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 This is not a single protein, but rather the whole genome as sequenced by Mao, D. and Grogan, D. for their hitherto unpublished paper “Genome diversification in the archaeon Sulfolobus acidocaldarius.”

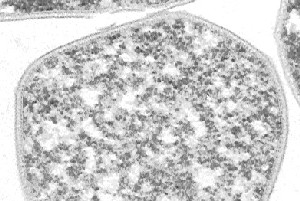


Figure 1. Behold! The mighty Sulfolobus acidocaldarius!

**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]](#footnote-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]](#footnote-2)

**5.**

1. **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.”

1. **What was the percent identity?**

100%!

1. **What organism does the matching Pfam protien query sequence come from?**

*Sulfolobus acidocaldarius!*

1. **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.

1. **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/](http://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]](#footnote-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.blastp

**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

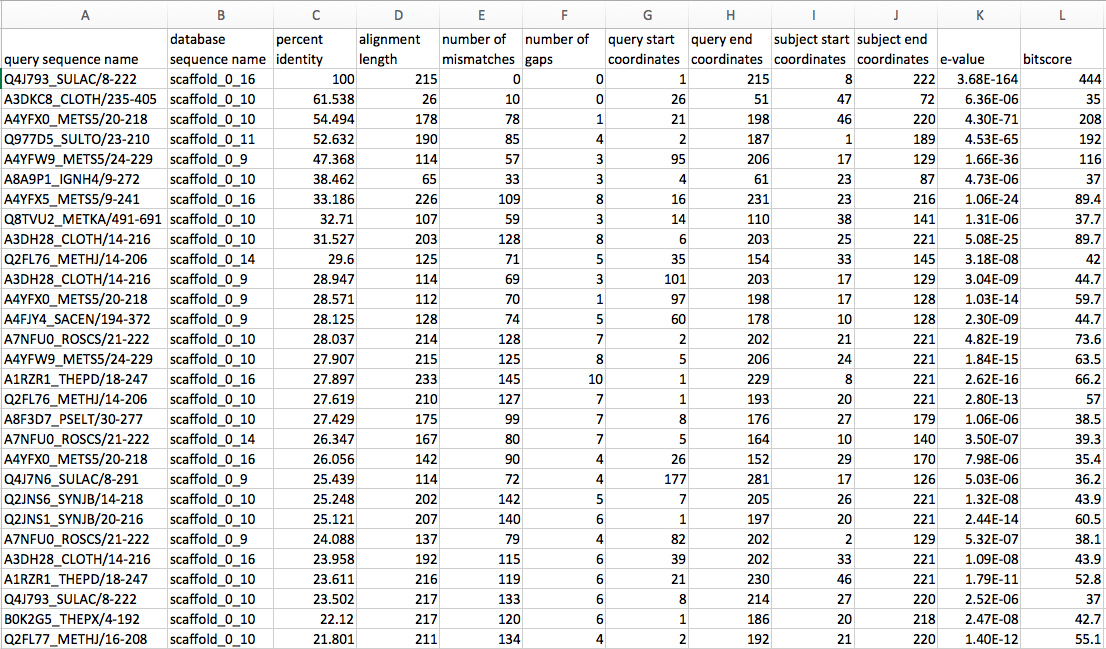
****

Figure 2. PF03787\_vs\_prodigal\_ORFs\_toy.blastp, in Xcel / with headers

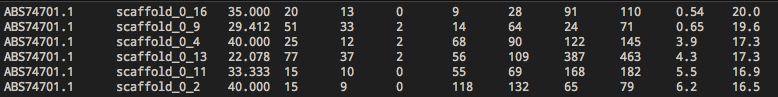


Figure 3. ZUR\_velezensis\_vs\_prodigal\_ORFs\_toy.blastp

# 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.

1. <https://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=176187> [↑](#footnote-ref-1)
2. <http://pfam.xfam.org/family/PF02538> [↑](#footnote-ref-2)
3. Make no mistake, contrary to popular belief, Bacillus velezensis is *not* a later heterotypic synonym of Bacillus amyloliquefaciens (Dunlap et al., 2015). [↑](#footnote-ref-3)