2007 7.013 Problem Set 3

Due before 5 PM on FRIDAY, March 16, 2007.
Turn answers in to the box outside of 68-120.
PLEASE WRITE YOUR ANSWERS ON THIS PRINTOUT.

Topic 1.

You have been studying transcription in a type of yeast that is frequently used in wine-making. TATA binding factor (also known as TFIID) in these yeast cells strictly recognizes the sequence 5'-TATAAT-3'. Upon binding, TFIID recruits other components of the transcription apparatus to assemble. Once RNA polymerase has bound to this region, it begins transcribing after the 25th nucleotide downstream of the TATAAT sequence.

Examine the following segments of chromosomal DNA from this species. In each case, write the first 10 nucleotides of the nascent messenger RNA that will be produced. Label the 5' and 3' ends of that molecule.

1a.

5′-GGCTATAATGCTCTACAGCAGACCCGGAAGCCCAGAGCAAACGCCAAGG-3′
3′-CCGATATTACGAGATGTCGTCTGGGCCTTCGGGTCTCGTTTGCGGTTCC-5′

1b.

5′-TAATGCTCTATAATCAGCAGACCTCTATTCGCTTAGCCCAGAAGCACGA-3′
3′-ATTACGAGATATTAGTCGTCTGGAGATAAGCGAATCGGGTCTTCGTGCT-5′

1c.

5′-GGCTATGATGCTCTACAGCAGACCCGGAAGCCCAGATTATACCGGATCG-3′
3′-CCGATATTACGAGATGTCGTCTGGGCCTTCGGGTCTAATATGGCCTAGC-5′

For each of the above sequences, circle the template strand of the DNA.

Having studied this yeast for some time, you've learned that it contains many genes that encode very small proteins, and the genes often contain tiny introns. The intron splice sites in this species always have the composition 5'-GUG-3' (for the 5' splice site) and 5'-UUG-3' (for the 3' splice site), meaning that these sequences and everything between them will be removed from the mature message. Following are three genes from this species. Only the non-template strands are shown. In each case, identify the coding portions and the introns of each gene. Assume that translation always begins at the first methionine codon encountered, and that the introns get spliced out. **In the space beneath each sequence, write out the primary sequence of the polypeptide that would be made.** Use the translation table on the last page of this problem set. Please use the single letter abbreviations for each amino acid.

Gene 1

5′-GACTATGATAACTATCGTGCAGCAGACCCGGAAGCCCAGAAAACCACTGGATTGTCTGGCAGAGAGGCCACATGATC-3′

1d.

Gene 2

5′-ACAATGTACTGCGCGACGCACGGTGACCACTTTCCCAGTTTTCCTTTCTTGCATCGTTCTTAGAAGCGTCGTAGAA-3′

1e.

Gene 3

5′-TAAATGTATACCCATATCAGCATTTCGGTGGAACGATATCACGCTAGAGATTAATTGGAAGCTAGTTACTGACCCA-3′

1f.

You've isolated several mutants from this species and found they each have subtle changes in the DNA sequences of the three genes listed above. Indicate what kind of mutation was involved in each case (Your choices are "missense," "nonsense," or "silent.")

1g. Nucleotide 10 in gene 1 was changed from "A" to "G": _____

1h. Nucleotide 16 in Gene 1 was changed from "C" to "A": _____

1i. Nucleotide 9 of Gene 2 was changed from "C" to "G": _____

1j. Nucleotide 3 in Gene 3 was changed from "A" to "G": _____

1k. Examining the rest of the sequence of the mutated Gene 3 reveals that two more changes lie further downstream within this same gene, and that these mutations are not even in the exons. One of these converts the sequence 5'-GGAAC-3' to 5'-CGAAC-3'. At the same time, the other mutation converts 5'-TTAATT-3' to 5'-TTAAAT-3'. Do you expect these changes to change the translation product. If not, *briefly* explain why not in the space below (10 words or less). If it does change the translation product, then show the modified product in the space below.

Topic 2

Towards the end of this problem set is the sequence of the alpha-tubulin gene from *Zea mays*. The sequence of the mature messenger RNA is also presented. Examine the sequence to answer the following questions.

2a. Where in the DNA sequence would you expect to find promoter sequences, including a TATA box? (Your choices are: "to the 5' side of the transcription start site," "to the 3' side of the transcription start site," or "within the first exon."

2b. How many introns are there in this gene? _____

2c. How long is the 5' UTR? _____

2d. How long is the 3' UTR? _____

2e. How long is the mRNA (excluding the polyA tail and the cap)?

2f. How long was the pre-mRNA primary transcript (assuming no processing occurred until after transcription had been completed; assume also that the primary transcript ended where the current polyA tail begins)?

2g. How many amino acids make up the protein that will be translated directly from this transcript (before any post-translational modifications)?

Topic 3

Having sequenced several alpha-tubulin genes in your laboratory, you notice that the alpha-tubulin gene from *Zea mays* differs from most other alpha-tubulins in one specific portion of the protein. The portion that differs most strongly is encoded by exon 2. You decide that to test the importance of these differences in the function of alpha-tubulin, you will transfer exon 2 from *Zea mays* into the alpha-tubulin gene from another species. The first step in doing this will require you to obtain a copy of exon 2 by PCR. Using the sequence information of the alpha-tubulin gene on subsequent pages, describe two primers, each 10 bases in length, that could be used to precisely amplify exon 2 of alpha-tubulin. Write out the sequences of those primers in the space below. Be sure to label their 5' and 3' ends.

3a. Primer 1:

3b. Primer 2:

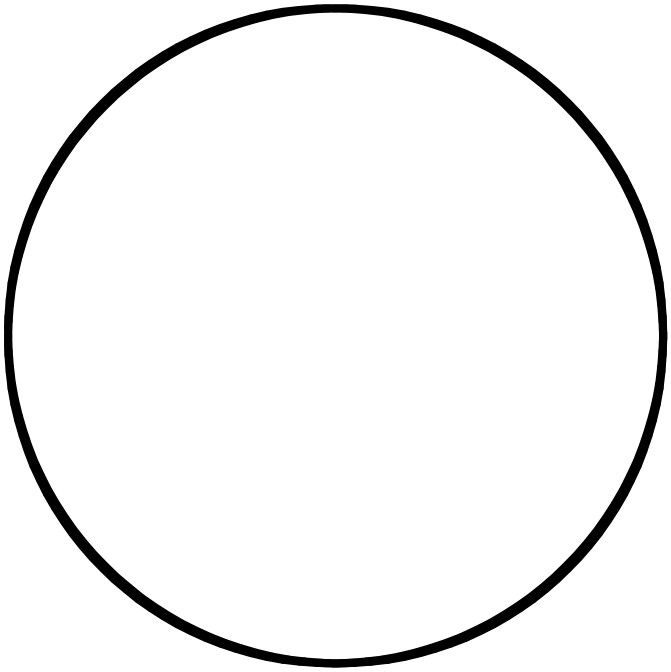
Topic 4

You have discovered a new plasmid in species of foul-smelling, soil dwelling bacteria. As a first step toward understanding the composition of this plasmid, you digest the plasmid with a series of restriction enzymes, either alone or in combination. Using agarose gel electrophoresis, you determine the size of each fragment produced in these reactions and list the results in the Table below:

Enzyme(s) used	*Afe*I	*Pst*I	*Sma*I	*Afe*I +*Pst*I	*Afe*I +*Sma*I	*Pst*I +*Sma*I	*Afe*I +*Pst*I +*Sma*I
DNA fragments produced	5000 bp	5000 bp	2900 bp 2100 bp	4000 bp 1000 bp	2300 bp 2100 bp 600 bp	2900 bp 1700 bp 400 bp	2300 bp 1700 bp 600 bp 400 bp

4a. How big is this plasmid (in basepairs)?

4b. Using the circle below, draw a restriction map of the plasmid. Be sure to indicate the distances between adjacent restriction sites (in basepairs).



Topic 5

Gaucher's disease is genetic disorder with a higher incidence among the Ashkenazi Jewish population. Affected individuals cannot metabolize certain lipids, primarily because of a deficiency in the enzyme glucocerebrosidase (GCA), causing glucocerebrosides to accumulate in the blood and other tissues. The deficiency is caused by a mutation in the gene encoding GCA.

In an effort to develop a gene therapy for this disorder, you have set out upon an ambitious strategy:

- you intend to produce a copy of the GCA gene that is specifically expressed in liver cells;
- You will introduce this gene into human liver cells by using the specialized, non-disease causing virus called HEP-RC, which can infect liver cells;
- you will then transplant the resulting cells into the liver of a patient, hopefully establishing a cluster of cells in the liver that allows the patient to metabolize glucocerebrosides.

5a. You have a cloned copy of the wild-type (functional) GCA coding sequence. As a fist task, select a promoter that, when joined to the GCA gene, is most likely to ensure proper expression of the GCA gene in liver cells:

_____ The promoter from the gene that encodes β-globin, an important blood-specific protein

_____ The promoter from the gene that encodes the CK-19 protein, which is normally transcribed only in the liver

_____ The promoter from the gene that encodes GFAP, a brain-specific protein

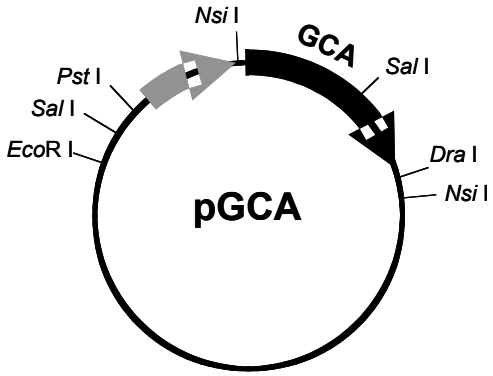
_____ The promoter from the gene that encodes α-tubulin, which is expressed in almost all cells

5b. In the space below, justify your choice in 15 words or less.

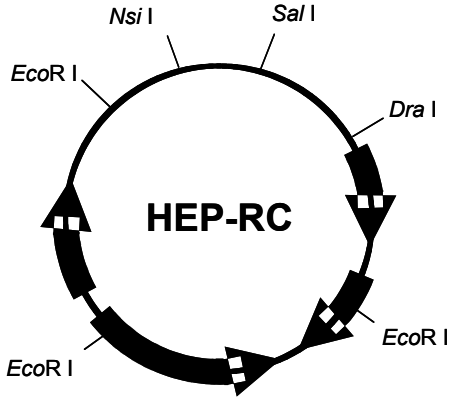
Having selected the best promoter, your assistant creates a plasmid in which the promoter and the GCA are fused together. This plasmid is called pGCA. You also have the HEP-RC genomic DNA (see below). You now need to excise the promoter-GCA region from pGCA, and you need to cut the HEP-RC DNA with restriction enzymes in a manner that allows you to ligate them together, to create a recombinant virus for use in your gene therapy strategy.

5c. Based on these diagrams, which two restriction enzyme(s) should you use to excise the promoter-GCA region from pGCA so that it can later be ligated into the viral genome. (Think about how you'll be cutting the HEP-RC DNA before you finalize your decision.)

5d. Which restriction enzymes would be suitable for preparing the HEP-RC genome to accept the resulting promoter-GCA fragment?



pGC contains the promoter you have selected (grey arrow) and the coding sequence from the GCA gene, (heavy black arrow). Also indicated are the relative positions of several restriction sites.



The HEP-RC genome is a circular double stranded DNA molecule and contains several genes (heavy arrows) that must not be disrupted if the viral DNA is to persist in liver cells.

Restriction Enzyme Recognition Sequences

Arrowheads indicate where the enzyme cleaves the phosphodiester bonds between DNA bases

DraI
5′...TTT▼AAA...3′
3′...AAA▲TTT...5′

EcoRI
5′...G▼AATTC...3′
3′...CTTAA▲G...5′

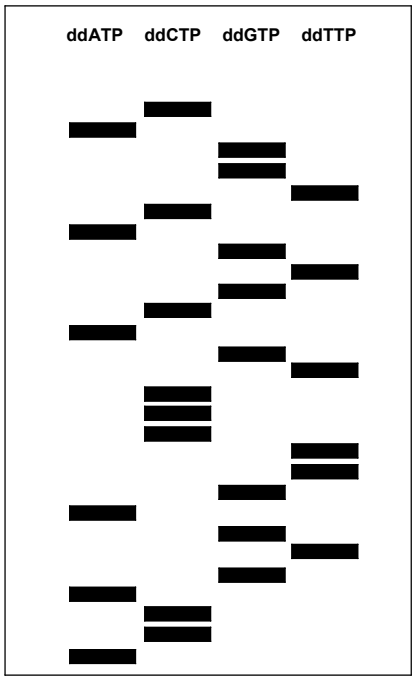
NsiI
5′...ATGCA▼T...3′
3′...T▲ACGTA...5′

PstI
5′...CTGCA▼G...3′
3′...G▲ACGTC...5′

SalI
5′...G▼TCGAC...3′
3′...CAGCT▲G...5′

Topic 6

6a. In your laboratory, you have discovered a gene that seems to be involved in cancer progression. As a first step toward understanding its function, you clone the gene and sequence it. Below is an image of the sequencing gel that you prepared user the Sanger dideoxy nucleotide sequencing technique. From this image determine the sequence of the DNA that was examined. Label the 5' and 3' ends of that DNA sequence. Remember that smaller DNA molecules migrate more quickly through a gel and will reach the bottom of the gel sooner than larger molecules. At the top of each lane is indicated the type of dideoxynucleotide that had been included in each polymerization reaction.



(You do not need to turn in this page)

One strand of the DNA sequence from the alpha-tubulin locus is shown below in **BOLD.** Where appropriate, the sequence of the mature mRNA is shown directly beneath the corresponding DNA sequence (NOT BOLD). Relative numbering of each is indicated to the left. "@" indicates the position of the 7-methyl-G cap. Exclamation points (!!!) are positioned beneath the nucleotides that constitute the start codon; Asterisks (***) are positioned beneath the nucleotides that constitute the stop codon.

```
   1 TACGGACGGG AGAAGCGCGC GGGCGGGCCG GCCGGCCTCA CGTGAACAAC GCCCGCCCAC
  61 GAAACGCGCT GGCTGTGCAA AAGCAACCGT TTCCAGTTTC CTCCCCAGTG CACCCCCTCC
 121 CCTCTCCCCT CCCTCTCTCT CCTCCGGCCC CCTTTTTTTA TGCCCGATCT CACACATTTT
 181 GGAAGGAGAG AGAGAGAGGC AGAGGGAGAG ATTGGAGGGA GGCCCCTGCC GTAGGCAAGA
   1                                                @CCCUGCC GUAGGCAAGA

 241 GAAACCGCGG CGCGCGGAGA GGGAGGAAGG GCAGACGCAG ACGCAGACGG GCGAACAAGA
  18 GAAACCGCGG CGCGCGGAGA GGGAGGAAGG GCAGACGCAG ACGCAGACGG GCGAACAAGA
                                                                   !

 301 TGAGGGAGAT CATCAGCATC CACATCGGCC AGGCCGGGAT CCAGGTCGGC AACGCCTGCT
  78 UGAGGGAGAU CAUCAGCAUC CACAUCGGCC AGGCCGGGAU CCAGGUCGGC AACGCCUGCU
     !!

 361 GGGAGCTCTA CTGCCTCGAG CACGGCATCG AGCACGATGG CACCATGCCC AGGTGAGTGC
 138 GGGAGCUCUA CUGCCUCGAG CACGGCAUCG AGCACGAUGG CACCAUGCCC AG

 421 TGCATGCCCT GTCGATCCGC CGGGACGCGC TCCTCCTGTT TACCCATTTT TTGCAGCTTA
 481 AAGCTGTTGC GCGTTCTATG GAGCCGGCGC ATCAAAATTC TGGCGTGCTC ATTCGTCGCA
 541 TTTCCTAGAT TCGCTGCGTT GTTTTCTTCC AATTCGTCGG AGCAAATGAC GGGGCGTGTA
 601 GGTGTAGTGA TGGTACTTCC ATATCCGTTT CGTATTTTTT GGTTCGTCTG TCCCTTTCCG
 661 TTGCACTCGC AGTTTTGGTT CCTTTGCGAC TGGATTTGTT TGTTGGGTGG GATGTGTTGT
 721 TGTTGTTGTC TGTTTGGGTT TTGGTGCATT TTGCCGGTTG TTCGAGAGAT TGCTATGAGC
 781 CGGGACCAGA TGCCACGTTT CCTAATTAAG AGTTTGCGCT AGCGTTTTTT TTTTGGACTA
 841 TTAATACTAG GCCTGTTAAT TAGGCAGGCG CCACAGAAGG ATGCGTTGAG CTCTGAAGAA
 901 TTGAATCTGC AGGGATGTCA TCTAAATAAA TAATATAACT GGGATGTTCT GATGAGTGTC
 961 AGACCATGGA TCTCCCTGGA TTCACGAGTT TCATCTGACC CTTTGTAAAA AGTAGCACCC
1021 TTCTCTCTGT TTGCCTGATC TTGTAGTACC ATAGAAAATT GCCCTAATGC TGCTCCGCCT
1081 GATAATAGAA AAAATCCTGT TTTAGGACCC CAAGGAGGCA GGTTTCGTAG GCTATAAATT
1141 CTCTTTAAGC GCATATGCCA ATGGTTCGTG CCTTATTGTT ATTGGAGTTC CATGCTATCC
1201 ATTTTTTATA ACACTGGTTT GATTGTGATG CAGTGATTCC TCGGTTGGCG TCGCACATGA
 190                                    UGAUUCC UCGGUUGGCG UCGCACAUGA

1261 TGCCTTCAAC ACGTTCTTCA GCGAGACTGG TTCCGGCAAG CATGTGCCCA GGGCCATCTT
 217 UGCCUUCAAC ACGUUCUUCA GCGAGACUGG UUCCGGCAAG CAUGUGCCCA GGGCCAUCUU

1321 CGTCGACCTT GAGCCCACTG TCATCGACGA GGTTCGCACT GGCTCGTACC GCCAGCTCTT
 277 CGUCGACCUU GAGCCCACUG UCAUCGACGA GGUUCGCACU GGCUCGUACC GCCAGCUCUU

1381 CCACCCAGAG CAGCTCATCT CGGGGAAGGA GGATGCAGCT AACAACTTTG CCCGTGGCCA
 337 CCACCCAGAG CAGCUCAUCU CGGGGAAGGA GGAUGCAGCU AACAACUUUG CCCGUGGCCA

1441 CTACACGGTA ATTCACCTTG TGTCTGAATA GATAATAGGA GTGGGTCCTA TGCTTCTGCA
 397 CUACACG

1501 ACTGTGGAAA TGTTCAATTC TAATTATGGT GCAATGCAGT TGGGAAGGAG ATTGTCGATC
 404                                       GU UGGGAAGGAG AUUGUCGAUC

1561 TATGCCTGGA CCGTGTGCGC AAGCTTGCAG ACAACTGCAC TGGGCTGCAG GGATTCTTGG
 426 UAUGCCUGGA CCGUGUGCGC AAGCUUGCAG ACAACUGCAC UGGGCUGCAG GGAUUCUUGG
```

(You do not need to turn in this page)

```
1621 TGTTCAATGC TGTTGGTGGT GGTACTGGCT CTGGACTTGG TTCACTGCTG TTGGAGCGCC
 486 UGUUCAAUGC UGUUGGUGGU GGUACUGGCU CUGGACUUGG UUCACUGCUG UUGGAGCGCC

1681 TTTCAGTTGA TTATGGCAAG AAGTCTAAGC TTGGTTTCAC CATTTATCCT TCCCCACAGG
 546 UUUCAGUUGA UUAUGGCAAG AAGUCUAAGC UUGGUUUCAC CAUUUAUCCU UCCCCACAG


1741 TATGACCTTA TGACAGAACC TTACATTAGC AACCTAGTTT CTGTGGATTT CACACAAAAA
1801 TATATTCATA TAATTATTCT TTCCAAAAGG AGTTTGGCTT AAGTATATGA GAATTGTAAA
1861 ACATACATGA AGTATTAGGT TGACAATTTA TATTTATACA CCAACATGAA ATTTTTTTAT
1921 ATGTAAATTA ATCTACGGAT TATGCAATTT GGTGGCACTG TCCTTGCCTT CAAAATTACA
1981 GACAGAGCAC TGATTATTGG ATCAGCCTAG CATTATATAA ATCATGTGTG GAACAAAACA
2041 AGCAACTGTT GTGAACTTAT GGTTAGTGAG ACCTTCTTGT ATATTGTTTA GACCATGGTT
2101 AGTACAGCTG TCACAGTCTT CAGGTCTTCT TTATGAGTGT CTTTAAATCA TGCATTGCAT
2161 ATGTGCCAGC ATTATCCTAT AGTGCTGCTC CTTCCTGTCC ATGTGATGTG GTCCATTCAT
2221 CTTGTTCATG TGAGTAGGTT TTCTGCATCT TAGCTGTTGT GGGGAAAAGT GGGTAGAGAT
2281 GTTCGATGAA TTTGGATCTG ATCCATCATT TAACTTTTCA TGTAACGCCA CTCTGGAATG
2341 CAGGTGTCAA CAGCTGTTGT AGAGCCATAC AACAGTGTCC TCTCGACCCA CTCCTTGCTC
 605    GUGUCAA CAGCUGUUGU AGAGCCAUAC AACAGUGUCC UCUCGACCCA CUCCUUGCUC

2401 GAGCACACTG ATGTTGCAGT CCTCCTGGAC AACGAGGCTA TCTATGACAT ATGCAGGAGG
 662 GAGCACACUG AUGUUGCAGU CCUCCUGGAC AACGAGGCUA UCUAUGACAU AUGCAGGAGG

2461 TCCCTTGACA TCGAAAGGCC AACCTATACC AACTTGAACA GGCTGATCTC ACAGATCATA
 722 UCCCUUGACA UCGAAAGGCC AACCUAUACC AACUUGAACA GGCUGAUCUC ACAGAUCAUA

2521 TCATCACTTA CCACCTCCCT GAGGTTTGAT GGGGCTATCA ATGTGGATGT TACTGAGTTC
 782 UCAUCACUUA CCACCUCCCU GAGGUUUGAU GGGGCUAUCA AUGUGGAUGU UACUGAGUUC

2581 CAGACCAACC TTGTTCCATA CCCACGTATT CATTTCATGC TTTCCTCATA TGCCCCTGTA
 842 CAGACCAACC UUGUUCCAUA CCCACGUAUU CAUUUCAUGC UUUCCUCAUA UGCCCCUGUA

2641 ATATCTGCTG AGAAGGCCTA CCATGAGCAG CTCTCTGTGC CTGAAATCAC CAATGCTGTA
 902 AUAUCUGCUG AGAAGGCCUA CCAUGAGCAG CUCUCUGUGC CUGAAAUCAC CAAUGCUGUA

2701 TTTGAGCCCT CAAGCATGAT GGCCAAGTGT GACCCAAGGC ACGGAAAGTA CATGGCTTGC
 962 UUUGAGCCCU CAAGCAUGAU GGCCAAGUGU GACCCAAGGC ACGGAAAGUA CAUGGCUUGC

2761 TGCTTGATGT ACCGCGGTGA TGTTGTTCCT AAGGATGTCA ACGCTGCAGT TGCAACCATC
1022 UGCUUGAUGU ACCGCGGUGA UGUUGUUCCU AAGGAUGUCA ACGCUGCAGU UGCAACCAUC

2821 AAGACCAAGA GAACTGTCCA GTTTGTGGAC TGGTAAGATG TCTTTGTCGC ATTTGAGTTG
1082 AAGACCAAGA GAACUGUCCA GUUUGUGGAC UG

2881 TGCTTATTCT ATGATATCAC ATGCTTGTCA CTTGATTCGC GATGGTAAGT GGGTAACTAA
2941 TTCAATTATG CTGTCGCCAC ACAGGTGCCC CACTGGATTC AAGTGTGGCA TCAACTATCA
1114                          GUGCCC CACUGGAUUC AAGUGUGGCA UCAACUAUCA

3001 GCCACCCTCT GTTGTCCCTG GAGGTGACCT GGCAAAGGTC CAGCGTGCCG TGTGCATGAT
1150 GCCACCCUCU GUUGUCCCUG GAGGUGACCU GGCAAAGGUC CAGCGUGCCG UGUGCAUGAU

3061 CAGCAACAAC ACCGCCGTTG CCGAGGTGTT TTCACGCATC GACCACAAGT TTGACCTTAT
1210 CAGCAACAAC ACCGCCGUUG CCGAGGUGUU UUCACGCAUC GACCACAAGU UUGACCUUAU

3121 GTACGCCAAG CGCGCATTCG TTCACTGGTA TGTCGGTGAG GGTATGGAAG AGGGTGAGTT
1270 GUACGCCAAG CGCGCAUUCG UUCACUGGUA UGUCGGUGAG GGUAUGGAAG AGGGUGAGUU

3181 CTCAGAAGCC CGTGAGGACT TGGCTGCTCT TGAGAAGGAC TATGAGGAAG TCGGTGCAGA
1330 CUCAGAAGCC CGUGAGGACU UGGCUGCUCU UGAGAAGGAC UAUGAGGAAG UCGGUGCAGA

3241 GGGTGCCGAT GACGAGGGTG ACGAGGGAGA CGACTATTGA ATAGCTGGCT AATAAGTAGT
```

(You do not need to turn in this page)

```
1390 GGGUGCCGAU GACGAGGGUG ACGAGGGAGA CGACUAUUGA AUAGCUGGCU AAUAAGUAGU
                                                ***

3301 TCTCTGGTGG TTAATGGTTG GACTATTTTG GAGCTATATA CTCTATGGCT CCACTCCATT
1450 UCUCUGGUGG UUAAUGGUUG GACUAUUUUG GAGCUAUAUA CUCUAUGGCU CCACUCCAUU

3361 GGATACTGCT GCTGGTTGTG TGTTTCCATT TTGTACTATG TAGTAAATTG TTCGTAACCC
1510 GGAUACUGCU GCUGGUUGUG UGUUUCCAUU UUGUACUAUG UAGUAAAUUG UUCGUAACCC

3421 CCCTATTGGC CATGATTGTT CCATATCATC CTTCTTTGCG TTTGCAACGC TATTCGTCCA
1570 CCCUAUUGGC CAUGAUUGUU CAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA

3481 ATTTCCGGTG TATATGCTAT AATGCTATTA TGTTGCAAGT GTGCTGTGCT CAACCATTGG
3541 GTTTTATTTA TATTTAGAAA CAGAAGGCAT TATTACTCTT CGCTGAAGCT TGTTTCTGAA
3601 TCAGGGGAAC ATTGAGATTT TGAGAAGTTT TGACAGCAGT CTACCACTGT GAGGGCCTAA
3661 GGATCCCACG GGAG
```

Codon Table

(You do not need to turn in this page)

	U	**C**	**A**	**G**
U	UUU Phe (F) UUC " UUA Leu (L) UUG "	UCU Ser (S) UCC " UCA " UCG "	UAU Tyr (Y) UAC " UAA ***Stop*** UAG ***Stop***	UGU Cys (C) UGC " UGA ***Stop*** UGG Trp (W)
C	CUU Leu (L) CUC " CUA " CUG "	CCU Pro (P) CCC " CCA " CCG "	CAU His (H) CAC " CAA Gln (Q) CAG "	CGU Arg (R) CGC " CGA " CGG "
A	AUU Ile (I) AUC " AUA " **AUG** Met (M)	ACU Thr (T) ACC " ACA " ACG "	AAU Asn (N) AAC " AAA Lys (K) AAG "	AGU Ser (S) AGC " AGA Arg (R) AGG "
G	GUU Val (V) GUC " GUA " GUG "	GCU Ala (A) GCC " GCA " GCG "	GAU Asp (D) GAC " GAA Glu (E) GAG "	GGU Gly (G) GGC " GGA " GGG "

The Genetic Code. Three-letter and one-letter abbreviations for each amino acid are listed alongside their corresponding codons. Ditto marks (") indicate the same amino acid as above.