Rucsanda Juncu
r0872274

Introduction:

Bacteria are known for their fast duplication rate and smaller than eukaryotic genomes. This makes them the ideal specimens to help develop tools for measuring genome evolution between species. BLAST, or the Basic Local Alignment Search Tool (Sayers et al., 2022), finds alignments between genes or proteins and returns e value, or a value measuring the probability of the matches occurring by random instead of through the conservation of sequences. Homologs are similar genes that share a common ancestor, but have been changed either by speciation to become orthologs, or duplication to become paralogs. *Thermophagus xiamenensis* is an anaerobic, gram positive bacteria that was first discovered in a hot spring in Xiamen, China. The bacteria is slightly thermophilic and shows positive catalase activity (Gao et al., 2013). Another anaerobic but gram positive bacteria, *Proteiniclasticum ruminis*, was isolated from a yak rumen**.** There, its main job was to proteolytically aid in digestion and fermentation of the grasses on the Qinghai–Tibetan Plateau, China (Zhang et al., 2010). This study explores the evolutionary relationship of these two bacteria through the identification of best bidirectional hits (BBH), the construction of phylogenetic and species trees, and the calculation of conserved sequences.

Materials and Methods:

Stand alone BLAST (*Standalone BLAST Setup for Unix - BLAST® Help - NCBI Bookshelf*, n.d.) was used to align *T. xiamenensis* proteins against *P. ruminis* and vice versa. Self BLASTs were also created to identify paralogs within BBHs. BBHs were identified by merging forward and reverse BLAST results together and matching along a common sequence ID. Results were sorted by e-value to identify best matches, which were searched against self-alignments to identify if orthologs were aligning with paralogs of the same gene duplication. A high scoring co-ortholog pair for oxaloacetate decarboxylase, which also contained paralogs in *P. ruminis*, was selected for phylogenetic analysis. 25 homologs from different species were combined with the selected ortholog pair. Clustalw2 (Sievers et al., 2011) was used to create a multiple sequence alignment of the homologs and a 1000x bootstrapped phylogenetic tree. A species tree was created by selecting 16S rRNA sequences for the unique species. Note that five species did not have 16S data available and were therefore omitted. These sequences were aligned using MUSCLE (Edgar, 2004) and trimmed with Gblocks (Castresana, 2000). The resulting consensus sequences were bootstrapped 100X into a species tree using PhyML (Guindon et al., n.d.). Identification of conserved sequences was calculated using Shannon Entropy (Torres-García et al., 2022) as shown below:

$$S(X) = -\sum_{i=1}^{N} p(x_i)\log_2(p(x_i)),$$

Only positions with an entropy of zero were considered to be conserved.

Results:

T. xiamenensis had 2,869 protein coding genes, while *P. ruminis* had 2,877 to 2,937 protein coding genes. A cutoff value of 1x10^{-5} was chosen for the e-value of meaningful BBHs, identifying 5964 orthologs. Ortholog pair WP_010528268.1 and WP_031577996.1 were identified with an e-value of 1.080000e-133. WP_031577996.1, a protein from *P. ruminis*, had a paralog identified with the self BLAST. This paralog was WP_031577489.1, and had a lower e-value of 0 and only one mismatch. The *T. xiamenensis* protein WP_010528268.1 also matched with the paralog with an e-value of 1.280000e-133. WP_031577996.1 was protein BLASTed and the first 25 homologs from unique species were selected for phylogenetic analysis (Appendix Table 1). The Clustalw2 phylogenetic tree is observed in **Figure 1**, and was visualized in ITOL.

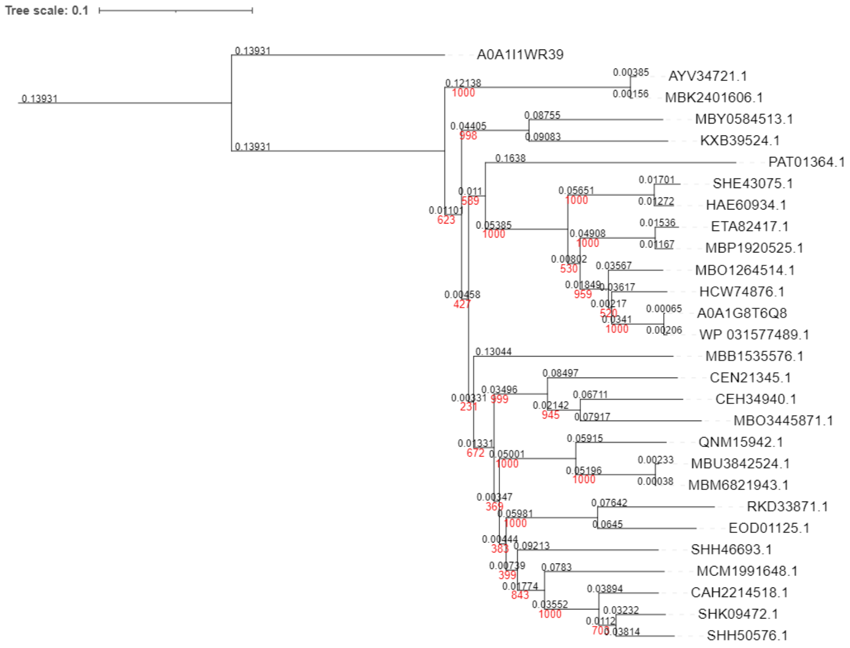

Figure 1: Phylogenetic Tree of 25 Species. This tree was built off of oxaloacetate homologs. Branch lengths are in black, boot values are in red. Sequence ID names can be found in Appendix Table 1.

The tree was midpoint rooted and showed strong boot values at tree leaves. The earliest split put A0A1I1WR39 (*T. xiamenensis*) in an outgroup of its own. Similar strains grouped together, such as MBK2401606.1 and AYV34721.1, which both belonged to the same genus of *Erysipelothrix*. *Youngiibacter* and *Tepidibacter* members also clustered together, however, members of *Clostridium* had several internal nodes separating each other. The *P. ruminis* paralogs had the shortest patristic distance between them. The species tree (**Figure 2**) generated by MUSCLE, Gblocks, and PhyML had slightly weaker but still acceptable boot values.

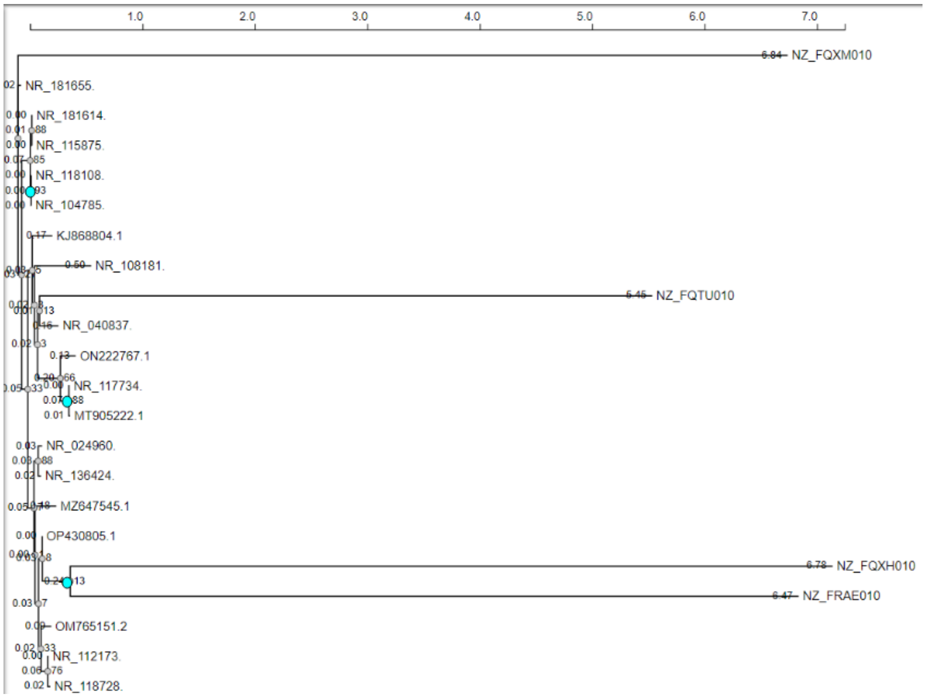

Figure 2: Species Tree of 20 Species. Three duplication events are shown (blue). Short branch lengths, a positive quality measure, are observed compared to **Figure 1**. Species names can be found in Appendix Table 1.

16S sequences showed 60% conservation in Gblocks, and most species show short distances between each other. Duplication events are marked in blue, while all other nodes should be attributed to speciation events. Duplication events were observed between *Youngiibacter fragilis 232.1* and *Youngiibacter multivorans*, *Fusobacterium mortiferum* and *Fusobacterium hominis*, and *Tepidibacter thalassicus DSM 15285* and *Tepidibacter formicigenes DSM 15518*. The the organisms with the largest branch lengths were *Tepidibacter thalassicus DSM 15285* (6.04), *Alkalibacter saccharofermentans DSM 14828* (5.45), *Tepidibacter thalassicus DSM 15285* (6.78), and *Tepidibacter formicigenes* DSM 15518 (6.47). Note that *Proteiniclasticum aestuarii* and *Thermophagus xiamenensis* had a shorter distance to each other than to *Proteiniclasticum ruminis*. Shannon entropy over the multiple sequence alignment (MSA) is shown in **Figure 3**.

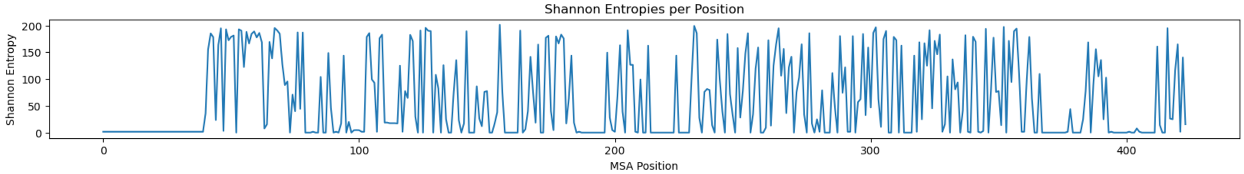

Figure 3: Shannon Entropy over MSA. Note that sequence conservation is mostly found in the middle and end of the sequence, with high sequence variability occurring at the beginning and interspaced between conserved regions.

See Appendix Table 2 for conserved sequences. Gaps ("-") in this table represent amino acids with non-zero entropies. Only longer conserved sequences are included, with short single bases having been omitted.

Discussion:

The two bacteria being compared have very little in common in terms of location and function within their micro-ecosystem. However, they had many BBHs and clustered together in both the species tree and the phylogenetic tree. More quality assessment should be performed on the identification of BBHs, such as adding annotations to confirm if BBHs have similar biological functions. The in-paralogs of oxaloacetate decarboxylase were more conserved with each other in *P. ruminis* than to the ortholog in *T. xiamenensis*, providing support that BBH detection was performed correctly. Additionally, oxaloacetate decarboxylase is a common enzyme used in the anaerobic metabolism of citrate (Labrou & Clonis, 1995). It makes logical sense that a basic enzyme for energy extraction would be conserved between two anaerobic species. The phylogenetic tree (**Figure 1**) showed *T. xiamenensis* having the furthest relation to the other bacteria, however, *T. xiamenensis* clustered more closely to *P. aestuarii* than *P. aestuarii* to *P.ruminis* in **Figure 2**. *P. aestuarii* and *P. ruminis* also did not cluster in a way to suggest a duplication event between them in the species tree. This may indicate that there has been an error in the computation of the species tree. How else could *T. xiamenensis*, a species that was suggested to have differentiated much earlier than the other bacteria, end up being more closely related to *P. ruminis* than its own paralog. Different alignment algorithms should be compared (Tcoffee to MUSCLE, etc…), as well as further literature exploration should be conducted to gain insight on the source of the error.

Rucsanda Juncu
r0872274

Appendix:
Appendix Table 1: Organism ID References

Organism	Phylogenetic Tree ID	Species Tree ID
Proteiniclasticum aestuarii	MBO1264514.1	NR_181614
Clostridiaceae bacterium	HCW74876.1	OP430805.1
Youngiibacter multivorans	MBP1920525.1	NR_104785.1
Youngiibacter fragilis 232.1	ETA82417.1	NR_118108.1
Eubacteriaceae bacterium	HAE60934.1	OM765151.2
Alkalibacter saccharofermentans DSM 14828	SHE43075.1	NZ_FQTU010
Oceanirhabdus seepicola	MCM1991648.1	NR_181655
Tepidibacter thalassicus DSM 15285	SHH50576.1	NZ_FQXH010
Clostridiales bacterium KA00274	KXB39524.1	KJ868804.1
Fusobacterium mortiferum	MBM6821943.1	NR_117734.1
Tepidibacter sp. 8C15b	CAH2214518.1	N/A
Candidatus Fusobacterium pullicola	MBU3842524.1	N/A
Fusobacterium hominis	QNM15942.1	MT905222.1
Caldisalinibacter kiritimatiensis	EOD01125.1	NR_136424
Tepidibacter formicigenes DSM 15518	SHK09472.1	NZ_FRAE010
Murdochiella sp. Marseille-P8839	MBY0584513.1	MZ647545.1
Clostridium grantii DSM 8605	SHH46693.1	NZ_FQXM01000085.1
Clostridium sp. CCUG 7971	MBO3445871.1	N/A
Thermohalobacter berrensis	RKD33871.1	NR_024960
Romboutsia lituseburensis	CEH34940.1	NR_118728
Erysipelothrix sp. strain 2 (EsS2-6-Brazil)	MBK2401606.1	N/A
Erysipelothrix rhusiopathiae	AYV34721.1	NR_040837
Candidatus Izimaplasma bacterium ZiA1	PAT01364.1	N/A

Leptotrichia sp.	MBB1535576.1	ON222767.1
Paeniclostridium sordellii	CEN21345.1	NR_112173
Proteiniclasticum ruminis	A0A1G8T6Q8	NR_108181
Thermophagus xiamenensis	A0A1I1WR39	NR_115875.1
Proteiniclasticum ruminis (paralog)	WP_031577489.1	N/A

Appendix Table 2: Conserved Sequences

Sequence	Position
L-----YEP-LL-PI--G-L--N-P	74-99
Y-G---G--P--IF---G--TDF-----NP----LGAAAQ-G------GA	124-174
I-IIGGADGPTA	186-198
A----G-IA-AAYSYMALVP-IQPPI	205-231
NP-IGAAGVSAVP--ARVV---G------N-LLMHAM-PN--GVIGSA--AG	365-417

Scripts:

```
#!/usr/bin/env python
# coding: utf-8


#import my libraries
import pandas as pd
import os
import matplotlib.pyplot as plt
from Bio import AlignIO
from math import log2

#set my working directory
os.chdir('/home/rucs/Comparative_genomics/ncbi-blast-
2.13.0+/bin/')

#read in my BLAST alignments into pandas dataframes
fwd_hits = pd.read_csv('forward_allign_10', header=None)
rev_hits = pd.read_csv('reverse_allign_10', header=None)
self_hits_therm = pd.read_csv('therm_self_allign_10',
header=None)
self_hits_prot = pd.read_csv('prot_self_allign_10', header=None)
```

Rucsanda Juncu
r0872274

```
#I chose output format 10 in my BLAST, so the files did not come
with headers and I had to assign them myself
headers = ["qseqid", "sseqid", "pident", "align_length",
"mismatch", "gapopen","qstart", "qend", "sstart", "send",
"evalue", "bitscore"]

fwd_hits.columns = headers
rev_hits.columns = headers
self_hits_therm.columns = headers
self_hits_prot.columns = headers

#perform inner join on the forward and reverse BLASTs
besthit = pd.merge(fwd_hits, rev_hits[['qseqid', 'sseqid']],
left_on='sseqid', right_on='qseqid', how='inner')


# Remove proteins that are not reciprocal
besthit = besthit.loc[besthit.qseqid_x == besthit.sseqid_y]

# Remove duplicates by sorting/keeping only maximum values
besthit = besthit.groupby(['qseqid_x', 'sseqid_x']).max()

#Sort by e value
besthit = besthit.sort_values(by = ['evalue'], ascending = True,
kind='quicksort')

#Keep matches above a threshold
besthit = besthit[besthit["evalue"] < 0.00001]

#Select an example ortholog for co-ortholog analysis
besthit.loc[besthit['sseqid_y']=='WP_010528268.1']

#See if the match is found also in the reverse hits
besthit.loc[besthit['qseqid_y']=='WP_031577489.1']

#Identify paralogs of the co-ortholog
self_hits_prot.loc[self_hits_prot['qseqid']=='WP_031577996.1']

#read in clustal alignment
#(The Module for Multiple Sequence Alignments, AlignIO ·
Biopython, n.d.)
alignment =
AlignIO.read(open("/home/rucs/Comparative_genomics/clustalw-2.1-
linux-x86_64-libcppstatic/seqdump_complete.aln"), "clustal")
```

Rucsanda Juncu
r0872274

```
print("Alignment length %i" % alignment.get_alignment_length())

#Get columns
sequences = []
for s in alignment:
    sequences.append("%s" % (s.seq))

#Calculate Shannon Entropy for each column in the alignment
all_entropies = []
for aindex in range(len(sequences[0])):
    column_bases = []
    column_entropy = []
    for bacteria in sequences:
        for aa in bacteria[aindex]:
            column_bases.append(aa)
        for aa in column_bases:
            p = column_bases.count(aa)/len(column_bases)
            entropy = p * log2(p)
            column_entropy.append(entropy)
    sh_entropy = -sum(column_entropy)
    all_entropies.append(sh_entropy)

#Plot the Shannon Entropy
fig = plt.figure(figsize=(20, 2))
ax = fig.add_subplot(111)
ax.plot(all_entropies)
plt.title('Shannon Entropies per Position')
plt.xlabel('MSA Position')
plt.ylabel('Shannon Entropy')

#Identify amino acids with a shannon entropy of 0 (complete
conservation)
seq1 = sequences[0]
conserved_seqs = ""
for i in range(len(all_entropies)):
    if all_entropies[i]==0:
        conserved_seqs = conserved_seqs + seq1[i]
    else:
        conserved_seqs = conserved_seqs + "-"
print(conserved_seqs)

#Append conserved sequences to the original alignment
sequences = sequences + [conserved_seqs]
final = open("final.txt", "w")
final.write(str(sequences))
final.close()
```

Rucsanda Juncu
r0872274

References:

Castresana, J. (2000). Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular Biology and Evolution*, *17*(4), 540–552. https://doi.org/10.1093/OXFORDJOURNALS.MOLBEV.A026334

Edgar, R. C. (2004). MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, *5*(1), 1–19. https://doi.org/10.1186/1471-2105-5-113/FIGURES/16

Gao, Z. M., Liu, X., Zhang, X. Y., & Ruan, L. W. (2013). Thermophagus xiamenensis gen. nov., sp. nov., a moderately thermophilic and strictly anaerobic bacterium isolated from hot spring sediment. *International Journal of Systematic and Evolutionary Microbiology*, *63*(1), 109–113. https://doi.org/10.1099/IJS.0.038547-0/CITE/REFWORKS

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., Pour, A., & Bioinformatique, L. (n.d.). *New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0*. Retrieved December 23, 2022, from http://www.lirmm.fr/~gascuel

Labrou, N. E., & Clonis, Y. D. (1995). Oxaloacetate Decarboxylase: On the Mode of Interaction with Substrate-Mimetic Affinity Ligands. *Archives of Biochemistry and Biophysics*, *321*(1), 61–70. https://doi.org/10.1006/ABBI.1995.1368

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., … Sherry, S. T. (2022). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, *50*(D1), D20–D26. https://doi.org/10.1093/NAR/GKAB1112

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., & Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, *7*, 539–539. https://doi.org/10.1038/MSB.2011.75

Standalone BLAST Setup for Unix - BLAST® Help - NCBI Bookshelf. (n.d.). Retrieved December 23, 2022, from https://www.ncbi.nlm.nih.gov/books/NBK52640/

The module for multiple sequence alignments, AlignIO · Biopython. (n.d.). Retrieved December 23, 2022, from https://biopython.org/wiki/AlignIO

Torres-García, A. A., Mendoza-Montoya, O., Molinas, M., Antelis, J. M., Moctezuma, L. A., & Hernández-Del-Toro, T. (2022). Pre-processing and feature extraction. *Biosignal Processing and Classification Using Computational Learning and Intelligence: Principles, Algorithms, and Applications*, 59–91. https://doi.org/10.1016/B978-0-12-820125-1.00014-2

Zhang, K., Song, L., & Dong, X. (2010). Proteiniclasticum ruminis gen. nov., sp. nov., a strictly anaerobic proteolytic bacterium isolated from yak rumen. *International Journal of Systematic and Evolutionary Microbiology*, *60*(9), 2221–2225. https://doi.org/10.1099/IJS.0.011759-0/CITE/REFWORKS