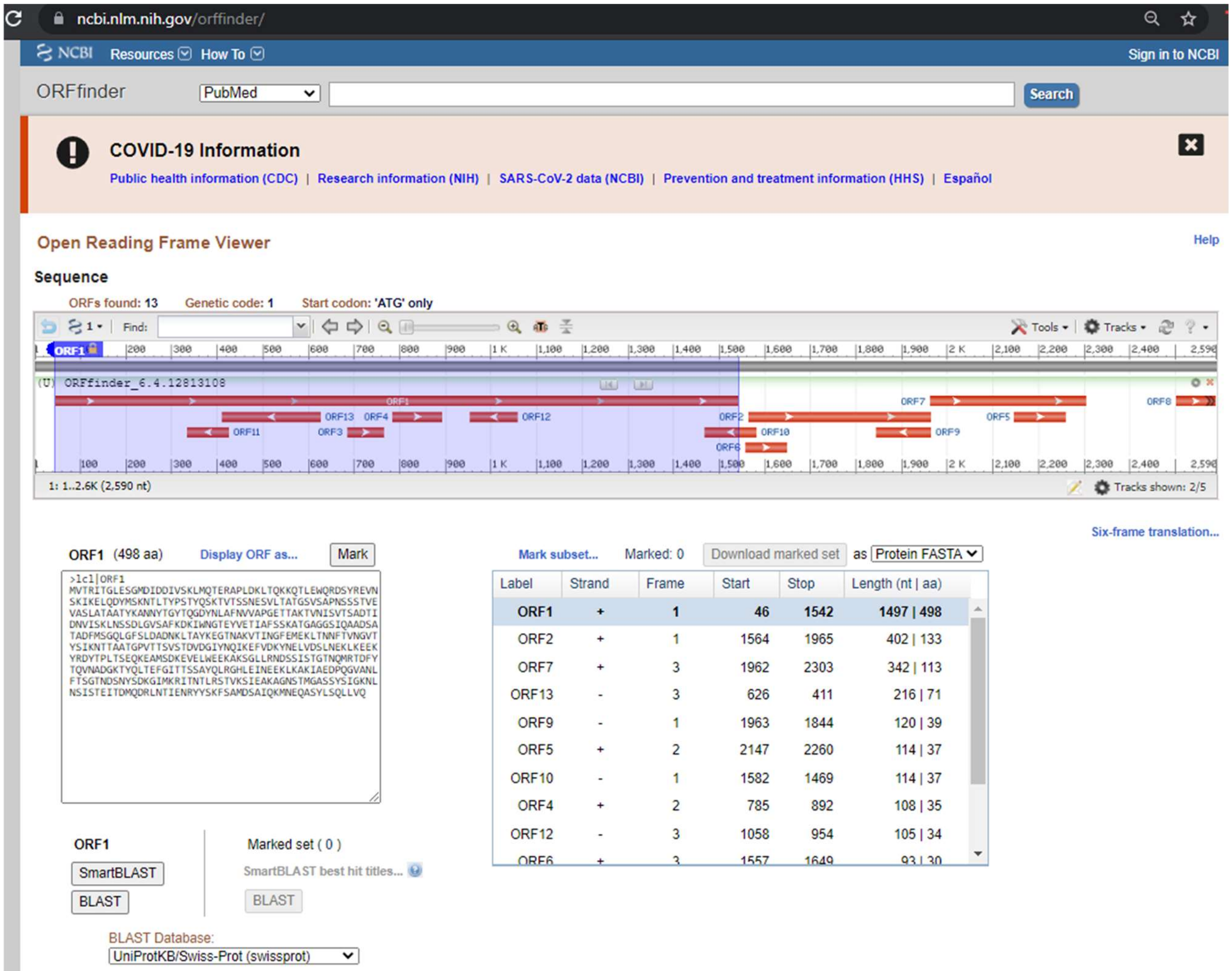
Jaimee Beckett
Unit 1-2 Graded Homework

1. Use ORF Finder to identify the locations of three coding regions (three longest ORFs) in the Bacillus subtilis genomic sequence (file:homework1.txt).

   ORF1, start is 46, stop is 1542
   ORF2, start is 1564, stop is 1965
   ORF7, start is 1962, stop 2303.



COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

Open Reading Frame Viewer

Help

Sequence

ORFs found: 13 Genetic code: 1 Start codon: 'ATG' only

1: 1..2.6K (2,590 nt)

ORF1 (498 aa) Display ORF as... Mark

```
>lcl|ORF1
MVTRITGLESGMDIDDIVSKLMQTERAPLDKLTQKKQTLEWQRDSYREVN
SKIKELQDYMSKNTLTYPSTYQSKTVTSSNESVLTATGSVSAPNSSSTVE
VASLATAATYKANNYTGYTQGDYNLAFNVVAPGETTAKTVNISVTSADTI
DNVISKLNSSDLGVSAFKDKIWNGTEYVETIAFSSKATGAGGSIQAADSA
TADFMSGQLGFSLDADNKLTAYKEGTNAKVTINGFEMEKLTNNFTVNGVT
YSIKNTTAATGPVTTSVSTDVDGIYNQIKEFVDKYNELVDSLNEKLKEEK
YRDYTPLTSEQKEAMSDKEVELWEEKAKSGLLRNDSSISTGTNQMRTDFY
TQVNADGKTYQLTEFGITTSSAYQLRGHLEINEEKLKAKIAEDPQGVANL
FTSGTNDSNYSDKGIMKRITNTLRSTVKSIEAKAGNSTMGASSYSIGKNL
NSISTEITDMQDRLNTIENRYYSKFSAMDSAIQKMNEQASYLSQLLVQ
```

Mark subset... Marked: 0 Download marked set as Protein FASTA

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
ORF13	-	3	626	411	216 | 71
ORF9	-	1	1963	1844	120 | 39
ORF5	+	2	2147	2260	114 | 37
ORF10	-	1	1582	1469	114 | 37
ORF4	+	2	785	892	108 | 35
ORF12	-	3	1058	954	105 | 34
ORF6	+	3	1557	1649	93 | 30

ORF1

Marked set (0)

BLAST Database: UniProtKB/Swiss-Prot (swissprot)

   a. On what reading frames are each of the genes in the Bacillus DNA based on ORF Finder?
      ORF1, reading frame is 1
      ORF2, reading frame is 1
      ORF7, reading frame is 3

2. Use the command line version of Glimmer to analyze CDSs in a partial sequence from Spiroplasma helicoides strain TABS-2, whose genome was submitted to GenBank on August 23, 2016 (file: sheliprt.fasta). The training set will be the full genome of S. helicoides strain TABS-2 (file: sheli.fasta).

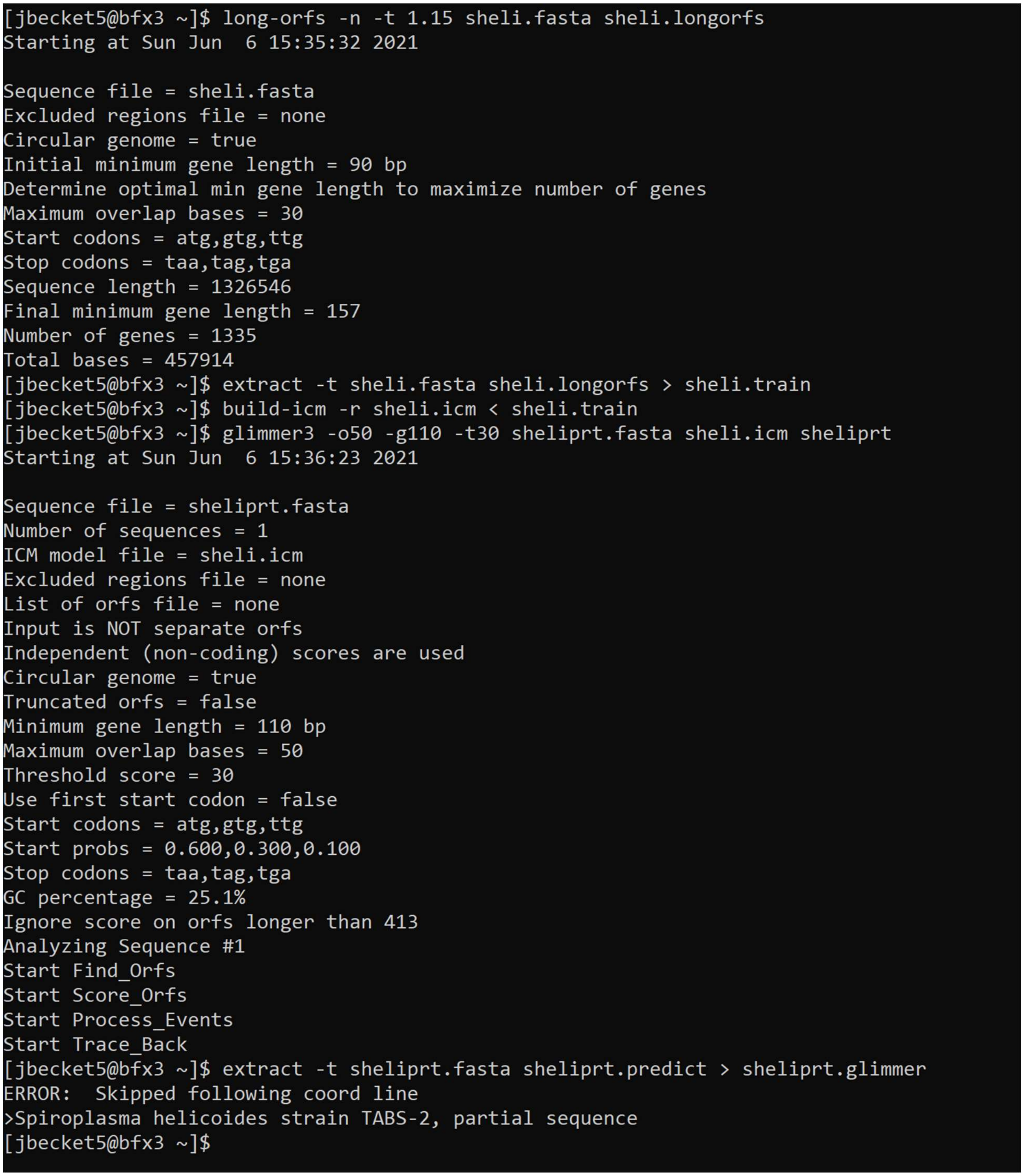
(i.e. full genome=> sheli.fasta   It is used to train.)
(i.e. partial genome => sheliprt.fasta You got the partial sequence. Predicting open reading frame for this file is the point of this particular homework question)

a. Either screen capture or copy & paste .predict file (command line).

```
>orf00001  635 991  len=354
ATGACTTATTCTTTTTCGTTTATAATTGAGGGAGTTCAAGAATACGATACCAGTAAATTT
TTAATCTCATCTATAGCTAGTTGTGCATTTATAATTGCACATTTATTATTTGAATATTTT
AGTCAATTGATTTTAAATCAATCTATTAAGTTAATTAACACAAAACTTAGAGTCATAACA
GCAAAAAACTTTTTTACAGAGAATTACAAAGTTAGTTTAGATACAGGAGAGTTTATAAAT
ATTAATTCAACTAAAATTAACCAATTGGCAGACAATTATTTTACATCAATTTTTGATATT
TCTAGATGCATCATAGCAATAATAATAAGTTATGGTTTTTTGTTATATATAAGT
>orf00002  998 1141  len=141
ATGTTGGCTGTGATGATTCTTTCATTACTAGTTTTAGTTATTCCGATGCTAATGTCTAAA
ATTGGACAAAAAAGAATAAATGTAGCTAATGAGGAAAATGATAAATTTTTGCAAACGACA
[jbecket5@bfx3 ~]$ cat sheliprt.glimmer
>orf00001  635 991  len=354
ATGACTTATTCTTTTTCGTTTATAATTGAGGGAGTTCAAGAATACGATACCAGTAAATTT
TTAATCTCATCTATAGCTAGTTGTGCATTTATAATTGCACATTTATTATTTGAATATTTT
AGTCAATTGATTTTAAATCAATCTATTAAGTTAATTAACACAAAACTTAGAGTCATAACA
GCAAAAAACTTTTTTACAGAGAATTACAAAGTTAGTTTAGATACAGGAGAGTTTATAAAT
ATTAATTCAACTAAAATTAACCAATTGGCAGACAATTATTTTACATCAATTTTTGATATT
TCTAGATGCATCATAGCAATAATAATAAGTTATGGTTTTTTGTTATATATAAGT
>orf00002  998 1141  len=141
ATGTTGGCTGTGATGATTCTTTCATTACTAGTTTTAGTTATTCCGATGCTAATGTCTAAA
ATTGGACAAAAAAGAATAAATGTAGCTAATGAGGAAAATGATAAATTTTTGCAAACGACA
AAAGATACTTACAACTCATAT
>orf00003  1154 1312  len=156
ATGAACCAAACAAATAAGCTTATTAACCAAATTGTTGAAGGATCAAAAAAGTTGGAAGTT
AAGAACCAGAAAATGAAAAACGTAATATCCACAACTAGGTTTTTAGATGAAATTGTTGTT
TTTCTTGGACAGGTTATTTTAATAATATTTTTTTGT
>orf00004  1334 1978  len=642
ATGAATATTGGTTTAATATTTACACTAAACATTTTATCAAGTGTTTATTGTTTTTTTAGT
AGTTCAAGTGCAAAAGCACTAATGAATATAATAAATCACAGAAAAGTTTATCTTTCAAAT
TATAAACAGGATAATAAAATCAATAATAATACTGTTATTGGAGAAGATTTAAAAACTATA
GAGTTTAAAAATGTTGATTTTAAATACAAAAATAGTTCTAATTTAATTATAGAAAAGTTC
AATTTAAAAATTAACAAGGGAGACAAGGTTCTTATTAAAGGTAAAAGTGGTATAGGAAAA
ACCACTTTATTAAAAACATTGTTTAATCCTTCTTTTAGAAGCAATGGTCAAGTGTATGTT
AATGAACAAGAAGTTGAAGCTTATGATATAAGATCTTTATGTTCATACATAAGTCAAGAT
ATTGTTTTTAGCAAAGGTAAATTGATAGATATGCTTAAAATAGCAAATGAATCTGCAGAA
GAAAAACAAGTATTAAGTTTATTTGAGTTACTTGGTCTAAATCAACTGTTAGAAAAATTA
CCCGAAGGGTTAAATACAAAAATTGATGATAATAGCTCAAATTTCTCTGGTGGTGAAAAA
CAAAGATTTTCGATTATAAGAGGATTGTTGGAAAATAAAAGT
>orf00006  2242 2463  len=219
ATGTTTGTTGATTTACTTGCAAGTACATCAGAAAAATTGACTGGAAATAGAATAGTTTTT
GCATTTGAAATAATTGCATTAGTAGTCTCAATTTTAATGATAACAGTTGGTATGATTCAA
AATAAAACTTCACAAACTGGACTGAGTGCATTAAATGGGGGTAATGATGAATTATTCTCA
```

```
AACTCTAAGGAAAGAGGAATGGACAGAACAATGTCTATT
>orf00008  2585 4003  len=1416
ATGGAAGAAAATATATTATCTCTAATAAAACAAAAACAAAAACTACATTTAAATGAATTA
CTTAAAACTTTTAAAGATGAAGAACTTTTAATGAGTTGTTTAAAAGAGCTACAAGATCAA
TATAAAATTAGTTGGTCAAAAGAAAATGTAGTTTATTTTATTGGGGAAAAATATAAAGTA
GGTTCAATTAAAATTAATGAAAAAGGCTTTGGTTTTGTAAAAGATTTAAATGATGTGGAA
CAAGATTATTTTGTACCACCAGATAGTCTTAATAAATCAATTACAACTGATGAAGTTGTT
TTTACAGTTTACAAAGAAAGCGAAGAAAGATATCGTGCAAATGTTGAAGATATTTCTTTA
AGGGTTAAATCTTTTTTAATAGGAGAAATTCAGCCATCAAGAGATGGTCGTTTTTTAGAT
TTTATCCCTAGTGAACCCGGTTTTAAAAATTACAGAATTGTAATGATTAATTCAAAGGAT
TTTAAATTAAAAAAAGATTTACTAGTTAAAGTCAAAATTTTGAATGTAAAAGAAAAAAAA
CTATTCACCAAAATTCAAAAAATAATTGGTGACTCAAATAAAGCTGTTGACAGAATTATT
TCAATTGCATATGAGTTTAATATAAACCCAGATTTTAATAGACAAACATTAGAGAATGCA
GACCAAGTTGCAATACCAATTAACTATGAAGATGAACAAGTAAAAAGAAGATTAAAAAAC
TCACTAGTAGATAAAAATTTAGTAACTATAGATGGTTCTGACTCAAAAGATTTAGATGAT
GCAATTTACGTGGAAAAAACTAAGGACGGATATAAATTATTTGTAGCAATTGCTGATGTA
AGTTATTATGTTTTACCTTTTTCACCTTTAGATAACACAGCTTTATATAGAGGTAATTCG
ACTTATCTTGCAAATAAAGTAATTCCAATGCTTCCAGAAAAACTTTCAAATGGAGTTTGT
AGTTTGAATCCAAATGAAGATAAACTTTGTATGGTTTCTGAAATGGATTTTGATAATAAT
GGAGTTATGAAAAACAAAAAAGTTTATGAATCAATCATGAATTCAAAAGCAAGACTAACA
TATAAAGAAGTAAATGATTTATTTGAAAAAAAATGTTTCAAATAGAGATAAAGAAATTGTT
GATATGCTTTTGGTTTCAAAAGAGCTACATGAATTAATTGATAAAGAAAGAGTATCAAGA
GGTTCAATCGATTTTGATGTTCCTGAACCAAAAATTGTTCTGGATAAAGAAAGCAATGTA
GTAGATATAGTTCCAAGAGATAGAGGAGTTAGTGAAAGACTAATCGAAAATTTTATGGTT
AGTGCTAATGAATCGGTTGCACAAATAATTTTTGAAAAAAATCTACCATATGTTTATAGA
AACCACGGTGCTCCTAAAGAAGAAAACTTGATTGAA
>orf00009  4010 4678  len=666
TTGATTAGAGCTTTGGGTATTAATGTGAAACTTACAGATTTAGAAAAAGTAAATCCCAAA
ACTATAAGAATGGCATTAGACCAAATTTCCAAACAGATTGAGGATCAAACAGAAAGAGAT
GTTATCAATGTTACATTGCTTAAGTTTATGGAAAAAGCTGCATATGAACTTGAAAATATA
GGTCACTTTGGTTTAGCTAGTGAATGCTACACCCACTTTACAAGTCCGATAAGAAGATAT
AGTGATTTAATGGTCCATAGATATTTAAAACAATATTTGATTGATAAAGATTTACGAGAT
TTCAAACTTGATTTAAATGAAAAATTTATAAATAAAGCTTGTAAAATAATTAATGAAACA
GAAAAAAACTCAGTTAATGCCGAAAGAGAAGTAAATAAAGTTTGTATGGCAGAGTTTATG
ACTAAACATATTGAGAAAGAGTATGAAGGGGTAGTTGCTGCTGTCTTGAAGTTTGGGTTA
TTTGTTCAGTTATCAAATTGCGTTGAAGGACTAATTCACATATCTGAACTTCCAGAATTT
ACTTTTGATCCCAAAACCAATATCTTGGTAAACAAACAAAATAAAGTGTTTAGACTTGGT
CAAAAAGTTAAAATAAAAGTTAAAAATGCTGATGTAAAAAAAAGAATTATTGACTTTGTG
CTAGTA
>orf00010  4880 5143  len=261
ATGAATATAAAAAAGTATGAGTATGCTAATTATGTTAAACAAGACCCAACAAGAACTAGA
AAACTATTGCTAAATAAAGATGAAATTAAAAAAATTTTAAAAAGAGTACAATTAGAAAAT
CTAACCATAATTCCATTAAAGTTGTATTTAAAGGGCAATTATGCAAAACTGGAAATCGGA
ATCGGTAAGGGTAAAAAACTTATAGATAAAAGAGAGACTATCAAAAAAAGAGATATAGAA
AGACGTTTAAATAAAATTAAG
```

b. Either screen capture or copy & paste all the necessary commands you used to obtain your results (you don't need to include basic commands such as "cd" or "ls").

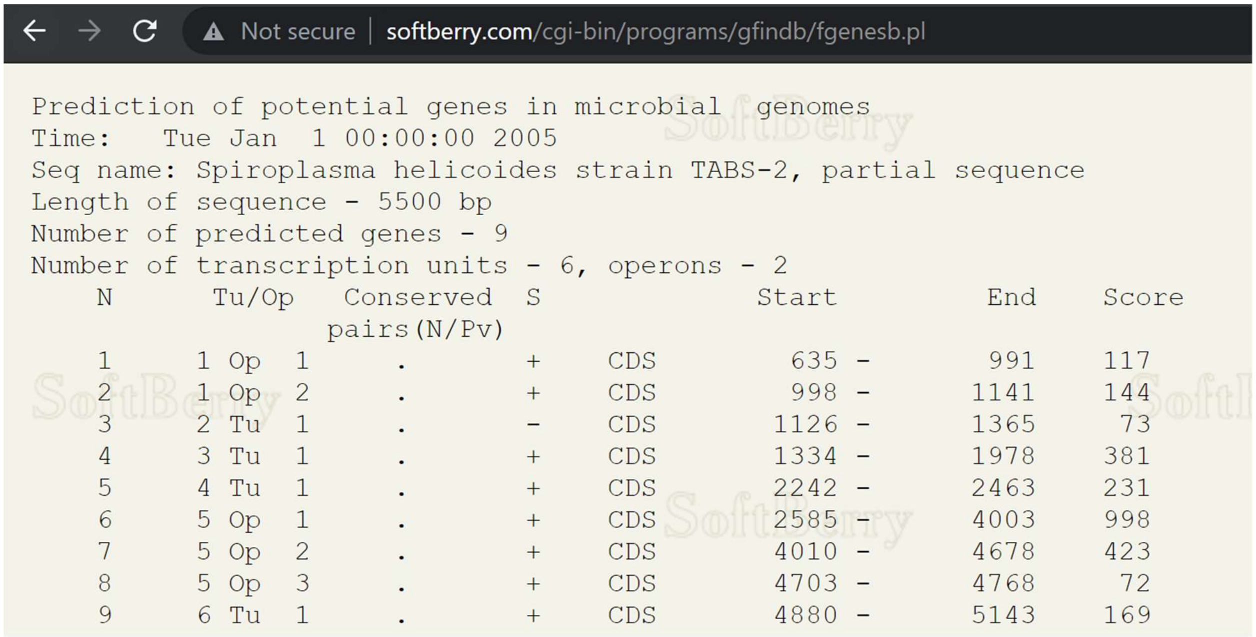


```
[jbecket5@bfx3 ~]$ long-orfs -n -t 1.15 sheli.fasta sheli.longorfs
Starting at Sun Jun  6 15:35:32 2021

Sequence file = sheli.fasta
Excluded regions file = none
Circular genome = true
Initial minimum gene length = 90 bp
Determine optimal min gene length to maximize number of genes
Maximum overlap bases = 30
Start codons = atg,gtg,ttg
Stop codons = taa,tag,tga
Sequence length = 1326546
Final minimum gene length = 157
Number of genes = 1335
Total bases = 457914
[jbecket5@bfx3 ~]$ extract -t sheli.fasta sheli.longorfs > sheli.train
[jbecket5@bfx3 ~]$ build-icm -r sheli.icm < sheli.train
[jbecket5@bfx3 ~]$ glimmer3 -o50 -g110 -t30 sheliprt.fasta sheli.icm sheliprt
Starting at Sun Jun  6 15:36:23 2021

Sequence file = sheliprt.fasta
Number of sequences = 1
ICM model file = sheli.icm
Excluded regions file = none
List of orfs file = none
Input is NOT separate orfs
Independent (non-coding) scores are used
Circular genome = true
Truncated orfs = false
Minimum gene length = 110 bp
Maximum overlap bases = 50
Threshold score = 30
Use first start codon = false
Start codons = atg,gtg,ttg
Start probs = 0.600,0.300,0.100
Stop codons = taa,tag,tga
GC percentage = 25.1%
Ignore score on orfs longer than 413
Analyzing Sequence #1
Start Find_Orfs
Start Score_Orfs
Start Process_Events
Start Trace_Back
[jbecket5@bfx3 ~]$ extract -t sheliprt.fasta sheliprt.predict > sheliprt.glimmer
ERROR:  Skipped following coord line
>Spiroplasma helicoides strain TABS-2, partial sequence
[jbecket5@bfx3 ~]$
```

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



```
Prediction of potential genes in microbial genomes
Time:   Tue Jan  1 00:00:00 2005
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. How many CDSs are listed?

There are 9 CDSs listed

b. How many mRNAs are predicted to code for those CDSs?

There are 6 mRNAs predicted

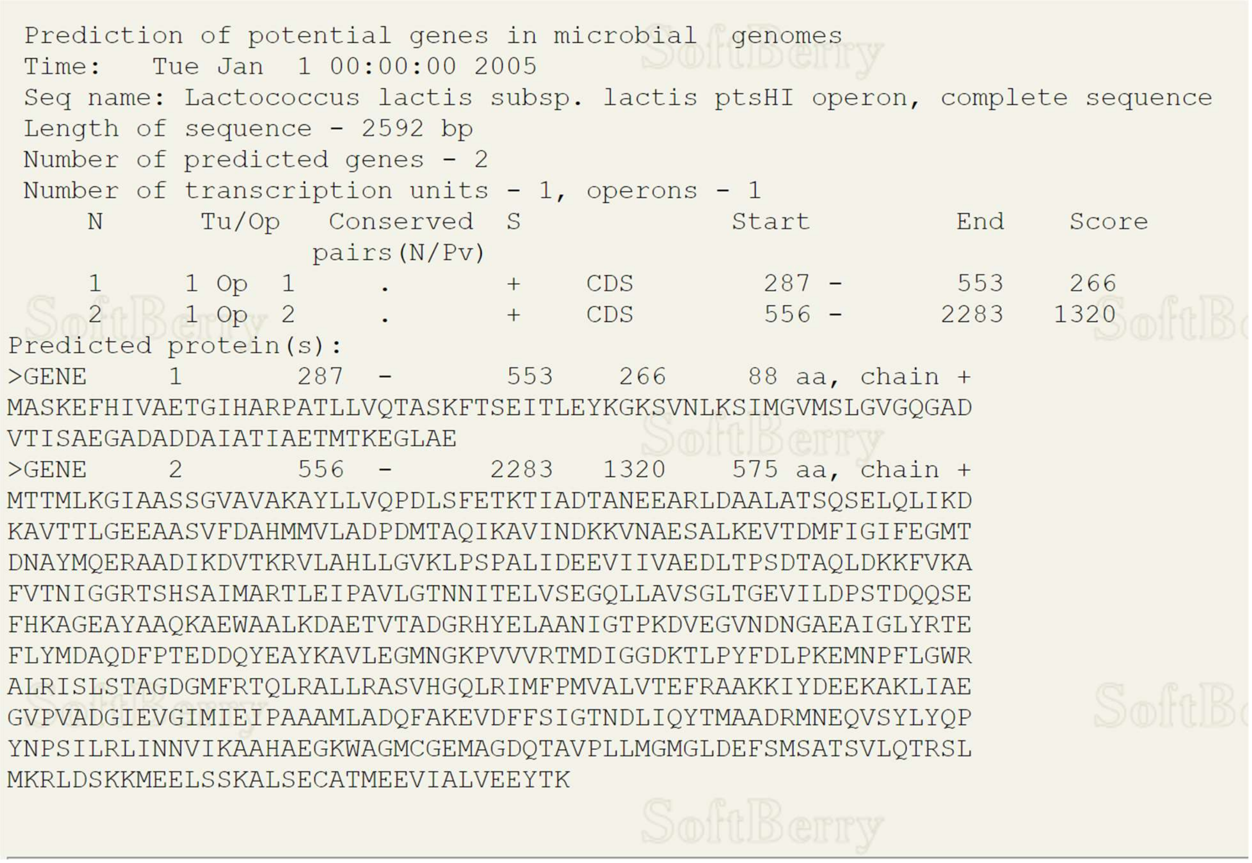
4. Use the attached lactococcus DNA sequence to identify the following genic features (file: lactococcus.txt).

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

FGENESB:

1st location, start is 287, end is 553

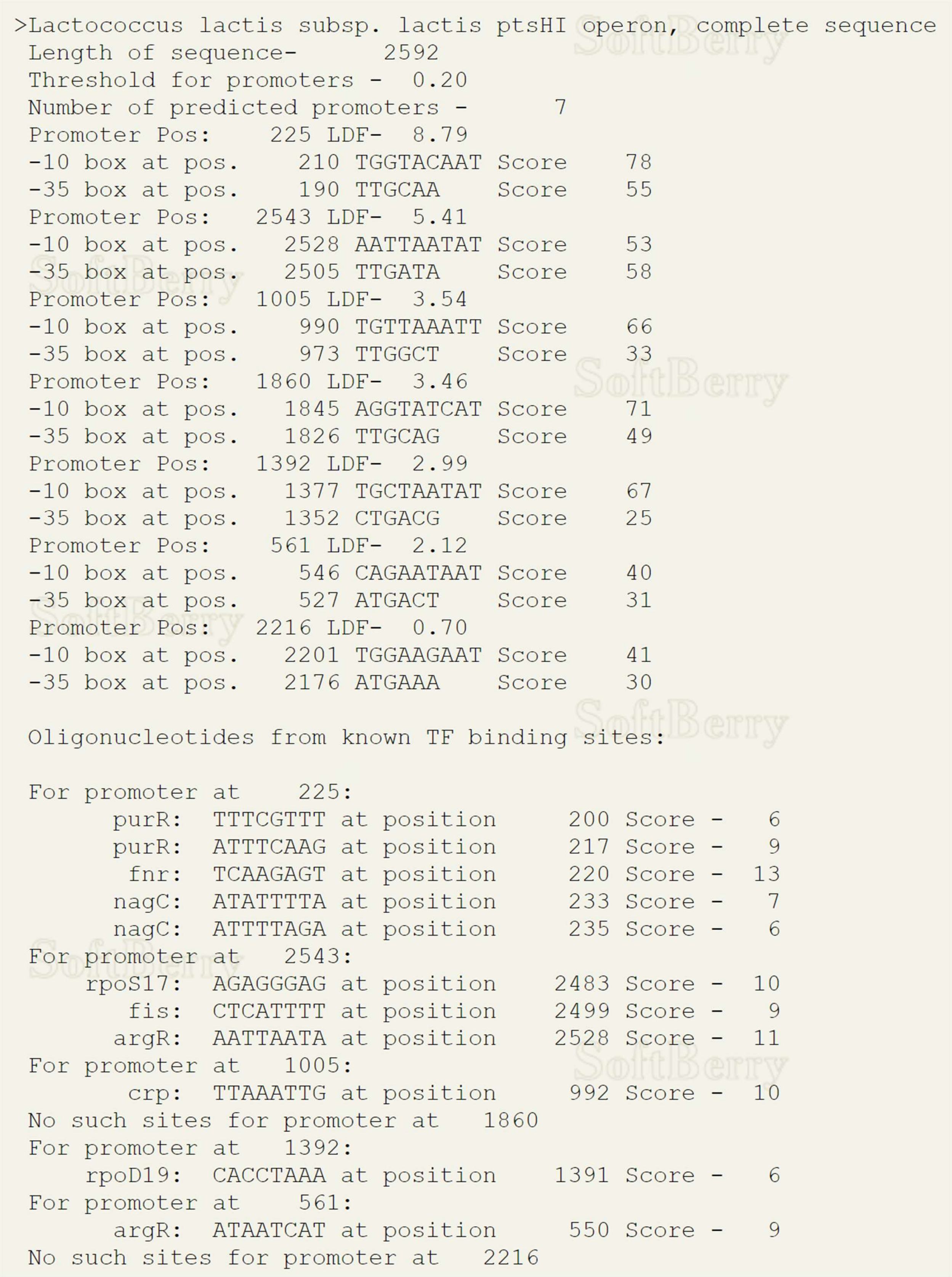
2nd location, start is 556, end is 2283



```
 Prediction of potential genes in microbial  genomes
 Time:   Tue Jan  1 00:00:00 2005
 Seq name: Lactococcus lactis subsp. lactis ptsHI operon, complete sequence
 Length of sequence - 2592 bp
 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
Predicted protein(s):
>GENE     1       287  -       553    266     88 aa, chain +
MASKEFHIVAETGIHARPATLLVQTASKFTSEITLEYKGKSVNLKSIMGVMSLGVGQGAD
VTISAEGADADDAIATIAETMTKEGLAE
>GENE     2       556  -      2283   1320    575 aa, chain +
MTTMLKGIAASSGVAVAKAYLLVQPDLSFETKTIADTANEEARLDAALATSQSELQLIKD
KAVTTLGEEAASVFDAHMMVLADPDMTAQIKAVINDKKVNAESALKEVTDMFIGIFEGMT
DNAYMQERAADIKDVTKRVLAHLLGVKLPSPALIDEEVIIVAEDLTPSDTAQLDKKFVKA
FVTNIGGRTSHSAIMARTLEIPAVLGTNNITELVSEGQLLAVSGLTGEVILDPSTDQQSE
FHKAGEAYAAQKAEWAALKDAETVTADGRHYELAANIGTPKDVEGVNDNGAEAIGLYRTE
FLYMDAQDFPTEDDQYEAYKAVLEGMNGKPVVVRTMDIGGDKTLPYFDLPKEMNPFLGWR
ALRISLSTAGDGMFRTQLRALLRASVHGQLRIMFPMVALVTEFRAAKKIYDEEKAKLIAE
GVPVADGIEVGIMIEIPAAAMLADQFAKEVDFFSIGTNDLIQYTMAADRMNEQVSYLYQP
YNPSILRLINNVIKAAHAEGKWAGMCGEMAGDQTAVPLLMGMGLDEFSMSATSVLQTRSL
MKRLDSKKMEELSSKALSECATMEEVIALVEEYTK
```

BPROM:

-35 signal is 210, -10 signal is 190



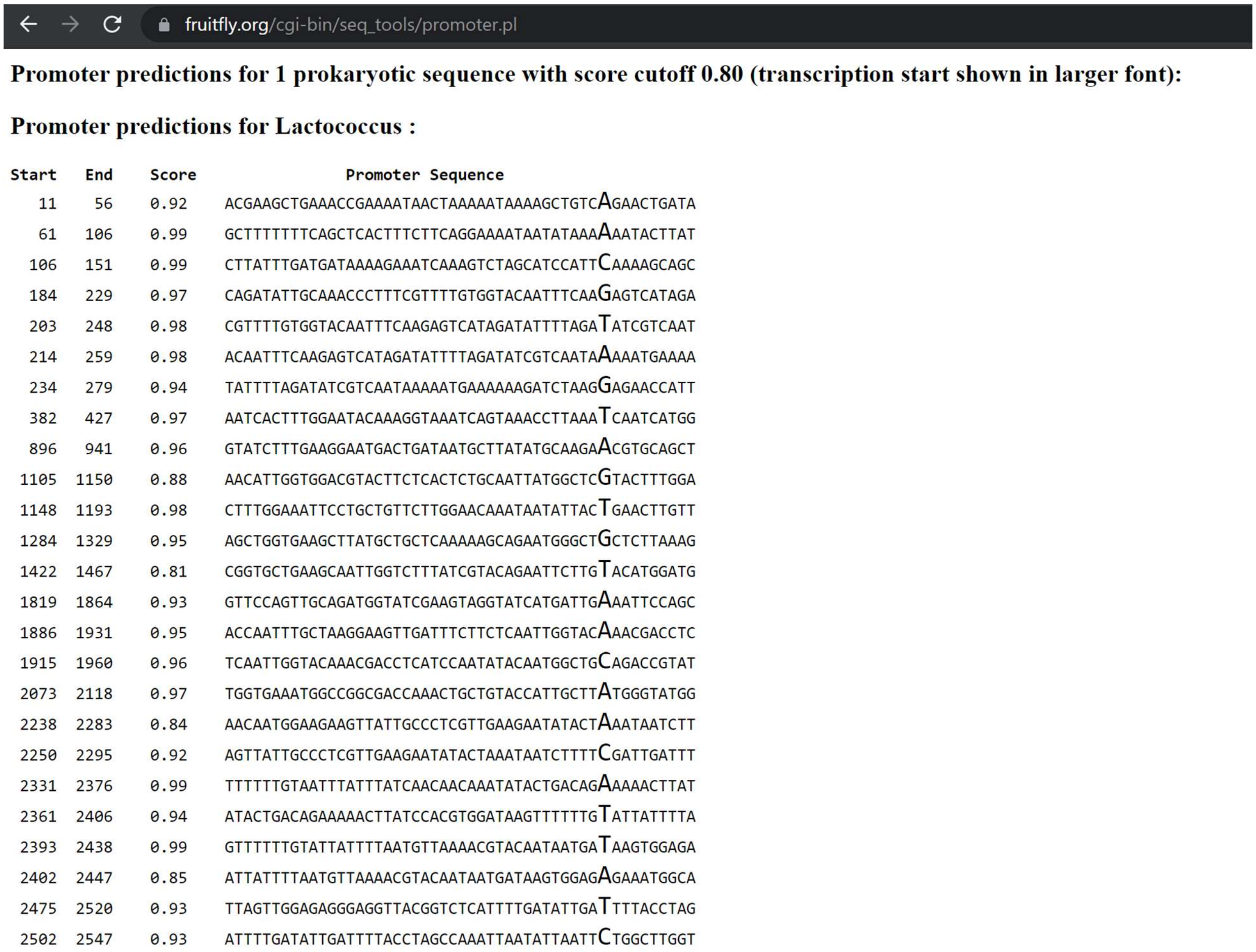
```
>Lactococcus lactis subsp. lactis ptsHI operon, complete sequence
 Length of sequence-      2592
 Threshold for promoters -  0.20
 Number of predicted promoters -       7
 Promoter Pos:     225 LDF-  8.79
 -10 box at pos.     210 TGGTACAAT Score    78
 -35 box at pos.     190 TTGCAA    Score    55
 Promoter Pos:    2543 LDF-  5.41
 -10 box at pos.    2528 AATTAATAT Score    53
 -35 box at pos.    2505 TTGATA    Score    58
 Promoter Pos:    1005 LDF-  3.54
 -10 box at pos.     990 TGTTAAATT Score    66
 -35 box at pos.     973 TTGGCT    Score    33
 Promoter Pos:    1860 LDF-  3.46
 -10 box at pos.    1845 AGGTATCAT Score    71
 -35 box at pos.    1826 TTGCAG    Score    49
 Promoter Pos:    1392 LDF-  2.99
 -10 box at pos.    1377 TGCTAATAT Score    67
 -35 box at pos.    1352 CTGACG    Score    25
 Promoter Pos:     561 LDF-  2.12
 -10 box at pos.     546 CAGAATAAT Score    40
 -35 box at pos.     527 ATGACT    Score    31
 Promoter Pos:    2216 LDF-  0.70
 -10 box at pos.    2201 TGGAAGAAT Score    41
 -35 box at pos.    2176 ATGAAA    Score    30

 Oligonucleotides from known TF binding sites:

 For promoter at     225:
       purR:  TTTCGTTT at position     200 Score -   6
       purR:  ATTTCAAG at position     217 Score -   9
        fnr:  TCAAGAGT at position     220 Score -  13
       nagC:  ATATTTTA at position     233 Score -   7
       nagC:  ATTTTAGA at position     235 Score -   6
 For promoter at    2543:
     rpoS17:  AGAGGGAG at position    2483 Score -  10
        fis:  CTCATTTT at position    2499 Score -   9
       argR:  AATTAATA at position    2528 Score -  11
 For promoter at    1005:
        crp:  TTAAATTG at position     992 Score -  10
 No such sites for promoter at    1860
 For promoter at    1392:
     rpoD19:  CACCTAAA at position    1391 Score -   6
 For promoter at     561:
       argR:  ATAATCAT at position     550 Score -   9
 No such sites for promoter at    2216
```

b. Run the prokaryotic promoter prediction at the Berkeley Drosophila Neural Network Prediction site. What is the most likely promoter to match the BPROM result? At what nucleotide is the transcription start site?

Most likely promoter start is 184, end is 229
Transcription start site is at nucleotide G


fruitfly.org/cgi-bin/seq_tools/promoter.pl

Promoter predictions for 1 prokaryotic sequence with score cutoff 0.80 (transcription start shown in larger font):

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
  203    248   0.98    CGTTTTGTGGTACAATTTCAAGAGTCATAGATATTTTAGATATCGTCAAT
  214    259   0.98    ACAATTTCAAGAGTCATAGATATTTTAGATATCGTCAATAAAAATGAAAA
  234    279   0.94    TATTTTAGATATCGTCAATAAAAATGAAAAAAGATCTAAGGAGAACCATT
  382    427   0.97    AATCACTTTGGAATACAAAGGTAAATCAGTAAACCTTAAATCAATCATGG
  896    941   0.96    GTATCTTTGAAGGAATGACTGATAATGCTTATATGCAAGAACGTGCAGCT
 1105   1150   0.88    AACATTGGTGGACGTACTTCTCACTCTGCAATTATGGCTCGTACTTTGGA
 1148   1193   0.98    CTTTGGAAATTCCTGCTGTTCTTGGAACAAATAATATTACTGAACTTGTT
 1284   1329   0.95    AGCTGGTGAAGCTTATGCTGCTCAAAAAGCAGAATGGGCTGCTCTTAAAG
 1422   1467   0.81    CGGTGCTGAAGCAATTGGTCTTTATCGTACAGAATTCTTGTACATGGATG
 1819   1864   0.93    GTTCCAGTTGCAGATGGTATCGAAGTAGGTATCATGATTGAAATTCCAGC
 1886   1931   0.95    ACCAATTTGCTAAGGAAGTTGATTTCTTCTCAATTGGTACAAACGACCTC
 1915   1960   0.96    TCAATTGGTACAAACGACCTCATCCAATATACAATGGCTGCAGACCGTAT
 2073   2118   0.97    TGGTGAAATGGCCGGCGACCAAACTGCTGTACCATTGCTTATGGGTATGG
 2238   2283   0.84    AACAATGGAAGAAGTTATTGCCCTCGTTGAAGAATATACTAAATAATCTT
 2250   2295   0.92    AGTTATTGCCCTCGTTGAAGAATATACTAAATAATCTTTTCGATTGATTT
 2331   2376   0.99    TTTTTTGTAATTTATTTATCAACAACAAATATACTGACAGAAAAACTTAT
 2361   2406   0.94    ATACTGACAGAAAAACTTATCCACGTGGATAAGTTTTTTGTATTATTTTA
 2393   2438   0.99    GTTTTTTGTATTATTTTAATGTTAAAACGTACAATAATGATAAGTGGAGA
 2402   2447   0.85    ATTATTTTAATGTTAAAACGTACAATAATGATAAGTGGAGAGAAATGGCA
 2475   2520   0.93    TTAGTTGGAGAGGGAGGTTACGGTCTCATTTTGATATTGATTTTACCTAG
 2502   2547   0.93    ATTTTGATATTGATTTTACCTAGCCAAATTAATATTAATTCTGGCTTGGT
```

5. Given the location of a CDS, explain why it is usually more difficult to predict a eukaryotic transcription start site (absent RNA-seq, cDNA data) than it is to predict a prokaryotic transcription start site. Your answer should address distance of a TSS from a start codon and differences in non-coding DNA frequency between eukaryotes and prokaryotes.

It is usually more difficult to predict a eukaryotic TSS than it is to predict a prokaryotic TSS because of alternative splicing. Eukaryotes have introns and exons which results in the average TSS being further from the stop coding. The distance is larger when there is an intron first.