Annotathon Report

Name: Darcy Jones **Student Number:** 17369904

Annotathon Code: GOS_4466010.1

Sample Information :

Sargasso Sea: Sargasso Sea, Station 3 (Bermuda (UK))

GPS: $32^\circ 10' 29.4n; 64^\circ 00' 36.6w$

Sampled on 02/25/03 at 01:00:00

Filtered to: 0.22-0.8 microns

Habitat: Open Ocean

Depth: 5m (Sea floor: > 4200m)

Temperature: 19.8 °C

Genomic Sequence :

```
>GOS_4466010 Genomic DNA (Sargasso Sea: Sargasso Sea, Station 3)
GGAAATTAAAAGAAGCAGTTATTAAATCTTGGTGTAAACCAGATAAAATTTCAAATAGAC
TTAGAAAAAGATGAGGTTGATAAATTGATTGCCCTATCTTCTTTTATAGCTGGAGGATCA
GGAGCAAGAGGAACGCCTTCTATGTTTGTAAATGAATTTTTTTATCCTGGATATCTATCA
AAGGATAGAATTGAAGGTCTTTTAAATCAATAAGCCCCTAATAAAATTATTAGGAGCTTA
AAAAAATTAAATTTTAAATCTTAAGCTTTTGAAACTTCTCTGGTGTTAGCATTGTAAGAC
TCAGTAAATGGTATAGAAACAACTTCAGCATCTACTGGTACACCTGGTGAGTCAGCATAT
TCATCGGGAAGATGAACTTTAAATTTTTGCCCAACTTTACCAATTGGAAAACCATTATTA
CTAAGAGTTCCATCAAACGGAACATAGCCCATAGCAATATTTCTTTTTTGTTCTGGATGG
TACCACGGTGAAGTAACATAACCACAAGGATCTCCTCCTTCAGCAGGAGAAATTAACCAA
AAATCAGGAGCATATTCCTCTATTGGTTTTCCTCCTAATACCATTCCTACTAATTGCAAT
TTGTAAGGCTTGTTTCCTGCAGTAATTTCCTTTTTCATTTTCTCTAAAGCTTCTTTACCA
ATATAATCAGTAGATTTTTTCCATTCTCCTACACCAGAAAGAGATACTTGATAGCCTAAG
TTACATTGAAAAGGATTATGTTGGTTATCCATATCCTGACCCCAAGATAAAATTCCAGCT
TGAATTCTTCTATGGTGAGCAGGAGCAATTACCATTAAATTATGTTTTTTACCTGCTTCC
AAAACAGCATTCCACATATCATCTGCATATAAAGTAGCATCATAAAGATATATTTCAAAA
ACCTGCTGCTCCTGAAAAGCCTGTTTGAGAAATACACATTTTCT
```

Translation Used :

The longest potential open reading frame (ORF) found was 681 nucleotides long (227 amino acids) in the negative (reverse) strand of the genomic sequence from 264–944 (inclusive). It is missing 5′ DNA sequence, and hence a start codon, which would extend past the 3′ end of the given genomic DNA.

Translated Sequence :

```
RKCVFLKQAFQEQQVFEIYLYDATLYADDMWNAVLEAGKKHNLMVIAPAHHRRIQAGILS
WGQDMDNQHNPFQCNLGYQVSLSGVGEWKKSTDYIGKEALEKMKKEITAGNKPYKLQLVG
MVLGGKPIEEYAPDFWLISPAEGGDPCGYVTSPWYHPEQKRNIAMGYVPFDGTLSNNGFP
IGKVGQKFKVHLPDEYADSPGVPVDAEVVSIPFTESYNANTREVSKA
```

Using the ORF selection criteria of >60 nucleotides with no stop codons there were 23 other potential open reading frames found in the sequence. One other potential ORF gave a single significant BLAST hit (Data not shown). The best candidate ORF, GOS_4466010_20 (shown above), was chosen because it was the longest of the potential ORF set and had the highest number of significant BLAST hits.

Initially, the complete (with both a start and stop codon) open reading frame on the negative strand from 857–264 was selected as the best candidate. However, after constructing the multiple sequence alignment it seemed likely that the start codon was upstream of the genomic sequence (See figure 1) and the stop to stop codon model of an ORF was adopted.

All ORF predictions were conducted using EMBOSS tools' 'getorf' program (Rice et al., 2000).

BLAST data

Homologues for the selected ORF, GOS_4466010_20, were found in the NCBI non-redundant (nr) protein database using the basic local alignment search tool (BLAST) protein algorithm (Altschul et al., 1990; Camacho et al., 2009). 239 sequences significant (e-value $\leq 10^{-8}$) matches were found, with the 10 most similar matches all being glycine cleavage system protein T partial matches (Table 1).

Table 1: Showing the 10 highest scoring (by evalue) BLASTp hits to the longest ORF in GOS_4466010.1

gi	accession	evalue	score	title
516678066	WP_018036634	7.0e-144	1082	glycine cleavage system protein T [alpha p ...
495821508	WP_008546087	2.0e-125	961	glycine cleavage system protein T [Candida ...
560891133	WP_023648742	4.0e-125	959	glycine cleavage system protein T [Candida ...
494055920	WP_006998017	1.0e-124	956	glycine cleavage system protein T [Candida ...
71083953	YP_266673	2.0e-124	954	glycine cleavage system protein T [Candida ...
406706486	YP_006756839	3.0e-123	946	glycine cleavage system T-protein-like,fol ...
519013695	WP_020169570	4.0e-123	945	glycine cleavage system protein T [Candida ...
564613018	WP_023854142	4.0e-122	938	glycine cleavage system T protein (aminome ...
516680935	WP_018039018	2.0e-115	895	glycine cleavage system protein T [alpha p ...
167042027	ABZ06763	3.0e-113	881	putative glycine cleavage T-protein (amino ...

The 10 highest scoring hits with unique taxonomic identifiers from the BLASTp search and the GOS_4466010_20 sequence were aligned using the T-Coffee algorithm (Notredame et al., 2000) to find conserved regions and evaluate the likelihood that the potential ORF is part of a protein coding gene (Figure 1). The alignment shows a high degree of conservation between predicted homologues and GOS_4466010_20, with few regions less than 50% similar. The alignment also indicates that it is likely that the ORF is missing 5′ DNA in the given genomic sequence and that the ORF is potentially another full length glycine cleavage protein. Regardless of whether the missing upstream DNA is the same as the presented homologue's, the ORF GOS_4466010_20 is highly likely to be part of a protein coding gene or pseudogene.

Given the large number of significant BLAST hits, and the low e-values, high bit-scores and consistency of protein function in the top hits, these BLAST results appear to be representative of the putative protein. The highest scoring BLAST hit, gi|516678066|ref|WP|018036634 had 85% sequence identity to GOS_4466010_20 which indicates strongly that the potential ORF is closely related to proteobacterium glycine cleavage system protein T.

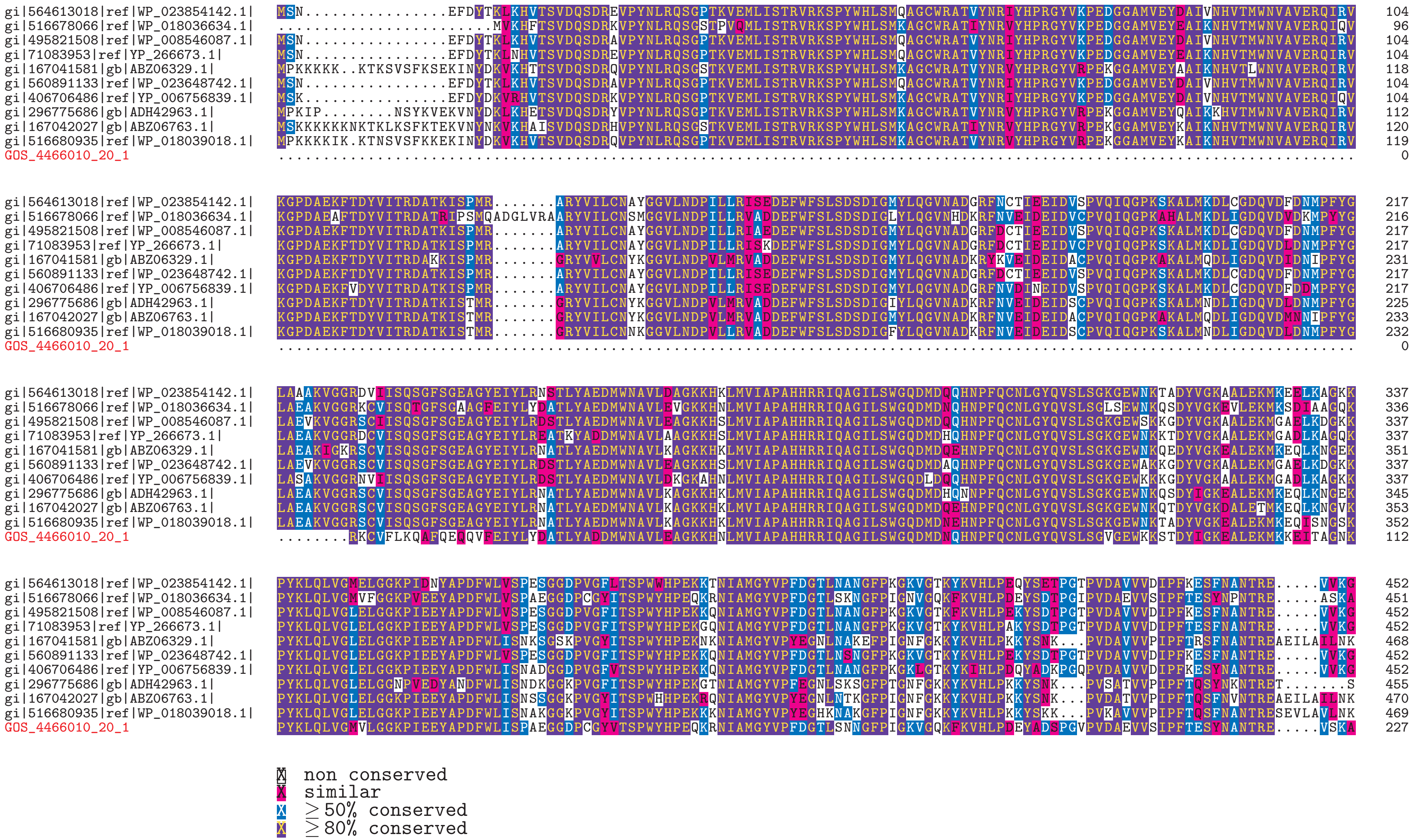
gi|564613018|ref|WP_023854142.1|
gi|516678066|ref|WP_018036634.1|
gi|495821508|ref|WP_008546087.1|
gi|71083953|ref|YP_266673.1|
gi|167041581|gb|ABZ06329.1|
gi|560891133|ref|WP_023648742.1|
gi|406706486|ref|YP_006756839.1|
gi|296775686|gb|ADH42963.1|
gi|167042027|gb|ABZ06763.1|
gi|516680935|ref|WP_018039018.1|
GOS_4466010_20_1
104
96
104
104
118
104
104
112
120
119
0
217
216
217
217
231
217
217
225
233
232
0
337
336
337
337
351
337
337
345
353
352
112
452
451
452
452
468
452
452
455
470
469
227
X non conserved
X similar
X ≥ 50% conserved
X ≥ 80% conserved

Figure 1: A multiple sequence alignment of the top 10 BLASTp hits with the query sequence GOS_4466010_20 (highlighted in red)

Biological Function

To predict the function of any potential protein product ofGOS_4466010_20, the NCBI conserved domain database was searched using the Reverse Position-Specific (RPS)-BLAST algorithm (Camacho et al., 2009; Marchler-Bauer et al., 2011). Eight significant (e-value $\leq 10^{-6}$) hits, were found with two single conserved domain matches: Glycine cleavage T-protein C-terminal barrel domain (pfam08669), and Aminomethyltransferase folate-binding domain (pfam01571) (Table 2). Six multi-domain conserved protein profiles were also detected with three biological functions: glycine cleavage (COG0404), sarcosine oxidation (TIGR01372) and sulphur flux regulation (PRK12486).

Aminomethyltransferase (AKA Glycine Cleavage System T protein, GCST protein) is a part of the glycine cleavage system, which catalyses the decarboxylation of glycine in bacteria and mitochondria (Lee et al., 2004). This enzyme contains both the GCST-protein C-terminal barrel domain and Aminomethyltransferase folate-binding domain found in GOS_4466010_20. Sarcosine oxidase catalyses the oxidative demethylation of sarcosine to glycine, which involves a folate-binding domain (Suzuki, 1994).

Dimethyl sulphoniopropionate demethylase, an enzyme involved in marine bacterial sulphur regulation, reversibly catalyses the conversion of dimethylsulphoniopropionate to sulphur and dimethylsulphide (Vila-Costa et al., 2006). Some bacterioplankton GCST-family proteins have been found to have Dimethyl sulphoniopropionate methyltransferase activity, which would explain the presence of this multi-domain match to GOS_4466010_20 (Howard et al., 2006).

Table 2: Showing all significant hits from an RPS-BLAST search of the NCBI conserved domain database.

id	evalue	score	title
gnl|CDD|223481	2.0e-23	239	COG0404, GcvT, Glycine cleavage system T protein (...
gnl|CDD|237113	2.0e-21	223	PRK12486, dmdA, putative dimethyl sulfoniopropiona ...
gnl|CDD|234742	2.0e-19	209	PRK00389, gcvT, glycine cleavage system aminomethy ...
gnl|CDD|254962	9.0e-12	142	pfam08669, GCV_T_C, Glycine cleavage T-protein C-t ...
gnl|CDD|233010	4.0e-09	131	TIGR00528, gcvT, glycine cleavage system T protein ...
gnl|CDD|233382	5.0e-09	131	TIGR01372, soxA, sarcosine oxidase, alpha subunit ...
gnl|CDD|177953	1.0e-08	128	PLN02319, PLN02319, aminomethyltransferase. ...
gnl|CDD|250713	2.0e-08	124	pfam01571, GCV_T, Aminomethyltransferase folate-bi ...

Given that the aminomethyltransferase protein contains both conserved single-domain matches to GOS_4466010_20, and that the dimethyl sulphoniopropionate methyltransferase multi-domain match appears to be related to a secondary function of aminomethyltransferase, it seems likely that the putative protein GOS_4466010_20 has aminomethyltransferase-like activity. Figure 2 shows the position of the aminomethyltransferase multi-domain patial match and the two single functional domain matches for the candidate ORF. The partial matches for COG0404 and pfam01571 conserved domains limited by the missing amino acid information toward the N-terminus. It is possible that these domains do exist in their complete form in the complete ORF if one exists.

To find more information about the likely structure and function of the putative protein product of GOS_4466010_20, the sequence was BLASTp searched against the curated Swiss-Prot database and the highest scoring homologue was used in place of the incomplete ORF. The highest scoring hit was Aminomethyltransferase (EC:2.1.2.10, ACC:Q67N36) from *Symbiobacterium thermophilum*, which is consistent with functional predictions from previous BLASTp and conserved domain analyses. The homologue is an 375 AA long cytosolic protein, and is part of the glycine cleavage system which catalyzes the degradation of glycine (Ueda et al., 2004). The protein has a predicted molecular weight (average mass) of 41.243 kDa and an isoelectric point (pI) of 5.51 (Predicted using ExPASy 'Compute pI/Mw tool' available at web.expasy.org/compute_pi/; Gasteiger et al., 2005).

An homology modelled protein structure (figure 3) is available for the Aminomethyltransferase homologue (pdb:1yx2A) which shows a globular protein with two functional domains; the Aminomethyltransferase folate-binding domain and the Glycine cleavage T-protein C-terminal barrel (Kiefer et al., 2009).

The probable function and identity of the GOS_4466010_20 sequences as an Aminomethyltransferase, is supported by high sequence homology from three separate protein databases (nr protein, CDD and Swiss-Prot). It is highly likely that the product of GOS_4466010_20 would be an Aminomethyltransferase

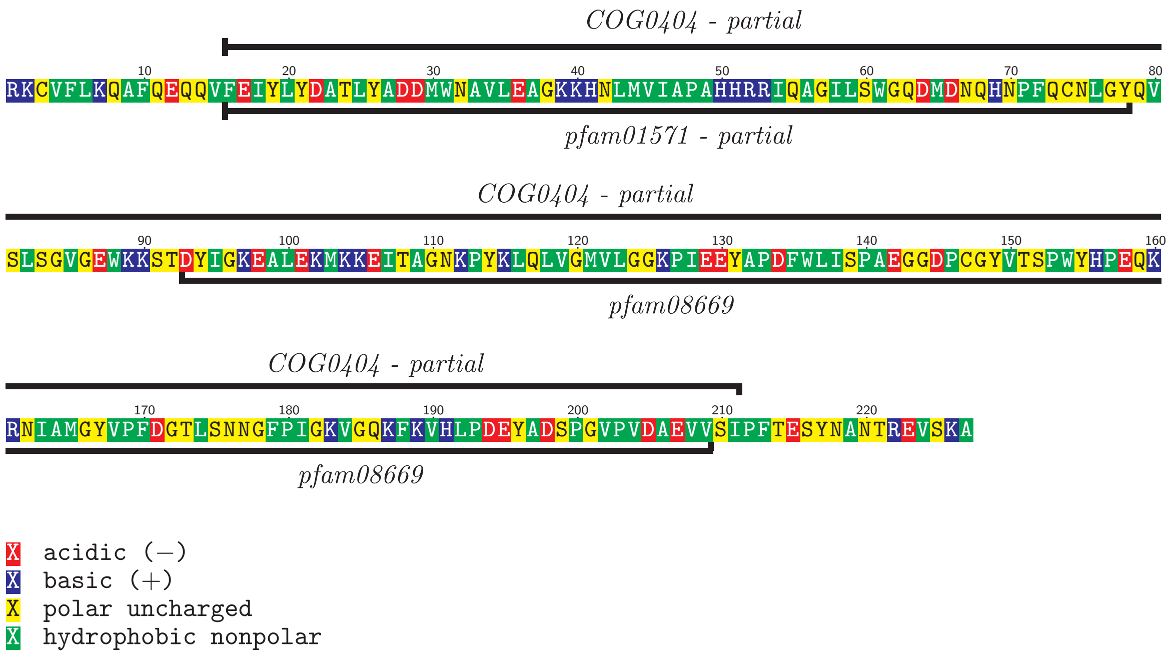
COG0404 - partial
10
20
30
40
50
60
70
80
RKCVFLKQAFQEQQVFEIYLYDATLYADDMWNAVLEAGKKHNLMVIAPAHHRRIQAGILSWGQDMDNQHNPFQCNLGYQV
pfam01571 - partial
COG0404 - partial
90
100
110
120
130
140
150
160
SLSGVGEWKKSTDYIGKEALEKMKKEITAGNKPYKLQLVGMVLGGKPIEEYAPDFWLISPAEGGDPCGYVTSPWYHPEQK
pfam08669
COG0404 - partial
170
180
190
200
210
220
RNIAMGYVPFDGTLSNNGFPIGKVGQKFKVHLPDEYADSPGVPVDAEVVSIPFTESYNANTREVSKA
pfam08669
X acidic (−)
X basic (+)
X polar uncharged
X hydrophobic nonpolar

Figure 2: The GOS_4466010_20 sequence showing matches to: Glycine cleavage T-protein C-terminal barrel domain (pfam08669), Aminomethyltransferase folate-binding domain superfamily (pfam01571), and multi-domain Glycine cleavage system T protein (COG0404). Full bar ends represent incomplete domain match boundaries, half bar ends represent complete domain match boundaries.

or a related protein, if the gene is complete.

Phylogenetics

To infer a phylogenetic tree for the ORF GOS_4466010_20, the results from the BLASTp search against the nr protein database was used. The highest scoring 10 BLAST hits and a random selection (from the $Beta(\alpha = 1, \beta = 2)$ distribution) of 10 from the remaining significant hits were used. These sequences and the putative ORF were aligned using the T-Coffee algorithm (Notredame et al., 2000). A maximum likelihood tree with bootstrapping was estimated using RAxML 8 (Stamatakis, 2014) from the multiple sequence alignment, using a CAT rate of homogeneity model and the BLOSUM62 substitution matrix.

The tree shows that GOS_4466010_20 is most closely related to bacterial species in the *Alphaproteobacterium* division, with greatest homology to Candidatus *Pelagibacter ubique* aminomethyltransferase proteins 4. The *Alphaproteobacteria* are a functionally diverse class of the phylum *Proteobacteria*, and predominantly consists of plant and animal pathogens, and mutualists, as well as marine dwelling bacteria (Williams et al., 2007). *Pelagibacter ubique* is a small-sized marine bacterial species that makes up a large proportion of the ocean surface bacterioplankton population (Sowell et al., 2008). The apparent phylogenetic closeness of GOS_4466010_20 with the *Alphaproteobacteria* and, more specifically, Candidatus *Pelagibacter ubique* makes sense in the context of the sampling methods (Marine surface).

Additional Resources

All scripts and commands used are included in a makefile and Sweave document at:
github.com/darcyabjones/BCH3BMA-annotathon.
Raw data and conclusions were added to the Annotathon project page for annotathon code: GOS_4466010.1.

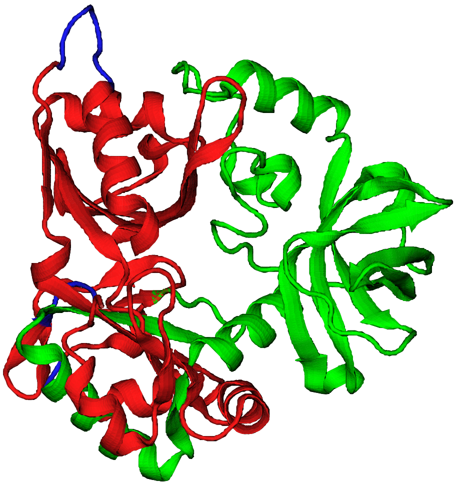

Figure 3: The predicted structure of Aminomethyltransferase ACC:Q67N36 showing a globular protein (pdb:1yx2A). The aligned region of GOS_4466010_20 corresponds to the region highlighted in green. The Aminomethyltransferase folate-binding domain is on the left, from residues 49–267. The Glycine cleavage T-protein C-terminal barrel is to the right, from residues 275–366.

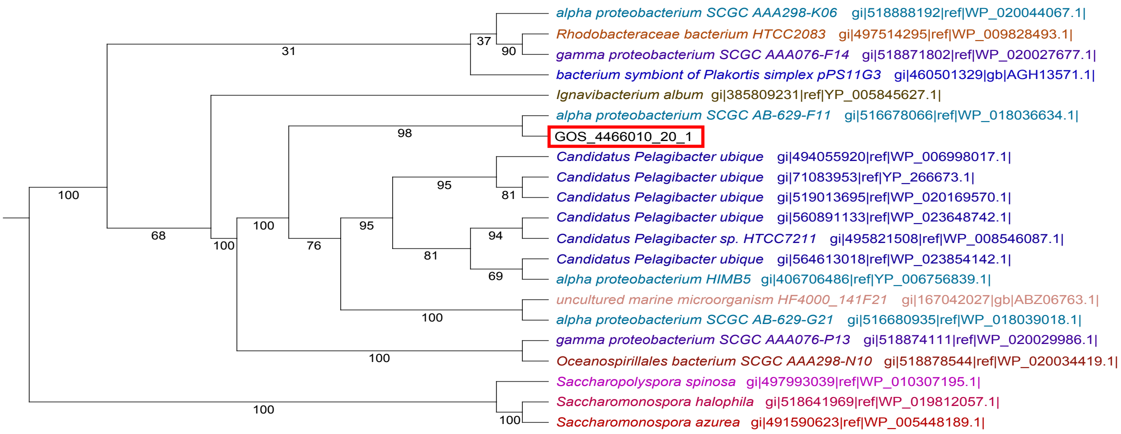
alpha proteobacterium SCGC AAA298-K06 gi|518888192|ref|WP_020044067.1|
Rhodobacteraceae bacterium HTCC2083 gi|497514295|ref|WP_009828493.1|
gamma proteobacterium SCGC AAA076-F14 gi|518871802|ref|WP_020027677.1|
bacterium symbiont of Plakortis simplex pPS11G3 gi|460501329|gb|AGH13571.1|
Ignavibacterium album gi|385809231|ref|YP_005845627.1|
alpha proteobacterium SCGC AB-629-F11 gi|516678066|ref|WP_018036634.1|
GOS_4466010_20_1
Candidatus Pelagibacter ubique gi|494055920|ref|WP_006998017.1|
Candidatus Pelagibacter ubique gi|71083953|ref|YP_266673.1|
Candidatus Pelagibacter ubique gi|519013695|ref|WP_020169570.1|
Candidatus Pelagibacter ubique gi|560891133|ref|WP_023648742.1|
Candidatus Pelagibacter sp. HTCC7211 gi|495821508|ref|WP_008546087.1|
Candidatus Pelagibacter ubique gi|564613018|ref|WP_023854142.1|
alpha proteobacterium HIMB5 gi|406706486|ref|YP_006756839.1|
uncultured marine microorganism HF4000_141F21 gi|167042027|gb|ABZ06763.1|
alpha proteobacterium SCGC AB-629-G21 gi|516680935|ref|WP_018039018.1|
gamma proteobacterium SCGC AAA076-P13 gi|518874111|ref|WP_020029986.1|
Oceanospirillales bacterium SCGC AAA298-N10 gi|518878544|ref|WP_020034419.1|
Saccharopolyspora spinosa gi|497993039|ref|WP_010307195.1|
Saccharomonospora halophila gi|518641969|ref|WP_019812057.1|
Saccharomonospora azurea gi|491590623|ref|WP_005448189.1|
37
90
31
98
95
81
100
95
94
68
100
76
81
69
100
100
100
100

Figure 4: Maximum likelihood cladogram of GOS_4466010_20 and a selection of BLAST results. Showing the genetic relationship of GOS_4466010_20-like genes with taxonomic information. Branch confidence numbers are bootstrap support values.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A (2005) Protein Identification and Analysis Tools on the ExPASy Server, in JM Walker (ed.), The Proteomics Protocols Handbook, pp. 571–608, Humana Press.

Howard EC, Henriksen JR, Buchan A, Reisch CR, Bürgmann H, Welsh R, Ye W, González J, Mace K, Joye SB, Kiene RP, Whitman WB, Moran MA (2006) Bacterial taxa that limit sulfur flux from the ocean. *Science*, **314**, 649–652.

Kiefer F, Arnold K, Künzli M, Bordoli L, Schwede T (2009) The SWISS-MODEL Repository and associated resources. *Nucleic acids research*, **37**, D387–92.

Lee HH, Kim DJ, Ahn HJ, Ha JY, Suh SW (2004) Crystal structure of T-protein of the glycine cleavage system. Cofactor binding, insights into H-protein recognition, and molecular basis for understanding nonketotic hyperglycinemia. *The Journal of biological chemistry*, **279**, 50514–23.

Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Lu F, Marchler GH, Mullokandov M, Omelchenko MV, Robertson CL, Song JS, Thanki N, Yamashita RA, Zhang D, Zhang N, Zheng C, Bryant SH (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Research*, **39**, D225–9.

Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, **302**, 205–17.

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics*, **16**, 276–7.

Sowell SM, Norbeck AD, Lipton MS, Nicora CD, Callister SJ, Smith RD, Barofsky DF, Giovannoni SJ (2008) Proteomic analysis of stationary phase in the marine bacterium "Candidatus Pelagibacter ubique". *Applied and environmental microbiology*, **74**, 4091–100.

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

Suzuki H (1994) Sarcosine oxidase: structure, function, and the application to creatinine determination. *Amino acids*, **7**, 27–43.

Ueda K, Yamashita A, Ishikawa J, Shimada M, Watsuji To, Morimura K, Ikeda H, Hattori M, Beppu T (2004) Genome sequence of Symbiobacterium thermophilum, an uncultivable bacterium that depends on microbial commensalism. *Nucleic acids research*, **32**, 4937–44.

Vila-Costa M, Simó R, Harada H, Gasol JM, Slezak D, Kiene RP (2006) Dimethylsulfoniopropionate uptake by marine phytoplankton. *Science*, **314**, 652–654.

Williams KP, Sobral BW, Dickerman AW (2007) A robust species tree for the alphaproteobacteria. *Journal of bacteriology*, **189**, 4578–86.