Foundations of Bioinformatics - Find a Gene Project Part I

Name: Mirte Ciz Marieke Kuijpers
Email: mkuijpers@ucsd.edu
PID: A59010989

Question 1:

Name of protein: bacteriorhodopsin
Species: *Halobacterium salinarum*
Accession number: WP_136361479
Function: photon powered proton pump

Bacteriorhodopsin is not unique to *H. salinarum*, being found in many species of the Archaea. However, the crystal structure of bacteriorhodopsin has been solved for *H. salinarum* (PDB 1FBB), thus it seemed a good choice to focus on the bacteriorhodopsin of this particular species.

Question 2:

I chose to search for sequences homologous to WP_136361479 in the nucleotide collection (nr/nt) using tblastn. With the exception of increasing the maximum number of hits to 250 and excluding results from the taxa *Halobacteria (taxid:183963)* (as this is the taxa in which my example protein is found) I left the search parameters at their default values.

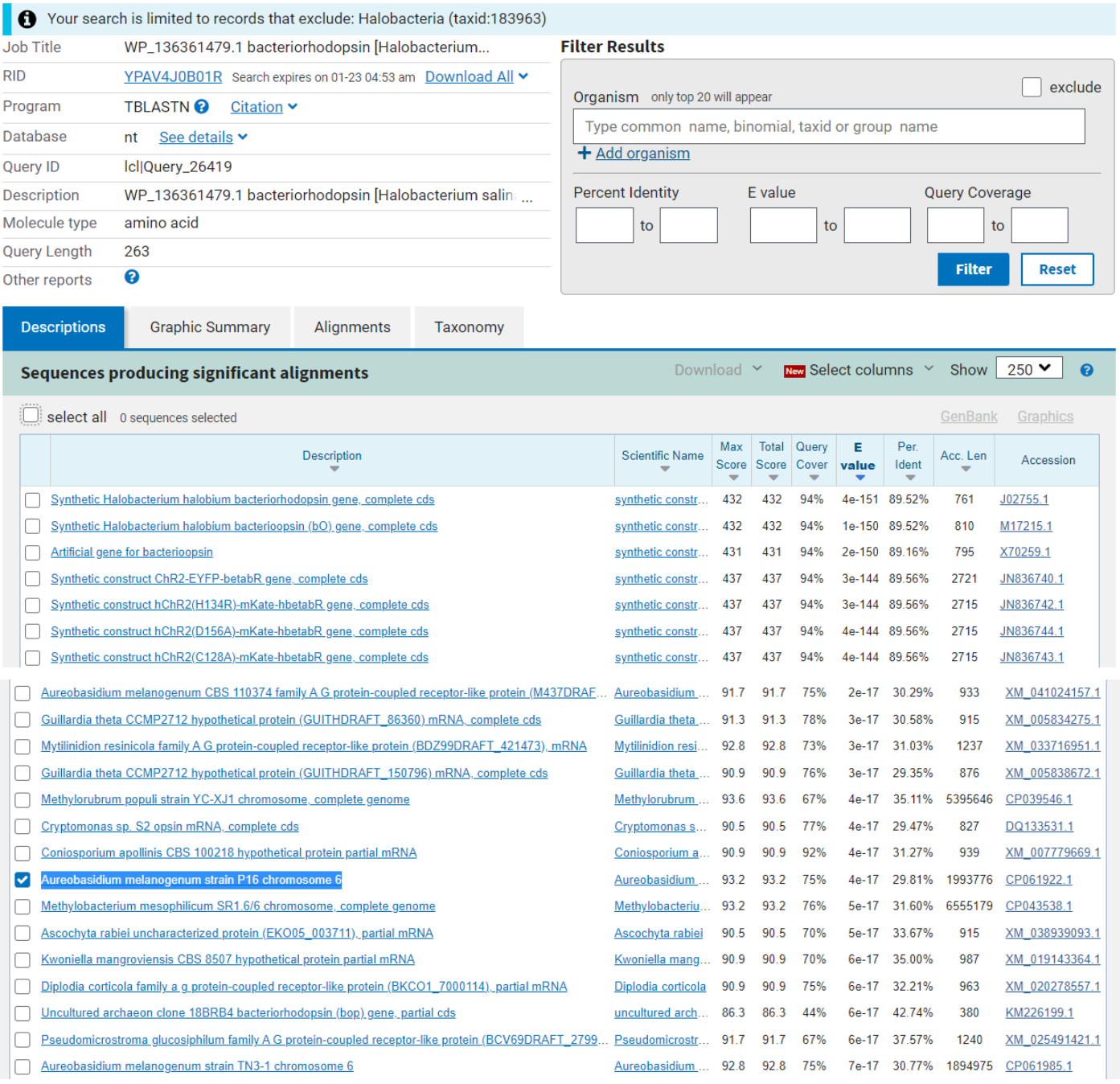
Search Parameters	
Program	tblastn
Word size	6
Expect value	0.05
Hitlist size	250
Gapcosts	11,1
Matrix	BLOSUM62
Low Complexity Filter	Yes
Filter string	L;
Genetic Code	1
Window Size	40
Threshold	21
Composition-based stats	2

Database	
Posted date	Jan 11, 2022 7:15 PM
Number of letters	632,502,643,815
Number of sequences	77,938,138
Entrez query	Excludes: Halobacteria (taxid:183963)

Karlin-Altschul statistics		
Lambda	0.323382	0.267
K	0.138507	0.041
H	0.430778	0.14
Alpha	0.7916	1.9
Alpha_v	4.96466	42.6028
Sigma		43.6362

Screenshot 1: Parameters for tblastn search used to find Gene X

Of the many hits that were found, one was to a region of *Aureobasidium melanogenum* strain P16's chromosome 6. This chromosome is unannotated and so the alignment may signify a novel protein. The hit can be seen highlighted in blue in the screenshot below (which has a break because I elected to omit some intervening sequences which were annotated and so not of interest).

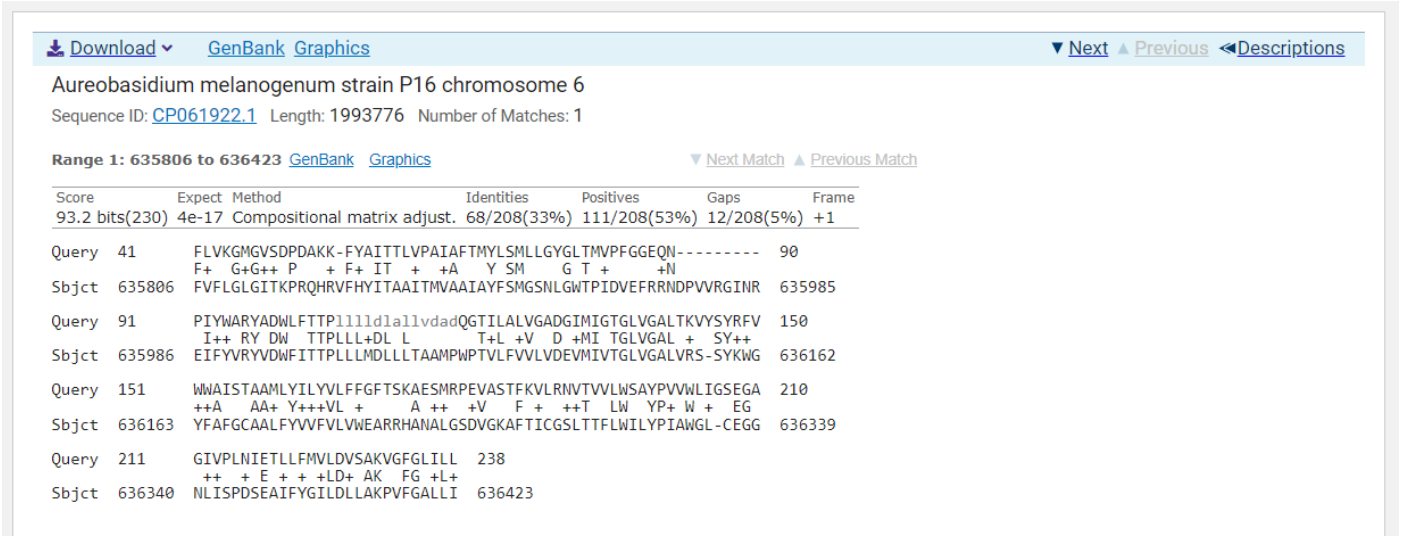


Screenshot 2: Search results for tblastn with parameters as specified in Screenshot 1.

The alignment is as follows:

```
Query  41      FLVKGMGVSDPDAKK-FYAITTLVPAIAFTMYLSMLLGYGLTMVPFGGEQN---------  90
               F+  G+G++ P   + F+ IT  +  +A   Y SM    G T +     +N
Sbjct  635806  FVFLGLGITKPRQHRVFHYITAAITMVAAIAYFSMGSNLGWTPIDVEFRRNDPVVRGINR  635985
```

```
Query  91      PIYWARYADWLFTTPllllldlallvdadQGTILALVGADGIMIGTGLVGALTKVYSYRFV  150
                 I++ RY DW  TTPLLL+DL L       T+L +V  D +MI TGLVGAL +  SY++
Sbjct  635986  EIFYVRYVDWFITTPLLLMDLLLTAAMPWPTVLFVVLVDEVMIVTGLVGALVRS-SYKWG  636162

Query  151     WWAISTAAMLYILYVLFFGFTSKAESMRPEVASTFKVLRNVTVVLWSAYPVVWLIGSEGA  210
                ++A   AA+ Y+++VL +     A ++  +V   F +  ++T  LW  YP+ W +  EG
Sbjct  636163  YFAFGCAALFYVVFVLVWEARRHANALGSDVGKAFTICGSLTTFLWILYPIAWGL-CEGG  636339

Query  211     GIVPLNIETLLFMVLDVSAKVGFGLILL  238
                 ++  + E + + +LD+ AK  FG +L+
Sbjct  636340  NLISPDSEAIFYGILDLLAKPVFGALLI  636423
```



Download GenBank Graphics ▼ Next ▲ Previous ◀◀ Descriptions

Aureobasidium melanogenum strain P16 chromosome 6

Sequence ID: CP061922.1 Length: 1993776 Number of Matches: 1

Range 1: 635806 to 636423 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
93.2 bits(230)	4e-17	Compositional matrix adjust.	68/208(33%)	111/208(53%)	12/208(5%)	+1

```
Query  41      FLVKGMGVSDPDAKK-FYAITTLVPAIAFTMYLSMLLGYGLTMVPFGGEQN---------  90
               F+  G+G++ P   + F+ IT  +  +A   Y SM    G T +     +N
Sbjct  635806  FVFLGLGITKPRQHRVFHYITAAITMVAAIAYFSMGSNLGWTPIDVEFRRNDPVVRGINR  635985

Query  91      PIYWARYADWLFTTPllllldlallvdadQGTILALVGADGIMIGTGLVGALTKVYSYRFV  150
                I++ RY DW  TTPLLL+DL L       T+L +V  D +MI TGLVGAL +  SY++
Sbjct  635986  EIFYVRYVDWFITTPLLLMDLLLTAAMPWPTVLFVVLVDEVMIVTGLVGALVRS-SYKWG  636162

Query  151     WWAISTAAMLYILYVLFFGFTSKAESMRPEVASTFKVLRNVTVVLWSAYPVVWLIGSEGA  210
               ++A   AA+ Y+++VL +     A ++  +V   F +  ++T  LW  YP+ W +  EG
Sbjct  636163  YFAFGCAALFYVVFVLVWEARRHANALGSDVGKAFTICGSLTTFLWILYPIAWGL-CEGG  636339

Query  211     GIVPLNIETLLFMVLDVSAKVGFGLILL  238
                ++  + E + + +LD+ AK  FG +L+
Sbjct  636340  NLISPDSEAIFYGILDLLAKPVFGALLI  636423
```

Screenshot 3: Alignment of hit of interest with Bacteriorhodopsin from the NCBI tblastn search results

The statistics for this alignment are as follows:

Score: 93.2 bits
Expect (e value): 4e-17
Method: Compositional matrix adjust.
Identities: 68/208 (33%)
Positives: 111/208 (53%)
Gaps: 12/208 (55%)
Frame: +1

While the percentage identity is poor, it is above 25% and the statistics seem sufficient for further investigation. From this point onwards this hit will be referred to as *Gene X* for convenience.

Question 3

Below is the full information on my chosen hit, "*Gene X*":

```
LOCUS       CP061922                618 bp    DNA     linear   PLN 15-APR-2021
DEFINITION  Aureobasidium melanogenum strain P16 chromosome 6.
ACCESSION   CP061922 REGION: 635806..636423
VERSION     CP061922.1
DBLINK      BioProject: PRJNA661658
            BioSample: SAMN16066980
KEYWORDS    .
SOURCE      Aureobasidium melanogenum
  ORGANISM  Aureobasidium melanogenum
            Eukaryota; Fungi; Dikarya; Ascomycota; Pezizomycotina;
            Dothideomycetes; Dothideomycetidae; Dothideales; Saccotheciaceae;
            Aureobasidium.
REFERENCE   1  (bases 1 to 618)
  AUTHORS   Jia,S.
  TITLE     Novel chromosomes and genomes provide new insights into evolution
            and adaptation of the whole genome duplicated yeast-like fungus
            TN3-1 isolated from natural honey
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 618)
  AUTHORS   Jia,S.
  TITLE     Direct Submission
  JOURNAL   Submitted (05-SEP-2020) College of Marine Life, Ocean University of
            China, No. 5 Yushan Road, Qingdao, Shandong 266003, China
COMMENT     ##Genome-Assembly-Data-START##
            Assembly Date           :: 20-JAN-2020
            Assembly Method         :: SOAPdenovo v. v.2.04.4; SMRT Link v.
                                       5.0.1
            Genome Representation   :: Full
            Expected Final Version  :: Yes
            Genome Coverage         :: 100.0x
            Sequencing Technology   :: Illumina HiSeq; PacBio
            ##Genome-Assembly-Data-END##
FEATURES             Location/Qualifiers
     source          1..618
                     /organism="Aureobasidium melanogenum"
                     /mol_type="genomic DNA"
                     /strain="P16"
                     /isolation_source="mangrove ecosystem"
                     /db_xref="taxon:46634"
                     /chromosome="6"
                     /country="China:Hainan"
                     /collection_date="2013-12-19"
```

Screenshot 4: Information about the hit, Gene X

The nucleotide sequence of the section of the genome which aligned in the search is provided below:

```
> Aureobasidium melanogenum strain P16’s chromosome 6 hit (Gene X)
TTTGTCTTCCTCGGTCTGGGTATCACCAAGCCTCGCCAGCACCGCGTCTTCCACTACATCACCGCTGCCATTACCAT
GGTCGCTGCTATTGCCTATTTCTCCATGGGTTCTAACCTCGGCTGGACTCCCATCGATGTTGAGTTCCGCCGCAACG
ATCCTGTTGTTCGCGGTATCAACCGTGAAATCTTCTACGTCCGCTACGTCGACTGGTTCATCACTACTCCTCTCCTC
CTCATGGATCTGTTGTTGACTGCCGCCATGCCTTGGCCCACTGTGCTCTTCGTCGTCTTGGTTGATGAAGTTATGAT
TGTCACTGGTCTCGTTGGTGCTCTCGTCCGTTCTTCTTACAAGTGGGGTTACTTCGCCTTCGGCTGCGCTGCTCTCT
TCTACGTTGTCTTTGTTCTCGTCTGGGAGGCTCGCCGTCACGCCAACGCTCTCGGCAGCGATGTTGGCAAGGCCTTC
ACTATCTGTGGCTCGCTCACCACCTTCCTCTGGATTCTTTACCCTATCGCCTGGGGTCTCTGTGAGGGTGGCAACCT
CATCTCTCCTGACTCCGAGGCTATCTTCTACGGTATCCTCGACCTGCTCGCTAAGCCTGTCTTTGGTGCTCTCCTCA
TC
```

Translation of just this section in all three frames using the ExPASy translation tool gives the following output:

Screenshot 5: ExPASy translation tool output showing predicted protein sequence for each frame with potential ORFs highlighted in pink

Encouragingly, the frame with the longest ORF is also the frame which was aligned in the Blast search. The sequence in this frame (frame 1) is below, with the first methionine in bold.

```
> Aureobasidium melanogenum strain P16’s chromosome 6 (frame 1)
FVFLGLGITKPRQHRVFHYITAAITMVAAIAYFSMGSNLGWTPIDVEFRRNDPVVRGINREIFYVRYVDWFITTPLL
LMDLLLTAAMPWPTVLFVVLVDEVMIVTGLVGALVRSSYKWGYFAFGCAALFYVVFVLVWEARRHANALGSDVGKAF
TICGSLTTFLWILYPIAWGLCEGGNLISPDSEAIFYGILDLLAKPVFGALLI
```

Considering that the match to the query sequence starts from it’s 41st codon, it is possible that the potential novel protein is encoded from a position earlier in the chromosome sequence than the start of the alignment. Lack of alignment in this region may be due to a lower conservation in that area. To test for this the sequence was expanded by 201 nucleotides in both directions (to keep the frame). The expanded sequence is shown below.

```
> Aureobasidium melanogenum strain P16’s chromosome 6 hit expanded
```

TCAAGATGGACTACCTCTCCAAGAGAAACGATGCCCTCAACGTCAATGGTAAGTGCATCTCTAACCACCATCGTGAT
CATGGGCAGCGCGCTGACCAGACTTGCAGGCAACACTGTCAACGGCAAGACCGTCGACATTGCCATCACCGTTCGCG
GTTCTGACTGGTACTGGACCGTCTGCGCTGTCATGACTACCTGCACCTTTGTCTTCCTCGGTCTGGGTATCACCAAG
CCTCGCCAGCACCGCGTCTTCCACTACATCACCGCTGCCATTACCATGGTCGCTGCTATTGCCTATTTCTCCATGGG
TTCTAACCTCGGCTGGACTCCCATCGATGTTGAGTTCCGCCGCAACGATCCTGTTGTTCGCGGTATCAACCGTGAAA
TCTTCTACGTCCGCTACGTCGACTGGTTCATCACTACTCCTCTCCTCCTCATGGATCTGTTGTTGACTGCCGCCATG
CCTTGGCCCACTGTGCTCTTCGTCGTCTTGGTTGATGAAGTTATGATTGTCACTGGTCTCGTTGGTGCTCTCGTCCG
TTCTTCTTACAAGTGGGGTTACTTCGCCTTCGGCTGCGCTGCTCTCTTCTACGTTGTCTTTGTTCTCGTCTGGGAGG
CTCGCCGTCACGCCAACGCTCTCGGCAGCGATGTTGGCAAGGCCTTCACTATCTGTGGCTCGCTCACCACCTTCCTC
TGGATTCTTTACCCTATCGCCTGGGGTCTCTGTGAGGGTGGCAACCTCATCTCTCCTGACTCCGAGGCTATCTTCTA
CGGTATCCTCGACCTGCTCGCTAAGCCTGTCTTTGGTGCTCTCCTCATCTGGGGTCACCGTGGCATTGACCCTGCTC
GTCTCGGCCTTTACATCCACGACTACAACGAGAAGGATCCCGCTGTTAAGGACAAGGTCGGCGCTCCTGGCCCCAAC
GTCCACCCTAACACCAACAACGCCAACAACGCCGCTGCCACCAACGATTCGACTCCCGAGACTGTCTAAATCTGCAG
TGCATGATATTCAGCACTT

Translation of this expanded sequence (performed as before), to observe whether the ORF could begin earlier is encouraging, as the same frame (1) has the longest complete ORF.

Screenshot 6: ExPASy translation tool output showing predicted protein sequence for each frame with potential ORFs highlighted in pink

The sequence for this longer (frame 1) ORF is shown below.

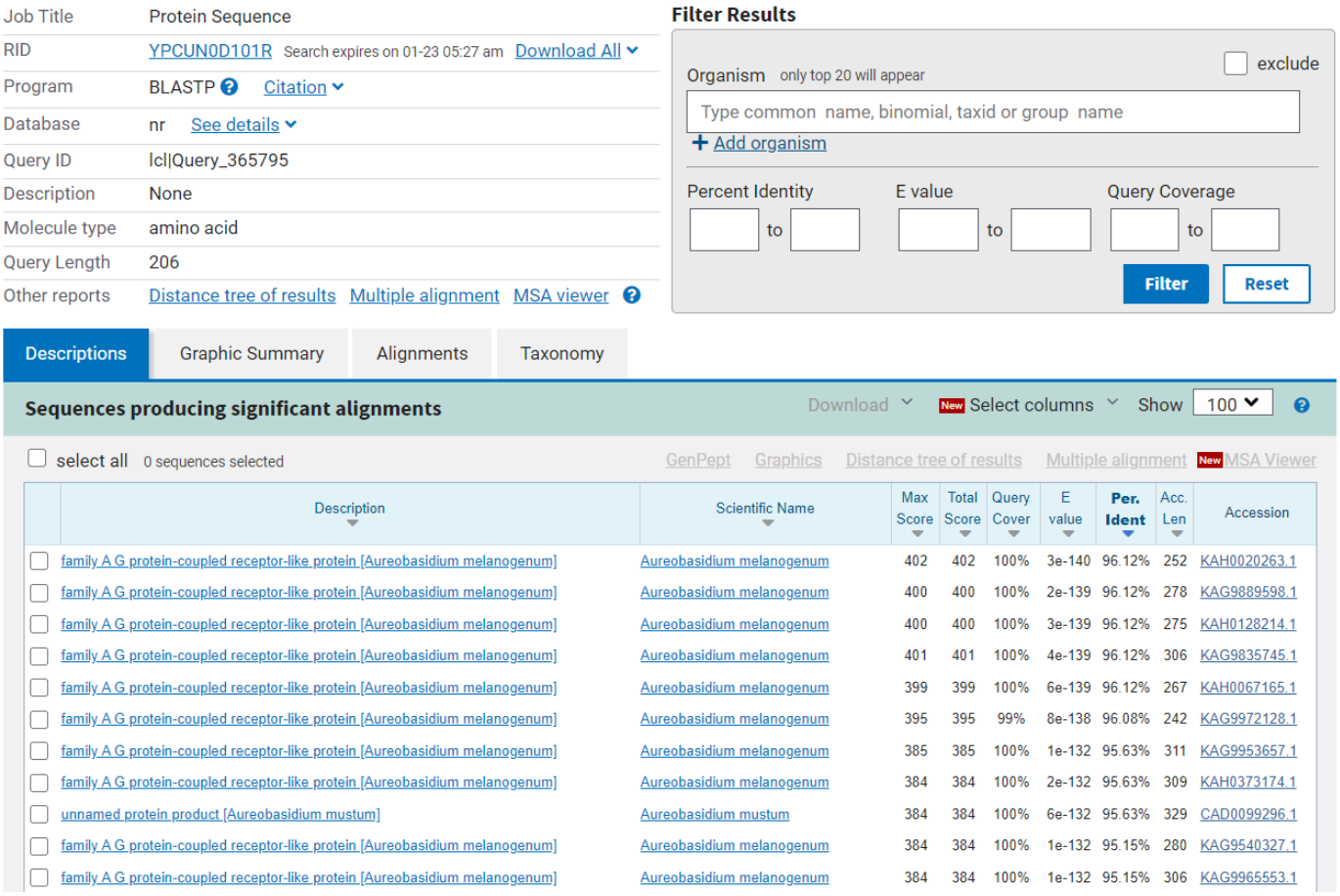
> Aureobasidium melanogenum strain P16's chromosome 6 hit expanded (longer predicted ORF)

MPSTSMVSASLTTIVIMGSALTRLAGNTVNGKTVDIAITVRGSDWYWTVCAVMTTCTFVFLGLGITKPRQHRVFHYI
TAAITMVAAIAYFSMGSNLGWTPIDVEFRRNDPVVRGINREIFYVRYVDWFITTPLLLMDLLLTAAMPWPTVLFVVL
VDEVMIVTGLVGALVRSSYKWGYFAFGCAALFYVVFVLVWEARRHANALGSDVGKAFTICGSLTTFLWILYPIAWGL
CEGGNLISPDSEAIFYGILDLLAKPVFGALLIWGHRGIDPARLGLYIHDYNEKDPAVKDKVGAPGPNVHPNTNNANN
AAATNDSTPETV

This may represent the entire protein, with the ends simply not aligning with query search. However, this remains speculative without mRNA-seq data. Nevertheless, I will consider both possible ORFs, from now on referred to as the longer and shorter predicted ORFs.

Question 4

Searching with the full sequence of the hit found in the initial tblastn search reveals no 100% percent identity hits in the blastp database (see image below).



Job Title	Protein Sequence
RID	YPCUN0D101R Search expires on 01-23 05:27 am Download All
Program	BLASTP Citation
Database	nr See details
Query ID	lcl|Query_365795
Description	None
Molecule type	amino acid
Query Length	206
Other reports	Distance tree of results Multiple alignment MSA viewer

Filter Results

Organism only top 20 will appear exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity to

E value to

Query Coverage to

Filter Reset

Descriptions Graphic Summary Alignments Taxonomy

Sequences producing significant alignments Download New Select columns Show 100

select all 0 sequences selected GenPept Graphics Distance tree of results Multiple alignment New MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
family A G protein-coupled receptor-like protein [Aureobasidium melanogenum]	Aureobasidium melanogenum	402	402	100%	3e-140	96.12%	252	KAH0020263.1
family A G protein-coupled receptor-like protein [Aureobasidium melanogenum]	Aureobasidium melanogenum	400	400	100%	2e-139	96.12%	278	KAG9889598.1
family A G protein-coupled receptor-like protein [Aureobasidium melanogenum]	Aureobasidium melanogenum	400	400	100%	3e-139	96.12%	275	KAH0128214.1
family A G protein-coupled receptor-like protein [Aureobasidium melanogenum]	Aureobasidium melanogenum	401	401	100%	4e-139	96.12%	306	KAG9835745.1
family A G protein-coupled receptor-like protein [Aureobasidium melanogenum]	Aureobasidium melanogenum	399	399	100%	6e-139	96.12%	267	KAH0067165.1
family A G protein-coupled receptor-like protein [Aureobasidium melanogenum]	Aureobasidium melanogenum	395	395	99%	8e-138	96.08%	242	KAG9972128.1
family A G protein-coupled receptor-like protein [Aureobasidium melanogenum]	Aureobasidium melanogenum	385	385	100%	1e-132	95.63%	311	KAG9953657.1
family A G protein-coupled receptor-like protein [Aureobasidium melanogenum]	Aureobasidium melanogenum	384	384	100%	2e-132	95.63%	309	KAH0373174.1
unnamed protein product [Aureobasidium mustum]	Aureobasidium mustum	384	384	100%	6e-132	95.63%	329	CAD0099296.1
family A G protein-coupled receptor-like protein [Aureobasidium melanogenum]	Aureobasidium melanogenum	384	384	100%	1e-132	95.15%	280	KAG9540327.1
family A G protein-coupled receptor-like protein [Aureobasidium melanogenum]	Aureobasidium melanogenum	384	384	100%	1e-132	95.15%	306	KAG9965553.1

Screenshot 7: Search results for blastp of the Gene X sequence

The top hit by E-value is a family A G protein-coupled receptor-like protein, also from Aureobasidium melanogenum, with 96% identity and an e value of 3e-140.

family A G protein-coupled receptor-like protein, partial [Aureobasidium melanogenum]

Sequence ID: KAH0020263.1 Length: 252 Number of Matches: 1

Range 1: 47 to 252 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
402 bits(1032)	3e-140	Compositional matrix adjust.	198/206(96%)	203/206(98%)	0/206(0%)

```
Query  1    FVFLGLGITKPRQHRVFHYITAAITMVAAIAYFSMGSNLGWTPIDVEFRRNDPVVRGINR  60
            FVFLGLGITKPRQHRVFHYITAAITMVAAIAYFSM SNLGWTPIDVEF RNDPVVRGINR
Sbjct  47   FVFLGLGITKPRQHRVFHYITAAITMVAAIAYFSMASNLGWTPIDVEFIRNDPVVRGINR  106

Query  61   EIFYVRYVDWFITTPLLLMDLLLTAAMPWPTVLFVVLVDEVMIVTGLVGALVRSSYKWGY  120
            EIFYVRY+DWFITTPLLLMDLLLTAAMPWPT+LFVVLVDEVMIVTGLVGALVRSSYKWGY
Sbjct  107  EIFYVRYIDWFITTPLLLMDLLLTAAMPWPTLLFVVLVDEVMIVTGLVGALVRSSYKWGY  166

Query  121  FAFGCAALFYVVFVLVWEARRHANALGSDVGKAFTICGSLTTFLWILYPIAWGLCEGGNL  180
            FAFGCAALFYVV+VLVWEARRHANALGSDVGKAFTICGSLTTFLWILYP+AWGLCEGGN+
Sbjct  167  FAFGCAALFYVVWVLVWEARRHANALGSDVGKAFTICGSLTTFLWILYPLAWGLCEGGNV  226

Query  181  ISPDSEAIFYGILDLLAKPVFGALLI  206
            ISPDSEAIFYGILDL AKPVFGALLI
Sbjct  227  ISPDSEAIFYGILDLFAKPVFGALLI  252
```

Screenshot 8: Alignment of the top hit to Gene X

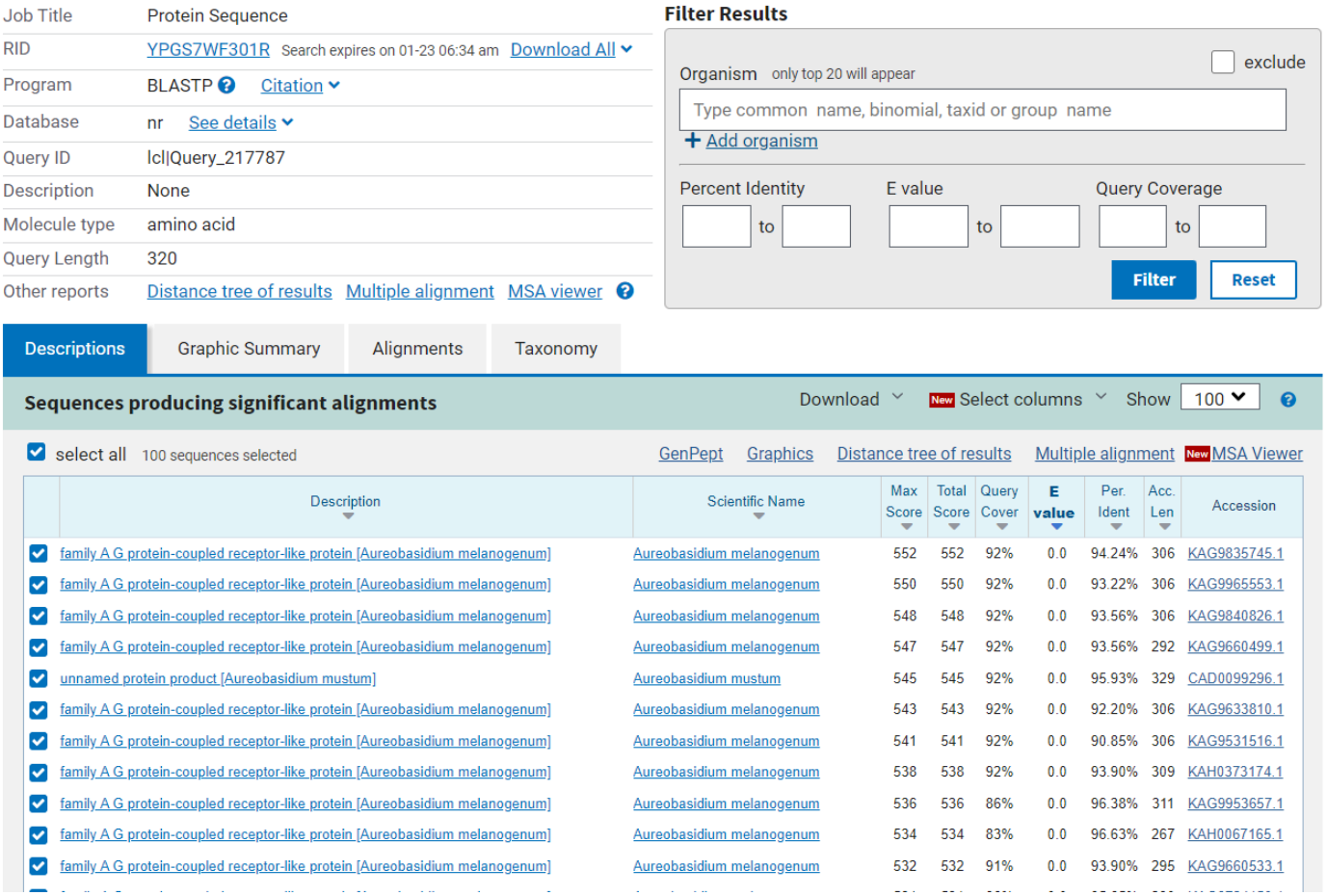
Note that both the top hits and bacteriorhodopsin are Class A G protein-coupled receptors (see below information from PDB on the original protein I began this search with), which is encouraging.

Domain Annotation: SCOP/SCOPe Classification — SCOP Database Homepage

Chains	Domain Info	Class	Fold	Superfamily	Family	Domain	Species	Provenance Source (Version)
A	d1fbba_	Membrane and cell surface proteins and peptides	Class A G protein-coupled receptor (GPCR)-like	Class A G protein-coupled receptor (GPCR)-like	Bacteriorhodopsin-like	Bacteriorhodopsin	(Halobacterium salinarum) [TaxId: 2242]	SCOPe (2.08)

Screenshot 9: Information about the Halobacterium salinarum bacteriorhodopsin from PDB

A Blast search for the longer potential *Gene X* ORF did not reveal any 100% percentage identity hits.



Job Title	Protein Sequence
RID	YPGS7WF301R Search expires on 01-23 06:34 am Download All
Program	BLASTP Citation
Database	nr See details
Query ID	lcl|Query_217787
Description	None
Molecule type	amino acid
Query Length	320
Other reports	Distance tree of results Multiple alignment MSA viewer

Filter Results

Organism only top 20 will appear — exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity to — E value to — Query Coverage to

Filter Reset

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments — Download — New Select columns — Show 100

select all 100 sequences selected — GenPept Graphics Distance tree of results Multiple alignment New MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
family A G protein-coupled receptor-like protein [Aureobasidium melanogenum]	Aureobasidium melanogenum	552	552	92%	0.0	94.24%	306	KAG9835745.1
family A G protein-coupled receptor-like protein [Aureobasidium melanogenum]	Aureobasidium melanogenum	550	550	92%	0.0	93.22%	306	KAG9965553.1
family A G protein-coupled receptor-like protein [Aureobasidium melanogenum]	Aureobasidium melanogenum	548	548	92%	0.0	93.56%	306	KAG9840826.1
family A G protein-coupled receptor-like protein [Aureobasidium melanogenum]	Aureobasidium melanogenum	547	547	92%	0.0	93.56%	292	KAG9660499.1
unnamed protein product [Aureobasidium mustum]	Aureobasidium mustum	545	545	92%	0.0	95.93%	329	CAD0099296.1
family A G protein-coupled receptor-like protein [Aureobasidium melanogenum]	Aureobasidium melanogenum	543	543	92%	0.0	92.20%	306	KAG9633810.1
family A G protein-coupled receptor-like protein [Aureobasidium melanogenum]	Aureobasidium melanogenum	541	541	92%	0.0	90.85%	306	KAG9531516.1
family A G protein-coupled receptor-like protein [Aureobasidium melanogenum]	Aureobasidium melanogenum	538	538	92%	0.0	93.90%	309	KAH0373174.1
family A G protein-coupled receptor-like protein [Aureobasidium melanogenum]	Aureobasidium melanogenum	536	536	86%	0.0	96.38%	311	KAG9953657.1
family A G protein-coupled receptor-like protein [Aureobasidium melanogenum]	Aureobasidium melanogenum	534	534	83%	0.0	96.63%	267	KAH0067165.1
family A G protein-coupled receptor-like protein [Aureobasidium melanogenum]	Aureobasidium melanogenum	532	532	91%	0.0	93.90%	295	KAG9660533.1

Screenshot 10: blastp results using the longer predicted ORF

Percentage identity is decreased in this test, with no 100% alignments. However, query coverage drops as well, suggesting that this expanded ORF may be erroneous. In either case, the sequence I have found appears to be that of a novel protein with homology to bacteriorhodopsin and family A G protein-coupled receptor-like proteins.

Question 5:

Without further information, it seems appropriate to continue investigating both the shorter and longer predicted potential ORFs for *gene X*, below are the sequences I will align these too. The sequences have been renamed for convenience in the following format *genus.species - accession-number protein-name*.

>Halobacterium.salinarum - WP_136361479.1 bacteriorhodopsin
MTPSLLPTAVEGVSQAQITGRPEWIWLALGTALMGLGTLYFLVKGMGVSDPDAKKFYAITTLVPAIAFTMYLSMLLG
YGLTMVPFGGEQNPIYWARYADWLFTTPLLLLDLALLVDADQGTILALVGADGIMIGTGLVGALTKVYSYRFVWWAI
STAAMLYILYVLFFGFTSKAESMRPEVASTFKVLRNVTVVLWSAYPVVWLIGSEGAGIVPLNIETLLFMVLDVSAKV
GFGLILLRSRAIFGEAEAPEPSAGDGAAATSD

>Aureobasidium.melanogenum CBS 110374 - XP_040878626.1 bacteriorhodopsin
MSWLEKRNDAIQVNPNTQNNKHVDIAITVRGSDFYFAICAVMGFVALGTIAASAMKPRTDRIFFYITAAINLTACIA
YFTMGSNLGWTPIDVEFPRTWSKVAGVNREIFYARYVDWFVTTPLLLMDLLLTAGLPWPTILYTIFLDEVMIVTGLI
GALVKSRYKWGFWTFGTVAMFAIFWNLAVEGRKHAKHLGSDVYRCYLMCGVLTLFVWLCYPICWGVSEGGNVIPPDS
EAVFYGVLDFLAKPCFSIALIAGHWNINPGRMGLKLRDYDEEPAYFGPKNGAEAAKERGRTDNAVDGVA

>Aureobasidium.pullulans EXF-150 - KEQ89910.1 bacteriorhodopsin
MSWLEKRNDAIQVNPNTQNNKHVDIAITVRGSDFYFAICAVMGFVALGVMAASAMKPRTDRIFFYITAAINTTACIA
YFAMGSNLGWTPIDVEWQRTWSQVAGVNREVFYVRYIDWFVTTPLLLMDLLLTAGLPWPTILWTIFLDEVMIVTGLV
GALVKSRYKWGFWTFGTVAMFAIFWNLAVEGRKHAKHLGSDIARTYTICGCLTLFIWLCYPICWGVSEGANVIPPDS
EAVFYGVLDFLAKPVFSIALIIGHWNINPGRMGLKLRDYDEDPDYFGPKNGAEAAKERSNGSSSGVDGGA

>Thermus.thermophilus - WP_165739012.1 bacteriorhodopsin
MRMLPELSFGEYWLVFNMLSLTIAGMFAAFVFFLLARSYVAPRYHIALYLSALIVFIAGYHYLRIFESWVGAYQLQG
GVYVPTGKPFNDFYRYADWLLTVPLLLLELILVLGLSPARTWNLGVKLVVASVLMLGLGYVGEANTEPGPRTLWGAL
SSVPFFYILYVLWVELGQAIRETRFGPRVLELLTAIRYVLLMSWGFYPIAYALGTWLPGGAAQEVVIQLGYSLADLI
AKPVYGLLIFAIARAKSLEEGFGEGVKAA

>Cyclobacterium.plantarum - WP_166144155.1 bacteriorhodopsin
MYFVLLAEIIGLDRIINSDPVAITFFIGYMAMFASAVFFFAERASVDGKWKTSLLVSGLITGIAAIHYYYMRDFYLQ
TGSSPTAFRYVDWTLTVPLMCVEFYLLTKPFGAKGATLTKLIIASLVMLVTGYIGETSGLDNNIFWGIVSTLGYLYI
VYEVFAGDVAKLSQSSDSPALKKAMFLLKIFITLGWSIYPIGYMVLPGNLLSGLFEVSSIDLFYNLADAINKIGFGL
VIYSVAIKESKKTTAQVA

>Ningiella.ruwaisensis - WP_168710959.1 bacteriorhodopsin
MENAMSSISVEGFEIVNHILTLGYATMAAALLFFILTRKDSLPKYQMSSILSVVVMVSALLLLYTQKISWTEAYAFD
GNEYTVREGADLFTNGYRYLNWLIDVPMLLIQILWVAQITGSQRTSYMFKFSFSGCLMILTGYIGQFYEPGRINEDV
TLWAVWGLISTAFFLHVLVLITRVIKEGSSKMSGGAKSVFSAILPLFLISWWLYPIAYLAPYFMTMGYSYETTIVSQ
QVIYTIADISSKVVYGVMLTVTATMLSKQEGMAESQA

>Halorubrum.sodomense - WP_211553476.1 bacteriorhodopsin
MDPIALQAGYDLLGDGRPETLWLGIGTLMMLIGTFYFIVKGWGVTDKEAREYYAITILVPGIASAAYLSMFFGIGLT
EVELVGGEVLDIYYARYADWLFTTPLLLLDLALLAKVDRVTIGTLVGVDALMIVTGLIGALSKTPLARYTWWLFSTI
AFLFVLYYLLTSLRSAAKQRSAEVQSTFNTLTALVAVLWTAYPILWIVGTEGAGVVGLGIETLAFMILDVTAKVGFG
FVLLRSRAILGDTEAPEPSAGADAQAAD

>Ktedonobacter.robiniae - WP_201374874.1 bacteriorhodopsin
MDSATVTVLWVTSLIMILCTLVFTYRSFRARIEIKHFYYLTALITLIAATLYMTMASGYGGIGLNGKVILFGRYIDW
VITTPLLLMNLALIALPRNFPSRFAVIGTMIAADVYMIVSGLGASLIRSNFRWAFFAVSCAGFLAVLYFIIVKLTPE
ANVRSGPVQRHYSTLAIMLIALWVCYPIVWILGTEGFGIISLLPEVILYAILDVLAKGAFGFVLLSKPGVLLEAERE
TAPINSVAAQW

>Bacillus.coahuilensis - WP_059282687.1 bacteriorhodopsin
MISILHYGYSFIMLLGALYFYLLSKDPKGVPASEYLIAMVIPLWSGAAYLSIALGQGLFQYDDTTIYYARYIDWVIS
TPLLLAALALTAMFGGKKNLTLLFSLVALDVFMIITGFVADLSIGTTKYIWYSLGVIALIIILVITFGPLRRIALSN
GTRLARHYTRVAIYLSALWVCYPTAWLLGPSGLGLAQELTEVLVFIILPIFSKVGFSIVDLHGLRKLHQSSYVHN

>Alkalihalobacillus.hwajinpoensis - WP_169525176.1 bacteriorhodopsin
MNSFEIFLYYFYFVVMISAAIYFFILSRKPKGVPLYEYVVAMMITAWSGVAYLSIALGQGFIERPEKTIYFARYLDW
VVSTPLLVLSLALTAMFYETKKNKVLIASIMATDVFMILTGLIADFSPDSLKYIWYSLGVIALFIILLITWIPLKRI
ADRHEQLSKHYKRVALYLTIFWLLYPTAWILGASGIGMTQGIIDTLAFVILPIFSKIGFGLLDLHGLRKLKTN

>Alteribacter.aurantiacus - WP_169720904.1 bacteriorhodopsin
MYEIEQQLLWIYVAFMGGGAVYFAYLAFHRKGVPRAEYLVAFIIPTWSGVAYASIALGQGLVEWGDRVIYFARYLDW
VVTTPLLLLALAMTAMYTISKDRVIIGGLIVADVFMVLTGLIAEFSPSPIKYVWYILGVVAFLIILWIIWWPLRAKA
KSQNHYVYRVFLIVAGYLSILWVGYPTVWLLGPSGLGVISQITDQALFVSLPIFSKVGFSILDLSCLRWLHVKHGQE
VTPQAT

>Halogranum.amylolyticum - WP_089823426.1 bacteriorhodopsin
MVTVGAESLWLWIGTLGMTIGTLYFVGRGRGVTDKKMQEFYIITIFITTIAAAMYLLMATGFGLTQVQVGNRTLDIY
WARYADWVFTTPLLLLDLALLAGANRNTIATLVGLDVFMIATGLIAALEPNATYRIMWWGISTGALLALLYILVGTL
SKQVETRDAEVQSLFSTLRNLTMVLWLLYPVVWILGTEGTIGILPLYWETAAFMVLDLSAKVGFGFLLLRSRAVLEK
ASTPTAAATA

>Salinigranum.halophilum - WP_136590783.1 bacteriorhodopsin
MATPGAESIWLWLGTAGMTLGTFYFIARGWGVEDEEQQRFYLITIFITAIASAAYFAMATGFGLTQVTVNGQVLDIY
WGRYADWLFTTPLLLLDLALLARASKNTIYTLVGLDVLMIGTGVIGALAASSAFIRIVWWAISTVFLLFLLYFLIRT
LSEAATRQSPEVRKLTTTLRNMLIVLWLAYPVVWILGTEGTIGIIPLYWETAAFMVLDLTAKVGFGFVLLRSHSVLE
AATQSTTAGATAD

>Natrinema.pallidum - WP_138652685.1 bacteriorhodopsin
MAATVGPESIWLWIGTIGMTLGTLYFVGRGRGVRDRKMQEFYIITIFITTIAAAMYFAMATGFGVTEVMVGDEALTI
YWARYADWLFTTPLLLLDLSLLAGANRNTIATLIGLDVFMIGTGAIAALSSTPGTRIAWWAISTGALLALLYVLVGT
LSENARNRAPEVASLFGRLRNLVIALWFLYPVVWILGTEGTFGILPLYWETAAFMVLDLSAKVGFGVILLQSRSVLE
RVATPTAAPT

>Salinigranum.salinum - WP_152039473.1 bacteriorhodopsin
MATPGAGLESISLWIGTIGMTLGTLYFVAQGWSVRDPDQQEYYIITIFIPAIAAASYFAMASGFGLVEVPVEGLGTL
DIYWARYADWLFTTPLLLLDLALLAGADRNTIYTLVGLDVFMIVTGLVGALAREGQVFRIIWWAISTGALLVLLYFL
LGSLSEQASRQAGEVGALFSRLRNLILVLWSAYPVVWILGTEGGFAIIPLGVETAAFMVLDLSAKVGFGFILLQSRD
VLSAAKSTGASATAD

>Haloplanus.rallus - WP_157687740.1 bacteriorhodopsin
MTQPGSESLWLWLGTAGMLIGMLYFIARGWGEKNRRRQEFYIVTIFITAIAFVNYLSMALGFGLTTIEIGGEELPIY
WARYTDWLFTTPLLLIDLGLLAGANRNQLSTLVGLDVLMIGTGAVATLSTAGVLLSPVGDRIIWWGVSTGFLLVLLY
FLFGTLTKEASQLSGAARSTFSTLRNLIVVVWLVYPVWWILGTEGLGVISLYSETAGFMVLDLVAKVGFGIILLSSR
DVLDAAGDTTGAALGDADPTD

>Natronomonas.gomsonensis - WP_178916035.1 bacteriorhodopsin
MADPGSEALWLWIGTAGMFLGMLYFIARGWGEENRRRQEFYIVTIFITAIAFVNYLMMALGFGLTTVTVAGEELPIY
WARYTDWLFTTPLLLIDLGLLAGANRNQIATLVGLDALMIGTGAVATLSTTGVLLSPVGDRLIWWGVSTGFLLVLLY
FLFGTLTEEANRLSDDAQSTFRTLRNLIVVVWLVYPVWWILGTEGLGTIGLYSETAGFMVLDLVAKVGFGIILLSSR
EVLDAAGDLAGSTAQPADD

>Halorubrum.trapanicum - WP_209546814.1 bacteriorhodopsin
MDPIALTAGYDLLGDGRPETLWLGIGTLLMLLGTFYFIARGWGVTDKEAREYYAITILVPGIASAAYLSMFFGIGLT
EVQVGGEMLDIYYARYADWLFTTPLLLLDLALLAKVDRVTIGTLVGVDALMIVTGLVGALSHTAVARYSWWLFSTIC
MIVVLYFLATSLRSAAKQRSADVQSTFNTLTALVLVLWTAYPILWIIGTEGAGVVGLGIETLLFMVLDVTAKVGFGF
ILLRSRAILGDTGAPEPSAGAEASAAD

>Halomicroarcula.salinisoli - WP_220588862.1 bacteriorhodopsin
MPQPGSEQIWLWLGTAGMFLGMLYFIGRGWGETDDRRQKFYIATILITAIAFVNYLAMALGFGLTIIELPNDPEAPI
YWARYTDWLFTTPLLLYDLALLAGADRNTISTLVSLDVLMIGTGVVATLSAGSGVLAAGAERLIWWGISTAFLLVLL
YFLFSSLSSRVTDLPSDTQGTFRTLRNLVAVVWLVYPVWWLIGTEGLALVGIFTETAGFMVIDLVAKVGFGFILLRS
HSVLDGAAQSQTTGASPADD

>Haloarcula.mannanilytica - WP_137682956.1 bacteriorhodopsin
MPEPGSEAIWLWLGTAGMFLGMLYFIGRGWGETDSRRQKFYIATILITAIAFVNYLAMALGFGLTIIEFGGSEHPIY
WARYTDWLFTTPLLLYDLGLLAGADRNTIASLVSLDVLMIGTGVVATLSAGSGVLSAGAERLVWWGISTAFLLVLLY
FLFSSLSGRVADLPSDTRSTFKTLRNLVTVVWLVYPVWWLVGTEGLGLVGIGIETAGFMVIDLTAKVGFGIILLRSH
GVLDGAAETTSTGATPADD

A multiple alignment for the shorter predicted ORF is below. The alignment was created using MUSCLE from EMBL-EBI with default parameters.

Results for job muscle-I20220204-171706-0870-8182256-p2m

Alignments | Result Summary | Phylogenetic Tree | Results Viewers | Submission Details

Program	Number of Sequences	Launched Date	Input Sequences
MUSCLE	21	Fri, Feb 04, 2022 at 17:17:07	muscle-I20220204-171706-0870-8182256-p2m.input
Version		End Date	Output Result
3.8.425		Fri, Feb 04, 2022 at 17:17:08	muscle-I20220204-171706-0870-8182256-p2m.output

Command

```
$APPBIN/muscle:3.8.425 /muscle -in muscle-I20220204-171706-0870-8182256-p2m.upfile -verbose -log muscle-I20220204-171706-0870-8182256-p2m.output -quiet -clw -out muscle-I20220204-171706-0870-8182256-p2m.clw
```

Input Parameters

Alignment format

clw

Output Tree

none

Screenshot 11: MUSCLE parameters for the shorter predicted Gene X ORF

Percent Identity Matrix - created by Clustal2.1

1: Ningiella.ruwaisensis	100.00	19.23	25.70	19.65	15.49	18.18	21.23	13.89	12.70	19.21	20.16	21.49	23.14	23.14	20.68	22.18	20.68	21.01	21.63	22.31	20.75
2: Cyclobacterium.plantarum	19.23	100.00	25.33	19.20	17.81	17.59	26.09	19.75	18.91	20.98	21.78	24.44	24.44	23.56	23.56	23.25	23.89	24.67	26.36	24.47	22.78
3: Thermus.thermophilus	25.70	25.33	100.00	22.71	21.88	22.94	25.99	18.62	19.03	22.71	23.21	24.15	23.73	22.88	27.23	27.43	25.11	25.96	25.94	25.32	24.58
4: Alteribacter.aurantiacus	19.65	19.20	22.71	100.00	51.53	52.21	27.98	25.86	23.71	30.08	30.00	30.00	32.61	31.30	33.19	32.76	33.62	32.89	30.47	33.91	32.19
5: Bacillus.coahuilensis	15.49	17.81	21.88	51.53	100.00	58.30	30.95	24.55	24.55	30.70	31.11	31.11	33.78	31.56	31.25	32.44	33.33	32.89	30.22	32.44	32.44
6: Alkalihalobacillus.hwajinpoensis	18.18	17.59	22.94	52.21	58.30	100.00	28.14	21.72	19.46	27.88	27.85	29.68	31.05	30.59	30.28	29.41	34.25	32.42	26.58	31.53	29.28
7: GeneX	21.23	26.09	25.99	27.98	30.95	28.14	100.00	64.09	64.09	36.90	30.64	31.98	31.98	31.98	34.88	36.42	33.72	33.72	33.72	36.42	36.63
8: Aureobasidium.melanogenum	13.89	19.75	18.62	25.86	24.55	21.72	64.09	100.00	89.67	28.39	24.69	26.03	27.46	27.27	29.17	28.51	28.57	28.15	26.05	28.79	29.30
9: Aureobasidium.pullulans	12.70	18.91	19.03	23.71	24.55	19.46	64.09	89.67	100.00	25.85	23.87	25.21	26.64	26.45	27.50	28.10	27.31	27.31	25.29	27.63	28.91
10: Ktedonobacter.robiniae	19.21	20.98	22.71	30.08	30.70	27.88	36.90	28.39	25.85	100.00	33.05	32.19	36.32	36.05	41.13	38.36	39.74	39.47	33.76	35.44	35.02
11: Halomicroarcula.salinisoli	20.16	21.78	23.21	30.00	31.11	27.85	30.64	24.69	23.87	33.05	100.00	87.60	67.20	69.20	60.08	55.14	55.83	54.81	48.77	47.35	49.59
12: Haloarcula.mannanilytica	21.49	24.44	24.15	30.00	31.11	29.68	31.98	26.03	25.21	32.19	87.60	100.00	69.20	70.40	59.26	54.55	56.25	55.65	50.82	47.95	50.00
13: Haloplanus.rallus	23.14	24.44	23.73	32.61	33.78	31.05	31.98	27.46	26.64	36.32	67.20	69.20	100.00	86.00	59.26	60.74	61.67	59.41	47.97	49.59	51.63
14: Natronomonas.gomsonensis	23.14	23.56	22.88	31.30	31.56	30.59	31.98	27.27	26.45	36.05	69.20	70.40	86.00	100.00	60.49	59.92	60.00	60.67	47.95	49.18	53.69
15: Salinigranum.halophilum	20.68	23.56	27.23	33.19	31.25	30.28	34.88	29.17	27.50	41.13	60.08	59.26	59.26	60.49	100.00	67.77	67.50	66.95	52.48	55.37	54.96
16: Salinigranum.salinum	22.18	23.25	27.43	32.76	32.44	29.41	36.42	28.51	28.10	38.36	55.14	54.55	60.74	59.92	67.77	100.00	68.88	70.42	53.09	54.10	55.56
17: Halogranum.amylolyticum	20.68	23.89	25.11	33.62	33.33	34.25	33.72	28.57	27.31	39.74	55.83	56.25	61.67	60.00	67.50	68.88	100.00	77.50	52.50	56.67	56.25
18: Natrinema.pallidum	21.01	24.67	25.96	32.89	32.89	32.42	33.72	28.15	27.31	39.47	54.81	55.65	59.41	60.67	66.95	70.42	77.50	100.00	52.08	51.88	51.46
19: Halobacterium.salinarum	21.63	26.36	25.94	30.47	30.22	26.58	33.72	26.05	25.29	33.76	48.77	50.82	47.97	47.95	52.48	53.09	52.50	52.08	100.00	56.59	56.98
20: Halorubrum.sodomense	22.31	24.47	25.32	33.91	32.44	31.53	36.42	28.79	27.63	35.44	47.35	47.95	49.59	49.18	55.37	54.10	56.67	51.88	56.59	100.00	89.15
21: Halorubrum.trapanicum	20.75	22.78	24.58	32.19	32.44	29.28	36.63	29.30	28.91	35.02	49.59	50.00	51.63	53.69	54.96	55.56	56.25	51.46	56.98	89.15	100.00

Screenshot 12: Percent Identity Matrix provided with MUSCLE output

```
CLUSTAL multiple sequence alignment by MUSCLE (3.8)


Ningiella.ruwaisensis                -------MENAMSSISVEGFE-------IVNHILTLGYATMAAALL-----FFILTRKDS
Cyclobacterium.plantarum             ---------MYFVLLAEIIGLDRIINSDPVAITFFIGYMAMFASAV-----FFFAERASV
Thermus.thermophilus                 --------MRMLPELS------------FGEYWLVFNMLSLTIAGMFAAFVFFLLARSYV
Alteribacter.aurantiacus             -------------------------MYEIEQQLLWIYVAFMGGGAVY----FAYLAFHRK
Bacillus.coahuilensis                ----------------------------MISILHYGYSFIMLLGALY----FYLLSKDPK
Alkalihalobacillus.hwajinpoensis     -------------------------MNSFEIFLYYFYFVVMISAAIY----FFILSRKPK
GeneX                                ------------------------------------------------------------
Aureobasidium.melanogenum            MSWLEKRNDAIQVNPNTQNNKHVDIAITVRGSDFYFAICAVMGFVAL----GTIAASAMK
Aureobasidium.pullulans              MSWLEKRNDAIQVNPNTQNNKHVDIAITVRGSDFYFAICAVMGFVAL----GVMAASAMK
Ktedonobacter.robiniae               -------------------------MDSATVTVLWVTSLIMILCTLV----FTYRSFRAR
Halomicroarcula.salinisoli           -----------MPQPG------------SEQIWLWLGTAGMFLGMLY----FIGRGWGET
Haloarcula.mannanilytica             -----------MPEPG------------SEAIWLWLGTAGMFLGMLY----FIGRGWGET
Haloplanus.rallus                    -----------MTQPG------------SESLWLWLGTAGMLIGMLY----FIARGWGEK
Natronomonas.gomsonensis             -----------MADPG------------SEALWLWIGTAGMFLGMLY----FIARGWGEE
Salinigranum.halophilum              -----------MATPG------------AESIWLWLGTAGMTLGTFY----FIARGWGVE
Salinigranum.salinum                 -----------MATPG----------AGLESISLWIGTIGMTLGTLY----FVAQGWSVR
Halogranum.amylolyticum              -----------MVTVG------------AESLWLWIGTLGMTIGTLY----FVGRGRGVT
Natrinema.pallidum                   ----------MAATVG------------PESIWLWIGTIGMTLGTLY----FVGRGRGVR
Halobacterium.salinarum              -------MTPSLLPTAVEGVSQAQITGRPEWIWLALGTALMGLGTLY----FLVKGMGVS
Halorubrum.sodomense                 -----------MDPIALQAGYDLLGDGRPETLWLGIGTLMMLIGTFY----FIVKGWGVT
Halorubrum.trapanicum                -----------MDPIALTAGYDLLGDGRPETLWLGIGTLLMLLGTFY----FIARGWGVT


Ningiella.ruwaisensis                LPKYQMSSILSVVVMVSALLLLYTQKI-----SWTEAYAFDGNEYTVREGADLFTNGYRY
Cyclobacterium.plantarum             DGKWKTSLLVSGLITGIAAIHYYYMRDFYLQTGSSPTAF-------------------RY
Thermus.thermophilus                 APRYHIALYLSALIVFIAGYHYLRIFE-----SWVGAYQLQGGVYVPTGKPFNDFY--RY
Alteribacter.aurantiacus             GVPRAE-YLVAFIIPTWSGVAYASIALGQ---GLVEWG-------------DRVIYFARY
Bacillus.coahuilensis                GVPASE-YLIAMVIPLWSGAAYLSIALGQ---GLFQYD-------------DTTIYYARY
Alkalihalobacillus.hwajinpoensis     GVPLYE-YVVAMMITAWSGVAYLSIALGQ---GFIERP-------------EKTIYFARY
GeneX                                --------------MVAAIAYFSMGSNL---GWTPIDVEFRRNDPVVRGINREIFYVRY
Aureobasidium.melanogenum            PRTDRIFFYITAAINLTACIAYFTMGSNL---GWTPIDVEFPRTWSKVAGVNREIFYARY
Aureobasidium.pullulans              PRTDRIFFYITAAINTTACIAYFAMGSNL---GWTPIDVEWQRTWSQVAGVNREVFYVRY
Ktedonobacter.robiniae               -IEIKHFYYLTALITLIAATLYMTMASGY---GGIGLN-------------GKVILFGRY
Halomicroarcula.salinisoli           DDRRQKFYIATILITAIAFVNYLAMALGF---GLTIIEL--------PNDPEAPIYWARY
Haloarcula.mannanilytica             DSRRQKFYIATILITAIAFVNYLAMALGF---GLTIIEF---------GGSEHPIYWARY
Haloplanus.rallus                    NRRRQEFYIVTIFITAIAFVNYLSMALGF---GLTTIEI---------GGEELPIYWARY
Natronomonas.gomsonensis             NRRRQEFYIVTIFITAIAFVNYLMMALGF---GLTTVTV---------AGEELPIYWARY
Salinigranum.halophilum              DEEQQRFYLITIFITAIASAAYFAMATGF---GLTQVTV---------NGQVLDIYWGRY
Salinigranum.salinum                 DPDQQEYYIITIFIPAIAAASYFAMASGF---GLVEVPV--------EGLGTLDIYWARY
Halogranum.amylolyticum              DKKMQEFYIITIFITTIAAAMYLLMATGF---GLTQVQV---------GNRTLDIYWARY
Natrinema.pallidum                   DRKMQEFYIITIFITTIAAAMYFAMATGF---GVTEVMV---------GDEALTIYWARY
Halobacterium.salinarum              DPDAKKFYAITTLVPAIAFTMYLSMLLGY---GLTMVPF---------GGEQNPIYWARY
Halorubrum.sodomense                 DKEAREYYAITILVPGIASAAYLSMFFGI---GLTEVEL--------VGGEVLDIYYARY
Halorubrum.trapanicum                DKEAREYYAITILVPGIASAAYLSMFFGI---GLTEVQV---------GGEMLDIYYARY
                                                          :             .                      **
```

```
Ningiella.ruwaisensis                LNWLIDVPMLLIQILWVAQITG-SQRTSYMFKFSFSGCLMILTGYIGQFYEPGRINEDVT
Cyclobacterium.plantarum             VDWTLTVPLMCVEFYLLTKPF--GAKGATLTKLIIASLVMLVTGYIGETSGLDN------
Thermus.thermophilus                 ADWLLTVPLLLLELILVLGLSP-ARTWNLGVKLVVASVLMLGLGYVGEANTEPGP-----
Alteribacter.aurantiacus             LDWVVTTPLLLLALAMTAMYTI-SKDRVIIGGLIVADVFMVLTGLIAEFSPSPI------
Bacillus.coahuilensis                IDWVISTPLLLAALALTAMFGG-KKNLTLLFSLVALDVFMIITGFVADLSIGTT------
Alkalihalobacillus.hwajinpoensis     LDWVVSTPLLVLSLALTAMFYETKKNKVLIASIMATDVFMILTGLIADFSPDSL------
GeneX                                VDWFITTPLLLMDLLLTA-----AMPWPTVLFVVLVDEVMIVTGLVGALVRSSY------
Aureobasidium.melanogenum            VDWFVTTPLLLMDLLLTA-----GLPWPTILYTIFLDEVMIVTGLIGALVKSRY------
Aureobasidium.pullulans              IDWFVTTPLLLMDLLLTA-----GLPWPTILWTIFLDEVMIVTGLVGALVKSRY------
Ktedonobacter.robiniae               IDWVITTPLLLMNLALIALPRNFPSRFAVIGTMIAADVYMIVSGLGASLIRSNF------
Halomicroarcula.salinisoli           TDWLFTTPLLLYDLALLA-----GADRNTISTLVSLDVLMIGTGVVATLSAGSGVLAAGA
Haloarcula.mannanilytica             TDWLFTTPLLLYDLGLLA-----GADRNTIASLVSLDVLMIGTGVVATLSAGSGVLSAGA
Haloplanus.rallus                    TDWLFTTPLLLIDLGLLA-----GANRNQLSTLVGLDVLMIGTGAVATLSTAGVLLSPVG
Natronomonas.gomsonensis             TDWLFTTPLLLIDLGLLA-----GANRNQIATLVGLDALMIGTGAVATLSTTGVLLSPVG
Salinigranum.halophilum              ADWLFTTPLLLLDLALLA-----RASKNTIYTLVGLDVLMIGTGVIGALAASSAFI----
Salinigranum.salinum                 ADWLFTTPLLLLDLALLA-----GADRNTIYTLVGLDVFMIVTGLVGALAREGQVF----
Halogranum.amylolyticum              ADWVFTTPLLLLDLALLA-----GANRNTIATLVGLDVFMIATGLIAALEPNATY-----
Natrinema.pallidum                   ADWLFTTPLLLLDLSLLA-----GANRNTIATLIGLDVFMIGTGAIAALSSTPGT-----
Halobacterium.salinarum              ADWLFTTPLLLLDLALLV-----DADQGTILALVGADGIMIGTGLVGALTKVYSY-----
Halorubrum.sodomense                 ADWLFTTPLLLLDLALLA-----KVDRVTIGTLVGVDALMIVTGLIGALSKTPLA-----
Halorubrum.trapanicum                ADWLFTTPLLLLDLALLA-----KVDRVTIGTLVGVDALMIVTGLVGALSHTAVA-----
                                      :* . .*::   :                     .  *:  *  .

Ningiella.ruwaisensis                LWAVWGLISTAFFLHVLVLITRVIKEGSS--KMSGGAKSVFSAILPLFLISWWLYPIAYL
Cyclobacterium.plantarum             -NIFWGIVSTLGYLYIVYEVFAGDVAKLSQSSDSPALKKAMFLLKIFITLGWSIYPIGYM
Thermus.thermophilus                 -RTLWGALSSVPFFYILYVLWVELGQAIRETRFGPRVLELLTAIRYVLLMSWGFYPIAYA
Alteribacter.aurantiacus             -KYVWYILGVVAFLIILWIIWWPLRAKAK--SQNHYVYRVFLIVAGYLSILWVGYPTVWL
Bacillus.coahuilensis                -KYIWYSLGVIALIIILVITFGPLRRIAL--SNGTRLARHYTRVAIYLSALWVCYPTAWL
Alkalihalobacillus.hwajinpoensis     -KYIWYSLGVIALFIILLITWIPLKRIAD---RHEQLSKHYKRVALYLTIFWLLYPTAWI
GeneX                                -KWGYFAFGCAALFYVVFVLVWEARRHAN--ALGSDVGKAFTICGSLTTFLWILYPIAWG
Aureobasidium.melanogenum            -KWGFWTFGTVAMFAIFWNLAVEGRKHAK--HLGSDVYRCYLMCGVLTLFVWLCYPICWG
Aureobasidium.pullulans              -KWGFWTFGTVAMFAIFWNLAVEGRKHAK--HLGSDIARTYTICGCLTLFIWLCYPICWG
Ktedonobacter.robiniae               -RWAFFAVSCAGFLAVLYFIIVKLTPEAN--VRSGPVQRHYSTLAIMLIALWVCYPIVWI
Halomicroarcula.salinisoli           ERLIWWGISTAFLLVLLYFLFSSLSSRVT--DLPSDTQGTFRTLRNLVAVVWLVYPVWWL
Haloarcula.mannanilytica             ERLVWWGISTAFLLVLLYFLFSSLSGRVA--DLPSDTRSTFKTLRNLVTVVWLVYPVWWL
Haloplanus.rallus                    DRIIWWGVSTGFLLVLLYFLFGTLTKEAS--QLSGAARSTFSTLRNLIVVVWLVYPVWWI
Natronomonas.gomsonensis             DRLIWWGVSTGFLLVLLYFLFGTLTEEAN--RLSDDAQSTFRTLRNLIVVVWLVYPVWWI
Salinigranum.halophilum              -RIVWWAISTVFLLFLLYFLIRTLSEAAT--RQSPEVRKLTTTLRNMLIVLWLAYPVVWI
Salinigranum.salinum                 -RIIWWAISTGALLVLLYFLLGSLSEQAS--RQAGEVGALFSRLRNLILVLWSAYPVVWI
Halogranum.amylolyticum              -RIMWWGISTGALLALLYILVGTLSKQVE--TRDAEVQSLFSTLRNLTMVLWLLYPVVWI
Natrinema.pallidum                   -RIAWWAISTGALLALLYVLVGTLSENAR--NRAPEVASLFGRLRNLVIALWFLYPVVWI
Halobacterium.salinarum              -RFVWWAISTAAMLYILYVLFFGFTSKAE--SMRPEVASTFKVLRNVTVVLWSAYPVVWL
Halorubrum.sodomense                 -RYTWWLFSTIAFLFVLYYLLTSLRSAAK--QRSAEVQSTFNTLTALVAVLWTAYPILWI
Halorubrum.trapanicum                -RYSWWLFSTICMIVVLYFLATSLRSAAK--QRSADVQSTFNTLTALVLVLWTAYPILWI
                                      .  :  ..    : :.                                  *  **  :
```

```
Ningiella.ruwaisensis                APYFMT--MGYS-YETTIVSQQVIYTIADISSKVVYGVML-------------TVTATML
Cyclobacterium.plantarum             ---VLP--GNLLSGLFEVSSIDLFYNLADAINKIGFGLV--------------IYSVAIK
Thermus.thermophilus                 ---LGTWLPG---GAAQEVVIQLGYSLADLIAKPVYGLLI-------------FAIARAK
Alteribacter.aurantiacus             ---LGP--SG-L-GVISQITDQALFVSLPIFSKVGFSILD-------------LSCLRWL
Bacillus.coahuilensis                ---LGP--SG-L-GLAQELTEVLVFIILPIFSKVGFSIVD-------------LHGLRKL
Alkalihalobacillus.hwajinpoensis     ---LGA--SG-I-GMTQGIIDTLAFVILPIFSKIGFGLLD-------------LHGLRKL
GeneX                                ---LCE--GG---NLISPDSEAIFYGILDLLAKPVFGALL-------------I------
Aureobasidium.melanogenum            ---VSE--GG---NVIPPDSEAVFYGVLDFLAKPCFSIALIAGHWNINPGRMGLKLRDYD
Aureobasidium.pullulans              ---VSE--GA---NVIPPDSEAVFYGVLDFLAKPVFSIALIIGHWNINPGRMGLKLRDYD
Ktedonobacter.robiniae               ---LGT--EG-F-GIISLLPEVILYAILDVLAKGAFGFVL-------------LSKPGVL
Halomicroarcula.salinisoli           ---IGT--EG-L-ALVGIFTETAGFMVIDLVAKVGFGFIL-------------LRSHSVL
Haloarcula.mannanilytica             ---VGT--EG-L-GLVGIGIETAGFMVIDLTAKVGFGIIL-------------LRSHGVL
Haloplanus.rallus                    ---LGT--EG-L-GVISLYSETAGFMVLDLVAKVGFGIIL-------------LSSRDVL
Natronomonas.gomsonensis             ---LGT--EG-L-GTIGLYSETAGFMVLDLVAKVGFGIIL-------------LSSREVL
Salinigranum.halophilum              ---LGT--EGTI-GIIPLYWETAAFMVLDLTAKVGFGFVL-------------LRSHSVL
Salinigranum.salinum                 ---LGT--EGGF-AIIPLGVETAAFMVLDLSAKVGFGFIL-------------LQSRDVL
Halogranum.amylolyticum              ---LGT--EGTI-GILPLYWETAAFMVLDLSAKVGFGFLL-------------LRSRAVL
Natrinema.pallidum                   ---LGT--EGTF-GILPLYWETAAFMVLDLSAKVGFGVIL-------------LQSRSVL
Halobacterium.salinarum              ---IGS--EG-A-GIVPLNIETLLFMVLDVSAKVGFGLIL-------------LRSRAIF
Halorubrum.sodomense                 ---VGT--EG-A-GVVGLGIETLAFMILDVTAKVGFGFVL-------------LRSRAIL
Halorubrum.trapanicum                ---IGT--EG-A-GVVGLGIETLLFMVLDVTAKVGFGFIL-------------LRSRAIL
                                        .                   :      *  :.

Ningiella.ruwaisensis                SKQEGMAESQA-------------------
Cyclobacterium.plantarum             ESKKTTAQVA--------------------
Thermus.thermophilus                 SLEEGFGEGVKAA-----------------
Alteribacter.aurantiacus             HVKHGQEVTPQAT-----------------
Bacillus.coahuilensis                HQSSYVHN----------------------
Alkalihalobacillus.hwajinpoensis     KTN---------------------------
GeneX                                ------------------------------
Aureobasidium.melanogenum            EEPAYFGPKNGAEAAKERGR-TDNAVDGVA
Aureobasidium.pullulans              EDPDYFGPKNGAEAAKERSNGSSSGVDGGA
Ktedonobacter.robiniae               LEAERETAPINSVAAQW-------------
Halomicroarcula.salinisoli           DGAAQSQTTGASPADD--------------
Haloarcula.mannanilytica             DGAAETTSTGATPADD--------------
Haloplanus.rallus                    DAAGDTTGAALGDADPTD------------
Natronomonas.gomsonensis             DAAGDLAGSTAQPADD--------------
Salinigranum.halophilum              -EAATQSTTAGATAD---------------
Salinigranum.salinum                 SAAKSTGASATAD-----------------
Halogranum.amylolyticum              EKASTPTAAATA------------------
Natrinema.pallidum                   ERVATPTAAPT-------------------
Halobacterium.salinarum              GEAEAPEPSAGDGAAATSD-----------
Halorubrum.sodomense                 GDTEAPEPSAGADAQAAD------------
Halorubrum.trapanicum                GDTGAPEPSAGAEASAAD------------
```

Screenshot 13: Multiple Sequence Alignment for shorter Gene X ORF provided by MUSCLE

The same process was repeated with exactly the same parameters for the longer predicted ORF.

Results for job muscle-I20220204-172321-0724-70550187-p1m

Alignments | Result Summary | Phylogenetic Tree | Results Viewers | Submission Details

Program	Number of Sequences	Launched Date	Input Sequences
MUSCLE	21	Fri, Feb 04, 2022 at 17:25:31	muscle-I20220204-172321-0724-70550187-p1m.input
Version		**End Date**	**Output Result**
3.8.425		Fri, Feb 04, 2022 at 17:25:33	muscle-I20220204-172321-0724-70550187-p1m.output

Command

```
$APPBIN/muscle:3.8.425 /muscle -in muscle-I20220204-172321-0724-70550187-p1m.upfile -verbose -log muscle-I20220204-172321-0724-70550187-p1m.output -quiet -clw -out muscle-I20220204-172321-0724-70550187-p1m.clw
```

Input Parameters

Alignment format

clw

Output Tree

none

Screenshot 14: MUSCLE parameters for the longer predicted Gene X ORF

Percent Identity Matrix - created by Clustal2.1

1: Ningiella.ruwaisensis	100.00	20.35	25.94	17.13	16.33	15.54	20.09	15.93	20.00	20.35	22.13	23.08	23.93	22.65	22.17	22.41	21.30	22.61	22.41	22.31	21.16
2: Cyclobacterium.plantarum	20.35	100.00	28.76	21.67	22.69	21.01	18.75	17.89	17.67	22.81	22.91	25.55	25.11	25.55	24.12	23.45	23.11	24.55	24.58	23.43	21.76
3: Thermus.thermophilus	25.94	28.76	100.00	24.39	19.92	19.92	22.57	23.32	22.58	23.58	23.18	24.57	24.57	24.57	28.02	25.97	24.45	25.44	26.29	25.86	25.86
4: GeneX	17.13	21.67	24.39	100.00	55.00	53.82	25.11	27.80	25.91	32.77	28.10	29.05	27.57	27.39	31.80	31.12	30.38	30.80	28.46	30.08	30.20
5: Aureobasidium.melanogenum	16.33	22.69	19.92	55.00	100.00	88.33	25.97	25.56	23.18	29.36	26.56	28.33	28.51	28.75	30.54	29.05	27.85	28.27	26.64	27.84	28.74
6: Aureobasidium.pullulans	15.54	21.01	19.92	53.82	88.33	100.00	23.81	25.56	21.36	28.09	26.03	27.39	27.57	27.80	28.87	28.22	26.16	27.43	26.15	25.39	27.06
7: Alteribacter.aurantiacus	20.09	18.75	22.57	25.11	25.97	23.81	100.00	51.97	52.21	30.47	30.00	30.43	33.48	31.74	31.74	33.33	34.20	32.47	31.17	33.77	32.03
8: Bacillus.coahuilensis	15.93	17.89	23.32	27.80	25.56	25.56	51.97	100.00	58.30	30.84	31.25	31.25	33.93	31.70	30.80	32.44	32.89	32.89	30.22	32.44	32.44
9: Alkalihalobacillus.hwajinpoensis	20.00	17.67	22.58	25.91	23.18	21.36	52.21	58.30	100.00	27.11	28.05	29.41	30.77	30.32	29.86	28.38	33.78	31.53	26.13	31.08	28.83
10: Ktedonobacter.robiniae	20.35	22.81	23.58	32.77	29.36	28.09	30.47	30.84	27.11	100.00	31.91	32.34	36.44	36.60	40.17	39.06	39.22	39.39	34.18	37.02	36.60
11: Halomicroarcula.salinisoli	22.13	22.91	23.18	28.10	26.56	26.03	30.00	31.25	28.05	31.91	100.00	87.60	67.60	69.20	60.08	54.96	56.07	55.04	49.59	49.17	51.24
12: Haloarcula.mannanilytica	23.08	25.55	24.57	29.05	28.33	27.39	30.43	31.25	29.41	32.34	87.60	100.00	69.60	70.40	59.26	54.36	56.49	55.88	52.05	50.00	52.07
13: Haloplanus.rallus	23.93	25.11	24.57	27.57	28.51	27.57	33.48	33.93	30.77	36.44	67.60	69.60	100.00	86.40	59.26	58.92	61.51	59.66	48.37	49.38	51.44
14: Natronomonas.gomsonensis	22.65	25.55	24.57	27.39	28.75	27.80	31.74	31.70	30.32	36.60	69.20	70.40	86.40	100.00	60.49	59.34	61.51	62.18	47.95	50.00	54.13
15: Salinigranum.halophilum	22.17	24.12	28.02	31.80	30.54	28.87	31.74	30.80	29.86	40.17	60.08	59.26	59.26	60.49	100.00	66.12	67.50	66.95	52.89	55.42	55.00
16: Salinigranum.salinum	22.41	23.45	25.97	31.12	29.05	28.22	33.33	32.44	28.38	39.06	54.96	54.36	58.92	59.34	66.12	100.00	68.05	69.29	53.50	54.81	56.07
17: Halogranum.amylolyticum	21.30	23.11	24.45	30.38	27.85	26.16	34.20	32.89	33.78	39.22	56.07	56.49	61.51	61.51	67.50	68.05	100.00	77.50	52.92	56.72	56.30
18: Natrinema.pallidum	22.61	24.55	25.44	30.80	28.27	27.43	32.47	32.89	31.53	39.39	55.04	55.88	59.66	62.18	66.95	69.29	77.50	100.00	52.50	52.32	51.90
19: Halobacterium.salinarum	22.41	24.58	26.29	28.46	26.64	26.15	31.17	30.22	26.13	34.18	49.59	52.05	48.37	47.95	52.89	53.50	52.92	52.50	100.00	57.36	57.75
20: Halorubrum.sodomense	22.31	23.43	25.86	30.08	27.84	25.39	33.77	32.44	31.08	37.02	49.17	50.00	49.38	50.00	55.42	54.81	56.72	52.32	57.36	100.00	89.15
21: Halorubrum.trapanicum	21.16	21.76	25.86	30.20	28.74	27.06	32.03	32.44	28.83	36.60	51.24	52.07	51.44	54.13	55.00	56.07	56.30	51.90	57.75	89.15	100.00

Screenshot 15: Percent Identity Matrix provided with MUSCLE output

```
CLUSTAL multiple sequence alignment by MUSCLE (3.8)


Ningiella.ruwaisensis                -----------MENAMSSISVEG---------FEIV------NHILTLGYATMA-----A
Cyclobacterium.plantarum             ---------MYFVLLAEIIGLDRI----INSDPVAI--------TFFIGYMAM-----FA
Thermus.thermophilus                 --------------------------------MRMLPELSFGEYWLVFNMLSLTIAGMFA
GeneX                                MPSTSMVSASLTTIVIMGSALTRLAGNTVNGKTVDIAITVRGSDWYWTVCAVMT----TC
Aureobasidium.melanogenum            -----------MSWLEKRNDAIQVNPNTQNNKHVDIAITVRGSDFYFAICAVMG----FV
Aureobasidium.pullulans              -----------MSWLEKRNDAIQVNPNTQNNKHVDIAITVRGSDFYFAICAVMG----FV
Alteribacter.aurantiacus             ----------------------------MYEIEQQ--------LLWIYVAFMG----GG
Bacillus.coahuilensis                --------------------------------MISI--------LHYGYSFIML----LG
Alkalihalobacillus.hwajinpoensis     ----------------------------MNSFEIF--------LYYFYFVVMI----SA
Ktedonobacter.robiniae               ---------------------------MDSATVTV---------LWVTSLIMI----LC
Halomicroarcula.salinisoli           ---------------------------MPQPGSEQI--------WLWLGTAGMF----LG
Haloarcula.mannanilytica             ---------------------------MPEPGSEAI--------WLWLGTAGMF----LG
Haloplanus.rallus                    ---------------------------MTQPGSESL--------WLWLGTAGML----IG
Natronomonas.gomsonensis             ---------------------------MADPGSEAL--------WLWIGTAGMF----LG
Salinigranum.halophilum              ---------------------------MATPGAESI--------WLWLGTAGMT----LG
Salinigranum.salinum                 ------------------------MATPGAGLESI--------SLWIGTIGMT----LG
Halogranum.amylolyticum              ---------------------------MVTVGAESL--------WLWIGTLGMT----IG
Natrinema.pallidum                   ---------------------------MAATVGPESI--------WLWIGTIGMT----LG
Halobacterium.salinarum              -----------MTPSLLPTAVEGVSQAQITGRPEWI--------WLALGTALMG----LG
Halorubrum.sodomense                 -----------MDPIALQAGYDLLG----DGRPETL--------WLGIGTLMML----IG
Halorubrum.trapanicum                -----------MDPIALTAGYDLLG----DGRPETL--------WLGIGTLLML----LG
                                                                                              .

 Ningiella.ruwaisensis               ALLFFILTRKDSLPKYQMSSILSVVVMVSALLLLYTQK--ISWTEAYAFDGNEYTVREGA
 Cyclobacterium.plantarum            SAVFFFAERASVDGKWKTSLLVSGLITGIAAIHYYYMR-DFYLQ-------------TGS
 Thermus.thermophilus                AFVFFLLARSYVAPRYHIALYLSALIVFIAGYHYLRIF--ESWVGAYQLQGGVYVP-TGK
 GeneX                               TFVFLGLG-ITKPRQHRVFHYITAAITMVAAIAYFSMGSNLGWTPIDVEFRRNDPVVRGI
 Aureobasidium.melanogenum           ALGTIAAS-AMKPRTDRIFFYITAAINLTACIAYFTMGSNLGWTPIDVEFPRTWSKVAGV
 Aureobasidium.pullulans             ALGVMAAS-AMKPRTDRIFFYITAAINTTACIAYFAMGSNLGWTPIDVEWQRTWSQVAGV
 Alteribacter.aurantiacus            AVYFAYLAFHRKGVPRAE-YLVAFIIPTWSGVAYASIALGQGLV-------------EWG
 Bacillus.coahuilensis               ALYFYLLSKDPKGVPASE-YLIAMVIPLWSGAAYLSIALGQGLF-------------QYD
 Alkalihalobacillus.hwajinpoensis    AIYFFILSRKPKGVPLYE-YVVAMMITAWSGVAYLSIALGQGFI-------------ERP
 Ktedonobacter.robiniae              TLVFTYRSFRAR-IEIKHFYYLTALITLIAATLYMTMASGYGGI-------------GLN
 Halomicroarcula.salinisoli          MLYFIGRGWGETDDRRQKFYIATILITAIAFVNYLAMALGFGLTIIELP--------NDP
 Haloarcula.mannanilytica            MLYFIGRGWGETDSRRQKFYIATILITAIAFVNYLAMALGFGLTIIEF---------GGS
 Haloplanus.rallus                   MLYFIARGWGEKNRRRQEFYIVTIFITAIAFVNYLSMALGFGLTTIEI---------GGE
 Natronomonas.gomsonensis            MLYFIARGWGEENRRRQEFYIVTIFITAIAFVNYLMMALGFGLTTVTV---------AGE
 Salinigranum.halophilum             TFYFIARGWGVEDEEQQRFYLITIFITAIASAAYFAMATGFGLTQVTV---------NGQ
 Salinigranum.salinum                TLYFVAQGWSVRDPDQQEYYIITIFIPAIAAASYFAMASGFGLVEVPVE--------GLG
 Halogranum.amylolyticum             TLYFVGRGRGVTDKKMQEFYIITIFITTIAAAMYLLMATGFGLTQVQV---------GNR
 Natrinema.pallidum                  TLYFVGRGRGVRDRKMQEFYIITIFITTIAAAMYFAMATGFGVTEVMV---------GDE
 Halobacterium.salinarum             TLYFLVKGMGVSDPDAKKFYAITTLVPAIAFTMYLSMLLGYGLTMVPF---------GGE
 Halorubrum.sodomense                TFYFIVKGWGVTDKEAREYYAITILVPGIASAAYLSMFFGIGLTEVEL--------VGGE
 Halorubrum.trapanicum               TFYFIARGWGVTDKEAREYYAITILVPGIASAAYLSMFFGIGLTEVQV---------GGE
                                                             :  :   :
```

```
Ningiella.ruwaisensis                DLFTNGYRYLNWLIDVPMLLIQILWVAQITG-SQRTSYMFKFSFSGCLMILTGYIGQFYE
Cyclobacterium.plantarum             SPTAF--RYVDWTLTVPLMCVEFYLLTKPF--GAKGATLTKLIIASLVMLVTGYIGETSG
Thermus.thermophilus                 PFNDFY-RYADWLLTVPLLLLELILVLGLSP-ARTWNLGVKLVVASVLMLGLGYVGEANT
GeneX                                NREIFYVRYVDWFITTPLLLMDLLLTA-----AMPWPTVLFVVLVDEVMIVTGLVGALVR
Aureobasidium.melanogenum            NREIFYARYVDWFVTTPLLLMDLLLTA-----GLPWPTILYTIFLDEVMIVTGLIGALVK
Aureobasidium.pullulans              NREVFYVRYIDWFVTTPLLLMDLLLTA-----GLPWPTILWTIFLDEVMIVTGLVGALVK
Alteribacter.aurantiacus             DRVIYFARYLDWVVTTPLLLLALAMTAMYTI-SKDRVIIGGLIVADVFMVLTGLIAEFSP
Bacillus.coahuilensis                DTTIYYARYIDWVISTPLLLAALALTAMFGG-KKNLTLLFSLVALDVFMIITGFVADLSI
Alkalihalobacillus.hwajinpoensis     EKTIYFARYLDWVVSTPLLVLSLALTAMFYETKKNKVLIASIMATDVFMILTGLIADFSP
Ktedonobacter.robiniae               GKVILFGRYIDWVITTPLLLMNLALIALPRNFPSRFAVIGTMIAADVYMIVSGLGASLIR
Halomicroarcula.salinisoli           EAPIYWARYTDWLFTTPLLLYDLALLA-----GADRNTISTLVSLDVLMIGTGVVATLSA
Haloarcula.mannanilytica             EHPIYWARYTDWLFTTPLLLYDLGLLA-----GADRNTIASLVSLDVLMIGTGVVATLSA
Haloplanus.rallus                    ELPIYWARYTDWLFTTPLLLIDLGLLA-----GANRNQLSTLVGLDVLMIGTGAVATLST
Natronomonas.gomsonensis             ELPIYWARYTDWLFTTPLLLIDLGLLA-----GANRNQIATLVGLDALMIGTGAVATLST
Salinigranum.halophilum              VLDIYWGRYADWLFTTPLLLLDLALLA-----RASKNTIYTLVGLDVLMIGTGVIGALAA
Salinigranum.salinum                 TLDIYWARYADWLFTTPLLLLDLALLA-----GADRNTIYTLVGLDVFMIVTGLVGALAR
Halogranum.amylolyticum              TLDIYWARYADWVFTTPLLLLDLALLA-----GANRNTIATLVGLDVFMIATGLIAALEP
Natrinema.pallidum                   ALTIYWARYADWLFTTPLLLLDLSLLA-----GANRNTIATLIGLDVFMIGTGAIAALSS
Halobacterium.salinarum              QNPIYWARYADWLFTTPLLLLDLALLV-----DADQGTILALVGADGIMIGTGLVGALTK
Halorubrum.sodomense                 VLDIYYARYADWLFTTPLLLLDLALLA-----KVDRVTIGTLVGVDALMIVTGLIGALSK
Halorubrum.trapanicum                MLDIYYARYADWLFTTPLLLLDLALLA-----KVDRVTIGTLVGVDALMIVTGLVGALSH
                                           ** :* . .*::   :                    .  *:  *  .

Ningiella.ruwaisensis                PGRINEDVTLWAVWGLISTAFFLHVLVLITRVIKEGSS--KMSGGAKSVFSAILPLFLIS
Cyclobacterium.plantarum             LDN-------NIFWGIVSTLGYLYIVYEVFAGDVAKLSQSSDSPALKKAMFLLKIFITLG
Thermus.thermophilus                 EPGP------RTLWGALSSVPFFYILYVLWVELGQAIRETRFGPRVLELLTAIRYVLLMS
GeneX                                SSY-------KWGYFAFGCAALFYVVFVLVWEARRHAN--ALGSDVGKAFTICGSLTTFL
Aureobasidium.melanogenum            SRY-------KWGFWTFGTVAMFAIFWNLAVEGRKHAK--HLGSDVYRCYLMCGVLTLFV
Aureobasidium.pullulans              SRY-------KWGFWTFGTVAMFAIFWNLAVEGRKHAK--HLGSDIARTYTICGCLTLFI
Alteribacter.aurantiacus             SPI-------KYVWYILGVVAFLIILWIIWWPLRAKAK--SQNHYVYRVFLIVAGYLSIL
Bacillus.coahuilensis                GTT-------KYIWYSLGVIALIIILVITFGPLRRIAL--SNGTRLARHYTRVAIYLSAL
Alkalihalobacillus.hwajinpoensis     DSL-------KYIWYSLGVIALFIILLITWIPLKRIAD---RHEQLSKHYKRVALYLTIF
Ktedonobacter.robiniae               SNF-------RWAFFAVSCAGFLAVLYFIIVKLTPEAN--VRSGPVQRHYSTLAIMLIAL
Halomicroarcula.salinisoli           GSGVLAAGAERLIWWGISTAFLLVLLYFLFSSLSSRVT--DLPSDTQGTFRTLRNLVAVV
Haloarcula.mannanilytica             GSGVLSAGAERLVWWGISTAFLLVLLYFLFSSLSGRVA--DLPSDTRSTFKTLRNLVTVV
Haloplanus.rallus                    AGVLLSPVGDRIIWWGVSTGFLLVLLYFLFGTLTKEAS--QLSGAARSTFSTLRNLIVVV
Natronomonas.gomsonensis             TGVLLSPVGDRLIWWGVSTGFLLVLLYFLFGTLTEEAN--RLSDDAQSTFRTLRNLIVVV
Salinigranum.halophilum              SSAFI-----RIVWWAISTVFLLFLLYFLIRTLSEAAT--RQSPEVRKLTTTLRNMLIVL
Salinigranum.salinum                 EGQVF-----RIIWWAISTGALLVLLYFLLGSLSEQAS--RQAGEVGALFSRLRNLILVL
Halogranum.amylolyticum              NATY------RIMWWGISTGALLALLYILVGTLSKQVE--TRDAEVQSLFSTLRNLTMVL
Natrinema.pallidum                   TPGT------RIAWWAISTGALLALLYVLVGTLSENAR--NRAPEVASLFGRLRNLVIAL
Halobacterium.salinarum              VYSY------RFVWWAISTAAMLYILYVLFFGFTSKAE--SMRPEVASTFKVLRNVTVVL
Halorubrum.sodomense                 TPLA------RYTWWLFSTIAFLFVLYYLLTSLRSAAK--QRSAEVQSTFNTLTALVAVL
Halorubrum.trapanicum                TAVA------RYSWWLFSTICMIVVLYFLATSLRSAAK--QRSADVQSTFNTLTALVLVL
                                           .  :  ..    : :.
```

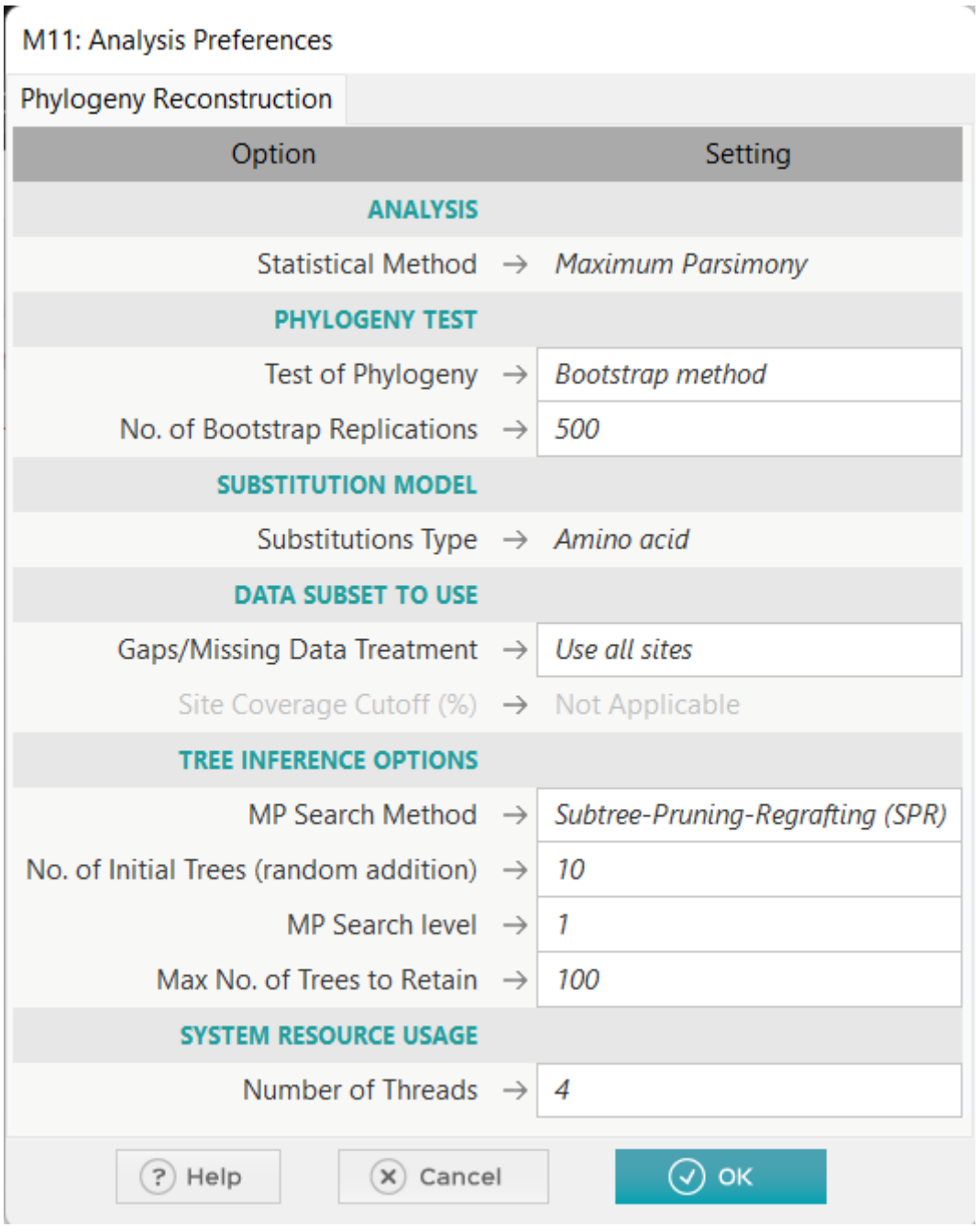
```
                                                    . .  ..   . ..

Ningiella.ruwaisensis                WWLYPIAYLAPYFMT--MGYS-YETTIVSQQVIYTIADISSKVVYGVMLTVTATMLSK--
Cyclobacterium.plantarum             WSIYPIGYM---VLP--GNLLSGLFEVSSIDLFYNLADAINKIGFGLVIYS---------
Thermus.thermophilus                 WGFYPIAYA---LGTWLPG---GAAQEVVIQLGYSLADLIAKPVYGLLIFAIARA-----
GeneX                                WILYPIAWG---LCE--GG---NLISPDSEAIFYGILDLLAKPVFGALLIWGHRGIDPAR
Aureobasidium.melanogenum            WLCYPICWG---VSE--GG---NVIPPDSEAVFYGVLDFLAKPCFSIALIAGHWNINPGR
Aureobasidium.pullulans              WLCYPICWG---VSE--GA---NVIPPDSEAVFYGVLDFLAKPVFSIALIIGHWNINPGR
Alteribacter.aurantiacus             WVGYPTVWL---LGP--SG-L-GVISQITDQALFVSLPIFSKVGFSILDLSCLRWLHV--
Bacillus.coahuilensis                WVCYPTAWL---LGP--SG-L-GLAQELTEVLVFIILPIFSKVGFSIVDLHGLRKLH---
Alkalihalobacillus.hwajinpoensis     WLLYPTAWI---LGA--SG-I-GMTQGIIDTLAFVILPIFSKIGFGLLDLHGLRKLK---
Ktedonobacter.robiniae               WVCYPIVWI---LGT--EG-F-GIISLLPEVILYAILDVLAKGAFGFVLLSKPGVLL---
Halomicroarcula.salinisoli           WLVYPVWWL---IGT--EG-L-ALVGIFTETAGFMVIDLVAKVGFGFILLRSHSVL----
Haloarcula.mannanilytica             WLVYPVWWL---VGT--EG-L-GLVGIGIETAGFMVIDLTAKVGFGIILLRSHGVL----
Haloplanus.rallus                    WLVYPVWWI---LGT--EG-L-GVISLYSETAGFMVLDLVAKVGFGIILLSSRDVL----
Natronomonas.gomsonensis             WLVYPVWWI---LGT--EG-L-GTIGLYSETAGFMVLDLVAKVGFGIILLSSREVL----
Salinigranum.halophilum              WLAYPVVWI---LGT--EGTI-GIIPLYWETAAFMVLDLTAKVGFGFVLLRSHSVL----
Salinigranum.salinum                 WSAYPVVWI---LGT--EGGF-AIIPLGVETAAFMVLDLSAKVGFGFILLQSRDVLS---
Halogranum.amylolyticum              WLLYPVVWI---LGT--EGTI-GILPLYWETAAFMVLDLSAKVGFGFLLLRSRAVLE---
Natrinema.pallidum                   WFLYPVVWI---LGT--EGTF-GILPLYWETAAFMVLDLSAKVGFGVILLQSRSVLE---
Halobacterium.salinarum              WSAYPVVWL---IGS--EG-A-GIVPLNIETLLFMVLDVSAKVGFGLILLRSRAIFG---
Halorubrum.sodomense                 WTAYPILWI---VGT--EG-A-GVVGLGIETLAFMILDVTAKVGFGFVLLRSRAILG---
Halorubrum.trapanicum                WTAYPILWI---IGT--EG-A-GVVGLGIETLLFMVLDVTAKVGFGFILLRSRAILG---
                                     *  **  :    .                  :      *  :.

                                                  .    .                    .          ..

Ningiella.ruwaisensis                    --------------------QEGMAESQA------------------
Cyclobacterium.plantarum                 ---------------------VAIKESKKTTAQVA------------
Thermus.thermophilus                     ---------------------KSLEEGFGEGVKAA------------
GeneX                                    LGLYIHDYNEKDPAVKDKVGAPGPNVHPNTNNANNAAATNDSTPETV
Aureobasidium.melanogenum                MGLKLRDYDEEPAYFGPKNGAEAAKERGRTDNAVDGVA---------
Aureobasidium.pullulans                  MGLKLRDYDEDPDYFGPKNGAEAAKERSNGSSSGVDGGA--------
Alteribacter.aurantiacus                 --------------------KHGQEVTPQAT----------------
Bacillus.coahuilensis                    ---------------------QSSYVHN-------------------
Alkalihalobacillus.hwajinpoensis         ---------------------TN-------------------------
Ktedonobacter.robiniae                   ---------------------EAERETAPINSVAAQW----------
Halomicroarcula.salinisoli               ---------------------DGAAQSQTTGASPAD--D--------
Haloarcula.mannanilytica                 ---------------------DGAAETTSTGATPAD--D--------
Haloplanus.rallus                        ---------------------DAAGDTTGAALGDADPTD--------
Natronomonas.gomsonensis                 ---------------------DAAGDLAGSTAQPAD--D--------
Salinigranum.halophilum                  ---------------------EAATQSTTAGATAD------------
Salinigranum.salinum                     ---------------------AAKSTGASATAD--------------
Halogranum.amylolyticum                  ---------------------KASTPTAAATA---------------
Natrinema.pallidum                       ---------------------RVATPTAAPT----------------
Halobacterium.salinarum                  ---------------------EAEAPEPSAGDGAAATSD--------
Halorubrum.sodomense                     ---------------------DTEAPEPSAGADAQA-AD--------
Halorubrum.trapanicum                    ---------------------DTGAPEPSAGAEASA-AD--------
```

Screenshot 16: Multiple Sequence Alignment for longer Gene X ORF provided by MUSCLE

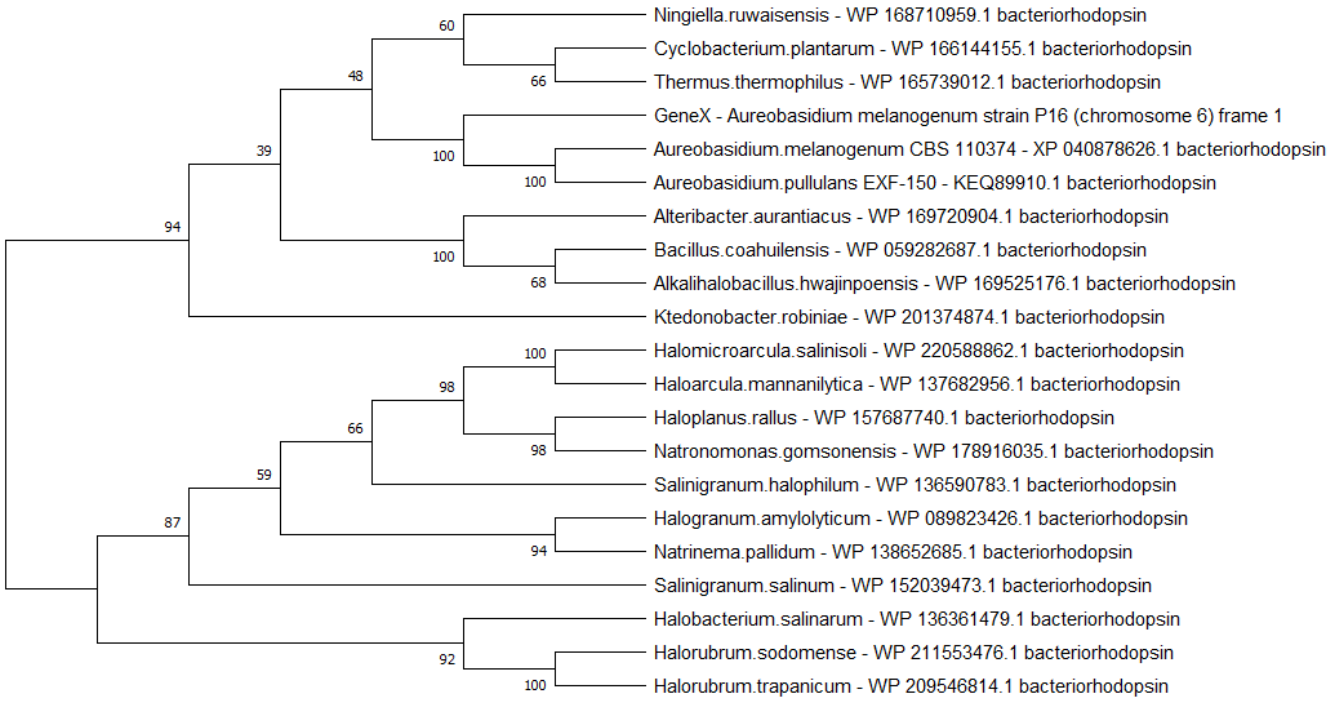
Question 6:

Note that within the multiple alignment there are organisms hailing from all three main domains of life. The phylogenies shown below do not follow the expected branching pattern for the relationships between the various species. However, it is not unprecedented for gene trees to differ from species trees, especially when prokaryotes are involved, given the extensive horizontal gene transfer that occurs within this domain.

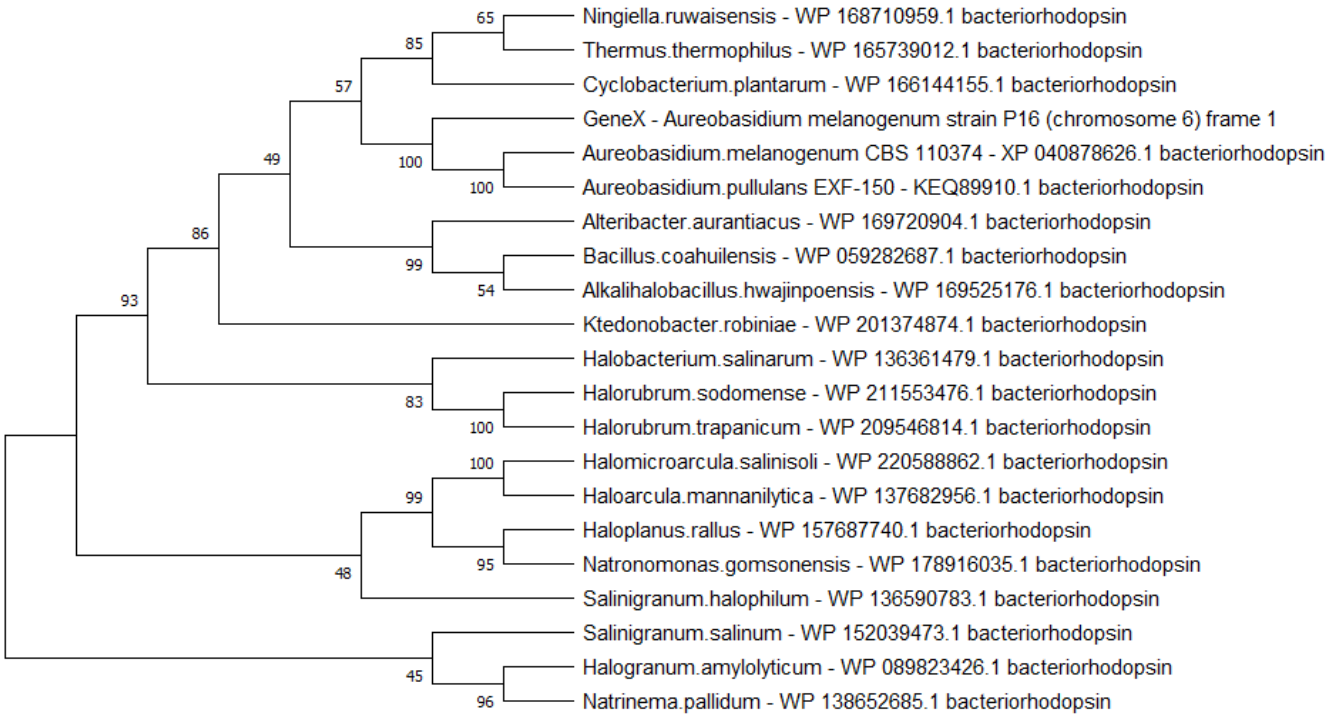
To create the gene phylogenies the MEGA11 program was used with the maximum parsimony method. Note however that the phylogeny produced was robust to the statistical method utilised, as maximum likelihood gave the same results.



Screenshot 17: MEGA parameters for cladogram estimation for both predicted ORFs



Screenshot 18: Bootstrap consensus cladogram for the longer predicted Gene X ORF and the other sequences introduced earlier

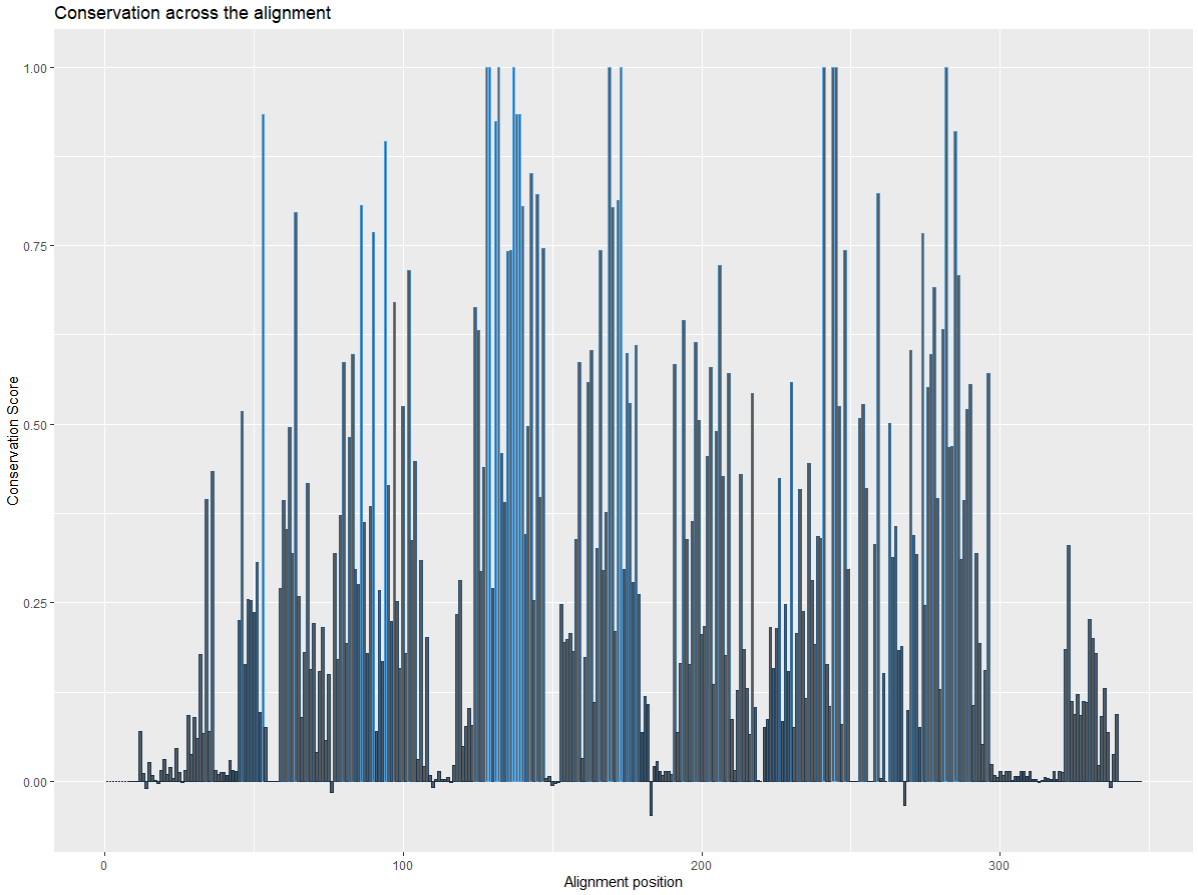


Screenshot 18: Bootstrap consensus cladogram for the shorter predicted Gene X ORF and the other sequences introduced earlier

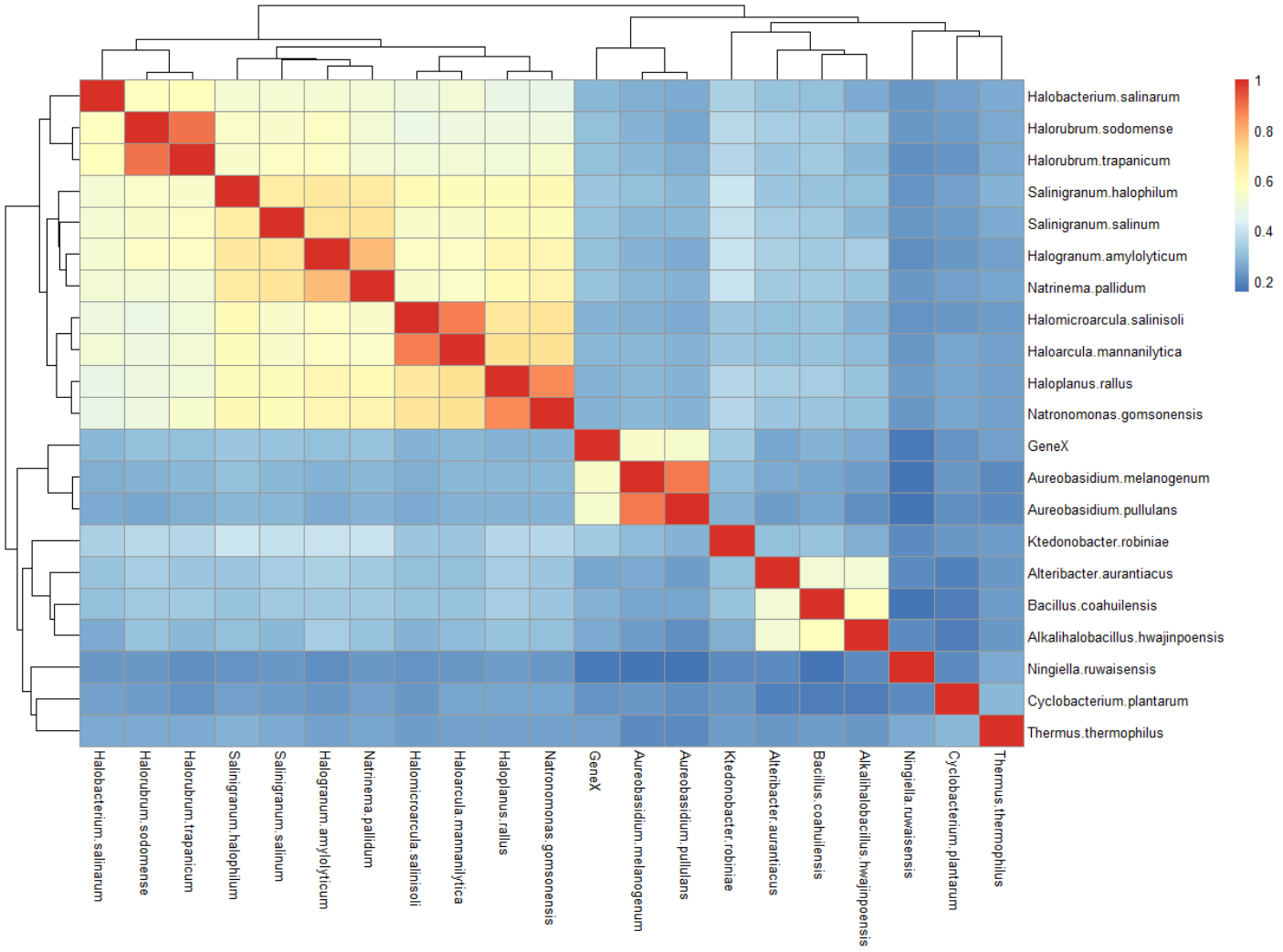
While there are a few differences between the two cladograms, they are largely identical, and both place the *Gene X* from *Aureobasidium melanogenum* as an outgroup to the bacteriorhodopsin genes of *Aureobasidium melanogenum* and *Aureobasidium pullulans*. This is important because *Gene X* shared a large amount of similarity with the *Aureobasidium melanogenum* bacteriorhodopsin, and its outgroup nature is therefore an important factor in considering it a novel gene, alongside its lack of annotation and the percentage identity < 100%.

Question 7:

Next it is interesting to consider conservation across the alignment. This can be achieved in R with some functions from the bio3D package and ggplot2 for plotting.

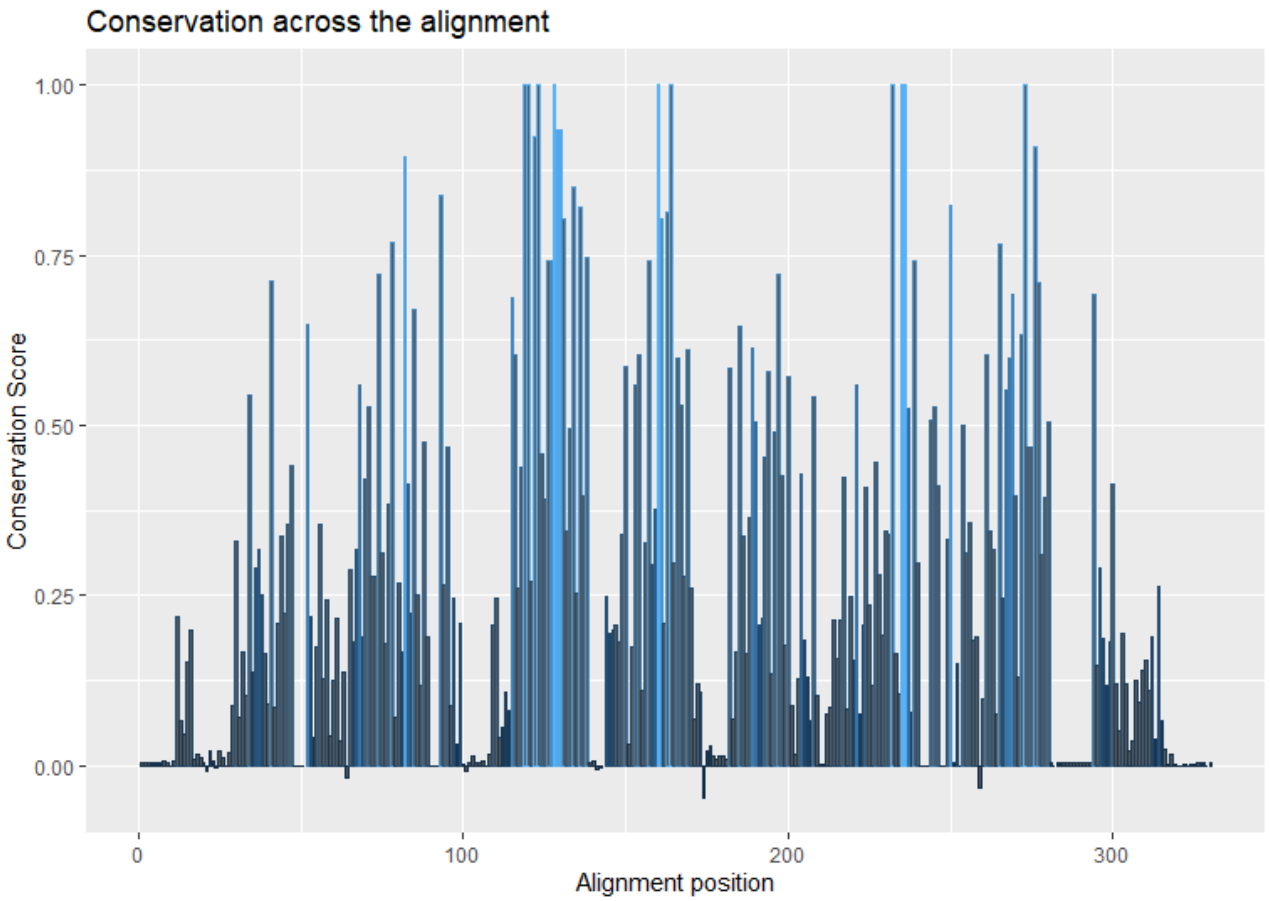


Plot 1: Plot of conservation scores across the multiple alignment (Long ORF)

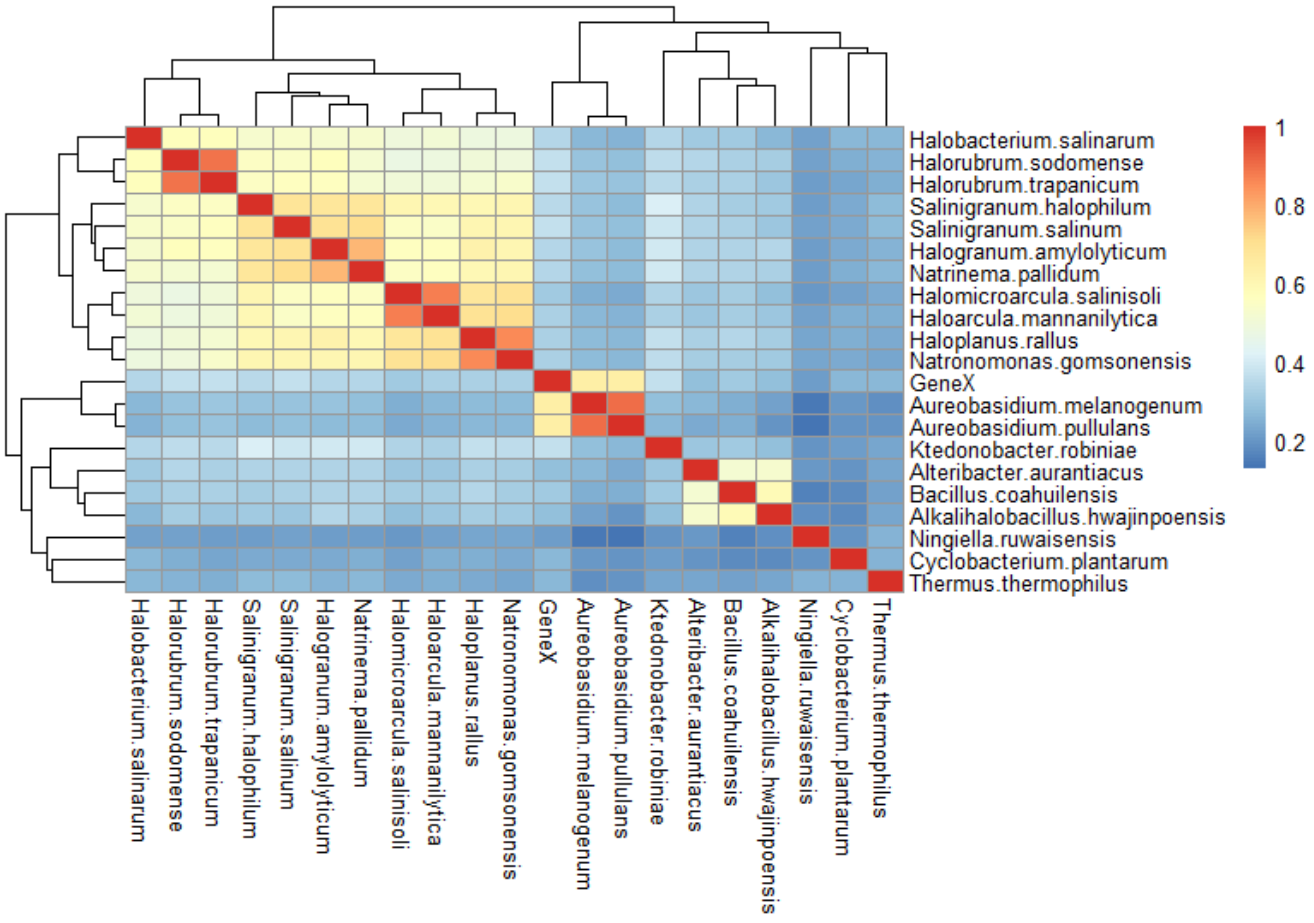


Plot 2: Heatmap of sequence similarity, where 1 = 100% identity (Long ORF)

It is interesting to note that this heatmap shows a strong clustering of Prokaryotes and Archaea, with the Eukaryotes inbetween, closer to the Prokaryotes. Also note that, again, Gene X is the outgroup to the two *Aureobasidium* bacteriorhodopsin sequences. The same patterns are observed for the shorter predicted ORF (see below).



Plot 3: Plot of conservation scores across the multiple alignment (Short ORF)



Plot 4: Heatmap of sequence similarity, where 1 = 100% identity (Short ORF)

Question 8:

Longer predicted ORF

Using the online server EMBL-EBI to get a results gives an uncharacterized protein as the top hit, followed by bacteriorhodopsin and a protein from an unplaced genomic scaffold.

Results for job fasta-I20220216-234158-0204-25994608-p1m

Summary Table | Tool Output | Visual Output | Functional Predictions | Submission Details

Selection: Select All | Invert | Clear

Apply to selection:

Annotations: Show | Hide

Alignments: Show | Hide

Entries: Download in fasta format

Tools: Launch Clustal Omega

Align.	DB:ID	Source	Length	Score (Bits)	Identities %	Positives %	E()
1	AFDB:AF-A0A1C1CZG1-F1	Uncharacterized protein	304	288.8	53.5	80.9	1.8E-76
2	AFDB:AF-A0A1C1CTS3-F1	Bacteriorhodopsin	301	266.1	52.0	81.7	1.2E-69
3	AFDB:AF-A0A0D2H0H7-F1	Unplaced genomic scaffold supercont1.5, whole genome shotgun sequence	298	259.3	50.3	79.0	1.4E-67
4	AFDB:AF-O74631-F1	Protein FDD123	283	189.6	38.3	72.5	1.2E-46
5	AFDB:AF-A0A1C1CDV2-F1	Opsin-1	302	158.5	38.7	68.4	3.0E-37
6	AFDB:AF-A0A0D2H7D9-F1	Unplaced genomic scaffold supercont1.4, whole genome shotgun sequence	300	148.6	36.2	70.1	2.7E-34
7	AFDB:AF-U7PWK4-F1	Uncharacterized protein	305	145.9	33.1	66.9	1.9E-33
8	AFDB:AF-P38079-F1	Protein YRO2	344	145.5	34.3	63.3	2.7E-33
9	AFDB:AF-Q12117-F1	Protein MRH1	320	145.3	34.2	61.1	2.9E-33

Screenshot 19: EMBL-EBI search for structures with similar protein sequences to the longer predicted ORF

Using R and Bio3D provides three hits above the threshold. Namely:

PDB Code	E value	Sequence Identity	Chain Length	Alignment Length	Experimental Technique	Resolution	Pfam Classification	Source
7BMH_A	2.13e-28	36.898	324	241	X-ray	2.20	Bacteriorhodopsin-like protein (Bac_rhodopsin)	Leptosphaeria maculans
5AWZ_A	4.29e-26	38.710	344	231	X-ray	1.57	Bacteriorhodopsin-like protein (Bac_rhodopsin)	Acetabularia acetabulum
6GYH_A	7.74e-30	47.222	236	222	X-ray	2.00	Bacteriorhodopsin-like protein (Bac_rhodopsin)	Coccomyxa subellipsoidea C-169

The use of bio3D and R allows more than simply blast searching, as shown below, in the inserted R markdown document.

Longer ORF PDB search and results

Mirte Ciz Marieke Kuijpers

02/03/2022

The code in this document is made to be useful with either the long or the short ORF, but in the set-up below the sequence to use is set to the long ORF.

```
# Set-up
library("bio3d")
library("ggplot2")
library("ggrepel")
library("msa")
library("bio3d.view")

# Load sequence of POI
seqL <- read.fasta("long.ORF.fa")
seqS <- read.fasta("short.ORF.fa")

#Choose which sequence to use
seq <- seqL
```

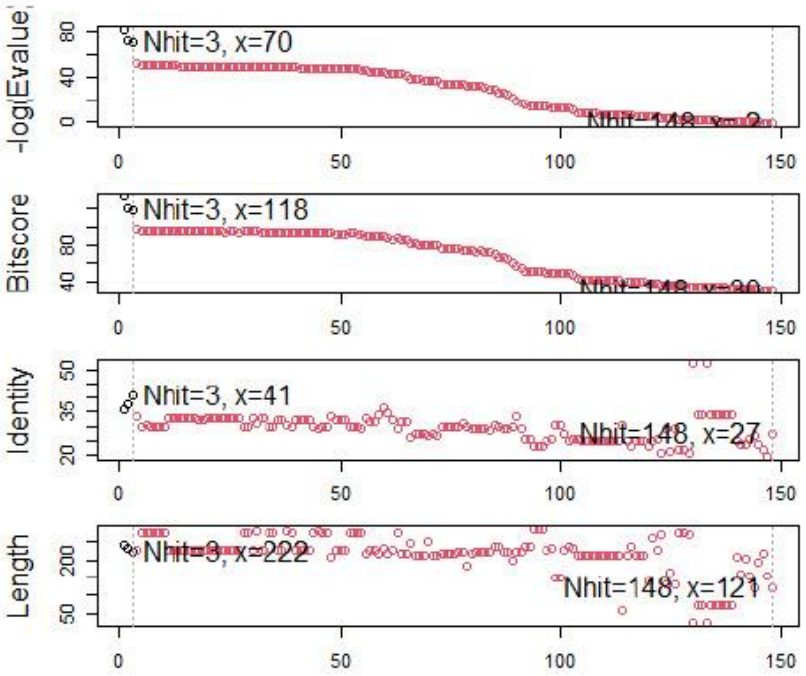
After set up the blast search can be completed and the summary statistics of this search can be plotted.

```
# Blast search
blast <- blast.pdb(seq, database = "pdb")
```

```
##  Searching ... please wait (updates every 5 seconds) RID = 1ZJ94FWD01R
##
.....................................................................
.....................................................................
.....................................................................
.....................................................................
##  Reporting 148 hits
```

```
# Plot summary statistics of results
hits <- plot(blast)
```

```
##   * Possible cutoff values:    70 -2
##             Yielding Nhits:    3 148
##
##   * Chosen cutoff value of:    70
##             Yielding Nhits:    3
```




```
# Print the IDs of the hits above the threshold
hit.IDs <- hits$pdb.id
hit.IDs

## [1] "7BMH_A" "5AWZ_A" "6GYH_A"
```

There are 3 hits that pass the statistical threshold, namely: 7BMH_A, 5AWZ_A, 6GYH_A. More information can be found on these by interrogating the blast results.

```
# Show the hit table for the top hits which pass the threshold
head(blast$hit.tbl, n = length(hit.IDs))

##        queryid subjectids identity alignmentlength mismatches gapopens
q.start
## 1 Query_540941     7BMH_A   35.685             241        149         3
42
## 2 Query_540941     5AWZ_A   37.662             231        115         5
39
## 3 Query_540941     6GYH_A   40.541             222        122         4
46
##   q.end s.start s.end   evalue bitscore positives mlog.evalue pdb.id
acc
## 1   276      51   291 6.05e-36      133     51.45    81.09301 7BMH_A
7BMH_A
## 2   258      16   228 6.87e-32      120     49.78    71.75556 5AWZ_A
5AWZ_A
```

```
## 3   266      14   226 2.73e-31      118     55.86    70.37584 6GYH_A
6GYH_A
```

We can also download these PDB files, annotate them for more information and align them with our sequence to get an overview of sequence alignment.

```
# Download related PDB files
files <- get.pdb(hits$pdb.id, path="pdbs", split=TRUE, gzip=TRUE)
```

```
##   |
|                                                                      |   0%
|
|=======================                                               |  33%
|
|===============================================                       |  67%
|
|======================================================================| 100%
```

```
# Align related PDBs
pdbs <- pdbaln(files, fit = TRUE, exefile="msa")
```

```
## Reading PDB files:
## pdbs/split_chain/7BMH_A.pdb
## pdbs/split_chain/5AWZ_A.pdb
## pdbs/split_chain/6GYH_A.pdb
##    PDB has ALT records, taking A only, rm.alt=TRUE
## .   PDB has ALT records, taking A only, rm.alt=TRUE
## .   PDB has ALT records, taking A only, rm.alt=TRUE
## .
##
## Extracting sequences
##
## pdb/seq: 1   name: pdbs/split_chain/7BMH_A.pdb
##    PDB has ALT records, taking A only, rm.alt=TRUE
## pdb/seq: 2   name: pdbs/split_chain/5AWZ_A.pdb
##    PDB has ALT records, taking A only, rm.alt=TRUE
## pdb/seq: 3   name: pdbs/split_chain/6GYH_A.pdb
##    PDB has ALT records, taking A only, rm.alt=TRUE
```

```
# Vector containing PDB codes for figure axis
ids <- basename.pdb(pdbs$id)

# Annotate hits for more information on the hits
anno <- pdb.annotate(ids)

# Find the organisms these PDB hits come from
unique(anno$source)
```

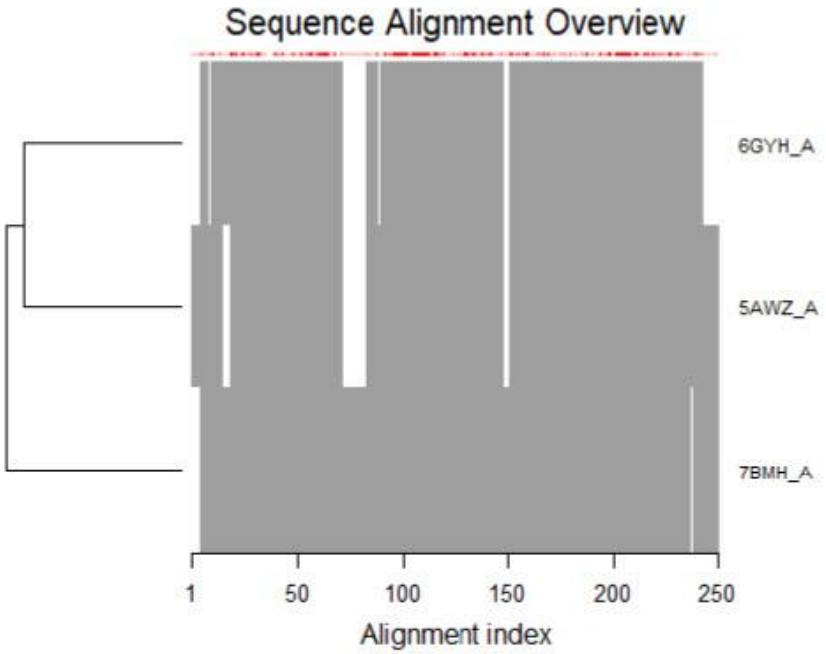
```
## [1] "Leptosphaeria maculans"         "Acetabularia acetabulum"
## [3] "Coccomyxa subellipsoidea C-169"
```

```
# View more information on the hits
anno
```

```
##        structureId chainId macromoleculeType chainLength
experimentalTechnique
## 7BMH_A        7BMH       A           Protein         324
X-ray
## 5AWZ_A        5AWZ       A           Protein         244
X-ray
## 6GYH_A        6GYH       A           Protein         236
X-ray
##        resolution scopDomain
pfam
## 7BMH_A       2.20       <NA> Bacteriorhodopsin-like protein
(Bac_rhodopsin)
## 5AWZ_A       1.57       <NA> Bacteriorhodopsin-like protein
(Bac_rhodopsin)
## 6GYH_A       2.00       <NA> Bacteriorhodopsin-like protein
(Bac_rhodopsin)
##                                  ligandId
## 7BMH_A                     LFA (22),OLA (3)
## 5AWZ_A RET,OLB,D12 (2),D10 (3),OCT (2),C14
## 6GYH_A                      RET,CLR,OLB (4)
##
ligandName
## 7BMH_A
EICOSANE (22),OLEIC ACID (3)
## 5AWZ_A RETINAL,(2S)-2,3-dihydroxypropyl (9Z)-octadec-9-enoate,DODECANE
(2),DECANE (3),N-OCTANE (2),TETRADECANE
## 6GYH_A                                    RETINAL,CHOLESTEROL,(2S)-2,3-
dihydroxypropyl (9Z)-octadec-9-enoate (4)
##                                source
## 7BMH_A         Leptosphaeria maculans
## 5AWZ_A        Acetabularia acetabulum
## 6GYH_A Coccomyxa subellipsoidea C-169
##
structureTitle
## 7BMH_A                        Crystal structure of a light-driven proton
pump LR (Mac) from Leptosphaeria maculans
## 5AWZ_A Crystal Structure of the Cell-Free Synthesized Membrane Protein,
Acetabularia Rhodopsin I, at 1.57 angstrom
## 6GYH_A                     Crystal structure of the light-driven proton
pump Coccomyxa subellipsoidea Rhodopsin CsR
##                                                                  citation
rObserved
## 7BMH_A                       Zabelskii, D., et al. Commun Biol (2021)
0.23840
## 5AWZ_A Furuse, M., et al. Acta Crystallogr D Biol Crystallogr (2015)
0.17760
## 6GYH_A                            Fudim, R., et al. Sci Signal (2019)
```

```
0.19398
##          rFree   rWork spaceGroup
## 7BMH_A 0.28470 0.23610 P 21 21 21
## 5AWZ_A 0.19410 0.17690    C 1 2 1
## 6GYH_A 0.22493 0.19231        H 3
```

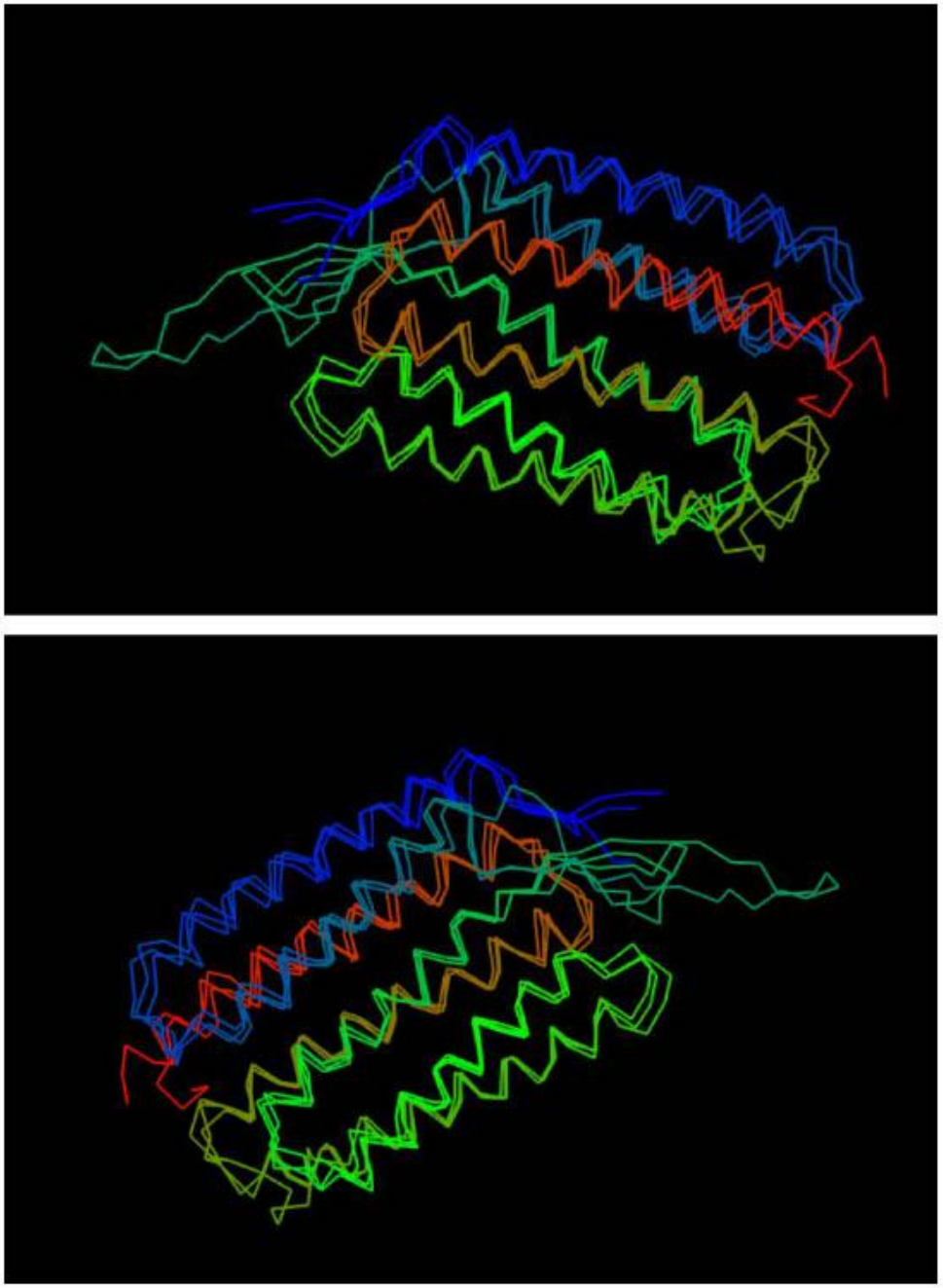
```
# Draw schematic alignment
plot(pdbs, labels=ids)
```



We can also plot the three structures we have found as follows:

```
# Set up
library(bio3d.view)
library(rgl)

# Plot
#view.pdbs(pdbs)
```

The `View.pdbs()` function brings up an interactive viewer, which cannot be directly viewed in the markdown document, so instead two screen-shots of this have been inserted.



With more proteins it could be interesting to plot variability, or even do PCA using the amino acid position data, but with only three proteins this is not useful.

Session Information

```
sessionInfo()

## R version 4.1.2 (2021-11-01)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 22000)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United Kingdom.1252
## [2] LC_CTYPE=English_United Kingdom.1252
## [3] LC_MONETARY=English_United Kingdom.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United Kingdom.1252
##
## attached base packages:
## [1] stats4    stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] rgl_0.108.3           bio3d.view_0.1.0.9000 msa_1.26.0
##  [4] Biostrings_2.62.0     GenomeInfoDb_1.30.1   XVector_0.34.0
##  [7] IRanges_2.28.0        S4Vectors_0.32.3      BiocGenerics_0.40.0
## [10] ggrepel_0.9.1         ggplot2_3.3.5         bio3d_2.4-3.9000
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.1.2       xfun_0.29              purrr_0.3.4
##  [4] colorspace_2.0-2       vctrs_0.3.8            generics_0.1.2
##  [7] htmltools_0.5.2        yaml_2.2.2             utf8_1.2.2
## [10] rlang_1.0.1            pillar_1.7.0           glue_1.6.1
## [13] withr_2.4.3            GenomeInfoDbData_1.2.7 lifecycle_1.0.1
## [16] stringr_1.4.0          zlibbioc_1.40.0        munsell_0.5.0
## [19] gtable_0.3.0           htmlwidgets_1.5.4      evaluate_0.15
## [22] knitr_1.37             extrafont_0.17         fastmap_1.1.0
## [25] curl_4.3.2             parallel_4.1.2         fansi_1.0.2
## [28] Rttf2pt1_1.3.10        highr_0.9              Rcpp_1.0.8
## [31] scales_1.1.1           jsonlite_1.8.0         digest_0.6.29
## [34] stringi_1.7.6          dplyr_1.0.8            grid_4.1.2
## [37] cli_3.2.0              tools_4.1.2            bitops_1.0-7
## [40] magrittr_2.0.2         RCurl_1.98-1.6         tibble_3.1.6
## [43] extrafontdb_1.0        crayon_1.5.0           pkgconfig_2.0.3
## [46] ellipsis_0.3.2         httr_1.4.2             rmarkdown_2.11
## [49] rstudioapi_0.13        R6_2.5.1               compiler_4.1.2
```

Print-out 1 - 8: R Markdown document of Bio3D investigation of the longer predicted ORF

Shorter predicted ORF

The same method can be used for the shorter predicted ORF.

PDB Code	E value	Sequence Identity	Chain Length	Alignment Length	Experimental Techniqe	Resolution	Pfam Classification	Source
7BMH_A		36.9	324	144	X-ray	2.20	Bacteriorhodopsin-like protein (Bac_rhodopsin)	Leptosphaeria maculans
5AWZ_A		38.7	344	187	X-ray	1.57	Bacteriorhodopsin-like protein (Bac_rhodopsin)	Acetabularia acetabulum
6GYH_A		47.2	236	186	X-ray	2.00	Bacteriorhodopsin-like protein (Bac_rhodopsin)	Coccomyxa subellipsoidea C-169

Longer ORF PDB search and results

Mirte Ciz Marieke Kuijpers

02/03/2022

The code in this document is made to be useful with either the long or the short ORF, but in the set-up below the sequence to use is set to the long ORF.

```
# Set-up
library("bio3d")
library("ggplot2")
library("ggrepel")
library("msa")
library("bio3d.view")

# Load sequence of POI
seqL <- read.fasta("long.ORF.fa")
seqS <- read.fasta("short.ORF.fa")

#Choose which sequence to use
seq <- seqS
```

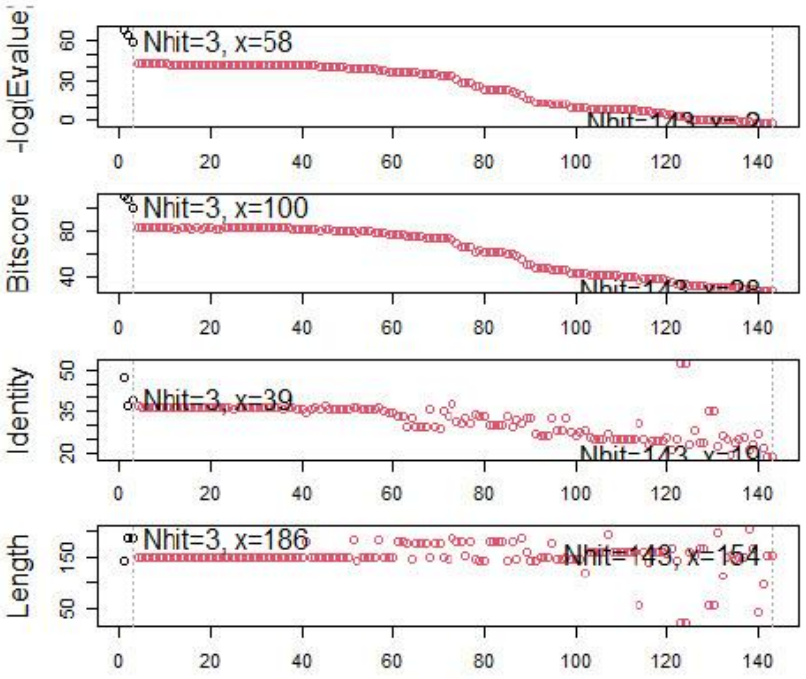
After set up the blast search can be completed and the summary statistics of this search can be plotted.

```
# Blast search
blast <- blast.pdb(seq, database = "pdb")
```

```
##  Searching ... please wait (updates every 5 seconds) RID = 1ZR0NWEN013
##  .
##  Reporting 143 hits
```

```
# Plot summary statistics of results
hits <- plot(blast)
```

```
##    * Possible cutoff values:    58 -3
##              Yielding Nhits:    3 143
##
##    * Chosen cutoff value of:    58
##              Yielding Nhits:    3
```




```
# Print the IDs of the hits above the threshold
hit.IDs <- hits$pdb.id
hit.IDs

## [1] "6GYH_A" "7BMH_A" "5AWZ_A"
```

There are 3 hits that pass the statistical threshold, namely: 6GYH_A, 7BMH_A, 5AWZ_A. More information can be found on these by interrogating the blast results.

```
# Show the hit table for the top hits which pass the threshold
head(blast$hit.tbl, n = length(hit.IDs))

##       queryid subjectids identity alignmentlength mismatches gapopens
q.start
## 1 Query_40791     6GYH_A   47.222             144         74         2
35
## 2 Query_40791     7BMH_A   36.898             187        112         2
1
## 3 Query_40791     5AWZ_A   38.710             186         85         5
2
##   q.end s.start s.end   evalue bitscore positives mlog.evalue pdb.id
acc
## 1   177      77   219 7.74e-30      110     61.81    67.03115 6GYH_A
6GYH_A
## 2   181      92   278 2.13e-28      108     52.94    63.71626 7BMH_A
7BMH_A
```

```
## 3   176      61   228 4.29e-26      100     51.08    58.41093 5AWZ_A
5AWZ_A
```

We can also download these PDB files, annotate them for more information and align them with our sequence to get an overview of sequence alignment.

```
# Download related PDB files
files <- get.pdb(hits$pdb.id, path="pdbs", split=TRUE, gzip=TRUE)
```

```
##   |
|                                                                      |   0%
|
|=======================                                               |  33%
|
|===============================================                       |  67%
|
|======================================================================| 100%
```

```
# Align related PDBs
pdbs <- pdbaln(files, fit = TRUE, exefile="msa")
```

```
## Reading PDB files:
## pdbs/split_chain/6GYH_A.pdb
## pdbs/split_chain/7BMH_A.pdb
## pdbs/split_chain/5AWZ_A.pdb
##    PDB has ALT records, taking A only, rm.alt=TRUE
## .   PDB has ALT records, taking A only, rm.alt=TRUE
## .   PDB has ALT records, taking A only, rm.alt=TRUE
## .
##
## Extracting sequences
##
## pdb/seq: 1   name: pdbs/split_chain/6GYH_A.pdb
##    PDB has ALT records, taking A only, rm.alt=TRUE
## pdb/seq: 2   name: pdbs/split_chain/7BMH_A.pdb
##    PDB has ALT records, taking A only, rm.alt=TRUE
## pdb/seq: 3   name: pdbs/split_chain/5AWZ_A.pdb
##    PDB has ALT records, taking A only, rm.alt=TRUE
```

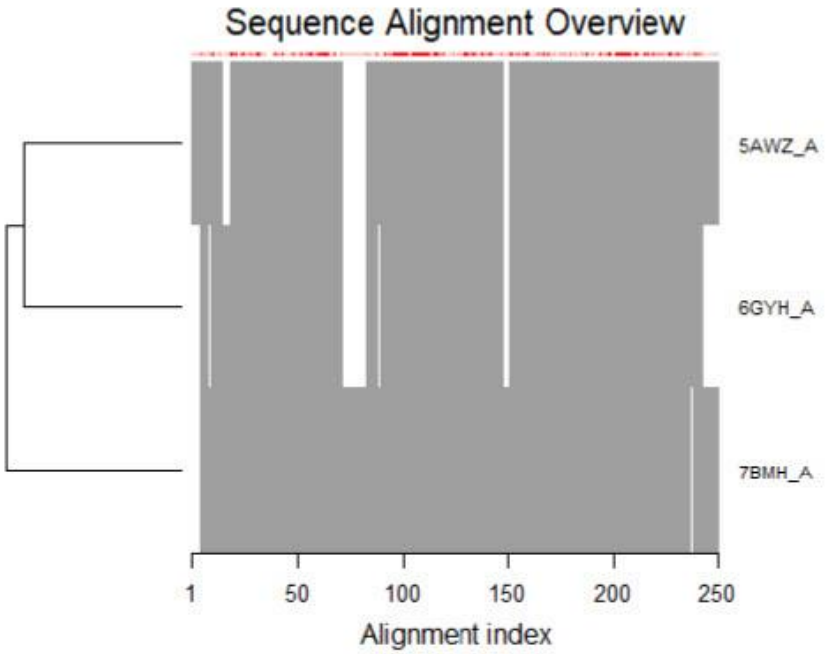
```
# Vector containing PDB codes for figure axis
ids <- basename.pdb(pdbs$id)

# Annotate hits for more information on the hits
anno <- pdb.annotate(ids)

# Find the organisms these PDB hits come from
unique(anno$source)
```

```
## [1] "Coccomyxa subellipsoidea C-169" "Leptosphaeria maculans"
## [3] "Acetabularia acetabulum"
```

```
# View more information on the hits
anno

##        structureId chainId macromoleculeType chainLength
experimentalTechnique
## 6GYH_A        6GYH       A           Protein         236
X-ray
## 7BMH_A        7BMH       A           Protein         324
X-ray
## 5AWZ_A        5AWZ       A           Protein         244
X-ray
##        resolution scopDomain
pfam
## 6GYH_A       2.00       <NA> Bacteriorhodopsin-like protein
(Bac_rhodopsin)
## 7BMH_A       2.20       <NA> Bacteriorhodopsin-like protein
(Bac_rhodopsin)
## 5AWZ_A       1.57       <NA> Bacteriorhodopsin-like protein
(Bac_rhodopsin)
##                                    ligandId
## 6GYH_A                      RET,CLR,OLB (4)
## 7BMH_A                     LFA (22),OLA (3)
## 5AWZ_A OCT (2),C14,RET,OLB,D12 (2),D10 (3)
##
ligandName
## 6GYH_A                                    RETINAL,CHOLESTEROL,(2S)-2,3-
dihydroxypropyl (9Z)-octadec-9-enoate (4)
## 7BMH_A
EICOSANE (22),OLEIC ACID (3)
## 5AWZ_A N-OCTANE (2),TETRADECANE,RETINAL,(2S)-2,3-dihydroxypropyl (9Z)-
octadec-9-enoate,DODECANE (2),DECANE (3)
##                                 source
## 6GYH_A Coccomyxa subellipsoidea C-169
## 7BMH_A         Leptosphaeria maculans
## 5AWZ_A        Acetabularia acetabulum
##
structureTitle
## 6GYH_A                   Crystal structure of the light-driven proton
pump Coccomyxa subellipsoidea Rhodopsin CsR
## 7BMH_A                       Crystal structure of a light-driven proton
pump LR (Mac) from Leptosphaeria maculans
## 5AWZ_A Crystal Structure of the Cell-Free Synthesized Membrane Protein,
Acetabularia Rhodopsin I, at 1.57 angstrom
##                                                                  citation
rObserved
## 6GYH_A                          Fudim, R., et al. Sci Signal (2019)
0.19398
## 7BMH_A                     Zabelskii, D., et al. Commun Biol (2021)
0.23840
## 5AWZ_A Furuse, M., et al. Acta Crystallogr D Biol Crystallogr (2015)
```

```
0.17760
##          rFree   rWork spaceGroup
## 6GYH_A 0.22493 0.19231        H 3
## 7BMH_A 0.28470 0.23610 P 21 21 21
## 5AWZ_A 0.19410 0.17690    C 1 2 1
```

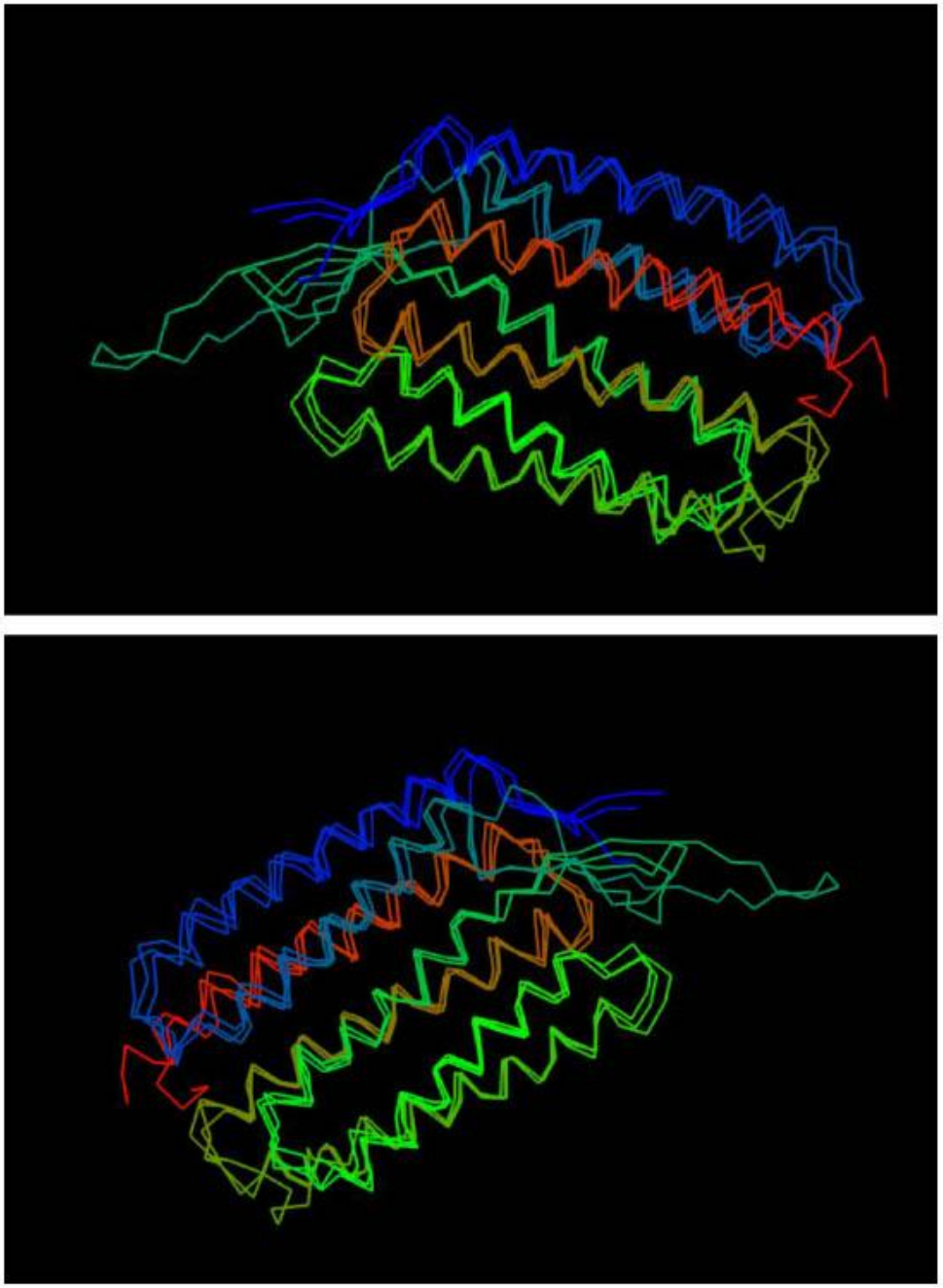
```
# Draw schematic alignment
plot(pdbs, labels=ids)
```



We can also plot the three structures we have found as follows:

```
# Set up
library(bio3d.view)
library(rgl)

# Plot
#view.pdbs(pdbs)
```

The `View.pdbs()` function brings up an interactive viewer, which cannot be directly viewed in the markdown document, so instead two screen-shots of this have been inserted.



With more proteins it could be interesting to plot variability, or even do PCA using the amino acid position data, but with only three proteins this is not useful.

Session Information

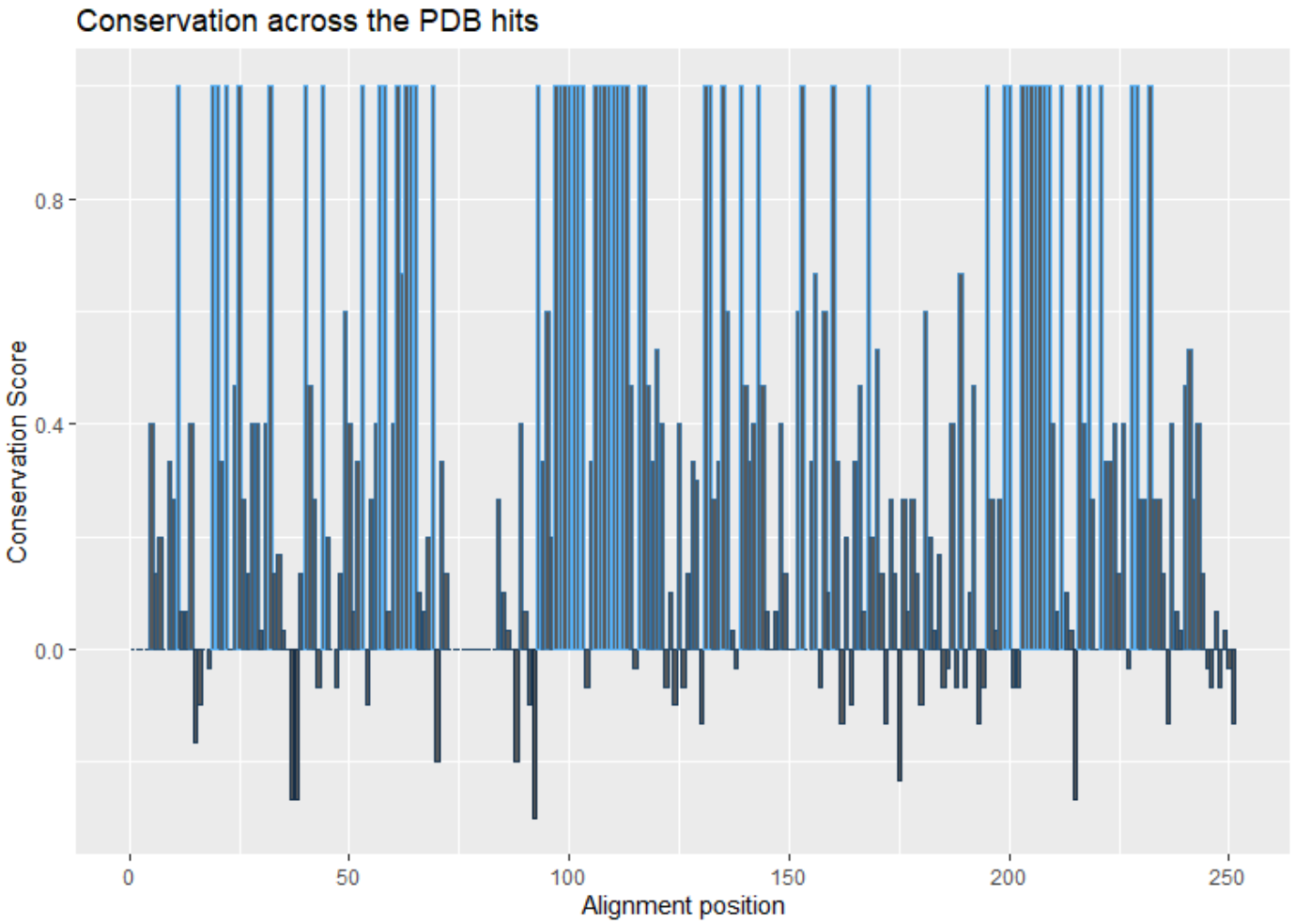
```
sessionInfo()

## R version 4.1.2 (2021-11-01)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 22000)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United Kingdom.1252
## [2] LC_CTYPE=English_United Kingdom.1252
## [3] LC_MONETARY=English_United Kingdom.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United Kingdom.1252
##
## attached base packages:
## [1] stats4    stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] rgl_0.108.3           bio3d.view_0.1.0.9000 msa_1.26.0
##  [4] Biostrings_2.62.0     GenomeInfoDb_1.30.1   XVector_0.34.0
##  [7] IRanges_2.28.0        S4Vectors_0.32.3      BiocGenerics_0.40.0
## [10] ggrepel_0.9.1         ggplot2_3.3.5         bio3d_2.4-3.9000
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.1.2       xfun_0.29              purrr_0.3.4
##  [4] colorspace_2.0-2       vctrs_0.3.8            generics_0.1.2
##  [7] htmltools_0.5.2        yaml_2.2.2             utf8_1.2.2
## [10] rlang_1.0.1            pillar_1.7.0           glue_1.6.1
## [13] withr_2.4.3            GenomeInfoDbData_1.2.7 lifecycle_1.0.1
## [16] stringr_1.4.0          zlibbioc_1.40.0        munsell_0.5.0
## [19] gtable_0.3.0           htmlwidgets_1.5.4      evaluate_0.15
## [22] knitr_1.37             extrafont_0.17         fastmap_1.1.0
## [25] curl_4.3.2             parallel_4.1.2         fansi_1.0.2
## [28] Rttf2pt1_1.3.10        highr_0.9              Rcpp_1.0.8
## [31] scales_1.1.1           jsonlite_1.8.0         digest_0.6.29
## [34] stringi_1.7.6          dplyr_1.0.8            grid_4.1.2
## [37] cli_3.2.0              tools_4.1.2            bitops_1.0-7
## [40] magrittr_2.0.2         RCurl_1.98-1.6         tibble_3.1.6
## [43] extrafontdb_1.0        crayon_1.5.0           pkgconfig_2.0.3
## [46] ellipsis_0.3.2         httr_1.4.2             rmarkdown_2.11
## [49] rstudioapi_0.13        R6_2.5.1               compiler_4.1.2
```

Print-out 9 - 16: R Markdown document of Bio3D investigation of the shorter predicted ORF

Question 9:

As **6GYH_A** has the smallest e-value for the longer ORF (and thus represents the best hit) this is the PDB structure I will generate a molecular figure for.

To get an idea of the most conserved residues the code used previously finding these values for the conservation plots of multiple sequence alignments can be used.



Plot 5: Conservation across the PDB hits

In VMD we can load the PDB files downloaded for all three proteins, and then use the MultiSeq extension to consider conservation. Unfortunately, I was not able to get this extension to work. Therefore, I simply listed all the 100% conserved residues and created a visualization state for them, namely, in the image below, all the residues with a color other than the main teal are 100% conserved.

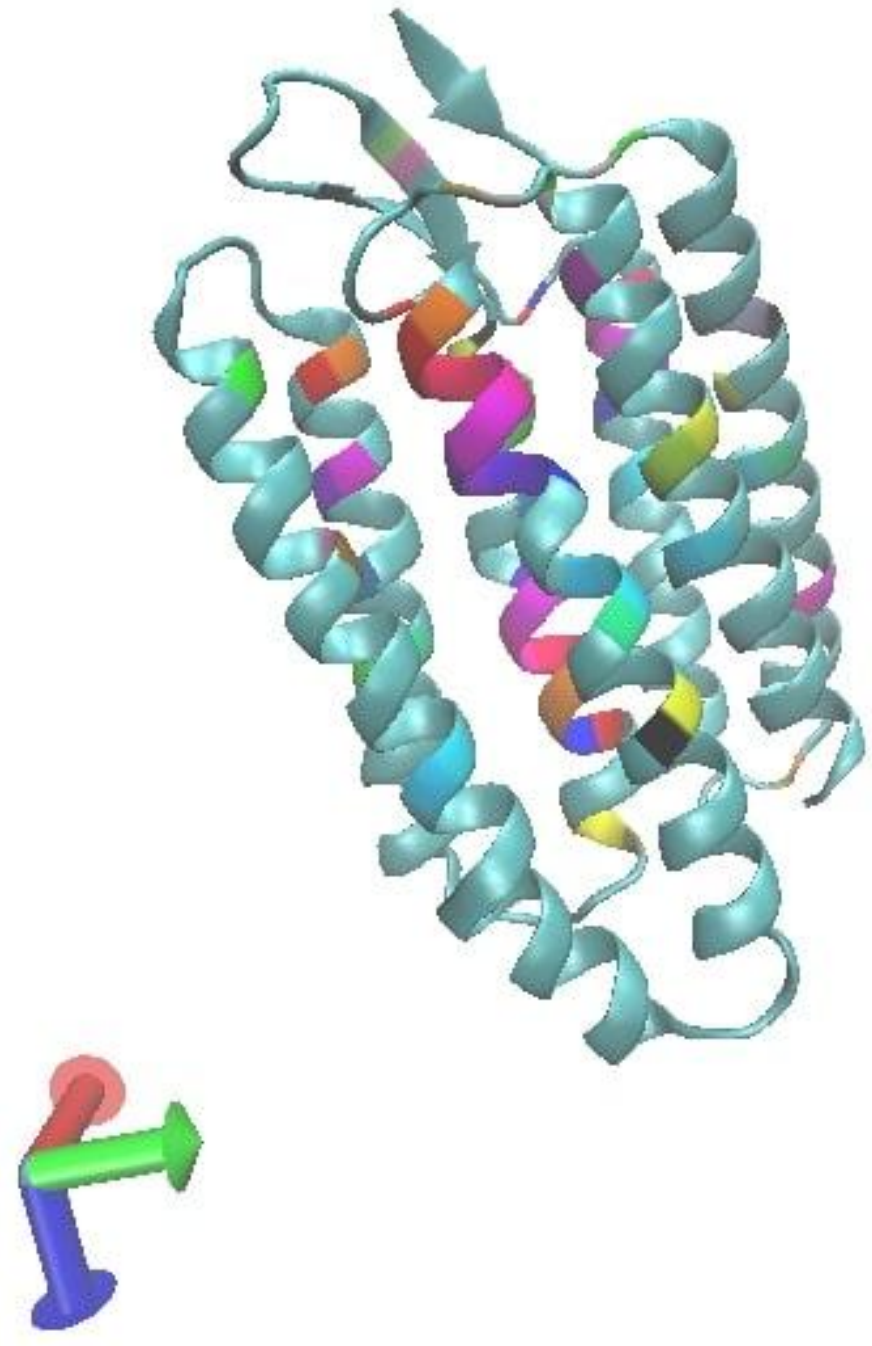


Image 1: VMD rendering of 6GHY, with 100% conserved residues (between the three hits) coloured by residue ID and represented in the licorice style, the rest of the protein is coloured teal and represented in the newCartoon style

Surprisingly, these residues are quite spread out, with no clear pattern, and some even existing in the loop region (a usually less conserved region). However, when the placing of the side chains of these residues is considered, most of them (though not all) are on the inside of the protein, where packing is tighter and the cofactor for bacteriorhodopsin and rhodopsin is found (at the center between the alpha helices). Both of these are expected to constrain change and so the conservation makes more sense in this light.

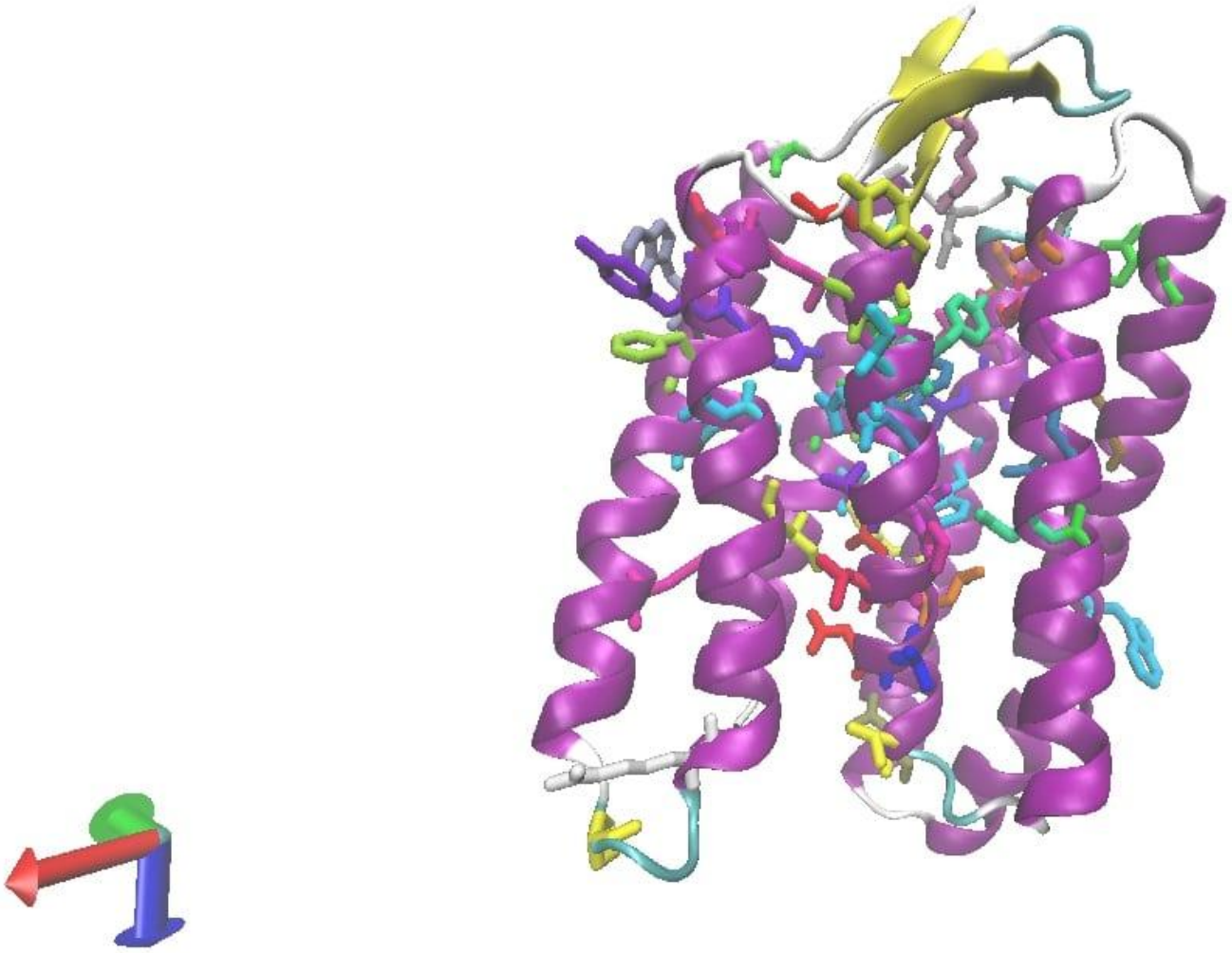


Image 2: VMD rendering of 6GHY, with 100% conserved residues (between the three hits) coloured by residue ID and represented in the licorice style, the rest of the protein is coloured by secondary structure and represented in the newCartoon style

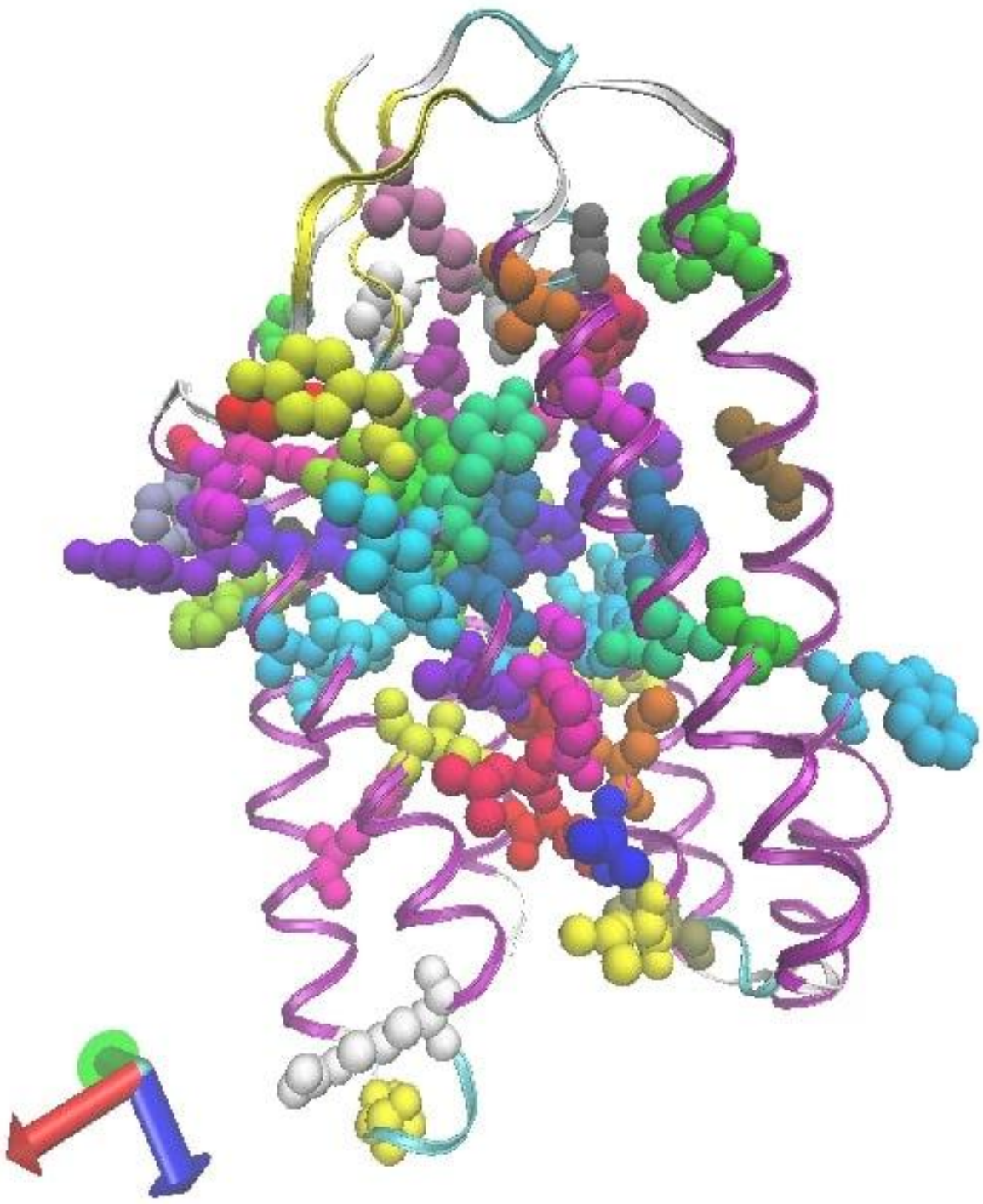
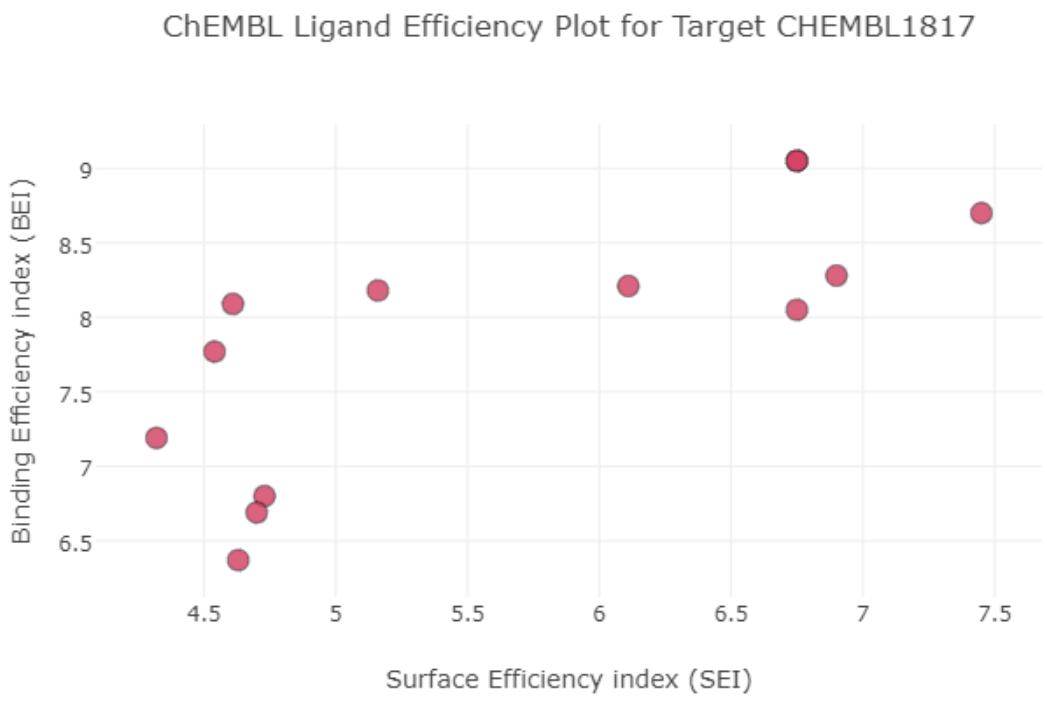


Image 3: VMD rendering of 6GHY, with 100% conserved residues (between the three hits) coloured by residue ID and represented in the VDW style, the rest of the protein is coloured by secondary structure and represented in the ribbon style. The size of the VDW spheres is enlarged for clarity.

Question 10:

The longer predicted ORF came up with 8 hits, however, the top one has an e value of 0.24, and the rest all have e values greater than 1. They are therefore, not very reliable hits. The top hit with the e value of 0.24 is an Erythropoietin receptor with ID CHEMBL1817. It has 9 drug and clinical candidates associated with it, and bioactivity assays data is present in the profile (binding, inhibition, IC50 etc). Furthermore, there are 36 compounds tested for ligand efficiency (see figure below) and 61 compounds are associated with it. However, while these may provide a useful starting point for investigation of the novel protein, it is unlikely given that the e value for this hit does not make the significance threshold.



Plot 6: Ligand Efficiency Plot from CHEMBL