Matt Demelo

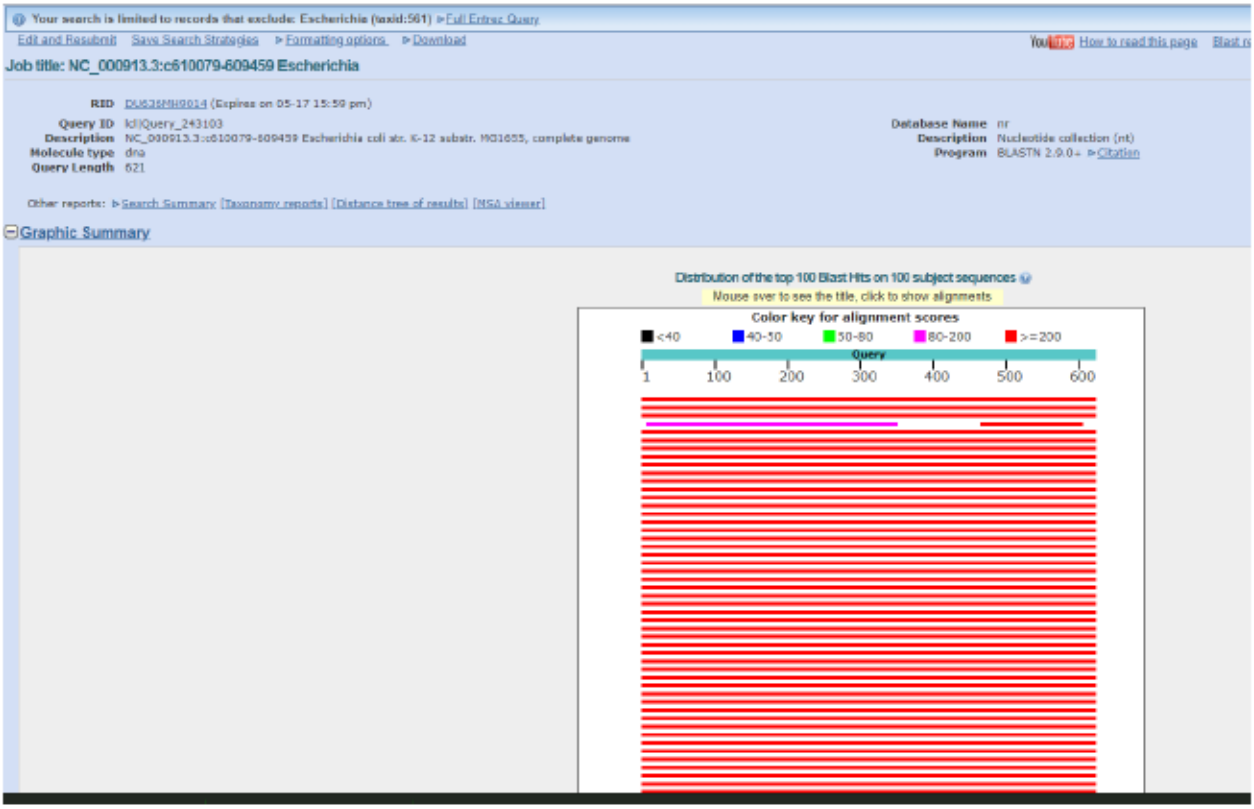
mdemelo@ucsd.edu

BIMM 143, 05/15/19

Find-A-Gene Project: An *entD* homolog in *Shigella dysenteriae* strain ATCC 12039

1) The gene I am interested in is the *ent*D gene, encoding the EntD protein (phosphopantetheinyl transferase) involved in enterobactin synthesis.
    a. Found in *Escherichia coli* str. K12, substr. MG1655. P
    b. Accession Number:
        i. Protein is NP_415115.2;
        ii. DNA sequence is NC_000913.3:c610079-609459.

2) BLAST Information:
    a. NCBI BLASTn
    b. Algorithm: Discontiguous megablast
    c. Limits: Exclusion for *Escherichia* (taxid: 561)
    d. Database: Nucleotide collection (nt)



Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
Triticum aestivum mRNA, clone: tplb0030f10, cultivar Chinese Spring	1121	1121	100%	0	100.00%	AK447339.1
Shigella sp. PAMC 28760, complete genome	1121	1121	100%	0	100.00%	CP014768.1
Uncultured bacterium Contigcl 1559 genomic sequence	1121	1121	100%	0	100.00%	KC246861.1
Sparus aurata clone contig01037 genomic sequence	249	249	22%	2.00E-61	99.29%	HQ021737.1
Shigella sonnei strain FC1653 chromosome, complete genome	1049	1049	100%	0	97.42%	CP037997.1
Shigella sonnei strain LC1477/18 chromosome, complete genome	1049	1049	100%	0	97.42%	CP035008.1

**This was shortened significantly since there were nearly 100 alignments from this BLAST search. Continued below:

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
Shigella flexneri enterobactin biosythesis (entD) gene, complete cds	958	958	100%	0	94.20%	U52684.1
Shigella dysenteriae strain ATCC 12039 chromosome, complete genome	945	945	100%	0	93.72%	CP026831.1
Otolemur crassicaudatus epsilon-, gamma-, delta-, and beta-globin genes, complete cds, and eta-globin pseudogene	530	530	59%	4.00E-146	91.73%	U60902.1

The following alignment was chosen as a potential novel gene, since the sequence was fairly similar, but was not mentioned to be a homologue of *ent*D, and the genome of the organism was only recently sequenced.

Shigella dysenteriae strain ATCC 12039 chromosome, complete genome
Sequence ID: CP026831.1 **Length:** 4880735 **Number of Matches:** 1

Range 1: 3239875 to 3240495 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
945 bits(1047)	0.0	582/621(94%)	0/621(0%)	Plus/Plus

```
Query  1        ATGAAAACTACGCATACCTCCCTCCCCTTTGCCGGACATACGCTGCATTTTGTTGAGTTC  60
               ||||||||||||||||| ||||||||||||||||||||||||||||||||||| |||||
Sbjct  3239875  ATGAAAACTACGCATACCGCCCTCCCCTTTGCCGGACATACGCTGCATTTTGTTAAGTTC  3239934

Query  61       GATCCGGCGAATTTTTGTGAGCAGGATTTACTCTGGCTGCCGCACTACGCACAACTGCAA  120
               |||||||||| |||||||||||||||||||||||||||| |||||||| ||||||||||
Sbjct  3239935  GATCCGGCGAGTTTTTGTGAGCAGGATTTACTCTGGCTGCCACACTACGCGCAACTGCAA  3239994

Query  121      CACGCTGGACGTAAACGTAAAACAGAGCATTTAGCCGGACGGATCGCTGCTGTTTATGCT  180
               ||||||||||||||||||||||||||||||||||||||||||||||||||| ||||||||
Sbjct  3239995  CACGCTGGACGTAAACGTAAAACAGAGCATTTAGCCGGACGGATCGCTGCAGTTTATGCG  3240054

Query  181      TTGCGGGAATATGGCTATAAATGTGTGCCCGCAATCGGCGAGCTACGCCAACCTGTCTGG  240
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||| |||
Sbjct  3240055  CTGCGGGAATATGGCTATAAATGTGTGCCCGCAATCGGCGAGCTACGCCAACCTGTGTGG  3240114

Query  241      CCTGCGGAGGTATACGGCAGTATTAGCCACTGTGGGACTACGGCATTAGCCGTGGTATCT  300
               ||||||| ||||||||||||| ||||| |||||||||||||||||||||||||||||||
Sbjct  3240115  CCTGCGGGGGTATACGGCAGCATTAGTCACTGTGGGACTACGGCATTAGCCGTGGTATCT  3240174

Query  301      CGTCAACCGATTGGCATTGATATAGAAGAAATTTTTCTGTACAAACCGCAAGAGAATTG  360
               |||||||| |||||||||||||| ||||| ||||| |||||||||||||||||||||||
Sbjct  3240175  CGTCAACCAATTGGCATTGATATCGAAGAGATTTTCTCTGTACAAACCGCAAGAGAATTG  3240234

Query  361      ACAGACAACATTATTACACCAGCGGAACACGAGCGACTCGCAGACTGCGGTTTAGCCTTT  420
               ||| |||||||||||||||| |||||||||||||||||| ||||||||| |||||
Sbjct  3240235  ACAAACAACATTATTACACCAGCAGAACACGAGCGACTCGCAGAATGCGGTTTAACCTTT  3240294

Query  421      TCTCTGGCGCTGACACTGGCATTTTCCGCCAAAGAGAGCGCATTTAAGGCAAGTGAGATC  480
               |||||||||||||||||||||||||||||||||||||||||||||||||||||| ||||
Sbjct  3240295  TCTCTGGCGCTGACACTGGCATTTTCCGCCAAAGAGAGCGCATTTAAGGCAAGCAAGATA  3240354

Query  481      CAAACTGATGCAGGTTTTCTGGACTATCAGATAATTAGCTGGAATAAACAGCAGGTCATC  540
               ||  | | |||||||||||| ||||||||||||||||||||||||||| |||||
Sbjct  3240355  CAGGCGGCTCAAGGTTTTCTGGATTATCAGATAATTAGCTGGAATAAACAGCAGATCATC  3240414

Query  541      ATTCATCGTGAGAATGAGATGTTTGCTGTGCACTGGCAGATAAAAGAAAAGATAGTCATA  600
               |||| |  ||| |||||  ||||||||||||||||||||||||||| ||||||||||
Sbjct  3240415  ATTCGACTAGAGGATGAGCAGTTTGCTGTGCACTGGCAGATAAAAGAAAAAATAGTCATA  3240474

Query  601      ACGCTGTGCCAACACGATTAA  621
               |||||||||||||||||||||
Sbjct  3240475  ACGCTGTGCCAACACGATTAA  3240495
```

*Note, this alignment came from much farther down the list of alignments. The BLAST search request ID was DU636MH9014.

3) The 'novel' gene was derived from Shigella dysenteriae strain ATCC 12039, and is not given any particular name; it is only indicated as a portion of a complete genome sequence. The protein amino acid sequence of the gene was generated using the EMBOSS tool from EBI, translated into six potential reading frames:

```
>3239875-3240495_1 Shigella dysenteriae strain ATCC 12039 chromosome, complete genome
MKTTHTALPFAGHTLHFVKFDPASFCEQDLLWLPHYAQLQHAGRKRKTEHLAGRIAAVYA
LREYGYKCVPAIGELRQPVWPAGVYGSISHCGTTALAVVSRQPIGIDIEEIFSVQTAREL
TNNIITPAEHERLAECGLTFSLALTLAFSAKESAFKASKIQAAQGFLDYQIISWNKQQII
IRLEDEQFAVHWQIKEKIVITLCQHD*
```

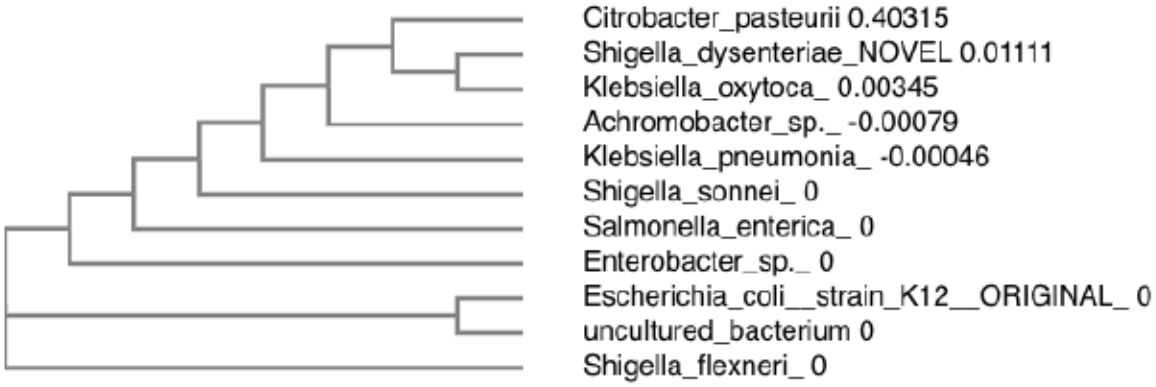
4) The gene is very likely a novel gene, as the BLASTp search brings up high identity alignments to *ent*D homologues in various species below 100%. Most notably, the top alignment is a non-annotated, unnamed protein product, and none of the top few alignments are aligned with sequences from *Shigella* species; one of the top alignments is an *E. coli* sequence, but falls below 100% identity. Several of the top matches are to *E. coli*-derived sequences, indicating that the gene is very likely homologue to my original *E. coli ent*D query. The top hits for the pblast nr-database search are listed below:

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
unnamed protein product	427	427	99%	2.00E-151	100.00%	WP_025210285.1
MULTISPECIES: enterobactin synthase subunit EntD [Enterobacteriaceae]	427	427	99%	2.00E-151	100.00%	WP_077768650.1
4'-phosphopantetheinyl transferase Npt [Escherichia coli]	425	425	99%	7.00E-150	98.54%	AOM68887.1
4'-phosphopantetheinyl transferase [Escherichia coli LAU-EC8]	425	425	99%	1.00E-149	98.54%	ETE13210.1
4'-phosphopantetheinyl transferase Npt [Escherichia coli]	424	424	99%	1.00E-149	98.54%	RDQ85810.1

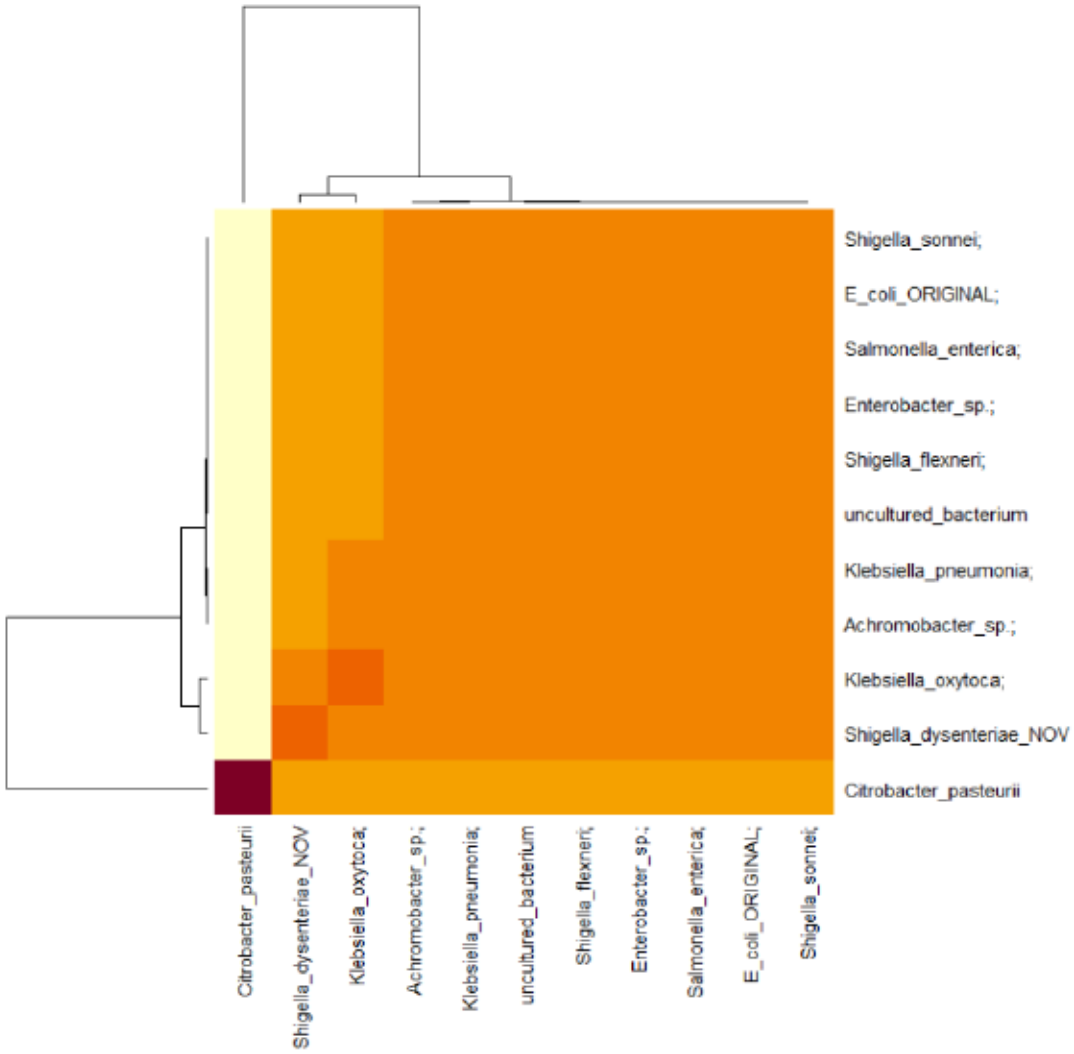
5) Multiple Sequence Alignment using CLUSTAL Omega:

```
CLUSTAL O(1.2.4) multiple sequence alignment

Citrobacter_pasteurii                    ----------------------------------------MMHTTHTPLTFADQTLHIVE          20
Escherichia_coli_(strain_K12)_ORIGINAL;   ----------------------------------------MKTTHTSLPFAGHTLHFVE          19
Shigella_sonnei;                         ----------------------------------------MKTTHTSLPFAGHTLHFVE          19
Salmonella_enterica;                     ----------------------------------------MKTTHTSLPFAGHTLHFVE          19
uncultured_bacterium                     -------------------------------------MVDMKTTHTSLPFAGHTLHFVE          22
Enterobacter_sp.;                        ----------------------------------------MKTTHTSLPFAGHTLHFVE          19
Shigella_flexneri;                       ----------------------------------------MKTTHTSLPFAGHTLHFVE          19
Achromobacter_sp.;                       MNALSGLQKSCQFNILQDHVGLISVAHQAVLRLSSVSNMVDMKTTHTSLPFAGHTLHFVE          60
Klebsiella_pneumonia;                    MNALSGLQKSCQFNILQDHVGLISVAHQAVLRLSSVSNMVDMKTTHTSLPFAGHTLHFVE          60
Shigella_dysenteriae_NOVEL               ----------------------------------------MKTTHTALPFAGHTLHFVK          19
Klebsiella_oxytoca;                      MNALSGLQKSCQFNILQDHVGLISVAHQAVLRLSSVSNMVDMKTTHTALPFVGHTLHFVE          60
                                                                                  *:**** * *..:***:*:

Citrobacter_pasteurii                    FDPNSFHEHDLLWLPHHAQLTAGARKRKAEHLAGRIAAIHALREYGIKTVPGIGEQRQPL          80
Escherichia_coli_(strain_K12)_ORIGINAL;   FDPANFCEQDLLWLPHYAQLQHAGRKRKTEHLAGRIAAVYALREYGYKCVPAIGELRQPV          79
Shigella_sonnei;                         FDPANFCEQDLLWLPHYAQLQHAGRKRKTEHLAGRIAAVYALREYGYKCVPAIGELRQPV          79
Salmonella_enterica;                     FDPANFCEQDLLWLPHYAQLQHAGRKRKTEHLAGRIAAVYALREYGYKCVPAIGELRQPV          79
uncultured_bacterium                     FDPANFCEQDLLWLPHYAQLQHAGRKRKTEHLAGRIAAVYALREYGYKCVPAIGELRQPV          82
Enterobacter_sp.;                        FDPANFCEQDLLWLPHYAQLQHAGRKRKTEHLAGRIAAVYALREYGYKCVPAIGELRQPV          79
Shigella_flexneri;                       FDPANFCEQDLLWLPHYAQLQHAGRKRKTEHLAGRIAAVYALREYGYKCVPAIGELRQPV          79
Achromobacter_sp.;                       FDPANFCEQDLLWLPHYAQLQHAGRKRKTEHLAGRIAAVYALREYGYKCVPAIGELRQPV          120
Klebsiella_pneumonia;                    FDPANFCEQDLLWLPHYAQLQHAGRKRKTEHLAGRIAAVYALREYGYKCVPAIGELRQPV          120
Shigella_dysenteriae_NOVEL               FDPASFCEQDLLWLPHYAQLQHAGRKRKTEHLAGRIAAVYALREYGYKCVPAIGELRQPV          79
Klebsiella_oxytoca;                      FDPASFCEQDLLWLPHYAQLQHAGRKRKTEHLAGRIAAVYALREYGYKCVPAIGELRQPV          120
                                         *** .* *:*******:***  ..****:*********::****** * **.*** ***:

Citrobacter_pasteurii                    WPHGLPGSISHSATTALAVVSCHPVGLDIEAIFSPQVAVELTDSIIDNTERQVLLHASLP          140
Escherichia_coli_(strain_K12)_ORIGINAL;   WPAEVYGSISHCGTTALAVVSRQPIGIDIEEIFSVQTARELTDNIITPAEHERLADCGLA          139
Shigella_sonnei;                         WPAEVYGSISHCGTTALAVVSRQPIGIDIEEIFSVQTARELTDNIITPAEHERLADCGLA          139
Salmonella_enterica;                     WPAEVYGSISHCGTTALAVVSRQPIGIDIEEIFSVQTARELTDNIITPAEHERLADCGLA          139
uncultured_bacterium                     WPAEVYGSISHCGTTALAVVSRQPIGIDIEEIFSVQTARELTDNIITPAEHERLADCGLA          142
Enterobacter_sp.;                        WPAEVYGSISHCGTTALAVVSRQPIGIDIEEIFSVQTARELTDNIITPAEHERLADCGLA          139
Shigella_flexneri;                       WPAEVYGSISHCGTTALAVVSRQPIGIDIEEIFSVQTARELTDNIITPAEHERLADCGLA          139
Achromobacter_sp.;                       WPAEVYGSISHCGTTALAVVSRQPIGIDIEEIFSVQTARELTDNIITPAEHERLADCGLA          180
Klebsiella_pneumonia;                    WPAEVYGSISHCGTTALAVVSRQPIGIDIEEIFSVQTARELTDNIITPAEHERLADCGLA          180
Shigella_dysenteriae_NOVEL               WPAGVYGSISHCGTTALAVVSRQPIGIDIEEIFSVQTARELTNNIITPAEHERLAECGLT          139
Klebsiella_oxytoca;                      WPAGVYGSISHCGTTALAVVSRQPIGIDIEEIFSVQTARELTDNIITPAEHERLAECGLT          180
                                         **  ::*****..******** :*:*:*** *** *.* ***:.**  :*:: * ...*

Citrobacter_pasteurii                    FPLALTLVFSAKESLYKAFSAHLTHLPGFSSANVIALTTTQITLQITPSFSQSLAGLSVN          200
Escherichia_coli_(strain_K12)_ORIGINAL;   FSLALTLAFSAKESAFKASEI--QTDAGFLDYQIISWNKQQVIIHR--------ENEMFA          189
Shigella_sonnei;                         FSLALTLAFSAKESAFKASEI--QTDAGFLDYQIISWNKQQVIIHR--------ENEMFA          189
Salmonella_enterica;                     FSLALTLAFSAKESAFKASEI--QTDAGFLDYQIISWNKQQVIIHR--------ENEMFA          189
uncultured_bacterium                     FSLALTLAFSAKESAFKASEI--QTDAGFLDYQIISWNKQQVIIHR--------ENEMFA          192
Enterobacter_sp.;                        FSLALTLAFSAKESAFKASEI--QTDAGFLDYQIISWNKQQVIIHR--------ENEMFA          189
Shigella_flexneri;                       FSLALTLAFSAKESAFKASEI--QTDAGFLDYQIISWNKQQVIIHR--------ENEMFA          189
Achromobacter_sp.;                       FSLALTLAFSAKESAFKASEI--QTDAGFLDYQIISWNKQQVIIHR--------ENEMFA          230
Klebsiella_pneumonia;                    FSLALTLAFSAKESAFKASEI--QTDAGFLDYQIISWNKQQVIIHR--------ENEMFA          230
Shigella_dysenteriae_NOVEL               FSLALTLAFSAKESAFKASKI--QAAQGFLDYQIISWNKQQIIIRL--------EDEQFA          189
Klebsiella_oxytoca;                      FSLALTLAFSAKESAFKASKI--QAAQGFLDYQIISWNKQQIIIRL--------EDEQFA          230
                                         * *****.****** :** .       ** . ::*: .. *: ::          .  .

Citrobacter_pasteurii                    VSWFRREENIITLCPAPAFSV          221
Escherichia_coli_(strain_K12)_ORIGINAL;   VHWQIKEKIVITLCQHD----          206
Shigella_sonnei;                         VHWQIKEKIVITLCQHD----          206
Salmonella_enterica;                     VHWQIKEKIVITLCQHD----          206
uncultured_bacterium                     VHWQIKEKIVITLCQHD----          209
Enterobacter_sp.;                        VHWQIKEKIVITLCQHD----          206
Shigella_flexneri;                       VHWQIKEKIVITLCQHD----          206
Achromobacter_sp.;                       VHWQIKEKIVITLCQHD----          247
Klebsiella_pneumonia;                    VHWQIKEKIVITLCQHD----          247
Shigella_dysenteriae_NOVEL               VHWQIKEKIVITLCQHD*---          206
Klebsiella_oxytoca;                      VHWQIKEKIVITLCQHD----          247
                                         * *  :*: :****
```

6) Simple Phylogenetic tree of sequences, generated with EBI's simple phylogeny:



7) Sequence identity heatmap, generated using the **bio3d** package in R:

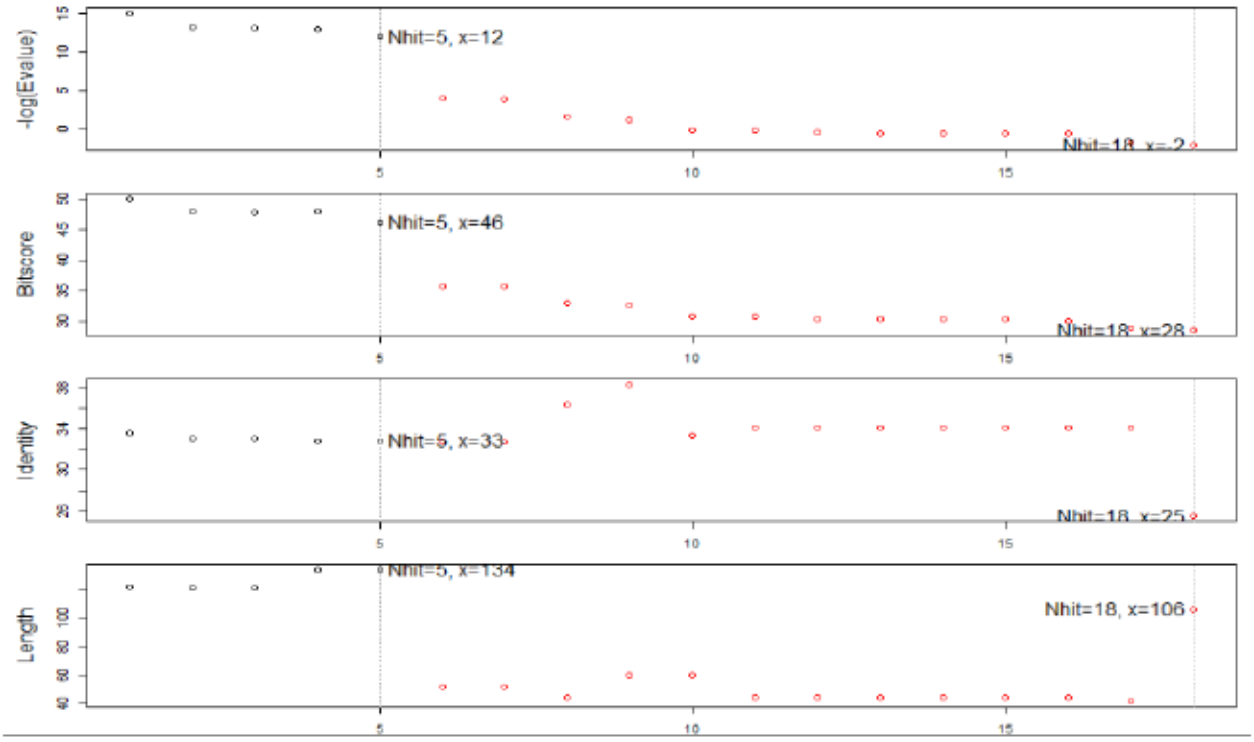


8) Search for homologous protein structures using the **bio3d** package, searching for similar sequences in the PDB database. Below are the top 3 hits for structure, as well a plot hits for a pdb.blast performed on the novel sequence against the PDB database.

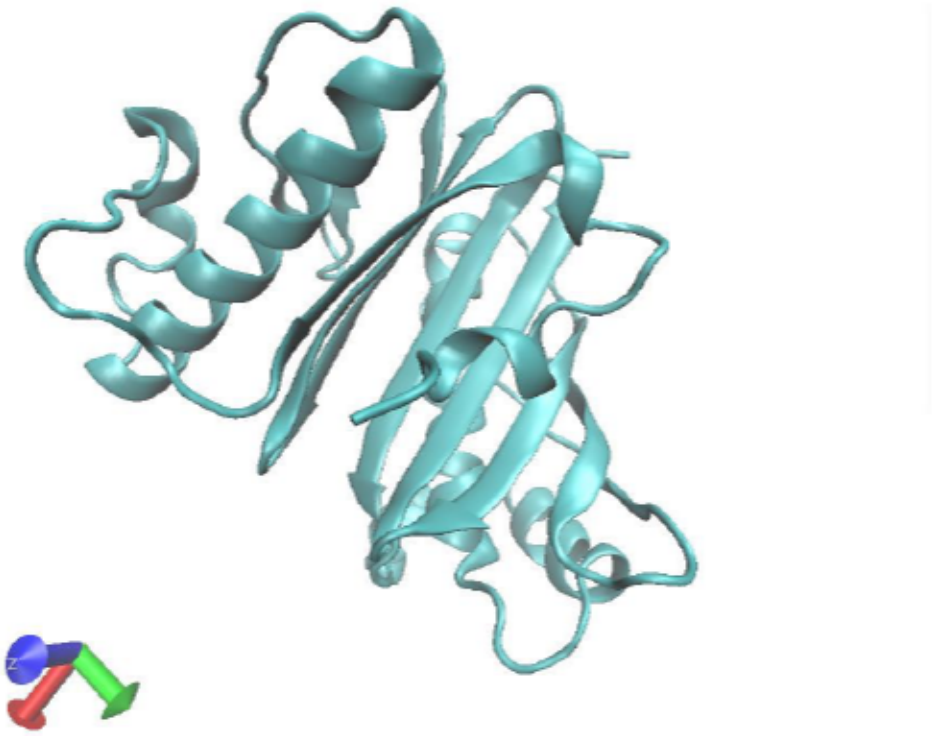
Top 3 sequence identity hits:

	structureId	experimentalTechnique	resolution	source	identity	evalue
1	4QJL	X-RAY DIFFRACTION	1.65	Mycobacterium ulcerans	33.607	3.06e-07
2	4U89	X-RAY DIFFRACTION	1.4	Mycobacterium tuberculosis	33.058	1.92e-06
3	6CT5	X-RAY DIFFRACTION	1.76	Mycobacterium tuberculosis	33.058	1.98e-06

Plot of sequence identity hits from pdb.blast for the novel sequence against the PDB database (x-axis is hit number):



9) 4QJL PDB structure generated using VMD:



It is probable that this structure homology model is similar to the structure of our novel protein, since its sequence similarity is approximately 33.6% of that novel protein query, which is above the sequence identity 25% cutoff that is traditionally used to determine if a protein is structurally homologous to a known structure. However, the fact that the structure comes from a completely separate genera of bacteria (*Mycobacterium*), might make this homology model somewhat questionable. Regardless, it is reasonable to believe that this model adequately represents our novel protein.

10) The initial screen of the ChEMBEL database yielded four compounds. Further screening showed several assays associated to this specific target for *Shigella sp.* found doing a ChEMBEL search: 15 assays for studying minimum inhibitory concentration of compounds and for assessing time-to-kill for compounds. Unfortunately, none of the identified compounds had any ligand efficiency data associated with them, which limits how much can be determined the ChEMBEL search. The list of assays would be very helpful for exploring inhibition of this novel protein, since they are basic anti-microbial compound screening assays. As a result, this ChEMBEL search would be of mild use for finding a potential inhibitor of the protein, as the assays in question would be very useful for screening inhibitors, but the lack of ligation efficiency data for the compounds searched makes it difficult to define a starting point.