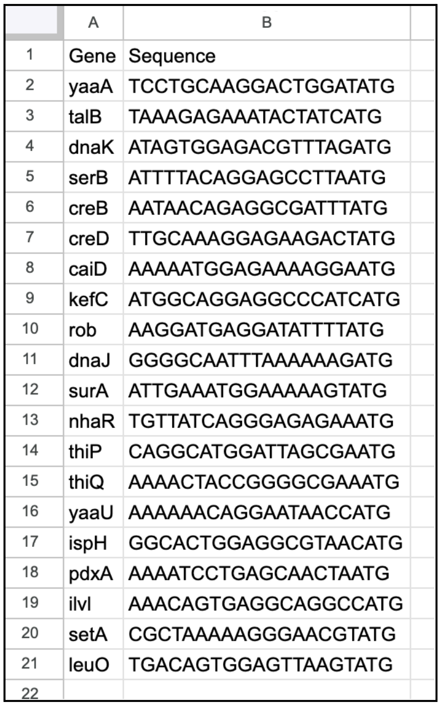
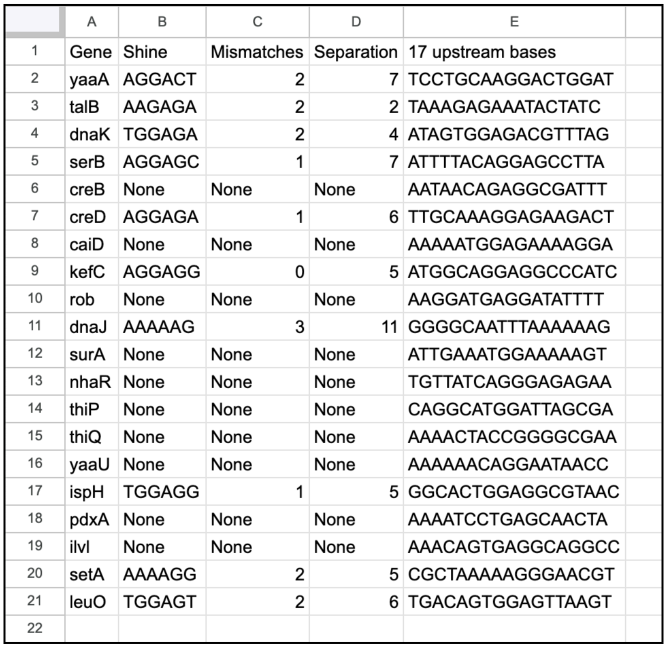
Shine-Dalgarno Sequence



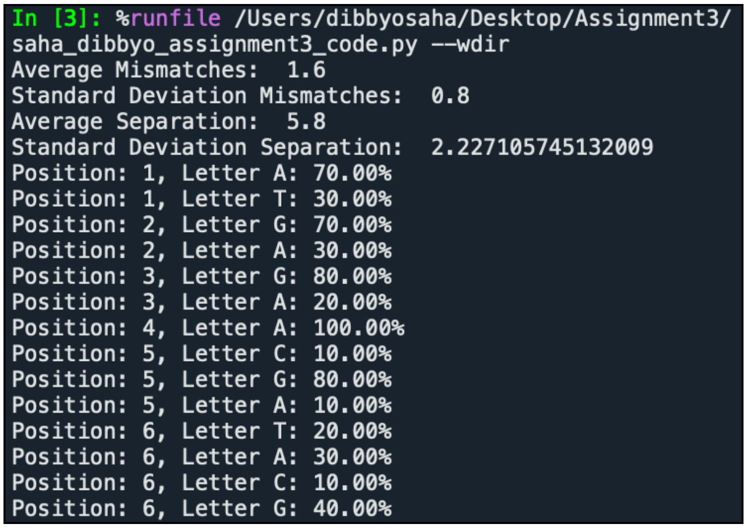
	A	B
1	Gene	Sequence
2	yaaA	TCCTGCAAGGACTGGATATG
3	talB	TAAAGAGAAATACTATCATG
4	dnaK	ATAGTGGAGACGTTTAGATG
5	serB	ATTTTACAGGAGCCTTAATG
6	creB	AATAACAGAGGCGATTTATG
7	creD	TTGCAAAGGAGAAGACTATG
8	caiD	AAAAATGGAGAAAAGGAATG
9	kefC	ATGGCAGGAGGCCCATCATG
10	rob	AAGGATGAGGATATTTTATG
11	dnaJ	GGGGCAATTTAAAAAAGATG
12	surA	ATTGAAATGGAAAAAGTATG
13	nhaR	TGTTATCAGGGAGAGAAATG
14	thiP	CAGGCATGGATTAGCGAATG
15	thiQ	AAAACTACCGGGGCGAAATG
16	yaaU	AAAAAACAGGAATAACCATG
17	ispH	GGCACTGGAGGCGTAACATG
18	pdxA	AAAATCCTGAGCAACTAATG
19	ilvI	AAACAGTGAGGCAGGCCATG
20	setA	CGCTAAAAAGGGAACGTATG
21	leuO	TGACAGTGGAGTTAAGTATG
22		

Figure 1: Table showing sequences for 20 different genes used as input from file *gene_sequence.csv*.



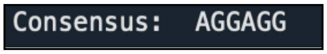
	A	B	C	D	E
1	Gene	Shine	Mismatches	Separation	17 upstream bases
2	yaaA	AGGACT	2	7	TCCTGCAAGGACTGGAT
3	talB	AAGAGA	2	2	TAAAGAGAAATACTATC
4	dnaK	TGGAGA	2	4	ATAGTGGAGACGTTTAG
5	serB	AGGAGC	1	7	ATTTTACAGGAGCCTTA
6	creB	None	None	None	AATAACAGAGGCGATTT
7	creD	AGGAGA	1	6	TTGCAAAGGAGAAGACT
8	caiD	None	None	None	AAAAATGGAGAAAAGGA
9	kefC	AGGAGG	0	5	ATGGCAGGAGGCCCATC
10	rob	None	None	None	AAGGATGAGGATATTTT
11	dnaJ	AAAAAG	3	11	GGGGCAATTTAAAAAAG
12	surA	None	None	None	ATTGAAATGGAAAAAGT
13	nhaR	None	None	None	TGTTATCAGGGAGAGAA
14	thiP	None	None	None	CAGGCATGGATTAGCGA
15	thiQ	None	None	None	AAAACTACCGGGGCGAA
16	yaaU	None	None	None	AAAAAACAGGAATAACC
17	ispH	TGGAGG	1	5	GGCACTGGAGGCGTAAC
18	pdxA	None	None	None	AAAATCCTGAGCAACTA
19	ilvI	None	None	None	AAACAGTGAGGCAGGCC
20	setA	AAAAGG	2	5	CGCTAAAAAGGGAACGT
21	leuO	TGGAGT	2	6	TGACAGTGGAGTTAAGT
22					

Figure 2: Table showing spreadsheet named *saha_dibbyo_assignment3_spreadsheet.csv* produced from the code.



```
In [3]: %runfile /Users/dibbyosaha/Desktop/Assignment3/
saha_dibbyo_assignment3_code.py --wdir
Average Mismatches:  1.6
Standard Deviation Mismatches:  0.8
Average Separation:  5.8
Standard Deviation Separation:  2.227105745132009
Position: 1, Letter A: 70.00%
Position: 1, Letter T: 30.00%
Position: 2, Letter G: 70.00%
Position: 2, Letter A: 30.00%
Position: 3, Letter G: 80.00%
Position: 3, Letter A: 20.00%
Position: 4, Letter A: 100.00%
Position: 5, Letter C: 10.00%
Position: 5, Letter G: 80.00%
Position: 5, Letter A: 10.00%
Position: 6, Letter T: 20.00%
Position: 6, Letter A: 30.00%
Position: 6, Letter C: 10.00%
Position: 6, Letter G: 40.00%
```

Figure 3: Image showing print output in the terminal for statistics calculated in *saha_dibbyo_assignment3_code.py*.



```
Consensus:  AGGAGG
```

Figure 4: Image showing print output in the terminal for *consensus sequence* determined in *saha_dibbyo_assignment3_code.py*.

Statistics

As can be observed from *Figure 3*, the average mismatches is 1.6, the standard deviation of mismatches is 0.8, the average separation is 5.8, and the standard deviation is 2.227105745132009. Respective positions can also be noticed with their respective letters and their percentages. Position 1 has 70% '*A*' and 30% '*T*', position 2 has 70% '*G*' and 30% '*A*', position 3 has 80% '*G*' and 20% '*A*', position 4 has 100% '*A*', position 5 has 80% '*G*', 10% '*C*' and 10% '*A*', and lastly position 6 has 40% 'G', 30% '*A*', 20% '*T*', and 10% '*C*'. Using the majority at each position, the determined consensus would be '*AGGAGG*' (as can be observed in *Figure 4*).

Comments

20 different genes and their sequences, each with 17 bases upstream from '*atg*' of *Escherichia coli K12 MG1655* were randomly collected from EcoCyc database (namely: *yaaA*, *talB*, *dnaK*, *serB*, *creB*, *creD*, *caiD*, *kefC*, *rob*, *dnaJ*, *surA*, *nhaR*, *thiP*, *thiQ*, *yaaU*, *ispH*, *pdxA*, *ilvl*, *setA*, *leuO*). The genes with their corresponding sequences were saved in the *gene_sequence.csv* file (as can be observed in *Figure 1*) in the same directory as *saha_dibbyo_assignment3_code.py*.

The *Python* code in *saha_dibbyo_assignment3_code.py* takes in the sequences from *gene_sequence.csv* (*Figure 1*) as inputs and analyzes to calculate mismatches between the sequences and the *Shine-Dalgarno* sequence of '*AGGAGG*'. The separation which is the distance between the start of the gene sequence and the *Shine-Dalgarno* sequence is also calculated. The data is saved in *saha_dibbyo_assignment3_spreadsheet.csv* (as can be observed in *Figure 2*) in the same directory as *saha_dibbyo_assignment3_code.py*. Required statistical values like average and standard deviation of the mismatches and separations are calculated using the *Python* code in *saha_dibbyo_assignment3_code.py* as discussed in the Statistics section above. The majority at each position suggests that consensus would be '*AGGAGG*'.