS74701.1      scaffold_0_9    29.412  51      33      2       14      64      24      71      0.65    19.6
ABS74701.1      scaffold_0_4    40.000  25      12      2       68      90      122     145     3.9     17.3
ABS74701.1      scaffold_0_13   22.078  77      37      2       56      109     387     463     4.3     17.3
ABS74701.1      scaffold_0_11   33.333  15      10      0       55      69      168     182     5.5     16.9
ABS74701.1      scaffold_0_2    40.000  15      9       0       118     132     65      79      6.2     16.5
```

Figure 3. ZUR_velezensis_vs_prodigal_ORFs_toy.blastp

Works Cited

Dunlap, C.A., Kim, S.-J., Kwon, S.-W., and Rooney, A.P. (2015). Bacillus velezensis is not a later heterotypic synonym of Bacillus amyloliquefaciens; Bacillus methylotrophicus, Bacillus amyloliquefaciens subsp plantarum and "Bacillus oryzicola" are later heterotypic synonyms of Bacillus velezensis based on phylogenomics. Int. J. Syst. Evol. Microbiol.