Used ORF Finder to identify the locations of three coding regions (three longest ORFs) in the Bacillus subtilis genomic sequence

b. Found on what reading frames are each of the genes in the Bacillus DNA based on ORF Finder.

1b. **ORF1 is reading frame 1.**
ORF2 is reading frame 1.
ORF7 is reading frame 3.

2. Used the command line version of Glimmer to analyze CDSs in a partial sequence from Spiroplasma helicoides strain TABS-2 (file: sheliprt.fasta). The training set used was the full genome of S. helicoides strain TABS-2 (file: sheli.fasta).

Sheli.fasta was used to train.

Predicted open reading frame from .predict file:

2a1.

>Spiroplasma helicoides strain TABS-2, partial sequence

orf00001	635	991	+2	4.13
orf00002	998	1141	+2	4.42
orf00003	1154	1312	+2	2.30
orf00004	1334	1978	+2	5.68
orf00006	2242	2463	+1	6.25
orf00008	2585	4003	+2	8.80
orf00009	4010	4678	+2	8.48
orf00010	4880	5143	+2	6.98

sheliprt.predict (END)

2b. Commands used:

```
long-orfs -n -t 1.15 sheli.fasta sheli.longorfs
extract -t sheli.fasta sheli.longorfs > sheli.train
build-icm -r sheli.icm < sheli.train
glimmer3 -o50 -g110 -t30 sheliprt.fasta sheli.icm sheliprt
extract -t sheliprt.fasta sheliprt.predict > sheliprt.glimmer
```

3. Used FGENESB to identify CDSs in the partial sequence from S. helicoides strain TABS-2 (file: sheliprt.fasta). Used 'bacterial generic' as the training set. (1 point)

3.

Label	Strand	Frame	Start	Stop	Length (nt | aa)
ORF1	+	1	46	1542	1497 | 498
ORF2	+	1	1564	1965	402 | 133
ORF7	+	3	1962	2303	342 | 113

```
Seq name: Spiroplasma helicoides strain TABS-2, partial sequence
Length of sequence - 5500 bp
Number of predicted genes - 9
Number of transcription units - 6, operons - 2
    N      Tu/Op   Conserved  S             Start          End     Score
                   pairs(N/Pv)
    1     1 Op  1      .       +    CDS        635 -        991     117
    2     1 Op  2      .       +    CDS        998 -       1141     144
    3     2 Tu  1      .       -    CDS       1126 -       1365      73
    4     3 Tu  1      .       +    CDS       1334 -       1978     381
    5     4 Tu  1      .       +    CDS       2242 -       2463     231
    6     5 Op  1      .       +    CDS       2585 -       4003     998
    7     5 Op  2      .       +    CDS       4010 -       4678     423
    8     5 Op  3      .       +    CDS       4703 -       4768      72
    9     6 Tu  1      .       +    CDS       4880 -       5143     169
```

a. There are **9** CDS listed.
b. **6** mRNAs

4. Used lactococcus DNA sequence to identify genetic features

a. Ran FGENESB to find the location of two genes on an operon, then ran BPROM to find the locations of the -35 signal and the -10 signal. CDS locations and the locations of the most appropriate -35 signal and -10 signal below.

4

a.

```
Number of predicted genes - 2
Number of transcription units - 1, operons - 1
    N      Tu/Op   Conserved  S             Start          End     Score
                   pairs(N/Pv)
    1     1 Op  1      .       +    CDS        287 -        553     266
    2     1 Op  2      .       +    CDS        556 -       2283    1320
```

```
Number of predicted promoters -      7
Promoter Pos:     225 LDF-  8.79
-10 box at pos.     210 TGGTACAAT Score     78
-35 box at pos.     190 TTGCAA    Score     55
```

b. Ran the prokaryotic promoter prediction at the [Berkeley Drosophila Neural Network Prediction] site to find most likely promoter and TSS.

b. **The most likely promoter at BDGP to match the BPROM result is the promoter starting at 184-229, with a guanine (G) as the nucleotide at the transcription start site**.

Promoter predictions for Lactococcus :

Start	End	Score	Promoter Sequence
11	56	0.92	ACGAAGCTGAAACCGAAAATAACTAAAAATAAAAGCTGTCAGAACTGATA
61	106	0.99	GCTTTTTTTCAGCTCACTTTCTTCAGGAAAATAATATAAAAAATACTTAT
106	151	0.99	CTTATTTGATGATAAAAGAAATCAAAGTCTAGCATCCATTCAAAAGCAGC
184	229	0.97	CAGATATTGCAAACCCTTTCGTTTTGTGGTACAATTTCAAGAGTCATAGA