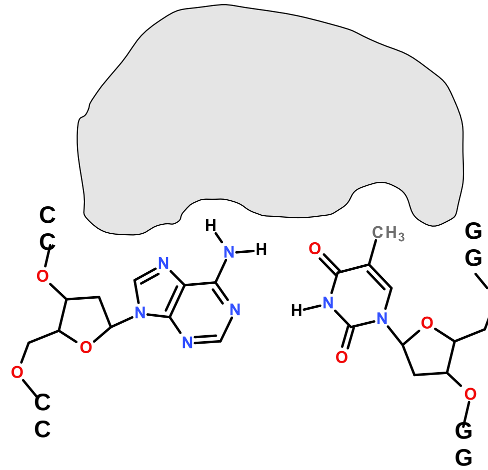
Note: This is an optional problem set that you can complete to replace a lower scoring problem set from earlier in the semester. If you do not submit it for grading, you are strongly encouraged to work through these problems in preparation for the final exam.

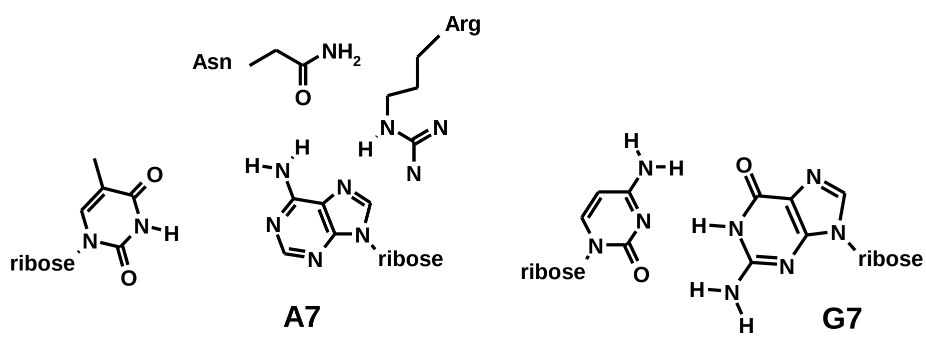


1. (10 pts, 15 min) An AT basepair is shown, contained within the sequence CCACC.
   i) Indicate the Watson-Crick hydrogen bonds (2 pts).
   ii) Identify all hydrogen bond donors and acceptors in both the major and minor groove on both (2 pts).
   iii) Label the major and minor groove (2 pts).
   iv) Assuming that a protein binds in the major groove, and contacts both the A and T bases, what will the DNA binding site on the protein look like, i.e. add the appropriate information to the gray area in the above diagram.

2. (8 pts, 20 min) The binding of proteins to short DNA fragments, such as restriction endonuclease sites, can be measured with equilibrium dialysis. The following set of experiments was performed using a **1 nM** solution of EcoRI inside the dialysis bag. The binding of two different double stranded DNA sequences was studied. The concentration of protein with DNA bound ([ML]) is shown below each DNA sequence.

Free [DNA]	**DNA Bound to Protein** $G_5 A_6 A_7 T_8 T_9 C_{10}$ $C_{10} T_9 T_8 A_7 A_6 G_5$	**DNA Bound to Protein** $G_5 A_6 G_7 T_8 T_9 C_{10}$ $C_{10} T_9 C_8 A_7 A_6 G_5$
10^{-9} M	0.50 nM	0.1 nM
10^{-8} M	0.90 nM	0.50 nM
10^{-7} M	0.99 nM	0.90 nM

   i) Determine the dissociation constant, K_D, for the binding of each DNA to EcoRI (2 pts)
   ii) Calculate the ΔG^o for the binding of Eco R1 to *each* DNA (2 pts) T=300K.
   iii) Explain, in terms of molecular interactions between Eco RI and the nucleotide bases, the difference in binding affinity of the two DNA sequences. Do the differences in the molecular interactions account for the ΔG^o difference that is measured in the binding experiment? (4 pts)



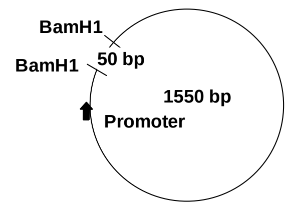
Note: The interaction of an Asn and Arg sidechains from EcoR1 with A_7 in the recognition sequence is shown to the right. You might find it useful to sketch the interactions between EcoR1 and the TA and CG basepairs in your answer. You should assume that the positions of these sidechains are the same in the wild-type and mutant protein.

3. (12 pts, 15 min) One of the Jmol pages associated with this problem set shows two DNA binding proteins - the methionine repressor and the lambda repressor. View this page and answer the following questions:
   i) Both of these proteins control gene expression by preventing RNA polymerase from synthesizing mRNA. However, they bind to DNA by inserting different secondary structure element into a groove in the DNA. View the Jmol page and state the secondary structure that is used by each protein and determine whether the protein binds in the major groove or the minor groove (2 pts).

ii) Based on your answer to *part i),* why do you think this groove is preferentially used by most DNA binding proteins? (2 pts).

iii) Click on the button marked "Lys" under the Jmol image of the methionine repressor. This will zoom in on the structure and highlight two Lys residues by placing a surface around their sidechain atoms. Briefly describe how each of these residues is interacting with the DNA and whether the interaction would be sequence specific or non-sequence specific (3 pts).

iv) Based on your answer to *part iii)*, would you expect the binding of DNA to the methionine repressor to be affected by salt or not? If so, how would increasing the salt affect the binding (5 pts).



4. (8 pts, 15 min) You are interested in expressing a protein from the influenza virus in E. coli for vaccine development. You only need to express part of the protein in bacteria since it will be used for a vaccine. You decide to use PCR to amplify a segment of DNA that begins with TTG and ends with GCG; the DNA to be present in the final PCR product is shown in uppercase. You want to add restriction endonuclease sites at both ends of the PCR product to facilitate cloning into your expression vector (shown below on the right). Note that your expression vector lacks a start codon and a stop codon, so these will have to be added during PCR.

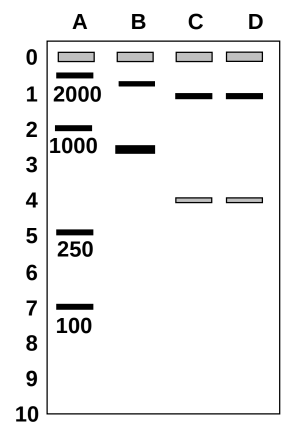
```
               |----------------------------400 bp ----------------------------|
ggcctaggtTTACGTTATGTTGCAGTATATAGGCGA-----------GGGTATAGCGCTAGCGATAGTTAATTGCGcgcgttagtagta
ccggatccaAATGCAATTCAACGTCATATATCCGCT-----------CCCATATCGCGATCGCTATCAATTAACGCgcgcaatcatcat
         LeuArgTyrValAlaValTyrArgArg...........GlyTyrSerAlaSerAspSer
```

Determine the sequence of both the right and left primers that you would use for the PCR reaction. Your primers should be 16 bases in length, including the additional nucleotide bases that comprise the restriction endonuclease sites, plus the start and stop codons (6 pts).

Considering the left primer only, what would the correct annealing temperature to use during the PCR reaction (2 pts) ? Use the following formula: **T_M = 81.5 +0.41(%GC)-675/N**

Example T_M calculation for the following primer: GGGAAACCCGTA

%GC=7/12=58.3%, N=12. Therefore, T_M = 81.5 + 0.41x58.3 - 675/12 = 49.1 C



5. (12 pts, 20 min) Continuing on from question 5. You digest the plasmid and the PCR product with BamH1, mix the two digestion, and then add DNA ligase. After ligating your PCR product into the vector, you transform cells, isolate plasmid DNA from two bacterial clones and perform a restriction digest and separate the DNA fragments by size using gel electrophoresis. An image of the gel is shown on the right. Standards of the indicated size are in lane A (units are basepairs) the uncut plasmid from clone 1 is in lane B, and the BamH1 digest of that clone is shown in lane C. Lane D contains the BamH1 digest of the second clone.

i) Why are there two bands in lane B (5 pts)?

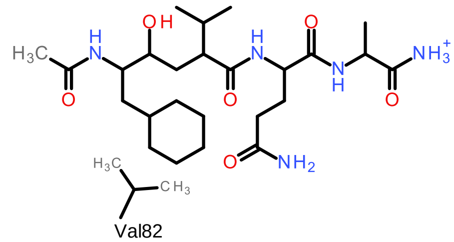
iii) The first clone does not produce any protein in bacteria, while the second clone does. Why (5 pts)? Hint, draw the possible ligation products.

6. (5 pts, 10 min) You want to express myoglobin from dinosaurs in bacteria to see how changes in the sequence of the myoglobin lead to efficient O_2 storage in bird muscle tissue. The first five amino acids of myoglobin are:

GlyTrpArgAsp…

Give the DNA sequence that you would use to code for the first amino acids in myoglobin. You should consult the following reference for E. coli codon usage when considering which codons to use. http://www.faculty.ucr.edu/~mmaduro/codonusage/codontable.htm (select the most frequent codon in class II, which is the codon usage in highly expressed genes).

7. (14 pts, 20 min) You have obtained the DNA that encodes a mutant HIV protease gene from a drug resistant patient. The sequencing reaction was performed on this DNA and the resultant trace is shown below. The wild-type sequence is also given to facilitate identifying the location of the mutation.
   - i) Determine the nucleotide change(s) in the mutant protease gene. Note, determine the reading frame by consulting the DNA and protein sequence of HIV protease (2 pts).
   - ii) Give the sequence of the primer that was used to generate the sequencing data, assume the primer is 10 bases in length and the first dideoxy base that is added can be detected in the sequencing data(2 pts).
   - iii) Determine the amino acid change induced by this mutation and comment on whether you expect this mutation to affect the substrate specificity or the catalytic efficiency of the protease (4 pts).
   - iv) How might you modify the original, cyclohexane containing drug (shown on the right) such that it would bind, and effectively inhibit the activity of the virus containing the mutant protease?(6 pts)



Region of HIV DNA Coding for HIV protease.

5'-ggagccgatagacaaggaactgtatcctttaacttcCCTCAGATCACTCTTTGGCAA 57
ProGlnIleThrLeuTrpGln 7

58 CGACCCCTCGTCACAATAAAGATAGGGGGGCAACTAAAGGAAGCTCTATTAGAT**ACAGGA** 117
8 ArgProLeuValThrIleLysIleGlyGlyGlnLeuLysGluAlaLeuLeuAspThrGly 27

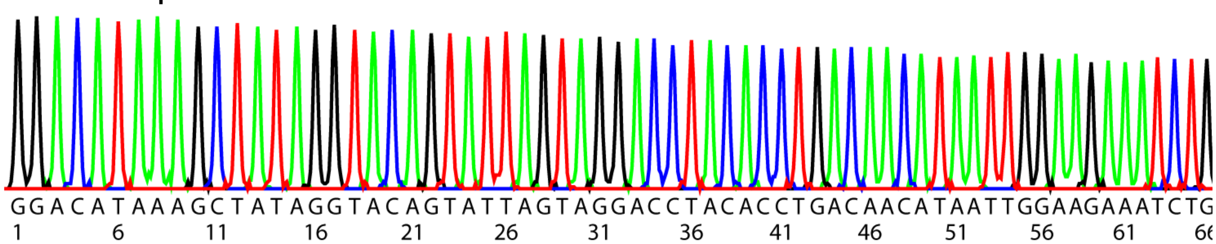
118 GCAGATG**A**TACAGTATTAGAAGAAATGAGTTTGCCAGGAAGATGGAAACCAAAAATGATA 177
28 AlaAspAspThrValLeuGluGluMetSerLeuProGlyArgTrpLysProLysMetIle 47

178 GGGGGAATTGGAGGTTTTATCAAAGTAAGACAGTATGATCAGATACTCATAGAAATCTGT 237
48 GlyGlyIleGlyGlyPheIleLysValArgGlnTyrAspGlnIleLeuIleGluIleCys 67

238 GGACATAAAGCTATAGGTACAGTATTAGTAGGACCTACACCTGTCAACATAATTGGAAGA 297
68 GlyHisLysAlaIleGlyThrValLeuValGlyProThrPro<u>Val</u>AsnIleIleGlyArg 87

298 AATCTGTTGACTCAGATTGGTTGCACTTTAAATTTTcccattagccctattgagact 354 -3'
88 AsnLeuLeuThrGlnIleGlyCysThrLeuAsnPhe

Mutant Sequence



Wild-type Sequence