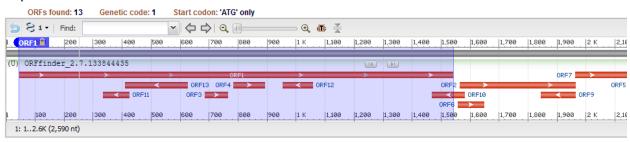
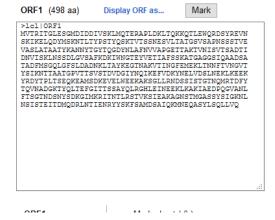
- 1. Use ORF Finder to identify the locations of three coding regions (three longest ORFs) in the Bacillus subtilis genomic sequence (file:homework1.txt). (1 point)
  - b. On what reading frames are each of the genes in the Bacillus DNA based on ORF Finder? (answer should be at the master pdf document)
    - 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





Mark subset Marked: 0				nload marke	ed 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	
	ODEC		2	1557	1610	03   30	٧

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

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
otal bases = 457914
agilson2@bfx3 ~]$
[agilson2@bfx3 ~]$ extract -t sheli.fasta sheli.longorfs > sheli.train
[agilson2@bfx3 ~]$ build-icm -r sheli.icm < sheli.train
[agilson2@bfx3 ~]$ glimmer3 -o50 -gll0 -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.

B		beque	iicc.				
Time: Tue Jan				oial geno			
Seq name: Spiro Length of seque			s strain	n TABS-2,	partial :	sequence	
Number of predi							
Number of trans			6. ope:	rons - 2			
	p Con	served S s(N/Pv)		Star	:t	End	Score
1 1 Op	1		CDS	63	35 -	991	125
			CDS		98 -	1141	130
$S_{3}^{2}$ $B_{2}^{1}$ $C_{Tu}$	1		CDS		26 -	1365	90
4 3 Tu		 . +			34 -	1978	375
					12 -		240
6 5 Op	1		CDS	C C 229	12 - 15		1026
-		. +		Solution	in Try	4678	420
<u>-</u> -	2		CDS CDS	401	0 - 0		72
					30 -	4768	
	_	. +	CDS	488	30 -	5143	179
Predicted protei >GENE 1		- 9	91 1	25 118	aa, chair	n +	
MTYSFSFIIEGVQEYD	TSKFLIS	SIASCAFII	AHLLFEY	FSQLILNQSI	KLINTKLR	TIV	
AKNFFTENYKVSLDTG	EFININS	TKINQLADN	YFTSIFD:	ISRCIIAIII	SYGFLLYI	3	
>GENE 2	998	- 11	.41 1:	30 47	aa, chair	n +	
MLAVMILSLLVLVIPM	LMSKIGQ	KRINVANEE	NDKFLQT'	TKDTYNSY			
>GENE 3	1126	- 13	65	90 79	aa, chair	n -	
MFSVNIKPIFIIYPAQ	YIQQKNI	IKITCPRKT	TISSKNL	VVDITFFIFW	FLTSNFFD	PST	
IWLISLFVWFMLQYTQ	YEL						
>GENE 4	1334	- 19	78 3	75 214	aa, chair	n +	
MNIGLIFTLNILSSVY	CFFSSSS	AKALMNIIN	HRKVYLSI	NYKQDNKINN			
MNIGLIFTLNILSSVY EFKNVDFKYKNSSNLI				_	NTVIGEDL	KTI	
	IEKFNLK	INKGDK <b>V</b> LI	KGKSGIG	KTTLLKTLFN	NTVIGEDLI IPSFRSNGQ	VY <b>V</b>	
EFKNVDFKYKNSSNLI	IEKFNLK ISQDI <b>V</b> F	INKGDK <b>V</b> LI SKGKLIDML	KGKSGIG KIANESA	KTTLLKTLFN	NTVIGEDLI IPSFRSNGQ	VY <b>V</b>	
EFKNVDFKYKNSSNLI NEQEVEAYDIRSLCSY PEGLNTKIDDNSSNFS	IEKFNLK ISQDI <b>V</b> F GGEKQRF	INKGDK <b>V</b> LI SKGKLIDML SIIRGLLEN	KGKSGIG KIANESAI KS	KTTLLKTLFN EEKQVLSLFE	INTVIGEDLI IPSFRSNGQ LLGLNQLLI	KTI VY <b>V</b> EKL	
EFKNVDFKYKNSSNLI NEQEVEAYDIRSLCSY PEGLNTKIDDNSSNFS >GENE 5	IEKFNLK ISQDIVF GGEKQRF 2242	INKGDK <b>V</b> LI SKGKLIDML SIIRGLLEN - 24	KGKSGIGI KIANESAI KS 63 2	KTTLLKTLFN EEKQVLSLFE 40 73	NTVIGEDLI IPSFRSNGQ LLGLNQLLI aa, chair	KTI VYV EKL n +	
EFKNVDFKYKNSSNLI NEQEVEAYDIRSLCSY PEGLNTKIDDNSSNFS >GENE 5 MFVDLLASTSEKLTGN	IEKFNLK ISQDIVF GGEKQRF 2242	INKGDK <b>V</b> LI SKGKLIDML SIIRGLLEN - 24	KGKSGIGI KIANESAI KS 63 2	KTTLLKTLFN EEKQVLSLFE 40 73	NTVIGEDLI IPSFRSNGQ LLGLNQLLI aa, chair	KTI VYV EKL n +	
EFKNVDFKYKNSSNLI NEQEVEAYDIRSLCSY PEGLNTKIDDNSSNFS >GENE 5 MFVDLLASTSEKLTGN NSKERGMDRTMSI	IEKFNLK ISQDIVF GGEKQRF 2242 RIVFAFE	INKGDKVLI SKGKLIDML SIIRGLLEN - 24 IIALVVSIL	KGKSGIGI KIANESAI KS 63 2 MITVGMI	KTTLLKTLFN EEKQVLSLFE 40 73 QNKTSQTGLS	INTVIGEDLI IPSFRSNGQ ELLGLNQLLI aa, chair SALNGGNDE	KTI VYV EKL n + LFS	
EFKNVDFKYKNSSNLI NEQEVEAYDIRSLCSY PEGLNTKIDDNSSNFS >GENE 5 MFVDLLASTSEKLTGN NSKERGMDRTMSI >GENE 6	IEKFNLK ISQDIVF GGEKQRF 2242 RIVFAFE 2585	INKGDKVLI SKGKLIDML SIIRGLLEN - 24 IIALVVSIL	KGKSGIGI KIANESAI KS 63 2 MITVGMI	KTTLLKTLFN EEKQVLSLFE 40 73 QNKTSQTGLS 26 472	INTVIGEDLI IPSFRSNGQ LLGLNQLLI aa, chair ALNGGNDE aa, chair	KTI VYV EKL n + LFS n +	
EFKNVDFKYKNSSNLI NEQEVEAYDIRSLCSY PEGLNTKIDDNSSNFS >GENE 5 MFVDLLASTSEKLTGN NSKERGMDRTMSI >GENE 6 MEENILSLIKQKQKLH	IEKFNLK ISQDIVF GGEKQRF 2242 RIVFAFE 2585 LNELLKT	INKGDKVLI SKGKLIDML SIIRGLLEN - 24 IIALVVSIL - 40 FKDEELLMS	KGKSGIGI KIANESAI KS 63 2 MITVGMI 03 10	KTTLLKTLFN EEKQVLSLFE 40 73 QNKTSQTGLS 26 472 QYKISWSKEN	INTVIGEDLI IPSFRSNGQ LLGLNQLLI aa, chair SALNGGNDE aa, chair aa, chair	KTI VYV EKL n + LFS n + YKV	
EFKNVDFKYKNSSNLI NEQEVEAYDIRSLCSY PEGLNTKIDDNSSNFS >GENE 5 MFVDLLASTSEKLTGN NSKERGMDRTMSI >GENE 6 MEENILSLIKQKQKLH GSIKINEKGFGFVKDL	IEKFNLK ISQDIVF GGEKQRF 2242 RIVFAFE 2585 LNELLKT NDVEQDY	INKGDKVLI SKGKLIDML SIIRGLLEN - 24 IIALVVSIL - 40 FKDEELLMS	KGKSGIG! KIANESA! KS 63 2: MITVGMI 03 10: CLKELQD( SITTDEV	KTTLLKTLFN EEKQVLSLFE 40 73 QNKTSQTGLS 26 472 QYKISWSKEN VFTVYKESEE	INTVIGEDLE IPSFRSNGQ LLGLNQLLE aa, chair SALNGGNDE aa, chair IVVYFIGEK RYRANVED	KTI VYV EKL n + LFS n + YKV ISL	
EFKNVDFKYKNSSNLI NEQEVEAYDIRSLCSY PEGLNTKIDDNSSNFS >GENE 5 MFVDLLASTSEKLTGN NSKERGMDRTMSI >GENE 6 MEENILSLIKQKQKLH GSIKINEKGFGFVKDL RVKSFLIGEIQPSRDG	IEKFNLK ISQDIVF GGEKQRF 2242 RIVFAFE 2585 LNELLKT NDVEQDY RFLDFIP	INKGDKVLI SKGKLIDML SIIRGLLEN - 24 IIALVVSIL - 40 FKDEELLMS FVPPDSLNK SEPGFKNYR	KGKSGIGI KIANESAI KS 63 24 MITVGMI 03 10: CLKELQDO SITTDEV IVMINSKI	KTTLLKTLFN EEKQVLSLFE 40 73 QNKTSQTGLS 26 472 QYKISWSKEN VFTVYKESEE	INTVIGEDLI IPSFRSNGQI LLGLNQLLI aa, chaii BALNGGNDE: aa, chaii IVVYFIGEK RYRANVED: VKVKILNVKI	KTI VYV EKL n + LFS n + YKV ISL EKK	
EFKNVDFKYKNSSNLI NEQEVEAYDIRSLCSY PEGLNTKIDDNSSNFS >GENE 5 MFVDLLASTSEKLTGN NSKERGMDRTMSI >GENE 6 MEENILSLIKQKQKLH GSIKINEKGFGFVKDL RVKSFLIGEIQPSRDG LFTKIQKIIGDSNKAV	IEKFNLK ISQDIVF GGEKQRF 2242 RIVFAFE 2585 LNELLKT NDVEQDY RFLDFIP DRIISIA	INKGDKVLI SKGKLIDML SIIRGLLEN - 24 IIALVVSIL - 40 FKDEELLMS FVPPDSLNK SEPGFKNYR YEFNINPDF	KGKSGIGI KIANESAI KS 63 2 MITVGMI 03 10: CLKELQDO SITTDEV IVMINSKI TNRQTLENI	KTTLLKTLFN EEKQVLSLFE 40 73 QNKTSQTGLS 26 472 QYKISWSKEN WFTVYKESEE DFKLKKDLLV ADQVAIPINY	INTVIGEDLI IPSFRSNGQ' LLGLNQLLI aa, chair ALNGGNDE: aa, chair IVVYFIGEK; IVVYFIGEK; IVVKLNVK; IVKVKLNVK; IEDEQVKRR;	KTI VYV EKL n + LFS n + YKV ISL EKK	
EFKNVDFKYKNSSNLI NEQEVEAYDIRSLCSY PEGLNTKIDDNSSNFS >GENE 5 MFVDLLASTSEKLTGN NSKERGMDRTMSI >GENE 6 MEENILSLIKQKQKLH GSIKINEKGFGFVKDL RVKSFLIGEIQPSRDG LFTKIQKIIGDSNKAV SLVDKNLVTIDGSDSK	IEKFNLK ISQDIVF GGEKQRF 2242 RIVFAFE 2585 LNELLKT NDVEQDY RFLDFIP DRIISIA DLDDAIY	INKGDKVLI SKGKLIDML SIIRGLLEN - 24 IIALVVSIL - 40 FKDEELLMS FVPPDSLNK SEPGFKNYR YEFNINPDF	KGKSGIGI KIANESAI KS 63 2 MITVGMI 03 10: CLKELQD CLKELQD SITTDEV IVMINSKI NRQTLEN	KTTLLKTLFN EEKQVLSLFE 40 73 QNKTSQTGLS 26 472 QYKISWSKEN VFTVKESEF VFTVKESLLV ANQVALPINY VSYYVLPFSE	INTVIGEDLI IPSFRSNGQ' LLGLNQLLI aa, chain ALNGGNDE: aa, chain IVVYFIGEK: EXYRANVED: EXYKALNVEN EDEQVKRE: EDEQVKRE: EDEQVKRE: EDEQVKRE:	KTI VYV EKL  1 + LFS  1 + YKV ISL EKK LKN ENK	
EFKNVDFKYKNSSNLI NEQEVEAYDIRSLCSY PEGLNTKIDDNSSNFS >GENE 5 MFVDLLASTSEKLTGN NSKERGMDRTMSI >GENE 6 MEENILSLIKQKQKLH GSIKINEKGFGFVKDL RVKSFLIGEIQPSRDA LFTKIQKIIGDSNKAV SLVDKNLVTIDGSDSK TYLANKVIPMLPEKLS	IEKFNLK ISQDIVF GGEKQRF 2242 RIVFAFE 2585 LNELLKT NDVEQDY RFLDFIP DRIISIA DLDDAIY NGVCSLN	INKGDKVLI SKGKLIDML SIIRGLLEN - 24 IIALVVSIL - 40 FKDEELLMS FVPPDSLNK SEPGFKNYR YEFNINPDF VEKTKDGYK PNEDKLCMV	KGKSGIGI KIANESAI KS 63 2: MITVGMI 03 10: CLKELQDE SITTDEVI SITTDEVI IVMINSKI IVMINSKI IVMINSKI IVMISKI IVMISKI SEMDFDNI	KTTLLKTLFN EEKQVLSLFE 40 73 QNKTSQTGLS 26 472 QYKISWSKEN WFTVYKESEE DFKLKKDLLV ADQVALPIN VSYYVLPFSE	INTVIGEDLI IPSFRSNGQ¹ ILIGLNQLLI aa, chain SALNGGNDE: aa, chain IVVYFIGEK: IRYRANVED: IRYRANVED: IEDEQVKRR: IEDEQVKRR: ILIDNTALYRG IESIMNSKAI	KTI VYV EKL  1 + LFS 1 + YKV ISL EKK LKN GNS RLT	
EFKNVDFKYKNSSNLI NEQEVEAYDIRSLCSY PEGLNTKIDDNSSNFS >GENE 5 MFVDLLASTSEKLTGN NSKERGMDRTMSI >GENE 6 MEENILSLIKQKQKLH GSIKINEKGFGFVKDL RVKSFLIGEIQPSRDG LFTKIQKIIGDSNKAV SLVDKNLVTIDGSDSK TYLANKVIPMLPEKLS YKEVNDLFEKNVSNRD	IEKFNLK ISQDIVF GGEKQRF 2242 RIVFAFE 2585 LNELLKT NDVEQDY RFLDFIP DRIISIA DLDDAIY NGVCSLN KEIVDML	INKGDKVLI SKGKLIDML SIIRGLLEN - 24 IIALVVSIL - 40 FKDEELLMS FVPPDSLNK SEPGFKNYR VEFNINPDF VEKTKDGYK PNEDKLCMV LVSKELHEL	KGKSGIGI KIANESAI KS 63 20 MITVGMI  03 100 CLKELQDO SITTDEVI IVMINSKI NRQTLENI KLEVAIADO SEMDFDNI IDKERVSI	KTTLLKTLFN EEKQVLSLFE 40 73 QNKTSQTGLS 26 472 QYKISWSKEN WFTWYKESEE DFKLKKDLLV ADQVALPFSE NGVMKNKKYY RGSIDFDVPE	INTVIGEDLI IPSFRSNGQ' LLGLNQLLI aa, chain SALNGGNDE aa, chain IVVYFIGEK: IRYRANVED IVKVKLINVKI IEDEQVKRRI LEDUTALYR IESIMNSKAI IPKIVLDKE	KTI VYV EKL  1 + LFS 1 + YKV ISL EKK LKN GNS RLT	
EFKNVDFKYKNSSNLI NEQEVEAYDIRSLCSY PEGLNTKIDDNSSNFS >GENE 5 MFVDLLASTSEKLTGN NSKERGMDRTMSI >GENE 6 MEENILSLIKQKQKLH GSIKINEKGFGFVKDL RVKSFLIGBIQPSRDG LFTKIQKIIGDSNKAV SLVDKNLVTIDGSDSK TYLANKVIPMLPEKLS YKEVNDLFEKNVSNRD VDIVPRDRGVSERLIE	IEKFNLK ISQDIVF GGEKQRF 2242 RIVFAFE 2585 LNELLKT NDVEQDY RFLDFIP DRIISIA DLDDAIY NGVCSLN KEIVDML	INKGDKVLI SKGKLIDML SIIRGLLEN - 24 IIALVVSIL - 40 FKDEELLMS FVPPDSLNK SEPGFKNYR YEFNINPDF VEKTKDGYK PNEDKLCMV LVSKELHEL ESVAQIIFE	KGKSGIGI KIANESAI KS 63 20 MITVGMI 03 100 CLKELQDO SITTDEV IVMINSKI NRQTLEN ILFVAIAD SEMDFDNI IDKERVSI KKLPYVY	KTTLLKTLFN EEKQVLSLFE 40 73 QNKTSQTGLS 26 472 QYKISWSKEN VFTVYKESEE DFKLKKDLLA ADQVALPINY VSYYVLPFSE NGYMKNKKY RGSIDFDVPE RNHGAPKEEN	INTVIGEDLI IPSFRSNGO ILIGLNQLLI aa, chaii BALNGGNDE: aa, chaii IVVYFIGEK: IRYRANVED: IVVKILNVKI IEDEQVKRR: ILDNTALYR( IESIMNSKAI IPKIVLDKE: ILIE	KTI VYV EKL n + LFS n + YKV ISL EKK LKN GNS RLT	
EFKNVDFKYKNSSNLI NEQEVEAYDIRSLCSY PEGLNTKIDDNSSNFS >GENE 5 MFVDLLASTSEKLTGN NSKERGMDRTMSI >GENE 6 MEENILSLIKQKQKLH GSIKINEKGFGFVKDL RVKSFLIGEIQPSRDG LFTKIQKIIGDSNKAV SLVDKNLVTIDGSDSK TYLANKVIPMLPEKLS YKEVNDLFEKNVSNRD VDIVPRDRGVSERLIE >GENE 7	IEKFNLK ISQDIVF GGEKQRF 2242 RIVFAFE 2585 LNELLKT NDVEQDY RFLDFIP DRIISIA DLDDAIY NGVCSLN KEIVDML NFWSAN 4010	INKGDKVLI SKGKLIDML SIIRGLLEN SIIRGLLEN IIALVVSIL - 40 FKDEELLMS FVPPDSLNK SEPGFKNYR YEFNINPDF VEKTKDGYK PNEDKLCMV LVSKELHEL LVSKELHEL 46	KGKSGIGI KIANESAI KS 63 2- MITVGMI 03 10: CLKELQDE SITTDEV' IVMINSKI NRQTLENI LFVAIAD SEMDFDNI JUDGERVSI KNLPYVI 178 4:	KTTLLKTLFN EEKQVLSLFE 40 73 QNKTSQTGLS 26 472 QYKISWSKEN VFTVYKESEF DFKLKKDLLV ADQVAIPINY VSYYVLPFSE NCVMKNKKVY RGSIDFDVPE RNHGAPKEEN 20 222	INTVIGEDLI IPSFRSNGQ¹ LLGLNQLLI aa, chain ALNGGNDE: aa, chain IVVYFIGEK: ERYRANVED: IVVKILNVKI IEDEQVKRRI ILDNTALYRC IESIMNSKAI EPKIVLDKE: LLIE aa, chain	KTI VYV EKL  n + LFS n + YKV ISL EKK LKN ENS RLT SNV	
EFKNVDFKYKNSSNLI NEQEVEAYDIRSLCSY PEGLNTKIDDNSSNFS >GENE 5 MFVDLLASTSEKLTGN NSKERGMDRTMSI >GENE 6 MEENILSLIKQKQKLH GSIKINEKFGFFVKDL RVKSFLIGEIQPSRDG LFTKIQKIIGDSNKAV SLVDKNLVTIDGSDSK TYLANKVIPMLPEKLS YKEVNLFEKNVSNRD VDIVPRDRGVSERLIE >GENE 7 LIRALGINVKLTDLEK	IEKFNLK ISQDIVF GGERQRF 2242 RIVFAFE 2585 LNELLKT NDVEQDY RFLDFIP DRIISIA DLDDAIY NGVCSLN KEIVDML NFMVSAN 4010 VNPKTIR	INKGDKVLI SKGKLIDML SIIRGLLEN - 24 IIALVVSIL - 40 FKDEELLMS FFVPPDSLNK SEPGFKNYR YEFNINPDF VEKTKDGYK PNEDKLCMV LVSKELHEL LVSKELHEL - 46 MALDQISKQ	KGKSGIGI KIANESAI KS 63 24 MITVGMI 03 10: CLKELQD CSITTDEV ILTVMINSK INRQTLEN ILFVAIAD SEMDFDNI IDKERVS INNLYYVI 178 4: EIEDQTERI	KTTLLKTLFN EEKQVLSLFE 40 73 ONKTSOTGLS 26 472 OYKISWSKEN VFTVYKESEE ADQVAIPINY VSYYVLPFSE NGVMKNKKVY RSYSTEPUPE RNHGAPKEEN 20 222 DVINVTLLKE	INTVIGEDLI IPSFRSNGQ' LLGLNQLLI aa, chair ALNGGNDE: EXPRANVED: EXP	KTI VYV EKL  n + LFS n + YKV YKV JISL EKK LKN GNS RLT SNV n +	
EFKNVDFKYKNSSNLI NEQEVEAYDIRSLCSY PEGLNTKIDDNSSNFS >GENE 5 MFVDLLASTSEKLTGN NSKERGMDRTMSI >GENE 6 MEENILSLIKQKQKLH GSIKINEKGFGFVKDL RVKSFLIGEIQPSRDG LFTKIQKIIGDSNKAV SLVDKNLVTIDGSDSK TYLANKVIPMLPEKLS YKEVNDLFEKNVSNRD VDIVPRDRGVSERLIE >GENE 7 LIRALGINVKLTDLEK GHFGLASECYTHFTSP	IEKFNLK ISQDIVF GGERQRF 2242 RIVFAFE 2585 LNELLKT NDVEQDY RFLDFIP DRIISIA DLDDAIY NGVCSLN KEIVDML NFMVSAN 4010 VNPKTIR IRRYSDL	INKGDKVLI SKGKLIDML SIIRGLLEN - 24 IIALVVSIL - 40 FKDEELLMS FVPPDSLNK SEPGFKNYR YEFNINPDF VEKTKDGYK PNEDKLCMV LVSKELHEL LESVAQIIFE - 46 MALDQISKQ MVHRYLKQY	KGKSGIGI KIANESAI KS 63 2: MITVGMI 03 10: CLKELQD SITTDEV LIVMINSKI NRQTLEN LIFVAIAD SEMDFDNI LIDKERVSI KNLPYVYI 178 4: LIEDQTERI LIDKDLRI	KTTLLKTLFN EEKQVLSLFE 40 73 QNKTSQTGLS 26 472 QYKISWSKEN PTVYKESEE DFKLKKDLLV ADQVAIPIN VSYYVLPFSE NGVMKNKKVY RGSIDFDVPE RNHGAPKEEN 20 222 DVINVTLLKE	INTVIGEDLI IPSFRSNGO' LLGLNQLLI aa, chain ALNGGNDE: aa, chain IVVYFIGEK: IVVYFIGEK: IVVKILNVKI IEDEQVKRR: IVLDNTALYRO IESIMNSKAI IPKIVLDKE: ILIE aa, chain IMEKAAYELI INKACKIII	KTI VYV EKL  1 + LFS  1 + YKV ISL EKK LKN ENS RLT SNV  1 + ENI	
EFKNVDFKYKNSSNLI NEQEVEAYDIRSLCSY PEGLNTKIDDNSSNFS >GENE 5 MFVDLLASTSEKLTGN NSKERGMDRTMSI >GENE 6 MEENILSLIKQKQKLH GSIKINEKGFGFVKDL RVKSFLIGEIQPSRDA LFTKIQKIIGDSNKAV SLVDKNLVTIDGSDSK TYLANKVIPMLPEKLS YKEVNDLFEKNVSNRD VDIVPRDRGVSERLIE >GENE 7 LIRALGINVKLTDLEK GHFGLASECYTHFTSP EKNSVNAEREVNKVCM	IEKFNLK ISQDIVF GGERQRF 2242 RIVFAFE 2585 LNELLKT NDVEQDY RFLDFIP DRIISIA DLDDAIY NGVCSLN KEIVDML NFMYSAN 4010 VNPKTIR IRRYSDL AEFMTKH	INKGDKVLI SKGKLIDML SIIRGLLEN - 24 IIALVVSIL - 40 FKDEELLMS FVPPDSLNK SEPGFKNYR YEFNINPDF VEKTKDGYK PNEDKLCMV LVSKELHEL ESVAQIIFE - 46 MALDQISKQ MVHRYLKQY IEKEYEGVV	KGKSGIGI KIANESAI KS 63 2: MITVGMI  03 10: CLKELQDE SITTDEVI SITTD	KTTLLKTLFN EEKQVLSLFE 40 73 QNKTSQTGLS 26 472 QYKISWSKEN WFTVYKESEE DFKLKKDLLV VSYYVLFFSE NGVMKNKKVY RGSIDFDVPE RNHGAPKEEN 20 20 QULVTLLKE DFKLDLNEKKLFVQLSNCVE	INTVIGEDLI IPSFRSNGO' LLGLNQLLI aa, chain ALNGGNDE: aa, chain IVVYFIGEK: IVVYFIGEK: IVVKILNVKI IEDEQVKRR: IVLDNTALYRO IESIMNSKAI IPKIVLDKE: ILIE aa, chain IMEKAAYELI INKACKIII	KTI VYV EKL  1 + LFS  1 + YKV ISL EKK LKN ENS RLT SNV  1 + ENI	
EFKNVDFKYKNSSNLI NEQEVEAYDIRSLCSY PEGLNTKIDDNSSNFS >GENE 5 MFVDLLASTSEKLTGN NSKERGMDRTMSI >GENE 6 MEENILSLIKQKQKLH GSIKINEKGFGFVKDL RVKSFLIGEIQPSRDG LFTKIQKIIGDSNKAV SLVDKNLVTIDGSDSK TYLANKVIPMLPEKLS VKEVNDLFEKNVSNRD VDIVPRDRGVSERLIE >GENE 7 LIRALGINVKLTDLEK GHFGLASECYTHFTSP EKNSVNAEREVNKVCM TFDPKTNILVNKQNKV	IEKFNLK ISQDIVF GGEKQRF 2242 RIVFAFE 2585 LNELLKT NDVEQDY RFLDFIP DRIISIA DLDDAIY NFUSLN KEIVDML NFMVSAN 4010 VNPKTIR IRRYSDL AEFMTKH FRLGQKV	INKGDKVLI SKGKLIDML SIIRGLLEN SIIRGLLEN 1IALVVSIL - 40 FKDEELLMS FVPPDSLNK SEPGFKNYR YEFNINPDF VEKTKDGYK PNEDKLCMV LVSKELHEL ESVACIIFE - 46 MALDQISKQ MWHRYLKQY IEKEYEGVV KIKVKNADV	KGKSGIGI KIANESAI KS 63 2- MITVGMI 03 10: CLKELQDE SITTDEV SITTDEV SIVMINSKI NRQTLEN: LIFVAIAD SEMDFDNI SIDKERVSI KNLPYVY 178 4: SIEDQTERI LIDKDLRI LIDKDLRI LAAVLKFG	KTTLLKTLFN EEKQVLSLFE 40 73 2NKTSQTGLS 26 472 22YKISWSKEN VFTVYKESEE DFKLKDLLV ADQVAIPINY VSYYVLPFSE RGYMKNKKVY RGSIDFDVPE RNHGAPKEEN 20 222 DVINVTLLKE DFKLDLNEKE LFVQLSNCVE	INTVIGEDLE IPSFRSNGO ILLGLNQLLE aa, chair BALNGGNDE aa, chair IVVYFIGEK IRYRANVED IVVKILNVKI IEDEQVKRRI IESIMNSKAI IPKIVLDKE ILLE aa, chair IMKAAYELE INKACKIIR ICGLIHISELE	KTI VYV EKL  n + LFS  n + YKV ISL EKK LKN GNS SNV  n + ENI	
EFKNVDFKYKNSSNLI NEQEVEAYDIRSLCSY PEGLNTKIDDNSSNFS >GENE 5 MFVDLLASTSEKLTGN NSKERGMDRTMSI >GENE 6 MEENILSLIKQKQKLH GSIKINEKGFGFVKDL RVKSFLIGEIQPSRDG LFTKIQKIIGDSNKAV SLVDKNLVTIDGSDSK TYLANKVIPMLPEKLS YKEVNDLFEKNVSNRD VDIVPRDRGVSERLIE >GENE 7 LIRALGINVKLTDLEK GHFGLASECYTHFTSP EKNSVNAEREVNKVCM TFDPKTNILVNKQNKV >GENE 8	IEKFNLK ISQDIVF GGERQRF 2242 RIVFAFE 2585 LNELLKT NDVEQDY RFLDFIP DRIISIA DLDDAIY NGVCSLN KEIVDML NFMYSAN 4010 VNPKTIR IRRYSDL AEFMTKH FRLGGKV 4703	INKGDKVLI SKGKLIDML SIIRGLLEN SIIRGLLEN 1IALVVSIL - 40 FKDEELLMS FVPPDSLNK SEPGFKNYR YEFNINPDF VEKTKDGYK PNEDKLCMV LVSKELHEL ESVACIIFE - 46 MALDQISKQ MWHRYLKQY IEKEYEGVV KIKVKNADV	KGKSGIGI KIANESAI KS 63 2: MITVGMI  03 10: CLKELQDE SITTDEVI SITTD	KTTLLKTLFN EEKQVLSLFE 40 73 2NKTSQTGLS 26 472 22YKISWSKEN VFTVYKESEE DFKLKDLLV ADQVAIPINY VSYYVLPFSE RGYMKNKKVY RGSIDFDVPE RNHGAPKEEN 20 222 DVINVTLLKE DFKLDLNEKE LFVQLSNCVE	INTVIGEDLI IPSFRSNGO' LLGLNQLLI aa, chain ALNGGNDE: aa, chain IVVYFIGEK: IVVYFIGEK: IVVKILNVKI IEDEQVKRR: IVLDNTALYRO IESIMNSKAI IPKIVLDKE: ILIE aa, chain IMEKAAYELI INKACKIII	KTI VYV EKL  n + LFS  n + YKV ISL EKK LKN GNS SNV  n + ENI	
EFKNVDFKYKNSSNLI NEQEVEAYDIRSLCSY PEGLNTKIDDNSSNFS >GENE 5 MFVDLLASTSEKLTGN NSKERGMDRTMSI >GENE 6 MEENILSLIKQKQKLH GSIKINEKGFGFVKDL RVKSFLIGEIQPSRDG LFTKIQKIIGDSNKAV SLVDKNLVTIDGSDSK TYLANKVIPMLPEKLS YKEVNDLFEKNVSNRD VDIVPRDRGVSERLIE >GENE 7 LIRALGINVKLTDLEK GHFGLASECYTHFTSP EKNSVNAEREVNKVCM TFDPKTNILVNKQNKV >GENE 8 MGEHILLKNKKAYFNY	IEKFNLK ISQDIVF GGERQRF 2242 RIVFAFE 2585 LNELLKT NDVEQDY RFLDFIP DRIISIA DLDDAIY NGVCSIN KEIVDML NFMVSAN 4010 VNPKTIR IRRYSDL AEFMTKH FRLGQKV 4703 EILDT	INKGDKVLI SKGKLIDML SIIRGLLEN 24 IIALVVSIL - 40 FKDEELLMS FFVPPDSLNK SEPGFKNYR YEFNINPDF VEKTKDGYK PNEDKLCMV LVSKELHEL ESVAQIIFE - 46 MALDQISKQ MVHRYLKQY IEKEYEGVV KIKVKNADV - 47	KGKSGIGI KIANESAI KS 63 2- MITVGMI 03 10: CLKELQD CSITTDEV' ILTVMINSKI NRQTLENI LIFVAIAD SEMDFDNI LIDKERVSI 78 4: LIEDQTERI LIDKDLRI LAVLKFG: KKRILDF 68	KTTLLKTLFN EEKQVLSLFE 40 73 ONKTSOTGLS 26 472 OYKISWSKEN VFTVYKESEE DFKLKKDLLV ADQVAIPINY VSYYVLPFSE NOVMKNKKVY RNHGAPKEEN 20 222 DVINVTLLKE DFKLDLNEKE LFVQLSNCVE VIV 72 21	INTVIGEDLI IPSFRSNGQ' LLGLNQLLI aa, chair CALNGGNDE: aa, chair IVVYFIGEK: IVV	KTI VYV EKL  n + LFS n + YKV ISL EKK LKN GNS RLT SNV n + ENI NET PEF	
EFKNVDFKYKNSSNLI NEQEVEAYDIRSLCSY PEGLNTKIDDNSSNFS >GENE 5 MFVDLLASTSEKLTGN NSKERGMRTMSI >GENE 6 MEENILSLIKQKQKLH GSIKINEKGFGFVKDL RVKSFLIGEIQPSRDG LFTKIQKIIGDSNKAV SLVDKNLVTIDGSDSK TYLANKVIPMLPEKLS YKEVNDLFEKNVSNRD VDIVPRDRGVSERLIE >GENE 7 LIRALGINVKLTDLEK GHFGLASECYTHFTSP EKNSVNAEREVNKVCM TFDPKTNILVNKQNKV >GENE 8 MGEHILLKNKKAYFNY >GENE 9	IEKFNLK ISQDIVF GGERQRF 2242 RIVFAFE 2585 LNELLKT NDVEQDY RFLDFIP DRIISIA DLDDAIY NGVCSLN KEIVDML NFMVSAN 4010 VNPKTIR IRRYSDL AEFMTKH FRLGQKV 4703 EILDT 4880	INKGDKVLI SKGKLIDML SIIRGLLEN - 24 IIALVVSIL - 40 FKDEELLMS FVPPDSLNK SEPGFKNYR YEFNINPDF VEKTKDGYK PNEDKLCMV LVSKELHEL ESVAQIIFE - 46 MALDQISKQ MVHRYLKQY IEKEYEGVV KIKVKNADV - 47	KGKSGIGI KIANESAI KS 63 2: MITVGMIC 03 10: CLKELQD SITTDEV' IVMINSK! INRQTLEN ILFVAIAD SEMDFDNI IDKERVSI KNLPYVYI 78 4: 1EDQTER ILIDKDLR IANVLKFG KKRILDF 68 43 1	KTTLLKTLFN EEKQVLSLFE 40 73 QNKTSQTGLS 26 472 QYKISWSKEN VFTVYKESEE VFTVYKESEE NGVMKNKKVY RGS1DFDVPE RNJGAPKEE 20 222 DVINVTLKE DFKLDLNEKE LFVQLSNCVE VLV 72 21 79 87	INTVIGEDLE IPSFRSNGO' LLGLNQLLI aa, chain ALNGGNDE: aa, chain IVVYFIGEK: RYRANVED: RYKVKILNVKI EDEQVKRR: PLDNTALYRO TESIMNSKAI PKIVLDKE: LILIE L	KTI VYV EKL  1 + LFS  1 + YKV LKN GNS RLT SNV  1 + ENI NET PEF	
EFKNVDFKYKNSSNLI NEQEVEAYDIRSLCSY PEGLNTKIDDNSSNFS >GENE 5 MFVDLLASTSEKLTGN NSKERGMDRTMSI >GENE 6 MEENILSLIKQKQKLH GSIKINEKGFGFVKDL RVKSFLIGEIQPSRDG LFTKIQKIIGDSNKAV SLVDKNLVTIDGSDSK TYLANKVIPMLPEKLS YKEVNDLFEKNVSNRD VDIVPRDRGVSERLIE >GENE 7 LIRALGINVKLTDLEK GHFGLASECYTHFTSP EKNSVNAEREVNKVCM TFDPKTNILVNKQNKV >GENE 8 MGEHILLKNKKAYFNY	IEKFNLK ISQDIVF GGERQRF 2242 RIVFAFE 2585 LNELLKT NDVEQDY RFLDFIP DRIISIA DLDDAIY NGVCSLN KEIVDML NFMVSAN 4010 VNPKTIR IRRYSDL AEFMTKH FRLGQKV 4703 EILDT 4880	INKGDKVLI SKGKLIDML SIIRGLLEN - 24 IIALVVSIL - 40 FKDEELLMS FVPPDSLNK SEPGFKNYR YEFNINPDF VEKTKDGYK PNEDKLCMV LVSKELHEL ESVAQIIFE - 46 MALDQISKQ MVHRYLKQY IEKEYEGVV KIKVKNADV - 47	KGKSGIGI KIANESAI KS 63 2: MITVGMIC 03 10: CLKELQD SITTDEV' IVMINSK! INRQTLEN ILFVAIAD SEMDFDNI IDKERVSI KNLPYVYI 78 4: 1EDQTER ILIDKDLR IANVLKFG KKRILDF 68 43 1	KTTLLKTLFN EEKQVLSLFE 40 73 QNKTSQTGLS 26 472 QYKISWSKEN VFTVYKESEE VFTVYKESEE NGVMKNKKVY RGS1DFDVPE RNJGAPKEE 20 222 DVINVTLKE DFKLDLNEKE LFVQLSNCVE VLV 72 21 79 87	INTVIGEDLE IPSFRSNGO' LLGLNQLLI aa, chain ALNGGNDE: aa, chain IVVYFIGEK: RYRANVED: RYKVKILNVKI EDEQVKRR: PLDNTALYRO TESIMNSKAI PKIVLDKE: LILIE L	KTI VYV EKL  1 + LFS  1 + YKV LKN GNS RLT SNV  1 + ENI NET PEF	
EFKNVDFKYKNSSNLI NEQEVEAYDIRSLCSY PEGLNTKIDDNSSNFS >GENE 5 MFVDLLASTSEKLTGN NSKERGMRTMSI >GENE 6 MEENILSLIKQKQKLH GSIKINEKGFGFVKDL RVKSFLIGEIQPSRDG LFTKIQKIIGDSNKAV SLVDKNLVTIDGSDSK TYLANKVIPMLPEKLS YKEVNDLFEKNVSNRD VDIVPRDRGVSERLIE >GENE 7 LIRALGINVKLTDLEK GHFGLASECYTHFTSP EKNSVNAEREVNKVCM TFDPKTNILVNKQNKV >GENE 8 MGEHILLKNKKAYFNY >GENE 9	IEKFNLK ISQDIVF GGERQRF 2242 RIVFAFE 2585 LNELLKT NDVEQDY RFLDFIP DRIISIA DLDDAIY NGVCSIN KEIUDML NFMVSAN 4010 VNPKTIR IRRYSDL AEFMTKH FRLGKV 4703 EILDT 4880 TRTRKLL	INKGDKVLI SKGKLIDML SIIRGLLEN - 24 IIALVVSIL - 40 FKDEELLMS FVPPDSLNK SEPGFKNYR YEFNINPDF VEKTKDGYK PNEDKLCMV LVSKELHEL ESVAQIIFE - 46 MALDQISKQ MVHRYLKQY IEKEYEGVV KIKVKNADV - 47 - 51 LNKDEIKKI	KGKSGIGI KIANESAI KS 63 2: MITVGMIC 03 10: CLKELQD SITTDEV IVMINSK: INRQTLEN ILFVAIAD IDKERVSI KNLPYVYI ITEDQTER ILIDKDLR IAVLKFG KKRILDF 68 43 1 LKRVQLE	KTTLLKTLFN EEKQVLSLFE 40 73 QNKTSQTGLS 26 472 QYKISWSKEN VFTVYKESEE VFTVYKESEE NGVMKNKKVY RGS1DFDVPE RNJGAPKEE 20 222 DVINVTLKE DFKLDLNEKE LFVQLSNCVE VLV 72 21 79 87	INTVIGEDLI IPSFRSNGO' LLGLNQLLI aa, chain ALNGGNDE: aa, chain IVVYFIGEK: ERYRANVED: ERYRANVED: ENTANVED: IVVYFIGEK: ENTANVED: IVVYFIGEK: IVVYFI	KTI VYV EKL  1 + LFS  1 + YKV LKN GNS RLT SNV  1 + ENI NET PEF	

- 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
         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
                                                             1337
Predicted protein(s):
>GENE 1 287 - 553 252 88 aa, chain +
MASKEFHIVAETGIHARPATLLVOTASKFTSEITLEYKGKSVNLKSIMGVMSLGVGOGAD
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 pr	romoters - 0.20		
Number of predic	cted promoters -	7	
	225 LDF- 8.79		
-10 box at pos.	210 TGGTACAAT	Score	78
-35 box at pos.	190 TTGCAA	Score	55
	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
	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 <u>Berkeley Drosophila Neural</u>
<u>Network Prediction</u> 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	acgaagctgaaaccgaaaataactaaaaataaaagctgtc $oldsymbol{A}$ gaactgata
61	106	0.99	gctttttttcagctcactttcttcaggaaaataatataaa $oldsymbol{A}$ aatacttat
106	151	0.99	${\tt CTTATTTGATGATAAAAGAAATCAAAGTCTAGCATCCATT} {\bf C} {\tt AAAAGCAGC}$
184	229	0.97	cagatattgcaaaccctttcgttttgtggtacaatttcaa $oldsymbol{G}$ agtcataga
203	248	0.98	cgttttgtggtacaatttcaagagtcatagatattttaga $T$ atcgtcaat
214	259	0.98	acaatttcaagagtcatagatattttagatatcgtcaata $oldsymbol{A}$ aaatgaaaa
234	279	0.94	tattttagatatcgtcaataaaaatgaaaaaagatctaag $oldsymbol{G}$ agaaccatt
382	427	0.97	аатсастттддаатасаааддтааатсадтааассттааа $T$ саатсатдд
896	941	0.96	${\tt GTATCTTTGAAGGAATGACTGATAATGCTTATATGCAAGA} {\tt ACGTGCAGCT}$
1105	1150	0.88	${\tt AACATTGGTGGACGTACTTCTCACTCTGCAATTATGGCTC} {\bf G} {\tt TACTTTGGA}$
1148	1193	0.98	$CTTTGGAAATTCCTGCTGTTCTTGGAACAAATAATATTAC \overline{T} GAACTTGTT$
1284	1329	0.95	${\tt AGCTGGTGAAGCTTATGCTGCTCAAAAAGCAGAATGGGCT} \textbf{G} {\tt CTCTTAAAG}$
1422	1467	0.81	${\tt cggtgctgaagcaattggtctttatcgtacagaattcttg} {\sf T}{\tt acatggatg}$
1819	1864	0.93	${\tt GTTCCAGTTGCAGATGGTATCGAAGTAGGTATCATGATTG} {\color{red}{\bf A}} {\tt AATTCCAGC}$
1886	1931	0.95	${\tt ACCAATTTGCTAAGGAAGTTGATTTCTTCTCAATTGGTAC} {\tt AACGACCTC}$
1915	1960	0.96	tcaattggtacaaacgacctcatccaatatacaatggctg ${f C}$ agaccgtat
2073	2118	0.97	${\tt TGGTGAAATGGCCGGCGACCAAACTGCTGTACCATTGCTT} {\color{red}\boldsymbol{A}} {\tt TGGGTATGG}$
2238	2283	0.84	aacaatggaagaagttattgccctcgttgaagaatatact $oldsymbol{A}$ aataatctt
2250	2295	0.92	${\tt AGTTATTGCCCTCGTTGAAGAATATACTAAATAATCTTTT} {\color{red}{\bf C}} {\tt GATTGATTT}$
2331	2376	0.99	ttttttgtaatttattatcaacaacaaatatactgacag $oldsymbol{A}$ aaaacttat
2361	2406	0.94	atactgacagaaaacttatccacgtggataagttttttg $T$ attatttta
2393	2438	0.99	${\tt GTTTTTTGTATTATTTTAATGTTAAAACGTACAATAATGA} {\tt T} {\tt AAGTGGAGA}$
2402	2447	0.85	attattttaatgttaaaacgtacaataatgataagtggag $oldsymbol{A}$ gaaatggca
2475	2520	0.93	ttagttggagagggaggttacggtctcattttgatattga $T$ tttacctag
2502	2547	0.93	attttgatattgattttacctagccaaattaatattaatt $\mathbf{C}$ tggcttggt

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.