(12 October 2018)

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?

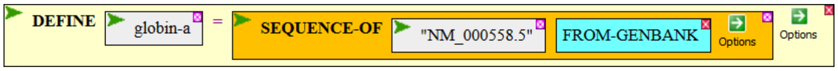
2b. How many if the code were overlapping triplet?

2c. Partially overlapping triplet?

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

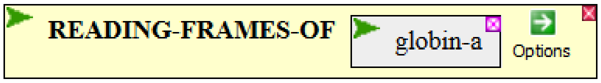
3. Fanconi anemia is an inherited disorder that leads to a variety of developmental defects and has many genetic causes. In one case, symptoms of Fanconi anemia were attributed to a deletion of a nucleotide within the gene encoding alpha globin, leading to a form of hemoglobin known as Hemoglobin Wayne.[1] In this form, the end of the amino acid sequence of alpha globin is markedly different from normal: instead of Val-Ala-Ser-Lys… (VASK…) Hemoglobin Wayne has Val-Ala-Ser-Asn… (VASN…). Your job is to seek a detailed molecular understanding of this mutation.

3a. Download the sequence of alpha globin mRNA from GenBank (accession number NM_000558.5) into BioBIKE as shown below:


DEFINE globin-a = SEQUENCE-OF "NM_000558.5" FROM-GENBANK Options Options

Use SEQUENCE-OF globin-a to see what you have (it displays U's as T's – do the mental substitution yourself). How long is the sequence? Can detect where the gene begins and ends?

3b. You probably couldn't find the beginning and end of the gene purely by inspection. The function READING-FRAMES-OF can help. Use it to display the mRNA (with T's instead of U's) in all possible reading frames as shown below:



READING-FRAMES-OF globin-a Options

3c. Explore. How long is the nucleotide sequence shown? How does it relate to the sequence you obtained in **3a**? But there are two nucleotide sequences for each cluster of lines. What is the second on (the one in the middle) about? How does it relate to the one on the top line?

3d. How many reading (translation) frames are displayed? Why that many? What is the significance of the first "T" in Translation Frame 1? How does relate to a nucleotide sequence?

3e. What is the significance of the first "E" in Translation Frame 4? How does relate to a nucleotide sequence?

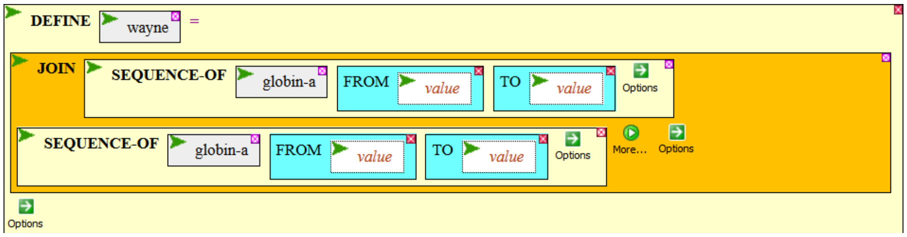
3f. Without looking too hard (yet), which of the translation frames might encode alpha globin? Which could not possibly encode alpha globin?

3g. Now go and get the amino acid sequence of wild-type human alpha-globin. You learned how to do this a couple of weeks ago. Where in the output of READING-FRAMES-OF globin-a do you find the alpha-globin protein sequence? Give coordinates.

3h. What is the position of the wild-type amino acid sequence of alpha-globin that differs from Hemoglobin Wayne? How do the lengths of the mutant and wild-type alpha-globins differ?

3i. What nucleotide was deleted to produce Hemoglobin Wayne (give its letter and coordinate)?

3j. (If you're so inclined) Construct the mRNA for Hemoglobin Wayne in the following way:



DEFINE wayne =
JOIN SEQUENCE-OF globin-a FROM value TO value Options
SEQUENCE-OF globin-a FROM value TO value Options
More... Options
Options

Fill in the first FROM/TO values with the coordinates 1 and the nucleotide before the deletion. Fill in the second FROM/TO values with the coordinates of the nucleotide after the deletion and the last nucleotide of the sequence. Execute the function. Execute READING-FRAMES-OF wayne to see the protein you produced.

4. The table to the right is derived from data from Speyer et al (1963),[2] who performed experiments very similar to those of Jones and Nirenberg (1962). RNA was made using ATP and CTP in the ratio of 5:1 or 1:5, resulting in RNA with randomly distributed A and C in the given proportion. The resulting random RNA polymers were translated in vitro, and the resulting protein analyzed for their amino acid content.

Incorporation of radioactivity directed by random polymer*

Amino acid	A:C=5:1	A:C=1:5
asparagine	1097	71
glutamine	1078	70
histidine	294	315
lysine	4555	14
proline	328	1342
threonine	1206	279

*Each value is the amount of radioactive amino acid incorporation in the presence of the random polymer minus incorporation in its absence. Data is from Speyer et al (1963).

4a. Calculate the expected values for each amino acid, given our present knowledge of the genetic code. The total amount of expected incorporation should be set so that it is the same as that observed.

4b. Are the results of Speyer et al (1963) compatible with what is now accepted as the genetic code? If you like, do a chi-squared analysis. Discuss any discrepancies, and imagine scenarios to explain them.

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?

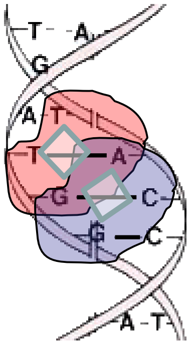
5e. Suppose that the overlapping triplet genetic code we're considering is degenerate, that is more than one triplet may encode the same amino acid. If the dipeptides shown below are found in nature, how many triplets, at minimum, must encode histidine (His)?

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?

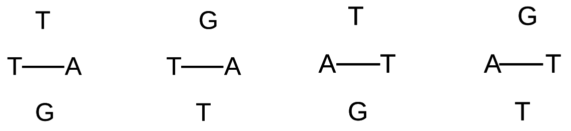
5g. You might enjoy reading Brenner (1957).[3]

6. In 1954, George Gamow published the first attempt to conceive a genetic code. You can read about it in Gamow (1954).[4] 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.

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.

REFERENCES

1. Seid-Akhavan M, Winter WP, Abramson RK, Rucknagel DL (1976). Hemoglobin Wayne: A frameshift mutation detected in human hemoglobin alpha chains. Proc Natl Acad Sci USA 73:882-886.

2. Speyer JF, Lengyel P, Basilio C, Wahba AJ, Gardner RS, Ochoa S (1963). Synthetic polynucleotides and the amino acid code. Cold Spring Harbor Symp Quant Biol 28:559-567.

3. 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.

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