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. Why were no suppressors of FC0 found in segments B1b2 and B2 (see Fig. 2 of Crick et al, 1961)? Let's find out. In doing so, I'll make use of two pieces of special knowledge. ViroBIKE calls the T4 rIIB gene `gene0288` and calls its protein `p-gene0288`.

   3a. Go to ViroBIKE (*not* CyanoBIKE) through the BioBIKE portal or the course Resources & Links page, and from the **All** menu, bring down the SEQUENCE-OF function. Display the amino acid sequence of the rIIB *protein* by executing the function below (double-click SEQUENCE-OF):
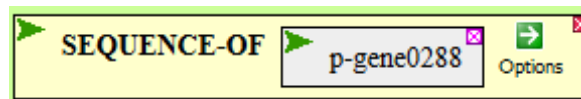
   Does the protein sequence begin with a methionine? How big is the protein? From the map of rIIB found in Crick (1962)[1] estimate the amino acid coordinates of the FC0 mutation and of the beginning of the B1b2 segment.

---

[1] Crick FHC (1962). The genetic code. Sci Amer 207:66-77 (October, 1962)

3b. From the **All** menu, bring down the READING-FRAMES-OF function. Display the reading frames of the *rIIB* **gene** by double-clicking READING-FRAMES-OF in:
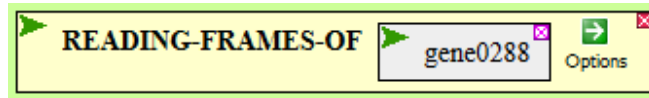

What is the length of the DNA sequence shown? How does that compare to the length of the amino acid sequence of the rIIB protein you found in **3a**?

3c. How many reading (translation) frames are displayed? Why that many? Which one gives an amino acid sequence that matches the amino acid sequence of the rIIB protein?

3d. What is the significance of the <u>second</u> reading frame? Hint: Note that the M of the first reading frame lies directly under the A of the first triplet. Where does the C of the second reading frame lie?

3e. Make an insertion mutant of rIIB, placing an extra A at position 15 of the gene's nucleotide sequence. Here's how. From the **All** menu, bring down the INSERT function. Drag the SEQUENCE-OF function into the *target* box and edit the entity within SEQUENCE-OF so that it's the gene `gene0288` (not the protein `p-gene0288`). Into the *insert-stuff* box, type "A" and press the Enter key. Finally, mouse over the Options icon, select AFTER, and then type 15. Close the box, producing:
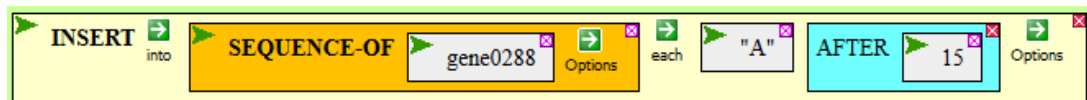
3e1. Execute the function. How does the resulting DNA sequence (in the result pane) compare with the sequence of `gene0288` you obtained from READING-FRAMES-OF?

Now bring down from the **All** menu the TRANSLATION-OF function. Drag the INSERT function (and all that's in it) into the *entity* box of TRANSLATION-OF.
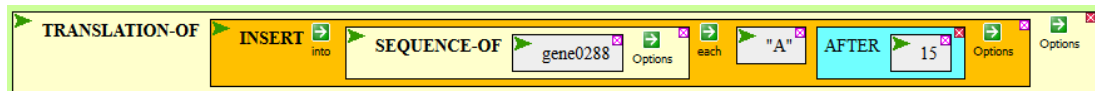
3e2. Execute the function. How does the resulting amino acid sequence (in the result pane) compare with the amino acid sequences in the different reading frames you found in **3b-3d**?

3e3. What is the significance of reading frame #2 relative to reading frame #1? Of reading frame #3 relative to reading frame #1?

3f. What is the significance of the asterisks (*) that occasionally appear in the amino acid sequences of some reading frames? (check to make sure your guess is correct) Do any appear in the first reading frame? Why not?

3g. The positions of the FC0 mutation and the FC9 mutation are very close together, according to Fig. 2 of Crick et al (1961), but the mutations that can suppress FC0 extend into B1b1 but no further, while the mutations that can suppress FC9 extend into B2. Considering your answers to the above questions (particularly **3a**, **3e2**, and **3f**), why were

no suppressors of FC0 found in segments B1b2 and B2? (and is FC0 an insertion or a deletion)?

4. Use the simulation of Crick et al (1961) to recreate their experiment

   4a. Create mutants that suppress FC0. List the steps you took, including the number of plaques you got at each step, the map positions of each mutant, and the phenotypes of each.

If you have the time and inclination:

   4b. [Not easy!] Use the mutants created in 4a to make double mutants and finally a wild-type triple mutant. Again, list the steps you took, including the number of plaques you got at each step, the map positions of each mutant, and the phenotypes of each.

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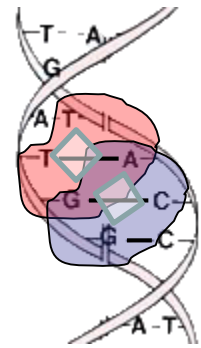
6a. Make up a genetic code that satisfies Gamow's criteria. The box shown below should be helpful. I suggest that you first make some arbitrary assignment (i.e. one of the 64 codons codes for some arbitrary amino acid), determine what other codons must code for the same amino acid, and then proceed along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

```
     T             G             T             G
  T—A          T—A          A—T          A—T
     G             T             G             T
```
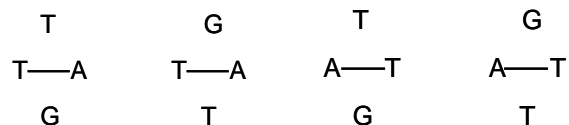
codon	aa	codon	aa	codon	aa	codon	aa
TTT		TCT		TAT		TGT	
TTC		TCC		TAC		TGC	
TTA		TCA		TAA		TGA	
TTG		TCG		TAG		TGG	
CTT		CCT		CAT		CGT	
CTC		CCC		CAC		CGC	
CTA		CCA		CAA		CGA	
CTG		CCG		CAG		CGG	
ATT		ACT		AAT		AGT	
ATC		ACC		AAC		AGC	
ATA		ACA		AAA		AGA	
ATG		ACG		AAG		AGG	
GTT		GCT		GAT		GGT	
GTC		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.

7. 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.