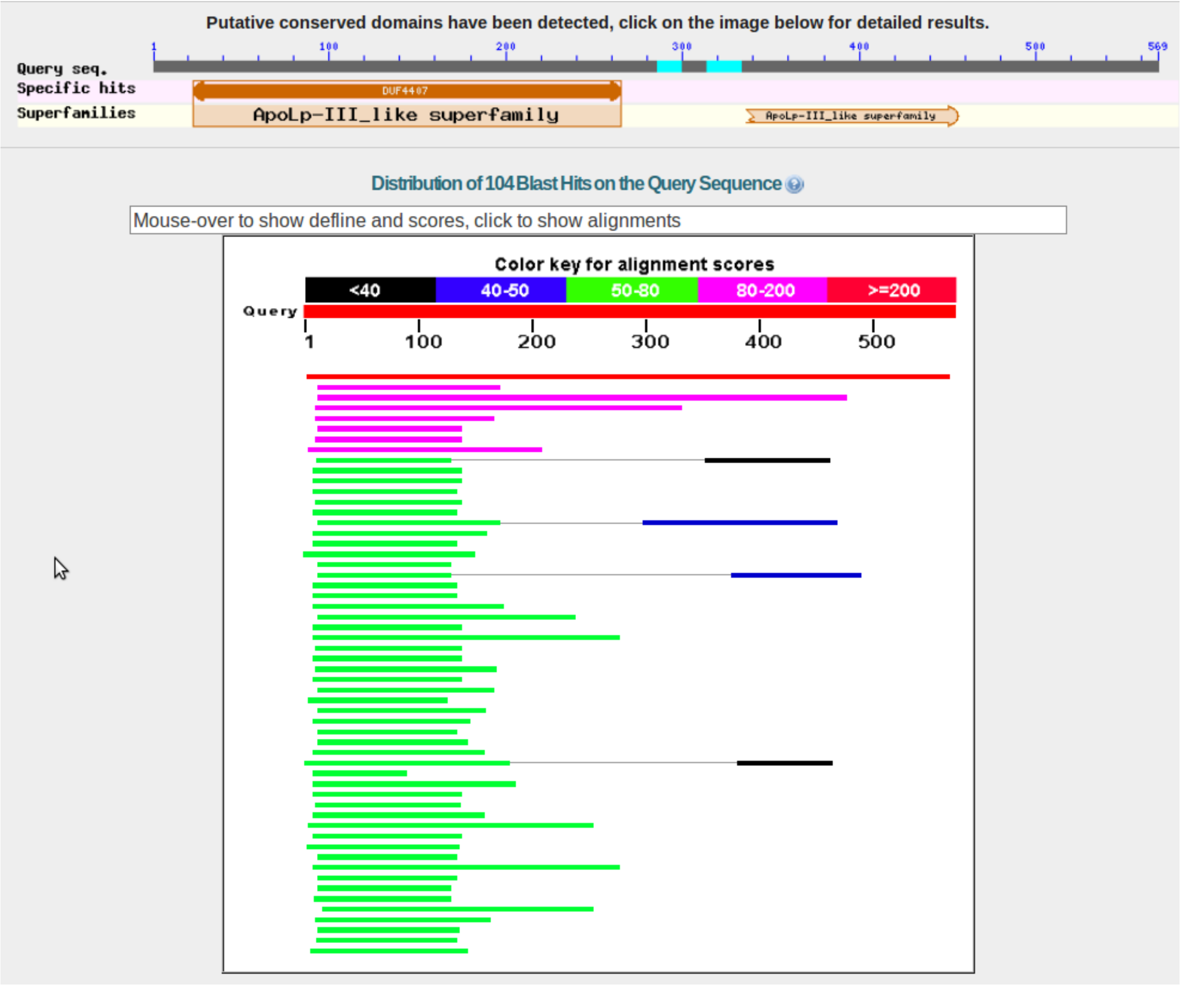
Lab Report 2

Zachary Foster

April 18, 2014

Dr. Muller suppiled us with contigs generated from a metagemoic assembly of DNA from a sample of marine water. The contigs are thought to be an undescribed lineage wihtin *verrucomicrobia*. We used BLAST, GLIMMER, and web-based tools to idenitify ORFs within one of the contigs and pridict possible functions. Using an interpolated markov model supplied by the instructors, we used GLIMMER to pridict ORFs. We used a custom `perl` script `parse_Glimmer.pl` to extract the amino acid sequence of the predicted ORFs. We then made a BLAST database, using sequences supplied by the instructors, and used the database to identify the ORFs. 4 ORFs were found, 2 of which return significant BLAST hits using an -e threshold of 0.01. Each of the 2 ORFs with siginificant hits had one identical hit and two very weak hits. We then used a custom `perl` script `grab_seq.pl` to exctract one of the ORFs with and identical hit to the reference database.

The extracted sequence was search aginst the genbank nr database using blastp with default settings 1. All of the hits were hypothetical protiens and only one was full length. Next, a PSI blast search with 3 iterations, useing default parameters, was used in order to discover more distaly realted putative homologs. This revealed many more full length matches, but all were stil hypothetical protiens 2. We used the COG and CCD databases to search for conserved domains using default parameters. The COG database search returned a hit to a chromosome segregation ATPase 3, but the CCD databasse search did not return any hits with known function 4 Finally, we searched pfam in order to detect any protien families that the unknown sequence may be a part of, but this to return no significant result with known function. The 2 hits that were found were of unknown function 5. From all of the analyses overall, it appears that this ORF is ether a novel protien that has not been assigned a faunction yet, or a conserved sequence suseptible to missaignment by automatic *ab inito* gene annotaing program, since there are many hits to multiple databases, but little functional information. The one signifiacnt hit with functional information was a chromosome segregation ATPase. but this alone is not suffiecnt evidence to predict the function of this unknown sequence.

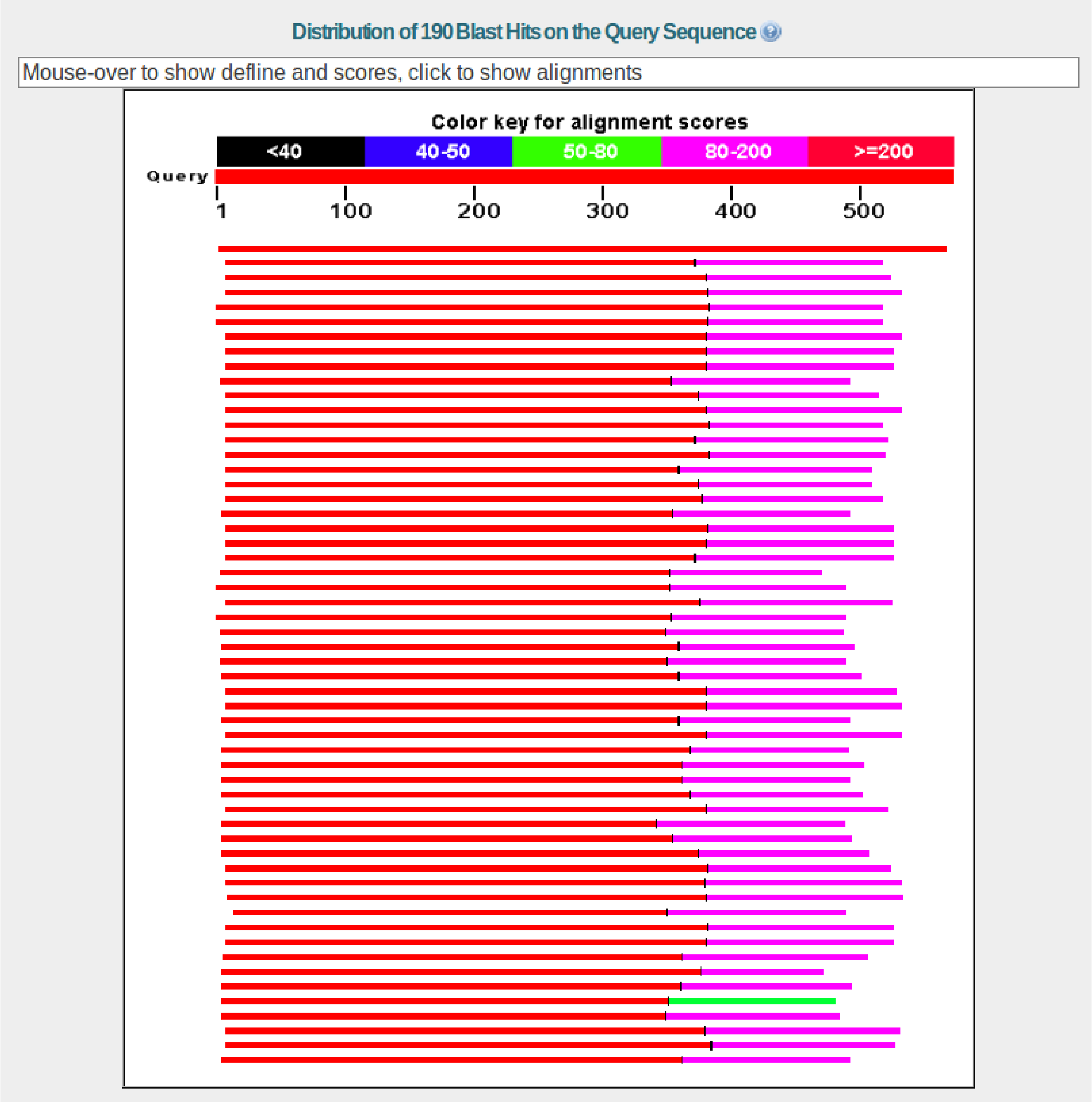


Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of 104 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments

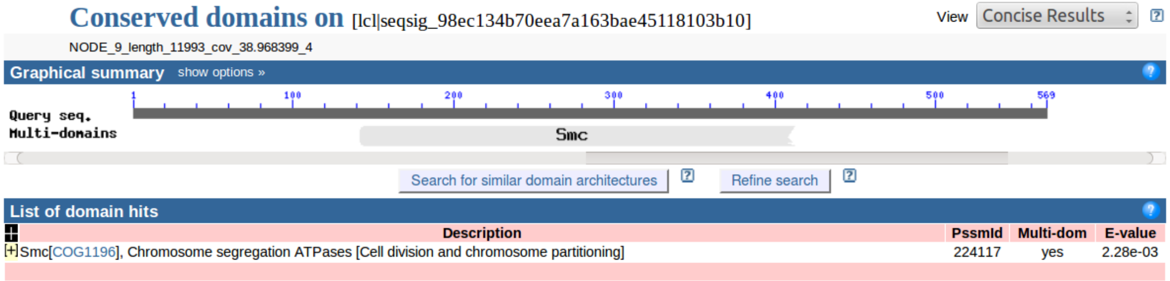
Figure 1: Screenshot of blastp results



Distribution of 190 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments

Figure 2: Screenshot of PSI-BLAST results



Conserved domains on [lcl|seqsig_98ec134b70eea7a163bae45118103b10]

NODE_9_length_11993_cov_38.968399_4

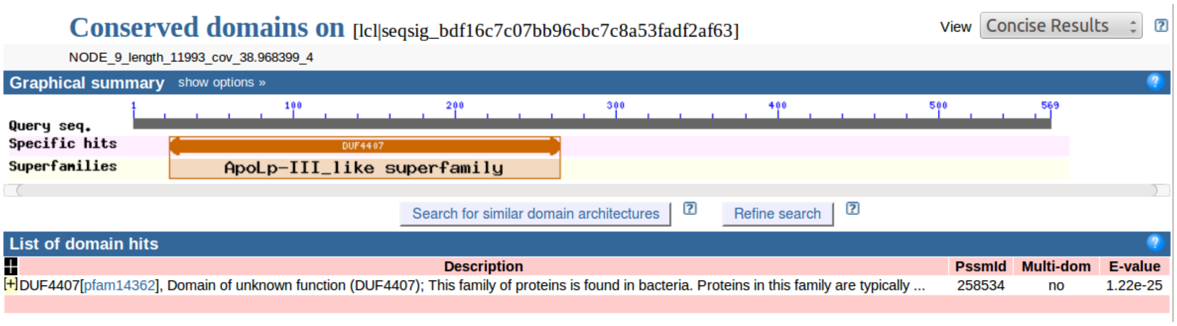
View Concise Results

Graphical summary show options »

List of domain hits

Description	PssmId	Multi-dom	E-value
[+]Smc[COG1196], Chromosome segregation ATPases [Cell division and chromosome partitioning]	224117	yes	2.28e-03

Figure 3: Screenshot of COG search results.

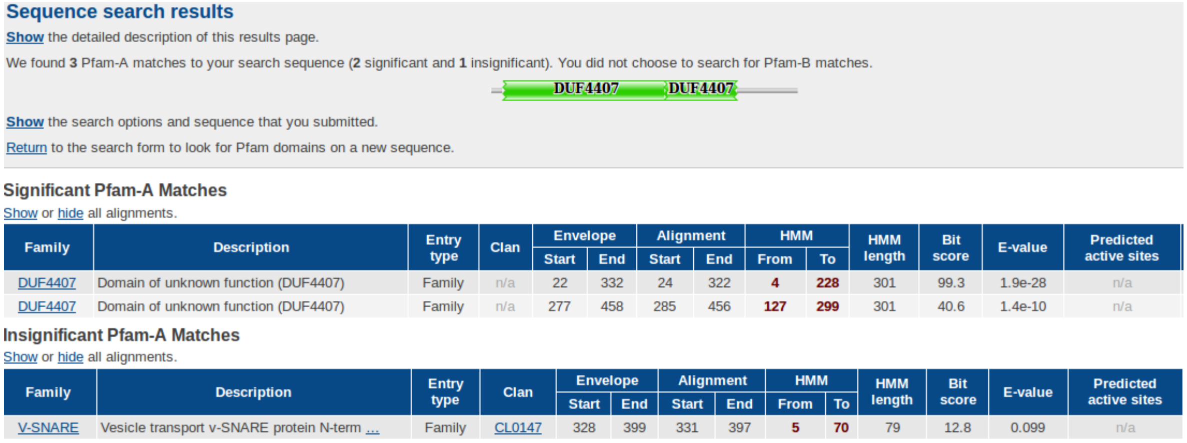


Conserved domains on [lcl|seqsig_bdf16c7c07bb96cbc7c8a53fadf2af63]

NODE_9_length_11993_cov_38.968399_4

Description	PssmId	Multi-dom	E-value
[+]DUF4407[pfam14362], Domain of unknown function (DUF4407); This family of proteins is found in bacteria. Proteins in this family are typically ...	258534	no	1.22e-25

Figure 4: Screenshot of CCD search results



Sequence search results

Show the detailed description of this results page.

We found **3** Pfam-A matches to your search sequence (**2** significant and **1** insignificant). You did not choose to search for Pfam-B matches.

Show the search options and sequence that you submitted.

Return to the search form to look for Pfam domains on a new sequence.

Significant Pfam-A Matches

Show or hide all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites
				Start	End	Start	End	From	To				
DUF4407	Domain of unknown function (DUF4407)	Family	n/a	22	332	24	322	4	228	301	99.3	1.9e-28	n/a
DUF4407	Domain of unknown function (DUF4407)	Family	n/a	277	458	285	456	127	299	301	40.6	1.4e-10	n/a

Insignificant Pfam-A Matches

Show or hide all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites
				Start	End	Start	End	From	To				
V-SNARE	Vesicle transport v-SNARE protein N-term ...	Family	CL0147	328	399	331	397	5	70	79	12.8	0.099	n/a

Figure 5: Screenshot of pfam results