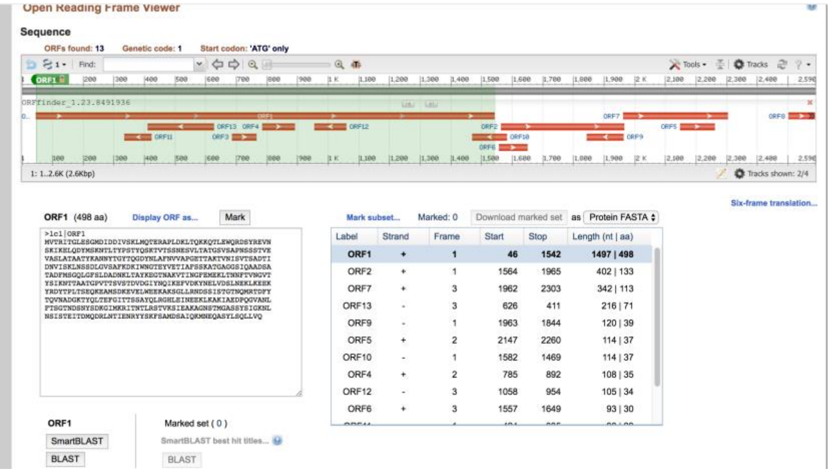
JULIE GARCIA
JANUARY 20, 2018

Genomics Spring 2018
Gene Prediction Homework

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

Below you can see the results of using the ORF Finder to find the longest ORFs in Bacillus subtillis. The three longest coding regions are ORF1 from 46-1542, ORF2 from 1564-1965 and ORF7 overlaps the last one at 1962-2303. These open reading frames are on the plus strand.

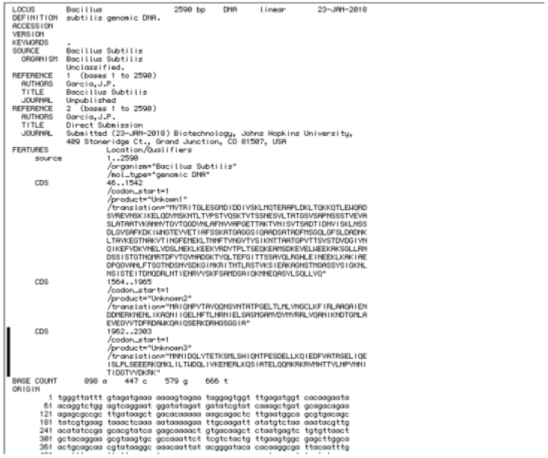


```
>lcl|ORF1
MVTRITGLESGMDIDDIVSKLMQTERAPLDKLTQKKQTLEWQRDSYREVN
SKIKELQDYMSKNTLTYPSTYQSKTVTSSNESVLTATGSVSAPNSSSTVE
VASLATAATYKANNYTGYTQGDYNLAFNVVAPGETTAKTVNISVTSADTI
DNVISKLNSSDLGVSAFKDKIWNGTEYVETIAFSSKATGAGGSIQAADSA
TADFMSGQLGFSLDADNKLTAYKEGTNAKVTINGFEMEKLTNNFTVNGVT
YSIKNTTAATGPVTTSVSTDVDGIYNQIKEFVDKYNELVDSLNEKLKEEK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

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

a. Using Sequin, annotate all three CDSs (one sequin file), validate, and then export the file to GenBank format. Submit the GenBank file. Screen capture the results in the pdf document.

Below is a screenshot of the Sequin results. I also submitted the exported GeneBank file.



```
LOCUS       Bacillus                 2590 bp    DNA     linear       23-JAN-2018
DEFINITION  subtilis genomic DNA.
ACCESSION
VERSION
KEYWORDS    .
SOURCE      Bacillus Subtilis
  ORGANISM  Bacillus Subtilis
            Unclassified.
REFERENCE   1  (bases 1 to 2590)
  AUTHORS   Garcia,J.P.
  TITLE     Baccillus Subtilis
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 2590)
  AUTHORS   Garcia,J.P.
  TITLE     Direct Submission
  JOURNAL   Submitted (23-JAN-2018) Biotechnology, Johns Hopkins University,
            409 Stoneridge Ct., Grand Junction, CO 81507, USA
FEATURES             Location/Qualifiers
     source          1..2590
                     /organism="Bacillus Subtilis"
                     /mol_type="genomic DNA"
     CDS             46..1542
                     /codon_start=1
                     /product="Unkown1"
                     /translation="MVTRITGLESGMDIDDIVSKLMQTERAPLDKLTQKKQTLEWQRD
                     SYREVNSKIKELQDVMSKNTLTVPSTVQSKTVTSSNESVLTATGSVSAPNSSSTVEVA
                     SLATAATVKANNVTGVTQGDVNLAFNVVAPGETTAKTVNISVTSADTIDNVISKLNSS
                     DLGVSAFKDKIWNGTEVVETIAFSSKATGAGGSIQAADSATADFMSGQLGFSLDADNK
                     LTAVKEGTNAKVTINGFEMEKLTNNFTVNGVTVSIKNTTAATGPVTTSVSTDVDGIVN
                     QIKEFVDKVNELVDSLNEKLKEEKVRDVTPLTSEQKEAMSDKEVELWEEKAKSGLLRN
                     DSSISTGTNQMRTDFVTQVNADGKTVQLTEFGITTSSAVQLRGHLEINEEKLKAKIAE
                     DPQGVANLFTSGTNDSNVSDKGIMKRITNTLRSTVKSIEAKAGNSTMGASSVSIGKNL
                     NSISTEITDMQDRLNTIENRVVSKFSAMDSAIQKMNEQASVLSQLLVQ"
     CDS             1564..1965
                     /codon_start=1
                     /product="Unknown2"
                     /translation="MAIQNPVTAVQQNSVNTATPGELTLMLVNGCLKFIRLAAQAIEN
                     DDMERKNENLIKAQNIIQELNFTLNRNIELSASMGAMVDVMVRRLVQANIKNDTGMLA
                     EVEGYVTDFRDAWKQAIQSERKDRHGSGGIA"
     CDS             1962..2303
                     /codon_start=1
                     /product="Unknown3"
                     /translation="MNNIDQLYTETKSMLSHIQNTPESDELLKQIEDFVATRSELIQE
                     ISLPLSEEERKQMKLILTWDQLIVKEMERLKQSIATELQQMKRKRVMHTTVLNPVNNI
                     TIDGTVVDKRK"
BASE COUNT      898 a    447 c    579 g    666 t
ORIGIN
        1 tgggttattt gtagatgaaa aaaagtagaa taggagtggt ttgagatggt cacaagaata
       61 acaggtctgg agtcaggaat ggatatagat gatatcgtat caaagctgat gcagacagaa
      121 agagcgccgc ttgataagct gacacaaaaa aagcagactc ttgaatggca gcgtgacagc
      181 tatcgtgaag taaactcaaa aataaaagaa ttgcaagatt atatgtctaa aaatacgttg
      241 acatatccga gcacgtatca gagcaaaact gtgacaagct ctaatgagtc tgtgttaact
      301 gctacaggaa gcgtaagtgc gccaaattct tcgtctactg ttgaagtggc gagcttggca
      361 actgcagcaa cgtataaggc aaacaattat acgggataca cacaaggcga ttacaatttg
```

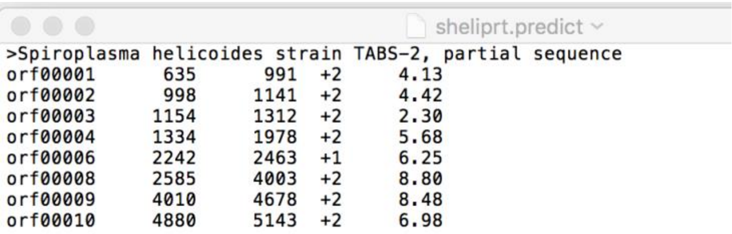
b. On what reading frames are each of the genes in the Bacillus DNA? (answer should be at the master pdf document)

ORF1 on Frame 1, ORF2 on Frame 1, ORF7 on Frame 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). (1 point)
(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).

Here are the predicted open reading frames for the partial sequence given, using Glimmer.



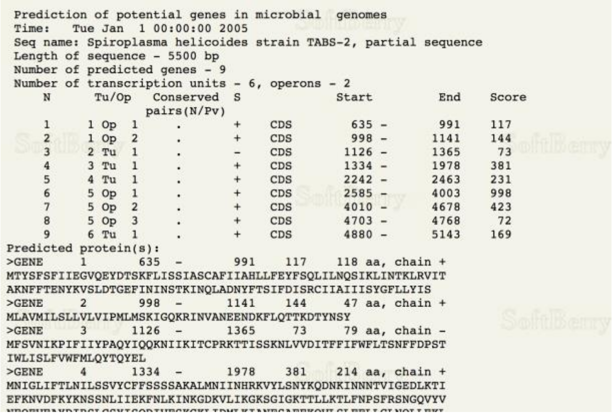
```
>Spiroplasma helicoides strain TABS-2, partial sequence
orf00001      635      991  +2     4.13
orf00002      998     1141  +2     4.42
orf00003     1154     1312  +2     2.30
orf00004     1334     1978  +2     5.68
orf00006     2242     2463  +1     6.25
orf00008     2585     4003  +2     8.80
orf00009     4010     4678  +2     8.48
orf00010     4880     5143  +2     6.98
```

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").

```
long-orfs 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. 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. (1 point)

Screenshot of FGENESB results below.



```
 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
Predicted protein(s):
>GENE     1        635   -        991     117     118 aa, chain +
MTYSFSFIIEGVQEYDTSKFLISSIASCAFIIAHLLFEYFSQLILNQSIKLINTKLRVIT
AKNFFTENYKVSLDTGEFININSTKINQLADNYFTSIFDISRCIIAIIISYGFLLYIS
>GENE     2        998   -       1141     144      47 aa, chain +
MLAVMILSLLVLVIPMLMSKIGQKRINVANEENDKFLQTTKDTYNSY
>GENE     3       1126   -       1365      73      79 aa, chain -
MFSVNIKPIFIIYPAQYIQQKNIIKITCPRKTTISSKNLVVDITFFIFWFLTSNFFDPST
IWLISLFVWFMLQYTQYEL
>GENE     4       1334   -       1978     381     214 aa, chain +
MNIGLIFTLNILSSVYCFFSSSSAKALMNIINHRKVYLSNYKQDNKINNNTVIGEDLKTI
EFKNVDFKYKNSSNLIIEKFNLKINKGDKVLIKGKSGIGKTTLLKTLFNPSFRSNGQVYV
```

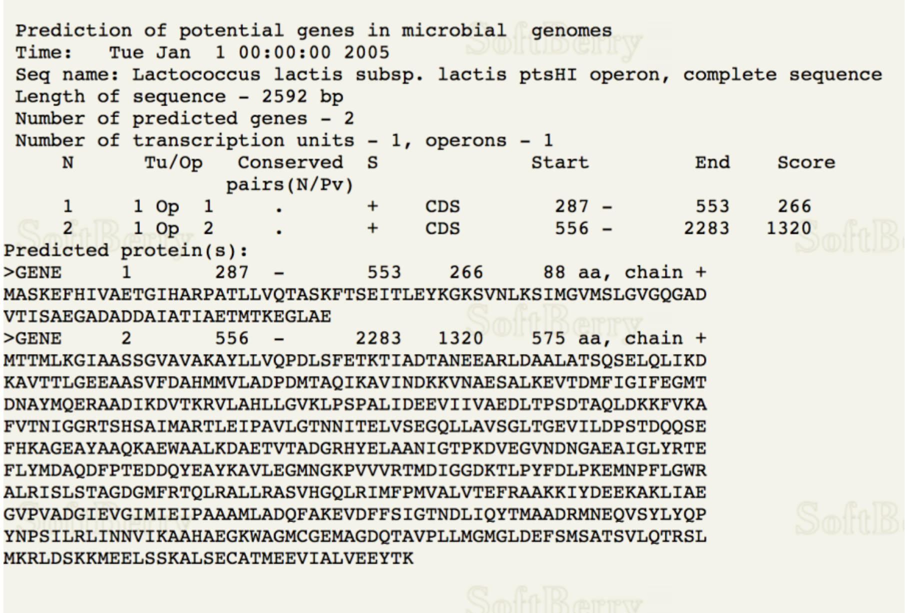
a. How many CDSs are listed? 9 CDSs are listed.

b. How many mRNAs are predicted to code for those CDSs? 6 mRNAs are predicted to code, 2 of them are operons and 4 are single gene transcription units.

4. Use the attached lactococcus DNA sequence to identify the following genic features (file: lactococcus.txt). (1 point)

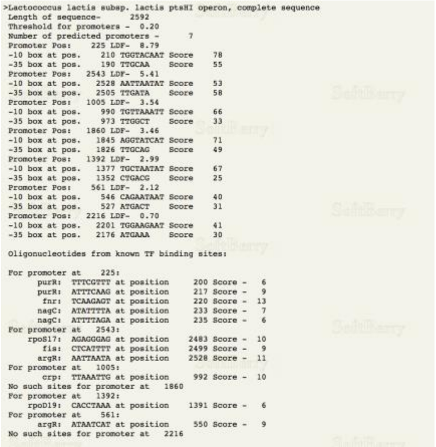
a. Run FGENESB to find the location of two genes on an operon, then run BPROM to find the locations of the -35 signal and the -10 signal. Report the CDS locations and the locations of the most appropriate -35 signal and -10 signal.

FGENESB Results: (Location of Gene 1 – 287-553, location of Gene 2 – 556-2283, The operon contains the two genes from 287 – 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 Results below: Since Gene 1 Starts at 287, the appropriate location for the -35 box in the results is 190, and the -10 box at location 210. This shows the promoter at 225.

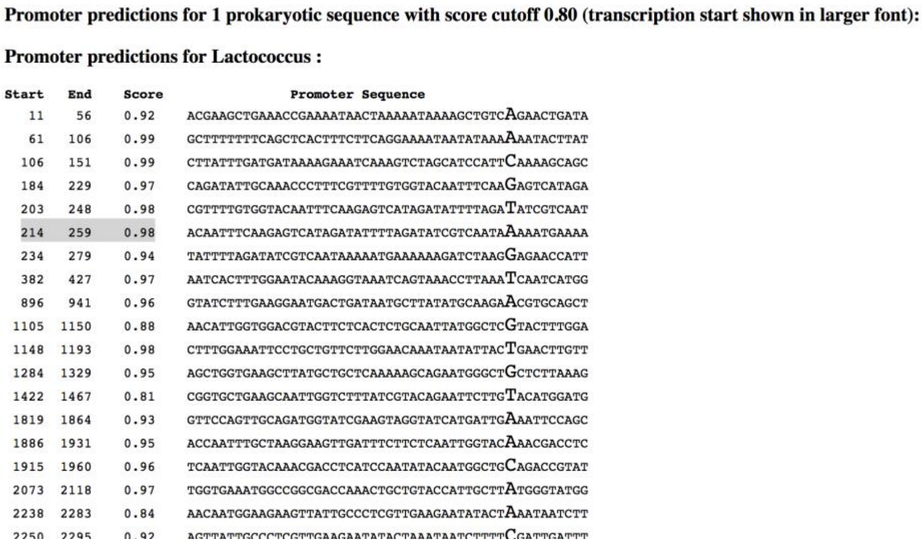


```
>Lactococcus lactis subsp. lactis ptsHI operon, complete sequence
 Length of sequence-      2592
 Threshold for promoters -  0.20
 Number of predicted promoters -       7
 Promoter Pos:     225 LDF-  8.79
 -10 box at pos.     210 TGGTACAAT Score    78
 -35 box at pos.     190 TTGCAA    Score    55
 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?

Results from Berkeley Drosophila Neural Network Prediction Site are below, with the likely promoter highlighted between 214 and 259. Since BPROM predicted the promoter at 225, this is the most likely location of the promoter. The transcription start site (TSS) is highlighted in that row as an "A" at position 254 (if you count from the left). I noticed that the sequence says 214-259, however if you count to the end there are more letters than there should be, so I figured counting from the left (starting at 214) would be best.



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

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. (1 point)

Eukaryotic transcription start sites are often harder to predict, because intergenic eukaryotic DNA generally contains introns in addition to exons, while prokaryotic DNA does not. Prokaryotic DNA is generally very compact, without a lot of excess non-coding regions or spaces between genes. Eukaryotic DNA may have introns between the TSS and the first exon of a gene, while prokaryotic DNA's TSS is generally very close to the start codon of the first gene, making it easy to find. It may also be common for eukaryotic genes to have several transcription start sites. More complicated organisms can have more complicated promoter regions, making it more difficult to find the TSS of a particular gene. This is why RNA-seq data is often used to compliment gene prediciton in the annotation process.