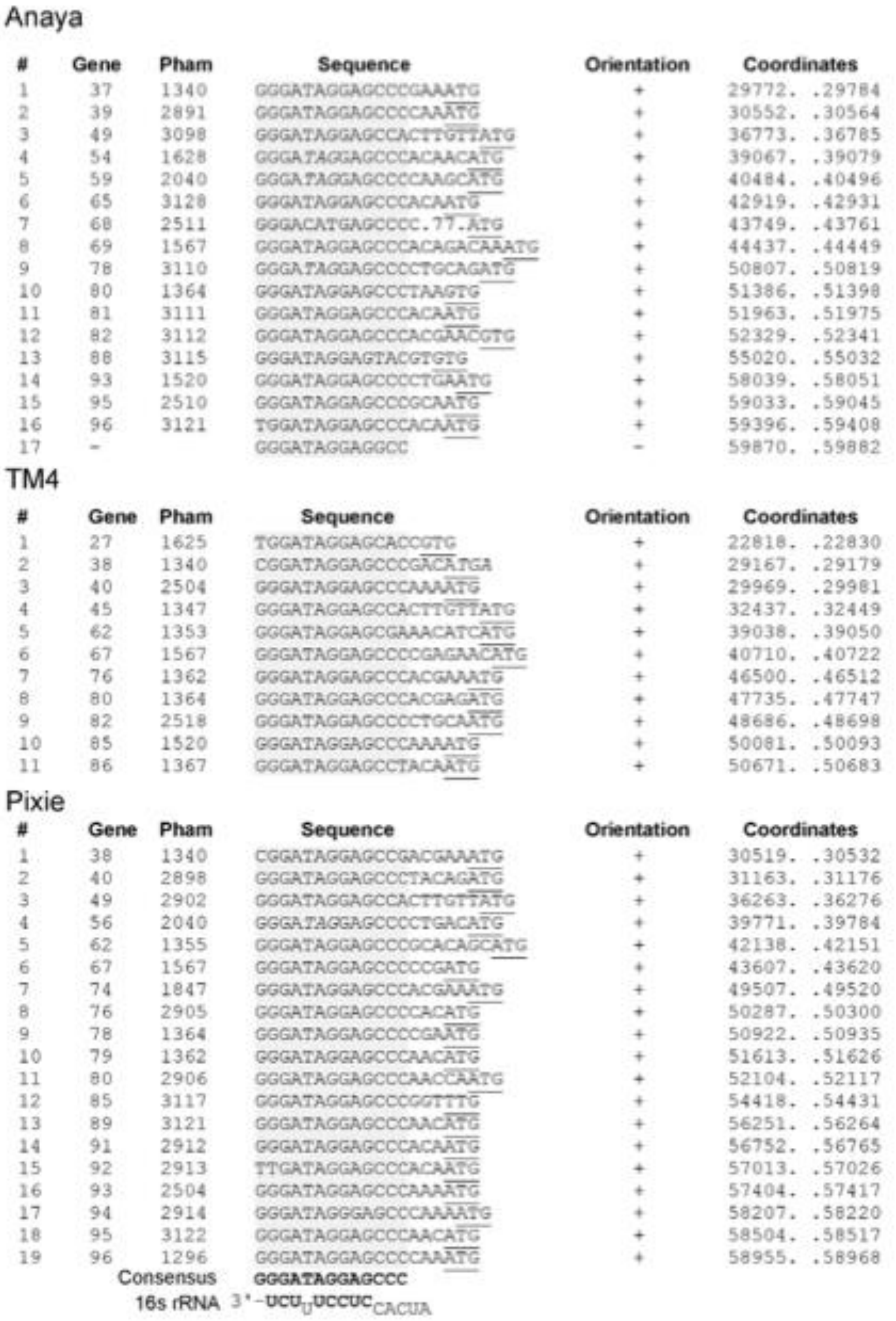
Conserved Sequence Repeats In Upstream Sequences Of Enterobacteriophage And Their Proposed Function.

Kaleigh Hedges

Introduction

In a recent article by Pope et. al, a conserved 13 bp sequence was found upstream of genes of mycobacteriophage cluster K, a taxonomic separation of mycobacteriophage dependent upon sequence similarity, genome length and structure. The frequency of the repeats and the containment of the Shine Dalgarno sequence suggested the sequence played a role in translation initiation. Since many untranslated regions contain important functional sequences such as promoters, terminators, and operons, Pope et. al are considering this sequence to be of importance for Cluster K mycobacteriophage. The intergenic regions of genomes are now being studied vigorously, resulting in the discovery that these spaces between genes actually carry functions that are important in gene expression.

Translating DNA requires ribosomes. Ribosomes need a sequence to recognize and attach to upstream of genes to begin translation, called a ribosome binding site, it promotes efficient and accurate translation of mRNA and is simply called the Shine-Dalgarno sequence.[1] The sequence is complementary to the 3' end of the rRNA and is a conserved sequence of “AGGA(G)” in most bacteria and their bacteriophages. A recent study of Cluster K mycobacteriophages found a 13 base pair repeat that is present 11-19 times (see Figure 1) in each Cluster K genome. The authors also noticed that the Shine Dalgarno sequence was inside of this repeat. Since the Shine Dalgarno sequence is important as a ribosomal binding site on mRNA, the authors suggest a role in translation initiation. The associated start codon with the repeat upstream of the gene seemed to have a high consensus of being ATG (86%), a rarer start codon for mycobacteriophage in general(55%).[2] Further speculation led to other repeats in Cluster M mycobacteriophage, also directly upstream from genes, but however without as many instances of the Shine Dalgarno sequence.[3] Conserved repeated sequences have also been spotted outside of mycobacteriophage, such as the haloarchaeal virus HF2[5] (also a bacteriophage, but family myoviridae meaning they have contractile tails; this is unlike mycobacteriophage, who are siphoviridae, or non-contractile tails), insisting that these conserved repeats seem to have a role in translation initiation, or in some way help the ribosome to recognize the site in which to bind on mRNA.



Anaya

#	Gene	Pham	Sequence	Orientation	Coordinates
1	37	1340	GGGATAGGAGCCCGAAATG	+	29772. .29784
2	39	2891	GGGATAGGAGCCCCAAATG	+	30552. .30564
3	49	3098	GGGATAGGAGCCACTTGTTATG	+	36773. .36785
4	54	1628	GGGATAGGAGCCCACAACATG	+	39067. .39079
5	59	2040	GGGATAGGAGCCCCAAGCATG	+	40484. .40496
6	65	3128	GGGATAGGAGCCCACAATG	+	42919. .42931
7	68	2511	GGGACATGAGCCCC.77.ATG	+	43749. .43761
8	69	1567	GGGATAGGAGCCCACAGACAAATG	+	44437. .44449
9	78	3110	GGGATAGGAGCCCCTGCAGATG	+	50807. .50819
10	80	1364	GGGATAGGAGCCCTAAGTG	+	51386. .51398
11	81	3111	GGGATAGGAGCCCACAATG	+	51963. .51975
12	82	3112	GGGATAGGAGCCCACGAACGTG	+	52329. .52341
13	88	3115	GGGATAGGAGTACGTGTG	+	55020. .55032
14	93	1520	GGGATAGGAGCCCCTGAATG	+	58039. .58051
15	95	2510	GGGATAGGAGCCCGCAATG	+	59033. .59045
16	96	3121	TGGATAGGAGCCCACAATG	+	59396. .59408
17	-		GGGATAGGAGGCC	-	59870. .59882

TM4

#	Gene	Pham	Sequence	Orientation	Coordinates
1	27	1625	TGGATAGGAGCACCGTG	+	22818. .22830
2	38	1340	CGGATAGGAGCCCGACATGA	+	29167. .29179
3	40	2504	GGGATAGGAGCCCAAAATG	+	29969. .29981
4	45	1347	GGGATAGGAGCCACTTGTTATG	+	32437. .32449
5	62	1353	GGGATAGGAGCGAAACATCATG	+	39038. .39050
6	67	1567	GGGATAGGAGCCCCGAGAACATG	+	40710. .40722
7	76	1362	GGGATAGGAGCCCACGAAATG	+	46500. .46512
8	80	1364	GGGATAGGAGCCCACGAGATG	+	47735. .47747
9	82	2518	GGGATAGGAGCCCCTGCAATG	+	48686. .48698
10	85	1520	GGGATAGGAGCCCAAAATG	+	50081. .50093
11	86	1367	GGGATAGGAGCCTACAATG	+	50671. .50683

Pixie

#	Gene	Pham	Sequence	Orientation	Coordinates
1	38	1340	CGGATAGGAGCCGACGAAATG	+	30519. .30532
2	40	2898	GGGATAGGAGCCCTACAGATG	+	31163. .31176
3	49	2902	GGGATAGGAGCCACTTGTTATG	+	36263. .36276
4	56	2040	GGGATAGGAGCCCCTGACATG	+	39771. .39784
5	62	1355	GGGATAGGAGCCCGCACAGCATG	+	42138. .42151
6	67	1567	GGGATAGGAGCCCCCGATG	+	43607. .43620
7	74	1847	GGGATAGGAGCCCACGAAATG	+	49507. .49520
8	76	2905	GGGATAGGAGCCCCACATG	+	50287. .50300
9	78	1364	GGGATAGGAGCCCCGAATG	+	50922. .50935
10	79	1362	GGGATAGGAGCCCAACATG	+	51613. .51626
11	80	2906	GGGATAGGAGCCCAACCAATG	+	52104. .52117
12	85	3117	GGGATAGGAGCCCGGTTTG	+	54418. .54431
13	89	3121	GGGATAGGAGCCCAACATG	+	56251. .56264
14	91	2912	GGGATAGGAGCCCACAATG	+	56752. .56765
15	92	2913	TTGATAGGAGCCCACAATG	+	57013. .57026
16	93	2504	GGGATAGGAGCCCAAAATG	+	57404. .57417
17	94	2914	GGGATAGGGAGCCCAAAATG	+	58207. .58220
18	95	3122	GGGATAGGAGCCCAACATG	+	58504. .58517
19	96	1296	GGGATAGGAGCCCCAAATG	+	58955. .58968

Consensus **GGGATAGGAGCCC**

16s rRNA 3'-**UCU**$_{U}$**UCCUC**$_{CACUA}$

Figure 1: Image from Pope et. al depicting the conserved repeats in upstream sequences in cluster K mycobacteriophage. The repeated sequence is highlighted in grey, containing the Shine Dalgarno sequence. The resulting start codon is underlined.

This project will search for conserved repeats in the sequences upstream of enterobacteriophage, specifically trying to locate those that include could include the Shine Dalgarno sequence, and analyze their possible function and purpose.

Methods

The web-based biological program BioBIKE[6] was used sort out all sequence repeats upstream of gene starts, or motifs in enterobacteriophage. This task was completed by creating a function within the program BioBIKE that uses a certain organism, or entity, and searches all upstream sequences of each gene for conserved repeats. The returned data, called a MEME, gives each conserved sequence a p-value, or how likely it is that the repeat occurs by chance, the actual repeated sequence and the overall consensus of the sequence, as well as the location of the repeat in terms of the area you are searching (for ours, simply upstream sequences).

BioBIKE's *motifs-in* function was used for data collected in this report. Inside the '*entity*' input box was a function within a function, *upstream sequences of [genes of]* while selecting the options from the *motifs-in* box for '*DNA'* and *'Return'*. This function was used for each enterobacteriophage currently existing in BioBIKE, 7 in total. Each of the 7 enterobacteriophage was ran through the function by placing it in the '*entity'* (grey) box and executing the function (Figure 2). The results were given in a separate window, or the MEME, and each MEME motif was considered.

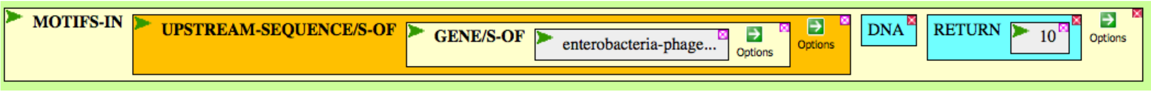


Figure 2: BioBIKE function used to obtain upstream sequenes from enterobacteriophage.

Results & Discussion

Throughout the conserved sequence search in upstream sequences of enterobacteriophage in the program BioBIKE, it became apparent that a conserved repeats containing the Shine Dalgarno sequence were not occurring in such high quantities in enterobacteriophage as they did in Cluster K mycobacteriophage. However, a different and striking conserved repeat was noted, having similarities in several entities of enterobacteriophage that were in BioBIKE (4 of the 7 phage – Min27, YYZ 2008, 2851, and SSL-2009a). The repeated sequences occur between 21-27 times, is approximately 13-35 bp in length, contains a Guanine rich area followed by a 4-7bp long Thymine repeat (See Supplemental Document, Figure 1). These repeats, being inside of intergenic regions within this genome, suggest their possible role as terminating sequences for the gene preceding these repeats. The T-tail present in terminating sequences is an easily identified factor in terminator sequences (Figure 3).

Terminating sequences consist of a general structure of a short stem-loop hairpin followed by a thymine rich region, or T-tail. The stem and loop structure generally consist of a GC rich region, which allows the hairpin to form by dyad symmetry (inverted repeats) that can base pair to each other and form a stem structure. As seen in E. coli, most hairpin stems vary from 5bp-17bp in length, and are show strong bias between CG pairs. The loop is non-base pairing, and thus needs no symmetry. The loop varies in size from 3-10bp[8]. The T-tail is transcribed to RNA to form the terminating hairpin and form the U-tail. The termination mechanism causes the RNA polymerase to stall over the T-tail, becoming more unstable as it pauses, and leads to a more likely disassociation or termination of translation[8]. Termination sequences such as these are part of rho-independent termination, which contain a hairpin structure on the elongating transcript which disrupts the mRNA-DNA-RNA polymerase complex.

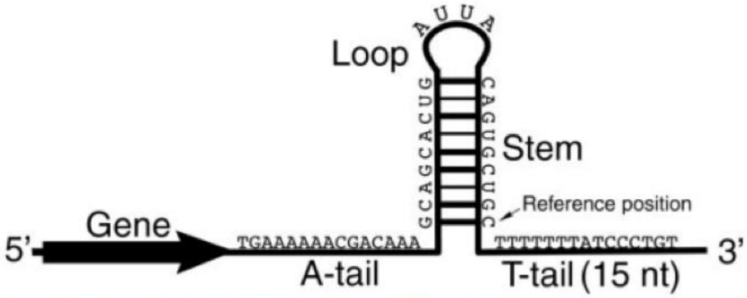


Figure 3: General schematic of hairpin loop with terminating sequences. The T-tail is noted.

The repeated sequences with T-tails found in the four enterobacteriophage show all the characteristics of a terminating sequence, including the GC rich area of the stem-loop structure, as well as the inverted repeats necessary for the sequence to form the stem structure by base pairing to itself. The location of these sequences also suggests they are termination sequences, being in intergenic regions of the genomes. Further data needs to be collected regarding the genes surrounding these repeats, and to determine if a correlation is found between the genes proceeding and preceeding the repeats.

Two of the seven enterobacteriophage (WV8 and phiEcoM-GJ1) interestingly did not contain the terminating sequences mentioned previously, but instead contain repeated sequences that contained Shine Dalgarno alternate sequences (Figure 4). Enterobacteriophage phiEcoM-GJ1 and WV8 showed alternate Shine Dalgarno sequences "GGAG" and "AGGAG" respectively (Supplemental, Figure 3) inside of repeats upstream of genes. The first repeat noted was in phiEcoM-GJ1, which occurred a shocking 33 times within its genome. Those repeats with the actual GGAG sequence (no deviations) all occur a few nucleotides before the start codons of the proceeding gene, giving evidence that these sequences could be associated with translation initiation as mentioned by Pope et. al in Cluster K mycobacteriophage. The repeats in WV8 all contain a perfect AGGA Shine Dalgarno sequence without deviations. All of these repeats also occur closely upstream of the genes, giving further evidence that these conserved sequences could assist in translation initiation. More evidence is needed to prove that this is indeed a factor that affects translation initiation, perhaps through BLAST using all known enterobacteriophage, and finding the resulting consensus.

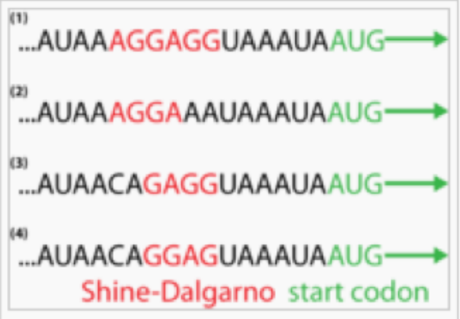


Figure 4: Shine Dalgarno sequence is listed at top, with alternate possible sequences for ribosomal binding sites.

It is being proven that intergenic regions of genomes are no longer being considered "junk DNA". With new discoveries using bioinformatics tools, we are discovering that this DNA plays roles in DNA replication. Whether the DNA is an operon, a terminator, or even the novel idea of a start associated sequence helping in translation initiation, it is clear that no part of DNA is merely "junk" but rather each nucleotide plays a key role in the way genes are expressed.

References

[1] Ribosome Binding Site Sequence Requirements. https://www.lifetechnologies.com/us/en/home/references/ambion-tech-support/translation-systems/general-articles/ribosomal-binding-site-sequence-requirements.html

[2] Cluster K mycobacteriophages: insights into the evolutionary origins of mycobacteriophage TM4. Pope et al. PLoS One. 2011;6(10):e26750. doi: 10.1371/journal.pone.0026750. Epub 2011 Oct 28.

[3] Cluster M mycobacteriophages Bongo, PegLeg, and Rey with unusually large repertoires of tRNA isotypes. Pope et al. J Virol. 2014 Mar;88(5):2461-80. doi: 10.1128/JVI.03363-13. Epub 2013 Dec 11.

[4] Transcriptional silencing by the mycobacteriophage L5 repressor. Brown KL1, Sarkis GJ, Wadsworth C, Hatfull GF. EMBO J. 1997 Oct 1;16(19):5914-21.

[5] HF2: a double-stranded DNA tailed haloarchaeal virus with a mosaic genome. Tang et. al. Molecular Microbiology (2002) 44(1), 283–296

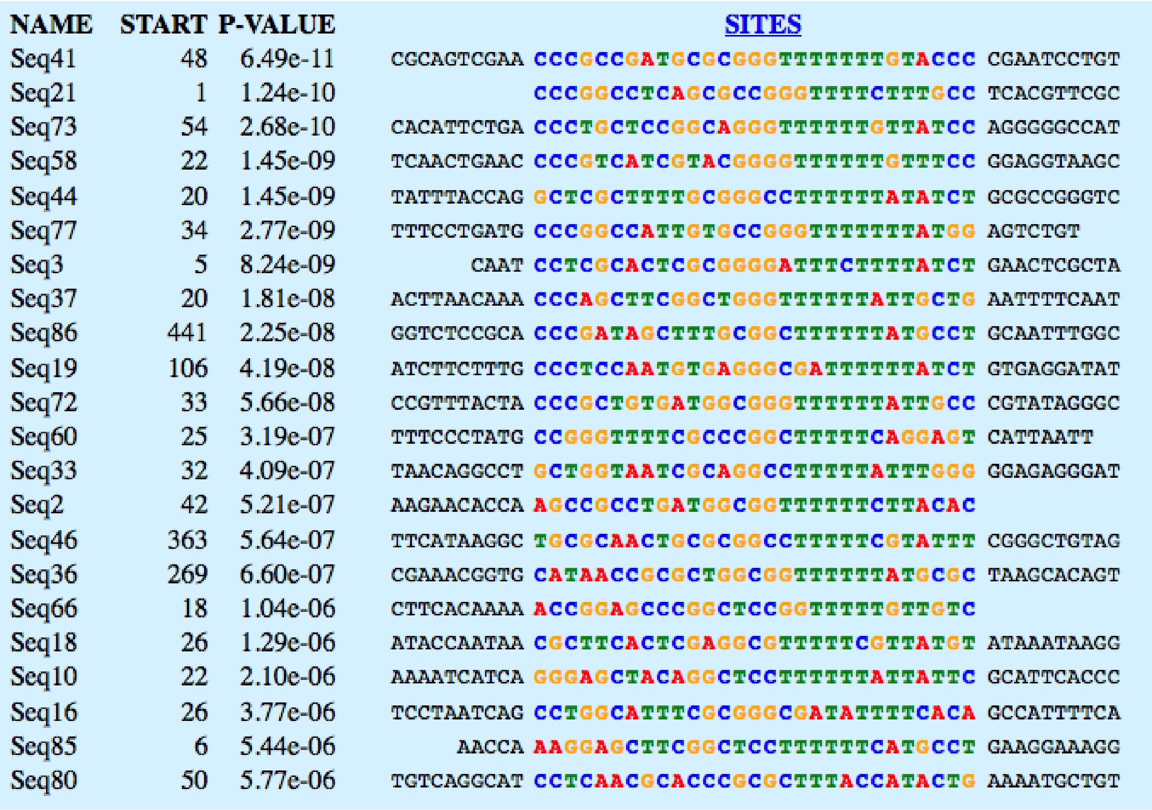
[6] BioBIKE. Web-based, programmable, integrated biological knowledge base. http://biobike.csbc.vcu.edu/

[7] Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. http://openi.nlm.nih.gov/detailedresult.php?img=1852404_gb-2007-8-2-r22-1&req=4

[8] Bacterial transcription terminators: the RNA 3'-end chronicles.Peters JM1, Vangeloff AD, Landick R. J Mol Biol. 2011 Oct 7;412(5):793-813. doi: 10.1016/j.jmb.2011.03.036. Epub 2011 Mar 23.

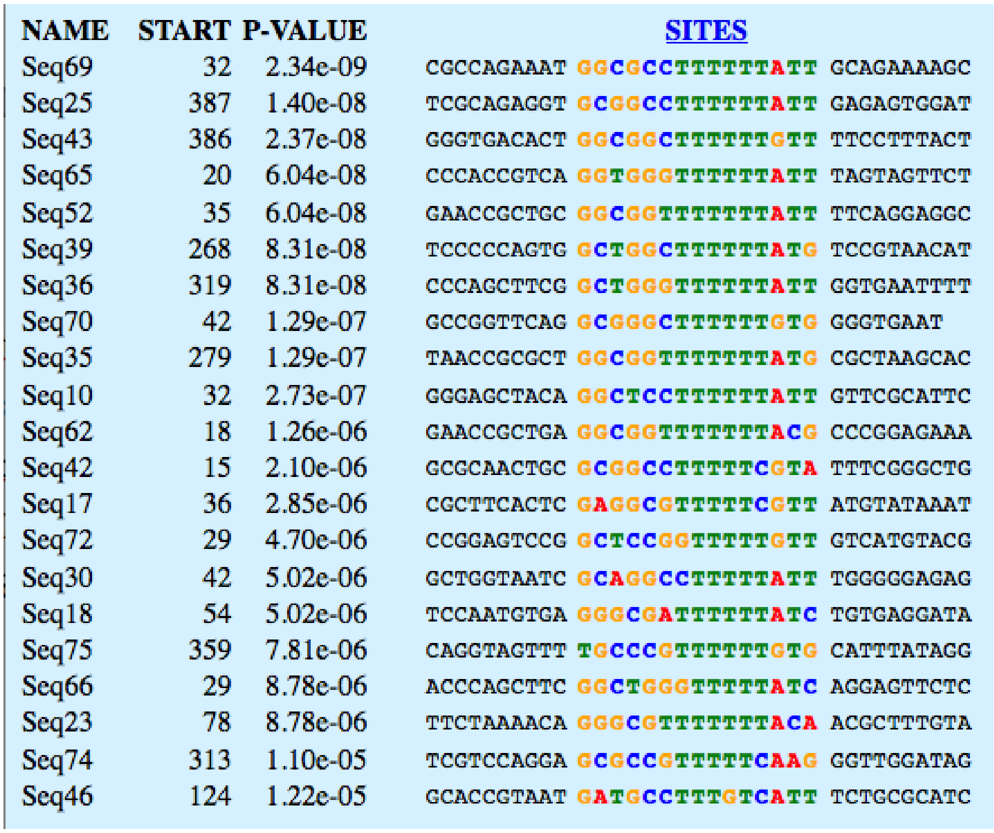
Figure 1: MEME outputs for Min27, YYZ 2008, 2851, and SSL-2009a. The outputs show the conserved T-tail, preceded by a guanine rich region.

Min27



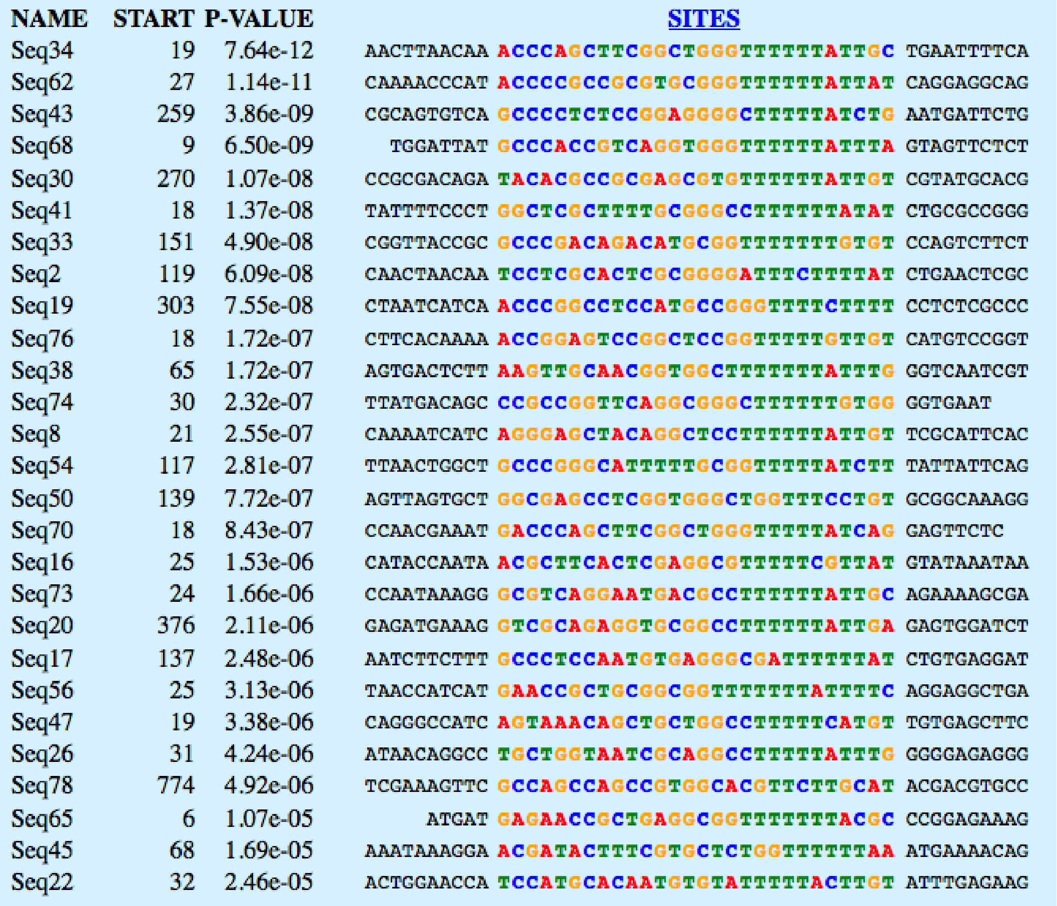
NAME	START	P-VALUE	SITES		
Seq41	48	6.49e-11	CGCAGTCGAA	CCCGCCGATGCGCGGGTTTTTTTGTACCC	CGAATCCTGT
Seq21	1	1.24e-10		CCCGGCCTCAGCGCCGGGTTTTCTTTGCC	TCACGTTCGC
Seq73	54	2.68e-10	CACATTCTGA	CCCTGCTCCGGCAGGGTTTTTTGTTATCC	AGGGGGCCAT
Seq58	22	1.45e-09	TCAACTGAAC	CCCGTCATCGTACGGGGTTTTTTGTTTCC	GGAGGTAAGC
Seq44	20	1.45e-09	TATTTACCAG	GCTCGCTTTTGCGGGCCTTTTTTATATCT	GCGCCGGGTC
Seq77	34	2.77e-09	TTTCCTGATG	CCCGGCCATTGTGCCGGGTTTTTTTATGG	AGTCTGT
Seq3	5	8.24e-09	CAAT	CCTCGCACTCGCGGGGATTTCTTTTATCT	GAACTCGCTA
Seq37	20	1.81e-08	ACTTAACAAA	CCCAGCTTCGGCTGGGTTTTTTATTGCTG	AATTTTCAAT
Seq86	441	2.25e-08	GGTCTCCGCA	CCCGATAGCTTTGCGGCTTTTTTATGCCT	GCAATTTGGC
Seq19	106	4.19e-08	ATCTTCTTTG	CCCTCCAATGTGAGGGCGATTTTTTATCT	GTGAGGATAT
Seq72	33	5.66e-08	CCGTTTACTA	CCCGCTGTGATGGCGGGTTTTTTATTGCC	CGTATAGGGC
Seq60	25	3.19e-07	TTTCCCTATG	CCGGGTTTTCGCCCGGCTTTTTCAGGAGT	CATTAATT
Seq33	32	4.09e-07	TAACAGGCCT	GCTGGTAATCGCAGGCCTTTTTATTTGGG	GGAGAGGGAT
Seq2	42	5.21e-07	AAGAACACCA	AGCCGCCTGATGGCGGTTTTTTCTTACAC	
Seq46	363	5.64e-07	TTCATAAGGC	TGCGCAACTGCGCGGCCTTTTTCGTATTT	CGGGCTGTAG
Seq36	269	6.60e-07	CGAAACGGTG	CATAACCGCGCTGGCGGTTTTTTATGCGC	TAAGCACAGT
Seq66	18	1.04e-06	CTTCACAAAA	ACCGGAGCCCGGCTCCGGTTTTTGTTGTC	
Seq18	26	1.29e-06	ATACCAATAA	CGCTTCACTCGAGGCGTTTTTCGTTATGT	ATAAATAAGG
Seq10	22	2.10e-06	AAAATCATCA	GGGAGCTACAGGCTCCTTTTTTATTATTC	GCATTCACCC
Seq16	26	3.77e-06	TCCTAATCAG	CCTGGCATTTCGCGGGCGATATTTTCACA	GCCATTTTCA
Seq85	6	5.44e-06	AACCA	AAGGAGCTTCGGCTCCTTTTTTCATGCCT	GAAGGAAAGG
Seq80	50	5.77e-06	TGTCAGGCAT	CCTCAACGCACCCGCGCTTTACCATACTG	AAAATGCTGT

YYZ 2008



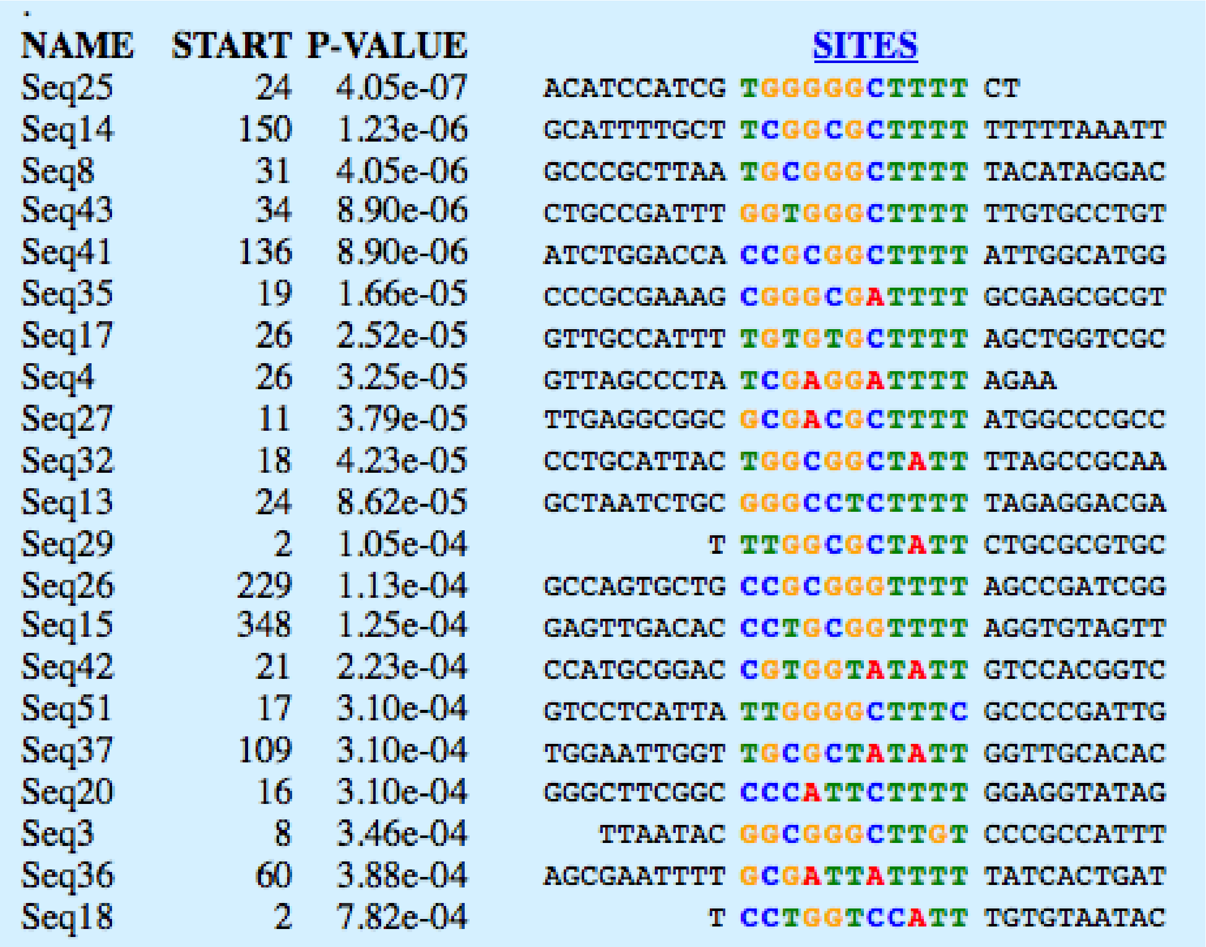
NAME	START	P-VALUE	SITES		
Seq69	32	2.34e-09	CGCCAGAAAT	GGCGCCTTTTTTATT	GCAGAAAAGC
Seq25	387	1.40e-08	TCGCAGAGGT	GCGGCCTTTTTTATT	GAGAGTGGAT
Seq43	386	2.37e-08	GGGTGACACT	GGCGGCTTTTTTGTT	TTCCTTTACT
Seq65	20	6.04e-08	CCCACCGTCA	GGTGGGTTTTTTATT	TAGTAGTTCT
Seq52	35	6.04e-08	GAACCGCTGC	GGCGGTTTTTTTATT	TTCAGGAGGC
Seq39	268	8.31e-08	TCCCCCAGTG	GCTGGCTTTTTTATG	TCCGTAACAT
Seq36	319	8.31e-08	CCCAGCTTCG	GCTGGGTTTTTTATT	GGTGAATTTT
Seq70	42	1.29e-07	GCCGGTTCAG	GCGGGCTTTTTTGTG	GGGTGAAT
Seq35	279	1.29e-07	TAACCGCGCT	GGCGGTTTTTTTATG	CGCTAAGCAC
Seq10	32	2.73e-07	GGGAGCTACA	GGCTCCTTTTTTATT	GTTCGCATTC
Seq62	18	1.26e-06	GAACCGCTGA	GGCGGTTTTTTTACG	CCCGGAGAAA
Seq42	15	2.10e-06	GCGCAACTGC	GCGGCCTTTTTCGTA	TTTCGGGCTG
Seq17	36	2.85e-06	CGCTTCACTC	GAGGCGTTTTTCGTT	ATGTATAAAT
Seq72	29	4.70e-06	CCGGAGTCCG	GCTCCGGTTTTTGTT	GTCATGTACG
Seq30	42	5.02e-06	GCTGGTAATC	GCAGGCCTTTTTATT	TGGGGGAGAG
Seq18	54	5.02e-06	TCCAATGTGA	GGGCGATTTTTTATC	TGTGAGGATA
Seq75	359	7.81e-06	CAGGTAGTTT	TGCCCGTTTTTTGTG	CATTTATAGG
Seq66	29	8.78e-06	ACCCAGCTTC	GGCTGGGTTTTTATC	AGGAGTTCTC
Seq23	78	8.78e-06	TTCTAAAACA	GGGCGTTTTTTTACA	ACGCTTTGTA
Seq74	313	1.10e-05	TCGTCCAGGA	GCGCCGTTTTTCAAG	GGTTGGATAG
Seq46	124	1.22e-05	GCACCGTAAT	GATGCCTTTGTCATT	TCTGCGCATC

2851



NAME	START	P-VALUE	SITES
Seq34	19	7.64e-12	AACTTAACAA ACCCAGCTTCGGCTGGGTTTTTATTGC TGAATTTTCA
Seq62	27	1.14e-11	CAAAACCCAT ACCCCGCCGCGTGCGGGTTTTTATTAT CAGGAGGCAG
Seq43	259	3.86e-09	CGCAGTGTCA GCCCCTCTCCGGAGGGGCTTTTTATCTG AATGATTCTG
Seq68	9	6.50e-09	TGGATTAT GCCCACCGTCAGGTGGGTTTTTATTTA GTAGTTCTCT
Seq30	270	1.07e-08	CCGCGACAGA TACACGCCGCGAGCGTGTTTTTATTGT CGTATGCACG
Seq41	18	1.37e-08	TATTTTCCCT GGCTCGCTTTTGCGGGCCTTTTTATAT CTGCGCCGGG
Seq33	151	4.90e-08	CGGTTACCGC GCCCGACAGACATGCGGTTTTTTGTGT CCAGTCTTCT
Seq2	119	6.09e-08	CAACTAACAA TCCTCGCACTCGCGGGGATTTCTTTTAT CTGAACTCGC
Seq19	303	7.55e-08	CTAATCATCA ACCCGGCCTCCATGCCGGGTTTTCTTTT CCTCTCGCCC
Seq76	18	1.72e-07	CTTCACAAAA ACCGGAGTCCGGCTCCGGTTTTTGTTGT CATGTCCGGT
Seq38	65	1.72e-07	AGTGACTCTT AAGTTGCAACGGTGGCTTTTTTTATTTG GGTCAATCGT
Seq74	30	2.32e-07	TTATGACAGC CCGCCGGTTCAGGCGGGCTTTTTTGTGG GGTGAAT
Seq8	21	2.55e-07	CAAAATCATC AGGGAGCTACAGGCTCCTTTTTTATTGT TCGCATTCAC
Seq54	117	2.81e-07	TTAACTGGCT GCCCGGGCATTTTTGCGGTTTTTATCTT TATTATTCAG
Seq50	139	7.72e-07	AGTTAGTGCT GGCGAGCCTCGGTGGGCTGGTTTCCTGT GCGGCAAAGG
Seq70	18	8.43e-07	CCAACGAAAT GACCCAGCTTCGGCTGGGTTTTTATCAG GAGTTCTC
Seq16	25	1.53e-06	CATACCAATA ACGCTTCACTCGAGGCGTTTTTCGTTAT GTATAAATAA
Seq73	24	1.66e-06	CCAATAAAGG GCGTCAGGAATGACGCCTTTTTTATTGC AGAAAAGCGA
Seq20	376	2.11e-06	GAGATGAAAG GTCGCAGAGGTGCGGCCTTTTTTATTGA GAGTGGATCT
Seq17	137	2.48e-06	AATCTTCTTT GCCCTCCAATGTGAGGGCGATTTTTTAT CTGTGAGGAT
Seq56	25	3.13e-06	TAACCATCAT GAACCGCTGCGGCGGTTTTTTTATTTTC AGGAGGCTGA
Seq47	19	3.38e-06	CAGGGCCATC AGTAAACAGCTGCTGGCCTTTTTCATGT TGTGAGCTTC
Seq26	31	4.24e-06	ATAACAGGCC TGCTGGTAATCGCAGGCCTTTTTATTTG GGGGAGAGGG
Seq78	774	4.92e-06	TCGAAAGTTC GCCAGCCAGCCGTGGCACGTTCTTGCAT ACGACGTGCC
Seq65	6	1.07e-05	ATGAT GAGAACCGCTGAGGCGGTTTTTTTACGC CCGGAGAAAG
Seq45	68	1.69e-05	AAATAAAGGA ACGATACTTTCGTGCTCTGGTTTTTTAA ATGAAAACAG
Seq22	32	2.46e-05	ACTGGAACCA TCCATGCACAATGTGTATTTTTACTTGT ATTTGAGAAG

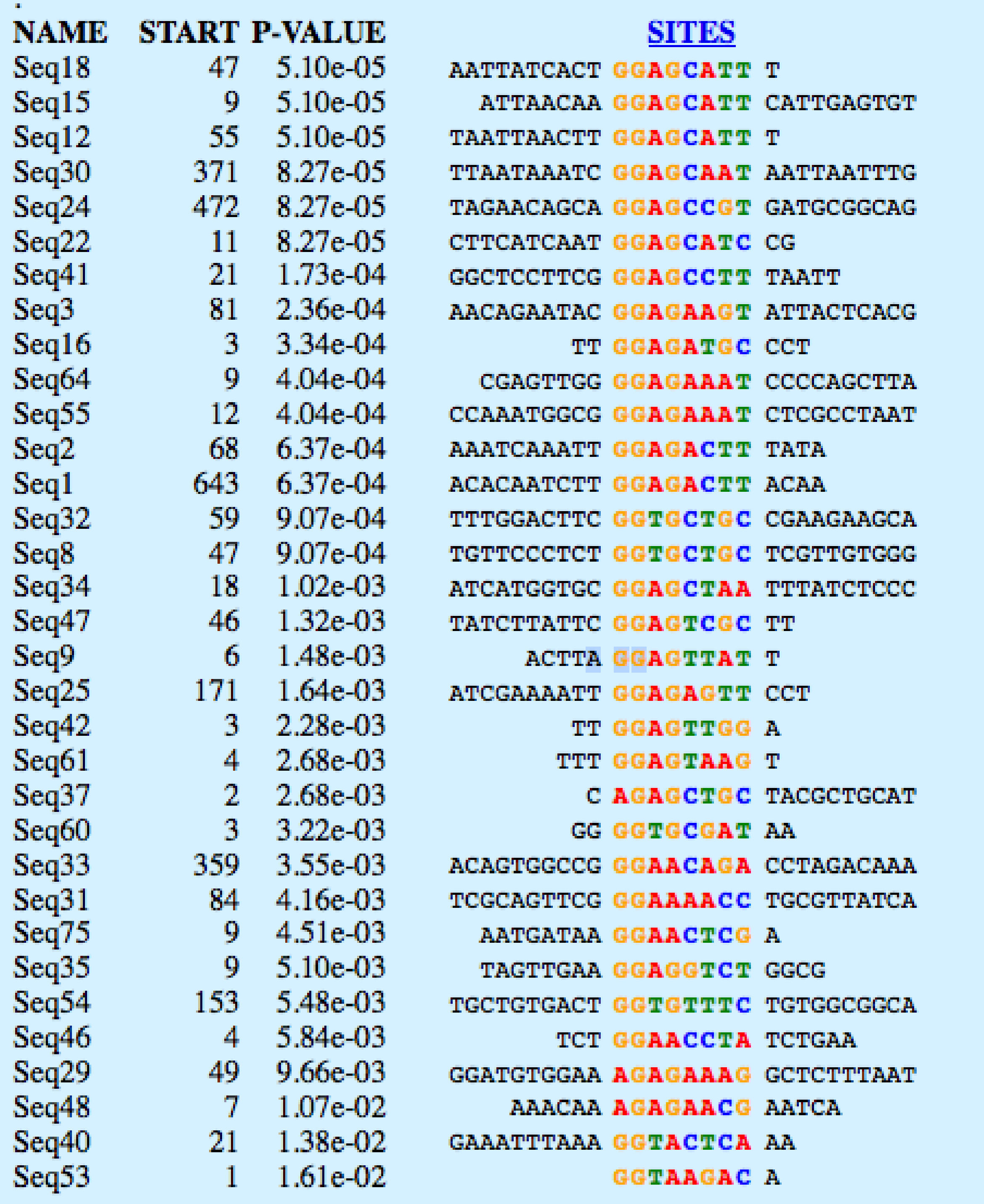
SSL-2009a



NAME	START	P-VALUE	SITES
Seq25	24	4.05e-07	ACATCCATCG TGGGGGCTTTT CT
Seq14	150	1.23e-06	GCATTTTGCT TCGGCGCTTTT TTTTTAAATT
Seq8	31	4.05e-06	GCCCGCTTAA TGCGGGCTTTT TACATAGGAC
Seq43	34	8.90e-06	CTGCCGATTT GGTGGGCTTTT TTGTGCCTGT
Seq41	136	8.90e-06	ATCTGGACCA CCGCGGCTTTT ATTGGCATGG
Seq35	19	1.66e-05	CCCGCGAAAG CGGGCGATTTT GCGAGCGCGT
Seq17	26	2.52e-05	GTTGCCATTT TGTGTGCTTTT AGCTGGTCGC
Seq4	26	3.25e-05	GTTAGCCCTA TCGAGGATTTT AGAA
Seq27	11	3.79e-05	TTGAGGCGGC GCGACGCTTTT ATGGCCCGCC
Seq32	18	4.23e-05	CCTGCATTAC TGGCGGCTATT TTAGCCGCAA
Seq13	24	8.62e-05	GCTAATCTGC GGGCCTCTTTT TAGAGGACGA
Seq29	2	1.05e-04	T TTGGCGCTATT CTGCGCGTGC
Seq26	229	1.13e-04	GCCAGTGCTG CCGCGGGTTTT AGCCGATCGG
Seq15	348	1.25e-04	GAGTTGACAC CCTGCGGTTTT AGGTGTAGTT
Seq42	21	2.23e-04	CCATGCGGAC CGTGGTATATT GTCCACGGTC
Seq51	17	3.10e-04	GTCCTCATTA TTGGGGCTTTC GCCCCGATTG
Seq37	109	3.10e-04	TGGAATTGGT TGCGCTATATT GGTTGCACAC
Seq20	16	3.10e-04	GGGCTTCGGC CCCATTCTTTT GGAGGTATAG
Seq3	8	3.46e-04	TTAATAC GGCGGGCTTGT CCCGCCATTT
Seq36	60	3.88e-04	AGCGAATTTT GCGATTATTTT TATCACTGAT
Seq18	2	7.82e-04	T CCTGGTCCATT TGTGTAATAC

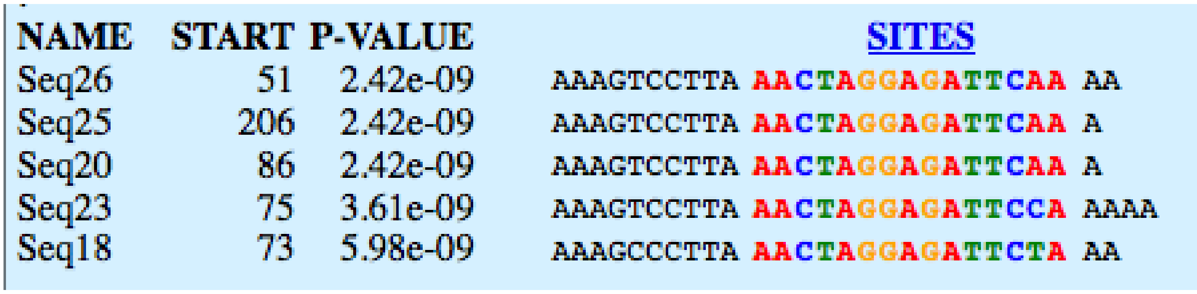
Figure 2: MEME outputs for enterobacteriophage phiEcoM-GJ1 and WV8, showing alternate Shine Dalgarno sequences "GGAG" and "AGGAG" respectively.

phiEcoM-GJ1



NAME	START	P-VALUE	SITES
Seq18	47	5.10e-05	AATTATCACT GGAGCATT T
Seq15	9	5.10e-05	ATTAACAA GGAGCATT CATTGAGTGT
Seq12	55	5.10e-05	TAATTAACTT GGAGCATT T
Seq30	371	8.27e-05	TTAATAAATC GGAGCAAT AATTAATTTG
Seq24	472	8.27e-05	TAGAACAGCA GGAGCCGT GATGCGGCAG
Seq22	11	8.27e-05	CTTCATCAAT GGAGCATC CG
Seq41	21	1.73e-04	GGCTCCTTCG GGAGCCTT TAATT
Seq3	81	2.36e-04	AACAGAATAC GGAGAAGT ATTACTCACG
Seq16	3	3.34e-04	TT GGAGATGC CCT
Seq64	9	4.04e-04	CGAGTTGG GGAGAAAT CCCCAGCTTA
Seq55	12	4.04e-04	CCAAATGGCG GGAGAAAT CTCGCCTAAT
Seq2	68	6.37e-04	AAATCAAATT GGAGACTT TATA
Seq1	643	6.37e-04	ACACAATCTT GGAGACTT ACAA
Seq32	59	9.07e-04	TTTGGACTTC GGTGCTGC CGAAGAAGCA
Seq8	47	9.07e-04	TGTTCCCTCT GGTGCTGC TCGTTGTGGG
Seq34	18	1.02e-03	ATCATGGTGC GGAGCTAA TTTATCTCCC
Seq47	46	1.32e-03	TATCTTATTC GGAGTCGC TT
Seq9	6	1.48e-03	ACTTA GGAGTTAT T
Seq25	171	1.64e-03	ATCGAAAATT GGAGAGTT CCT
Seq42	3	2.28e-03	TT GGAGTTGG A
Seq61	4	2.68e-03	TTT GGAGTAAG T
Seq37	2	2.68e-03	C AGAGCTGC TACGCTGCAT
Seq60	3	3.22e-03	GG GGTGCGAT AA
Seq33	359	3.55e-03	ACAGTGGCCG GGAACAGA CCTAGACAAA
Seq31	84	4.16e-03	TCGCAGTTCG GGAAAACC TGCGTTATCA
Seq75	9	4.51e-03	AATGATAA GGAACTCG A
Seq35	9	5.10e-03	TAGTTGAA GGAGGTCT GGCG
Seq54	153	5.48e-03	TGCTGTGACT GGTGTTTC TGTGGCGGCA
Seq46	4	5.84e-03	TCT GGAACCTA TCTGAA
Seq29	49	9.66e-03	GGATGTGGAA AGAGAAAG GCTCTTTAAT
Seq48	7	1.07e-02	AAACAA AGAGAACG AATCA
Seq40	21	1.38e-02	GAAATTTAAA GGTACTCA AA
Seq53	1	1.61e-02	GGTAAGAC A

WV8



NAME	START	P-VALUE	SITES
Seq26	51	2.42e-09	AAAGTCCTTA AACTAGGAGATTCAA AA
Seq25	206	2.42e-09	AAAGTCCTTA AACTAGGAGATTCAA A
Seq20	86	2.42e-09	AAAGTCCTTA AACTAGGAGATTCAA A
Seq23	75	3.61e-09	AAAGTCCTTA AACTAGGAGATTCCA AAAA
Seq18	73	5.98e-09	AAAGCCCTTA AACTAGGAGATTCTA AA