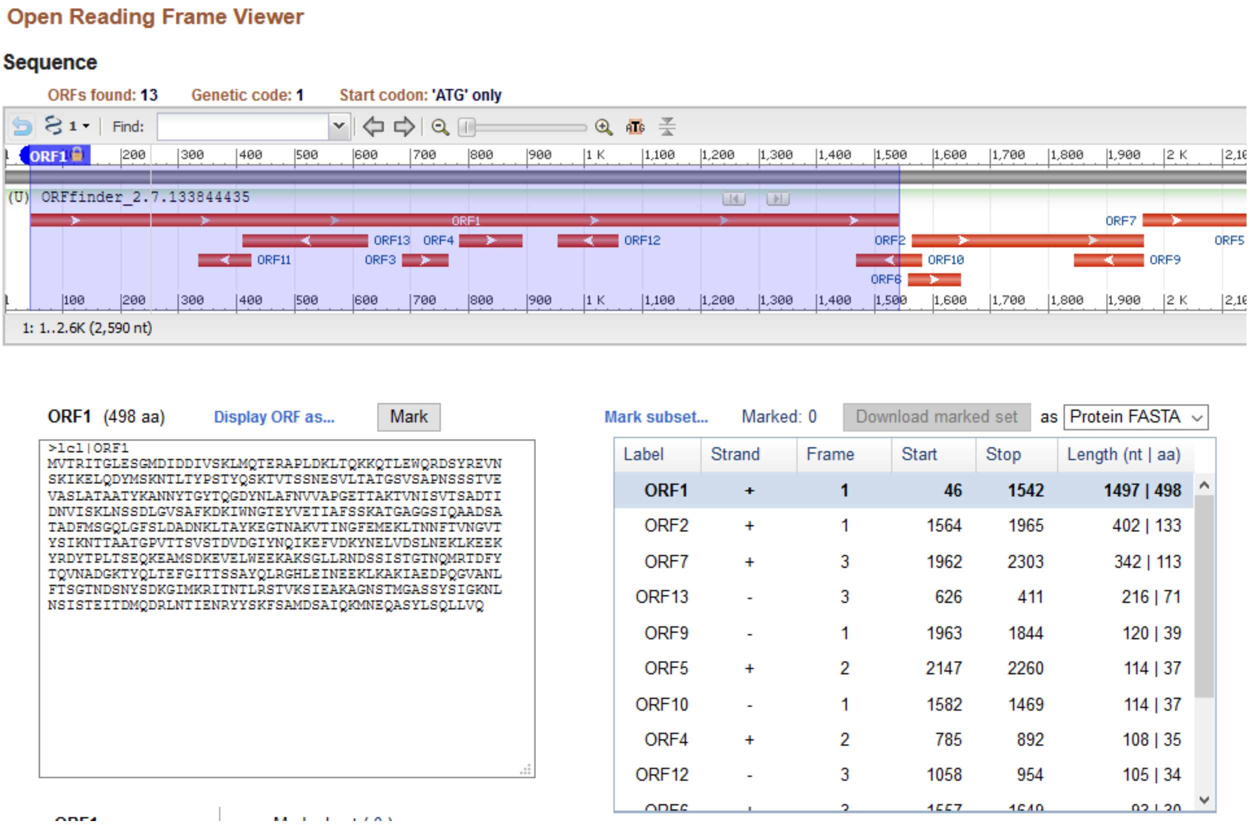
Alexander Gilson
AS.410.635.81
Unit 1-2 Graded Assignment

1. Use ORF Finder to identify the locations of three coding regions (three longest ORFs) in the Bacillus subtilis genomic sequence (file:homework1.txt). (1 point)
    b. On what reading frames are each of the genes in the Bacillus DNA based on ORF Finder? (answer should be at the master pdf document)
        i. The three longest ORF’s of this sequence that most likely contain genes are located in frame 1 (ORF 1: 46..1542)(ORF 2: 1564..1965) and frame 3 (ORF 7: 1962..2303). All of these sequences are on the plus strand.


Open Reading Frame Viewer

Sequence

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

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

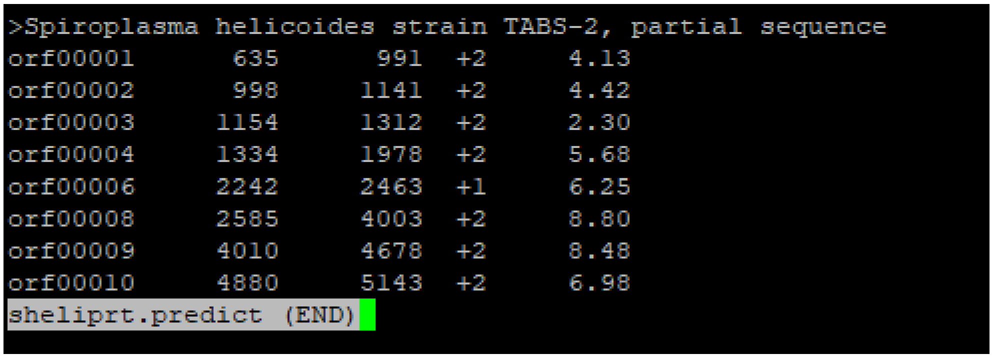
ORF1 (498 aa) Display ORF as... Mark

```
>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

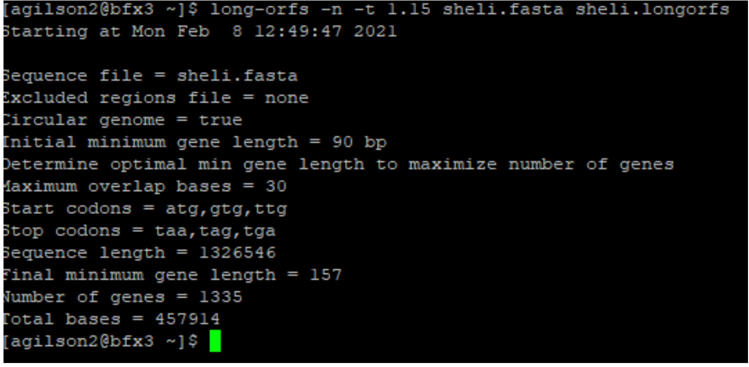
2. Use the command line version of Glimmer to analyze CDSs in a partial sequence from Spiroplasma helicoides strain TABS-2, whose genome was submitted to GenBank on August 23, 2016 (file: sheliprt.fasta). The training set will be the full genome of S. helicoides strain TABS-2 (file: sheli.fasta). (1 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).



```
>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)
```

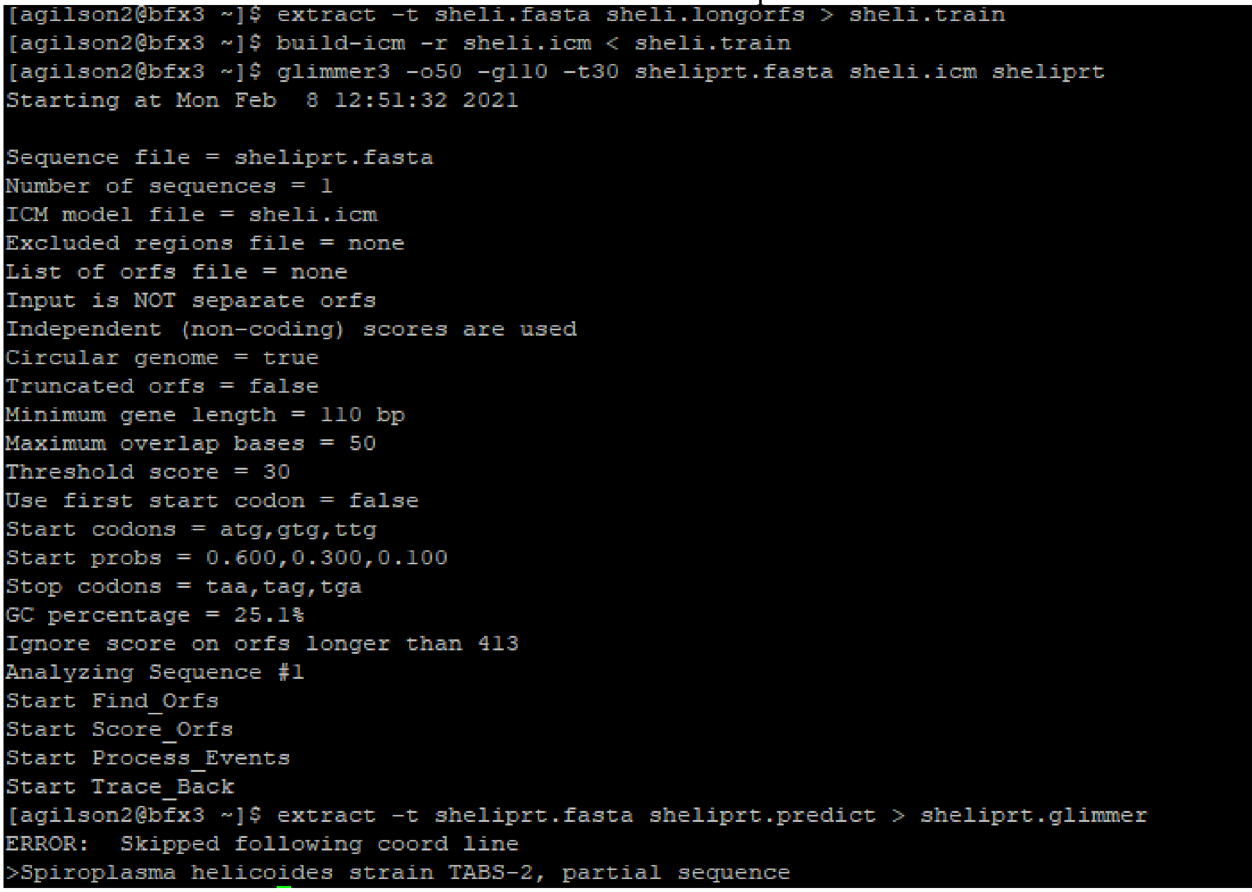
-

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"



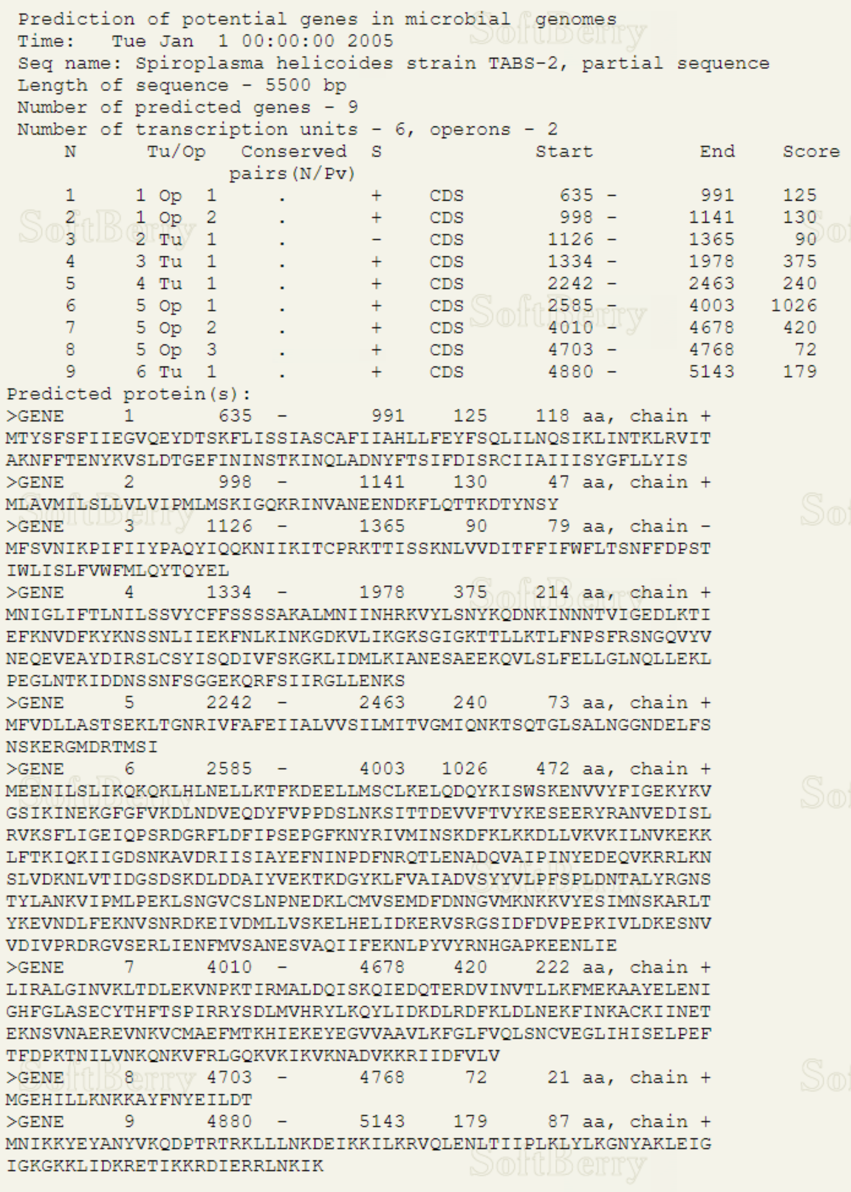
```
[agilson2@bfx3 ~]$ long-orfs -n -t 1.15 sheli.fasta sheli.longorfs
Starting at Mon Feb  8 12:49:47 2021

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
[agilson2@bfx3 ~]$
```

q



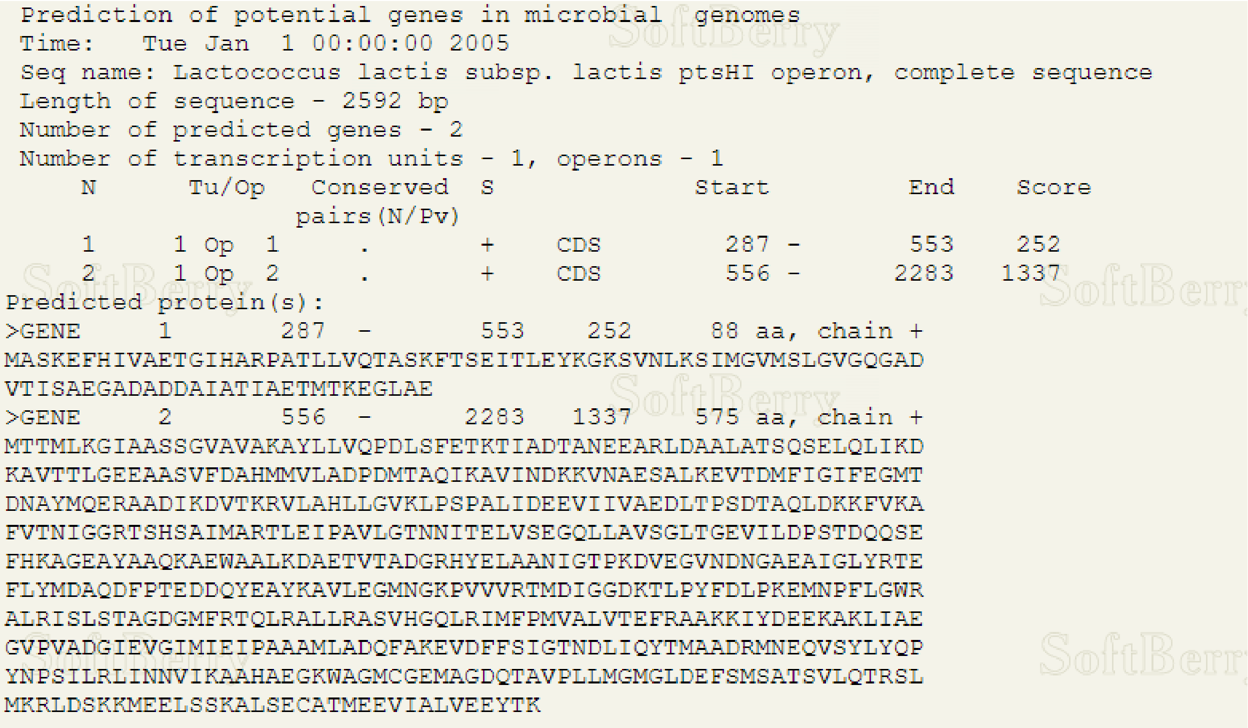
```
[agilson2@bfx3 ~]$ extract -t sheli.fasta sheli.longorfs > sheli.train
[agilson2@bfx3 ~]$ build-icm -r sheli.icm < sheli.train
[agilson2@bfx3 ~]$ glimmer3 -o50 -g110 -t30 sheliprt.fasta sheli.icm sheliprt
Starting at Mon Feb  8 12:51:32 2021

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
[agilson2@bfx3 ~]$ extract -t sheliprt.fasta sheliprt.predict > sheliprt.glimmer
ERROR:  Skipped following coord line
>Spiroplasma helicoides strain TABS-2, partial sequence
```

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)
    a. How many CDSs are listed?
        - There are nine total CDSs listed in the results for this file.
    b. How many mRNAs are predicted to code for those CDSs
        - The program determined a total of 6 “transcription units” found in the sequence.



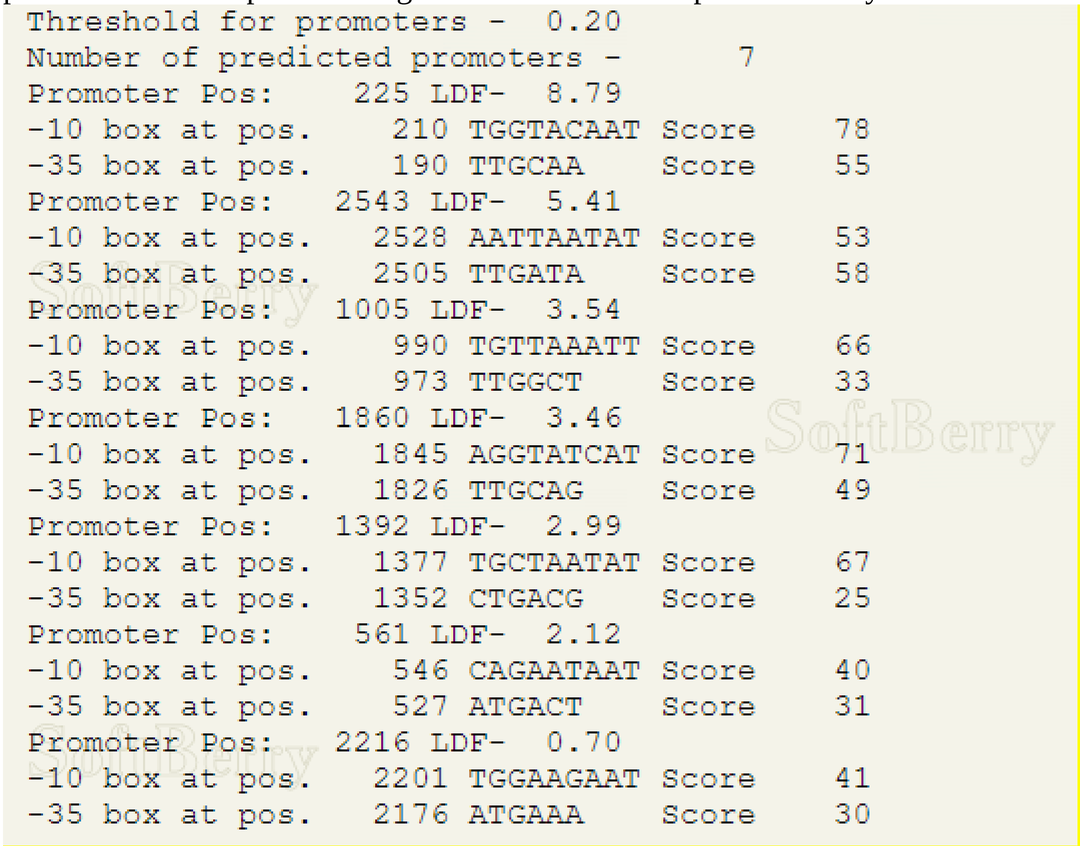
```
 Prediction of potential genes in microbial  genomes
 Time:   Tue Jan  1 00:00:00 2005
 Seq name: Spiroplasma helicoides strain TABS-2, partial sequence
 Length of sequence - 5500 bp
 Number of predicted genes - 9
 Number of transcription units - 6, operons - 2
     N      Tu/Op   Conserved  S            Start         End    Score
                   pairs(N/Pv)
     1     1 Op  1     .       +    CDS        635 -       991    125
     2     1 Op  2     .       +    CDS        998 -      1141    130
     3     2 Tu  1     .       -    CDS       1126 -      1365     90
     4     3 Tu  1     .       +    CDS       1334 -      1978    375
     5     4 Tu  1     .       +    CDS       2242 -      2463    240
     6     5 Op  1     .       +    CDS       2585 -      4003   1026
     7     5 Op  2     .       +    CDS       4010 -      4678    420
     8     5 Op  3     .       +    CDS       4703 -      4768     72
     9     6 Tu  1     .       +    CDS       4880 -      5143    179
Predicted protein(s):
>GENE     1       635  -       991    125    118 aa, chain +
MTYSFSFIIEGVQEYDTSKFLISSIASCAFIIAHLLFEYFSQLILNQSIKLINTKLRVIT
AKNFFTENYKVSLDTGEFININSTKINQLADNYFTSIFDISRCIIAIIISYGFLLYIS
>GENE     2       998  -      1141    130     47 aa, chain +
MLAVMILSLLVLVIPMLMSKIGQKRINVANEENDKFLQTTKDTYNSY
>GENE     3      1126  -      1365     90     79 aa, chain -
MFSVNIKPIFIIYPAQYIQQKNIIKITCPRKTTISSKNLVVDITFFIFWFLTSNFFDPST
IWLISLFVWFMLQYTQYEL
>GENE     4      1334  -      1978    375    214 aa, chain +
MNIGLIFTLNILSSVYCFFSSSSAKALMNIINHRKVYLSNYKQDNKINNNTVIGEDLKTI
EFKNVDFKYKNSSNLIIEKFNLKINKGDKVLIKGKSGIGKTTLLKTLFNPSFRSNGQVYV
NEQEVEAYDIRSLCSYISQDIVFSKGKLIDMLKIANESAEEKQVLSLFELLGLNQLLEKL
PEGLNTKIDDNSSNFSGGEKQRFSIIRGLLENKS
>GENE     5      2242  -      2463    240     73 aa, chain +
MFVDLLASTSEKLTGNRIVFAFEIIALVVSILMITVGMIQNKTSQTGLSALNGGNDELFS
NSKERGMDRTMSI
>GENE     6      2585  -      4003   1026    472 aa, chain +
MEENILSLIKQKQKLHLNELLKTFKDEELLMSCLKELQDQYKISWSKENVVYFIGEKYKV
GSIKINEKGFGFVKDLNDVEQDYFVPPDSLNKSITTDEVVFTVYKESEERYRANVEDISL
RVKSFLIGEIQPSRDGRFLDFIPSEPGFKNYRIVMINSKDFKLKKDLLVKVKILNVKEKK
LFTKIQKIIGDSNKAVDRIISIAYEFNINPDFNRQTLENADQVAIPINYEDEQVKRRLKN
SLVDKNLVTIDGSDSKDLDDAIYVEKTKDGYKLFVAIADVSYYVLPFSPLDNTALYRGNS
TYLANKVIPMLPEKLSNGVCSLNPNEDKLCMVSEMDFDNNGVMKNKKVYESIMNSKARLT
YKEVNDLFEKNVSNRDKEIVDMLLVSKELHELIDKERVSRGSIDFDVPEPKIVLDKESNV
VDIVPRDRGVSERLIENFMVSANESVAQIIFEKNLPYVVYRNHGAPKEENLIE
>GENE     7      4010  -      4678    420    222 aa, chain +
LIRALGINVKLTDLEKVNPKTIRMALDQISKQIEDQTERDVINVTLLKFMEKAAYELENI
GHFGLASECYTHFTSPIRRYSDLMVHRYLKQYLIDKDLRDFKLDLNEKFINKACKIINET
EKNSVNAEREVNKVCMAEFMTKHIEKEYEGVVAAVLKFGLFVQLSNCVEGLIHISELPEF
TFDPKTNILVNKQNKVFRLGQKVKIKVKNADVKKRIIDFVLV
>GENE     8      4703  -      4768     72     21 aa, chain +
MGEHILLKNKKAYFNYEILDT
>GENE     9      4880  -      5143    179     87 aa, chain +
MNIKKYEYANYVKQDPTRTRKLLLNKDEIKKILKRVQLENLTIIPLKLYLKGNYAKLEIG
IGKGKKLIDKRETIKKRDIERRLNKIK
```

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.



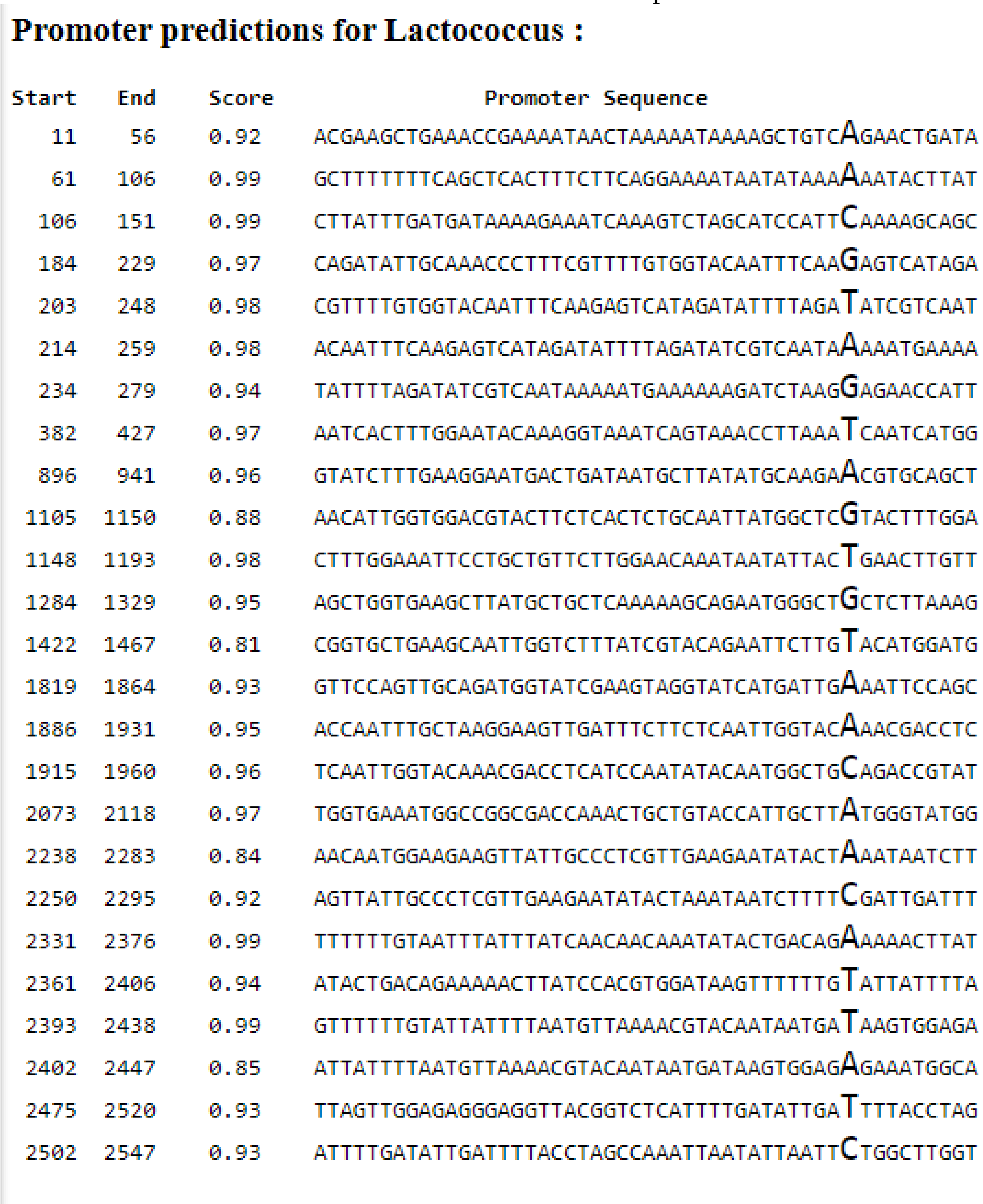
```
 Prediction of potential genes in microbial  genomes
 Time:   Tue Jan  1 00:00:00 2005
 Seq name: Lactococcus lactis subsp. lactis ptsHI operon, complete sequence
 Length of sequence - 2592 bp
 Number of predicted genes - 2
 Number of transcription units - 1, operons - 1
     N      Tu/Op   Conserved  S             Start         End    Score
                 pairs(N/Pv)
     1     1 Op  1      .       +    CDS        287 -       553    252
     2     1 Op  2      .       +    CDS        556 -      2283   1337
Predicted protein(s):
>GENE     1        287  -      553    252     88 aa, chain +
MASKEFHIVAETGIHARPATLLVQTASKFTSEITLEYKGKSVNLKSIMGVMSLGVGQGAD
VTISAEGADADDAIATIAETMTKEGLAE
>GENE     2        556  -     2283   1337    575 aa, chain +
MTTMLKGIAASSGVAVAKAYLLVQPDLSFETKTIADTANEEARLDAALATSQSELQLIKD
KAVTTLGEEAASVFDAHMMVLADPDMTAQIKAVINDKKVNAESALKEVTDMFIGIFEGMT
DNAYMQERAADIKDVTKRVLAHLLGVKLPSPALIDEEVIIVAEDLTPSDTAQLDKKFVKA
FVTNIGGRTSHSAIMARTLEIPAVLGTNNITELVSEGQLLAVSGLTGEVILDPSTDQQSE
FHKAGEAYAAQKAEWAALKDAETVTADGRHYELAANIGTPKDVEGVNDNGAEAIGLYRTE
FLYMDAQDFPTEDDQYEAYKAVLEGMNGKPVVVRTMDIGGDKTLPYFDLPKEMNPFLGWR
ALRISLSTAGDGMFRTQLRALLRASVHGQLRIMFPMVALVTEFRAAKKIYDEEKAKLIAE
GVPVADGIEVGIMIEIPAAAMLADQFAKEVDFFSIGTNDLIQYTMAADRMNEQVSYLYQP
YNPSILRLINNVIKAAHAEGKWAGMCGEMAGDQTAVPLLMGMGLDEFSMSATSVLQTRSL
MKRLDSKKMEELSSKALSECATMEEVIALVEEYTK
```

The genes predicted from the sequence are located at 287..553 & 556..2283, with the BPROM predictions for the promoter regions and the -10 / -35 postions likely to be found at



```
 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
```

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?



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
```

The most likely promoter that matches to the BPROM results found in the previous part of this question would be the prediction of Start end 214..259 with a score of 0.98, with the starting nucleotide being “A”.

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)
    a. A eukaryotic start site is more difficult to predict than that of a prokaryotic start site as eukaryotic gene promotor and regulatory regions are more complex than that of prokaryotes. A prokaryotic promoter consists of the locations of -10 & -35 before the TSS. A eukaryotic gene on the other hand can have a promoter region directly upstream of a gene, but have regulatory elements from multiple other locations within the entire genome, sometimes thousands of bases away. This complicates the ability to predict eukaryotic genes without first having RNA-seq or cDNA data as predicting the location of TSSs that have these longer distance regulatory elements becomes quite the hassle.