Molecular Biology Through Discovery

Problem Set 5: The Coding Problem

1. Using a convenient genetic code table, complete the following:

DNA double helix						A		G					A	
					T				G			T		
mRNA transcribed	5'				A						U			
Appropriate tRNA anticodon							U			G				5'
Amino acids incorporated into protein		met												

(Table available in DOCX format by clicking here)

2. Consider the RNA sequence below. Suppose that the fourth base, C, were mutated to a U.

GAGCGUGCGAACC

2a. How many amino acids might be affected if the code were nonoverlapping triplet?

2b. How many if the code were overlapping triplet?

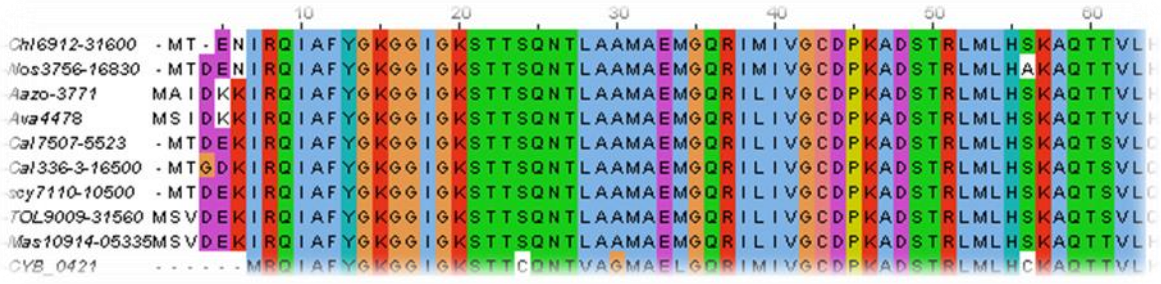
2c. Partially overlapping triplet?

2d. How would your answers be affected if the mutation were a deletion of the C?

3. You want to determine whether nitrogen fixation is taking place in a lake deep below the Antarctic permafrost. You could try to measure nitrogen fixation directly, but it is impossible to get to the lake with the necessary equipment, and a laboratory measurement of fixation in lake water sampled from the lake would be fraught with uncertainty. You therefore decide an indirect approach: isolating DNA from a lake water sample and determining whether there are any genes present that encode nitrogenase, the enzyme responsible for nitrogen fixation.

To do this, you need to amplify the genes using PCR… but how do you design PCR primers to amplify a gene whose DNA sequence is not known? There is hope. Nitrogenase is a highly conserved protein. Below you'll find an alignment of the amino acid sequences of nitrogenase subunits from 10 bacteria. You can see that there is unanimity at most positions. It's reasonable to expect that even weird Antarctic bacteria will have nitrogenase proteins that are similar to all others.

Protein **Amino acid position** (only beginning of alignment shown)



From an alignment of the sequences from more nitrogenase subunit proteins, a consensus sequence was determined, part of which is shown below. Letters are given when all sequences have the same amino acid at that position. A hyphen indicates that the position shows variability amongst the sequences.

```
         10        20        30        40        50
RQIAFYGKGGIGKSTT-QNT-A--A-----RI-IVGCDPKADSTRL-L—K
AQ---L--AAE-G-VED-EL--V---G-----CVESGGPEPGVGCAGRGI
IT-INFLEE-GAY-D--FV-YDVLGDVVCGGFAMPIRE-KAQEIYIV-SG
EMMAMYAANNIARG-LKYA--GGVRLGGLICNSR
```

Your goal is to design a forward primer and a backward primers to amplify a part of the nitrogenase genes of any nitrogen-fixing cyanobacteria that happen to be in the lake water. The primers should:

- be based on the consensus amino acid sequence shown above
- each be at least 14 nucleotides in length
- amplify a DNA fragment at least 100 nucleotides in length
- guarantee amplification of any target DNA that encodes a nitrogenase that matches the consensus amino acid sequence

You will find that no one DNA sequence will fit the last criterion, so you're permitted to specify ***degenerate*** primers with ambiguous positions. For example, AG[CG] is a degenerate sequence, because the third position can be either C or G. This leads to the fourth condition

- The degenerate primer sequence should minimize ambiguity as much as possible. For example AG[CG] (two possibilities) is less ambiguous than [GT]C[AT] (four possibilities)

Provide the two degenerate primer sequences that fulfill these conditions and calculate how many possible sequences each degenerate primer matches.

4. Suppose that every Virginia resident is to be assigned an ID number, except that it will be in the form of a DNA sequence. How long would the DNA sequence need to be to allow for a unique sequence for every resident? ***Provide details of your calculation plus any assumptions you made.*** Extra credit: Choose the sequence that would be your own ID.

5. We live in a world in which genes determine the linear sequence of amino acids that comprise a protein. There are only 20 possible amino acids that may be encoded (putting aside some specialized cases), and there are no restrictions as to what amino acid sequences are possible to encode.

5a. How many possible dipeptides are there? In other words, if you chop up all possible proteins (every conceivable sequence) into two amino acid-segments, how many different kinds of amino acid pairs would you get?

The remaining questions concern an alternate universe in which the genetic code consists of overlapping triplets, each codon overlapping the next by two nucleotides.

5b. Consider the triplet codon CAG. How many pairs of adjacent codons are possible in which the first codon of the pair is CAG? What is the maximum number of dipeptides that can be encoded by all of those pairs?

5c. How many possible triplet codons are there?

5d. How many possible pairs of adjacent triplet codons are there? What is the maximum number of dipeptides that can be encoded by all of those pairs?

5e. Suppose that the overlapping triplet genetic code we're considering is degenerate, that is more than one triplet may encode the same amino acid. If the dipeptides shown below are found in nature, how many triplets, at minimum, must encode histidine (His)?

His-Lys, His-Ser, His-Leu, His-Thr, His-Phe, His-Pro
Lys-His, Ser-His, Cys-His, Arg-His, Val-His, Phe-His, Glu-His, Gln-His, Ile-His

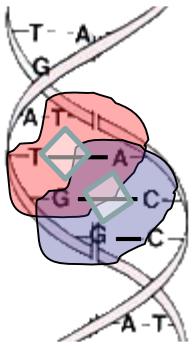
5f. It is 1957. There are many partial amino acid sequences of proteins known, but DNA sequencing is 20 years in the future. Can you think of a way to use known protein sequences to test the proposition that the genetic code consists of overlapping triplets?

5g. You might enjoy reading the following article:

Brenner S (1957). On the impossibility of all overlapping triplet codes in information transfer from nucleic acid to proteins. Proc Natl Acad Sci USA 43:687-694.

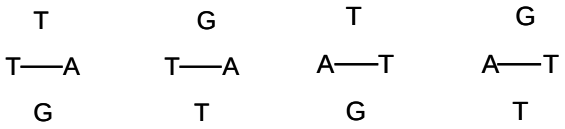
6. In 1954, George Gamow published the first attempt to conceive a genetic code. You can read about it in this very short article:

Gamow G (1954). Possible relation between deoxyribonucleic acid and protein structures. Nature 173:318.



Although Gamow was a fine artist (he illustrated his own popular science books), it may be difficult for you to interpret his rendition of the double helix. Each diamond in the article's figure lies within four nucleotides, defined by a basepair, one nucleotide above it, and one nucleotide below it. I've tried to clarify its message in the figure to the right. Each diamond is an amino acid binding site, surrounded by four nucleotides (highlighted in red for one diamond and blue for the other).

6a. Make up a genetic code that satisfies Gamow's criteria. The box shown below should be helpful. I suggest that you first make some arbitrary assignment (i.e. one of the 64 codons codes for some arbitrary amino acid), determine what other codons must code for the same amino acid, and then proceed along the same vein. Note that in his scheme, there is no concept of 5' and 3' and so the following diamonds are all equivalent:



codon aa	codon aa	codon aa	codon aa
TTT	**TCT**	**TAT**	**TGT**
TTC	**TCC**	**TAC**	**TGC**
TTA	**TCA**	**TAA**	**TGA**
TTG	**TCG**	**TAG**	**TGG**
CTT	**CCT**	**CAT**	**CGT**
CTC	**CCC**	**CAC**	**CGC**
CTA	**CCA**	**CAA**	**CGA**
CTG	**CCG**	**CAG**	**CGG**
ATT	**ACT**	**AAT**	**AGT**
ATC	**ACC**	**AAC**	**AGC**
ATA	**ACA**	**AAA**	**AGA**
ATG	**ACG**	**AAG**	**AGG**
GTT	**GCT**	**GAT**	**GGT**
GTC	**GCC**	**GAC**	**GGC**
GTA	**GCA**	**GAA**	**GGA**
GTG	**GCG**	**GAG**	**GGG**

(Table available in DOCX format by clicking here)

6b. Based on the genetic code you just made up, what is the DNA sequence that would encode Gly-Ala-Gly? Phe-Ala-Phe?

6c. Show that no code that follows Gamow's criteria could ever encode both of these tripeptides.