Chapter 4:

Integration Problems

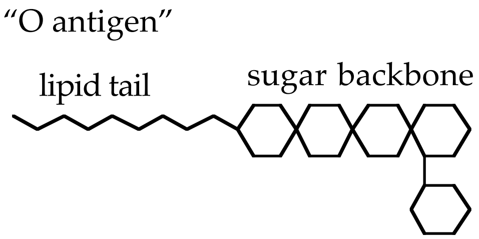
In the previous chapters, we have asked you to think about biological concepts from the view of a geneticist, a biochemist, or a molecular biologist. In this chapter, we offer problems that draw from the ideas found in all three chapters. By relating ideas from these three areas, you will have the chance to practice the familiar steps in a new context. This will provide new insights into the connections between these subject areas and deepen your understanding of these important concepts.

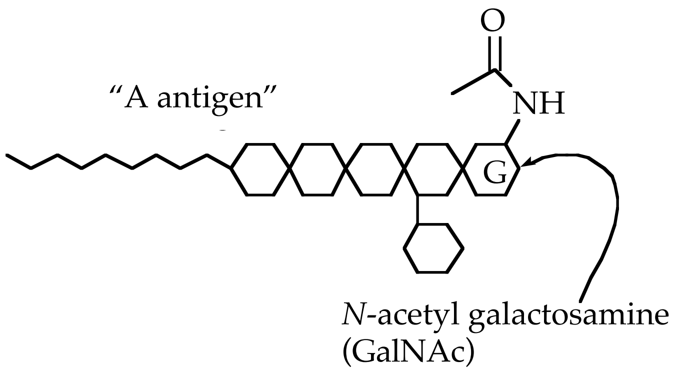
Integration Problems:

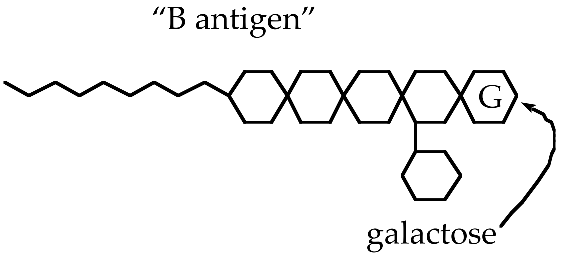
(1) For a description of the inheritance of blood type, see your textbook and section 1.3 in the genetics section of this book.

The genes involved in production of the blood types have been studied extensively. Blood type is determined by one gene with three alleles. This gene encodes an enzyme that is involved in the synthesis of a polysaccharide on the surface of red blood cells. This enzyme is called a glycosyltransferase.

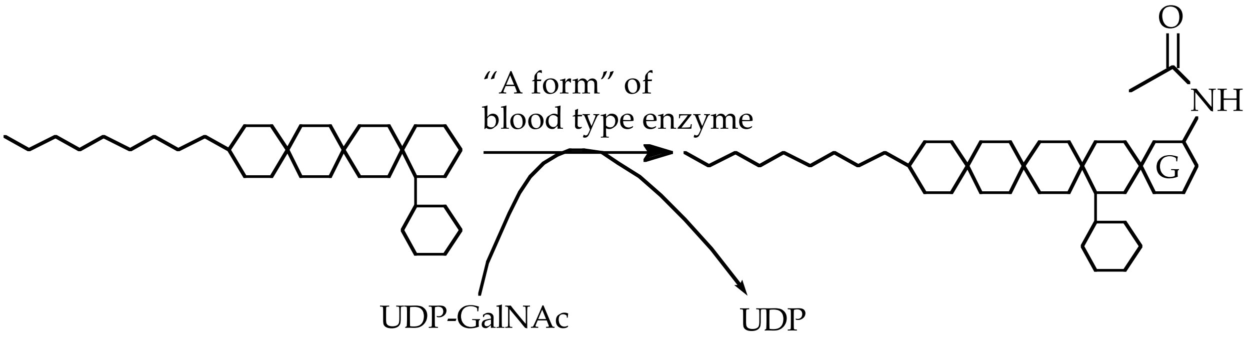
The structures of the blood type antigens (the molecules that the immune system responds to when rejecting blood of an incompatible type) are shown below:





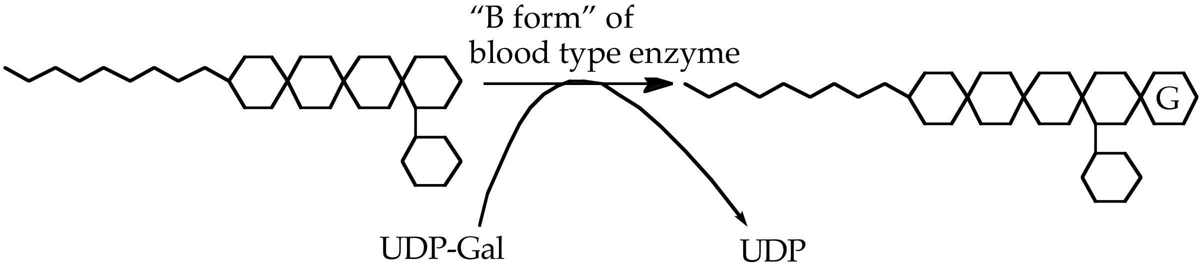


• The I^A allele of the blood type gene encodes a glycosyltransferase enzyme that catalyzes the following reaction:



(Note: UDP is uridine diphosphate, a relative of ADP.)

• The I^B allele of the blood type gene encodes a glycosyltransferase enzyme that catalyzes the following reaction:



• The i allele of the blood type gene encodes a glycosyltransferase enzyme that is inactive.

a) An individual with genotype ii would not have any active glycosyltransferase. Explain in biochemical terms why an ii individual would have type O blood.

b) Explain in biochemical terms why the blood type-A phenotype of the I^A allele and the blood type-B phenotype of the I^B allele are dominant to the blood type-O phenotype of the i allele. That is, why do people with genotypes I^Ai and I^Bi have type A and type B blood (respectively) and not type O blood?

c) Explain in biochemical terms why the blood type-A phenotype of the I^A allele and the blood type-B phenotype of the I^B allele are codominant to each other. That is, why do people with genotype I^AI^B have type AB blood (both A and B) and not A, B, or something else?

The i allele, which confers the recessive phenotype of type O blood, differs from the I^A allele by a frameshift mutation in the coding region of the gene for the blood type-determining enzyme. The DNA sequence of the coding strand (the DNA strand that has the same sequence as the mRNA, except that T's are replaced by U's) in the appropriate region of the I^A and i alleles is shown below:

Sequence of I^A allele: CGTGGTGACCCCTT…

Sequence of i allele: CGTGGTACCCCTT…

The relevant part of the sequence of the protein produced by the I^A and i alleles is shown below (the differences are shown in bold):

Sequence of protein encoded by I^A allele:
84 85 86 87 88 89
H_3N^+...Leu-Val-Val-Thr-Pro-Trp-Leu...COO^-

Sequence of protein encoded by i allele:
H_3N^+...Leu-Val-Val-**Pro-Leu-Gly-Trp...**COO^-

d) Based on the protein sequence data, indicate the reading frame of the DNA sequences above. That is, match the DNA sequences with their respective protein sequences. Note that the beginning of the reading frame must be the same in both sequences, starting from the left.

The I^A and I^B alleles differ by several point mutations, resulting in four amino acid changes in the encoded proteins. These changes are listed below:

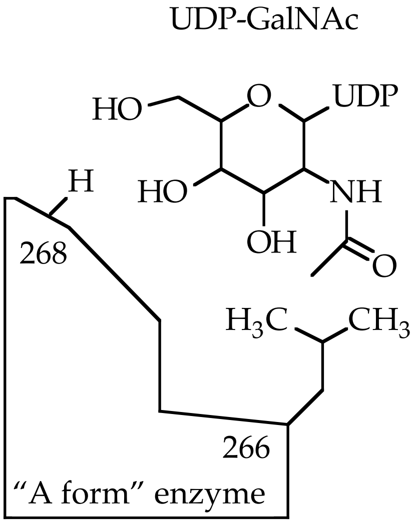
Position in	Amino Acid	Amino Acid
Polypeptide Chain	in I^A Allele	in I^B Allele
176	Arg	Gly
235	Gly	Ser
266	Leu	Met
268	Gly	Ala

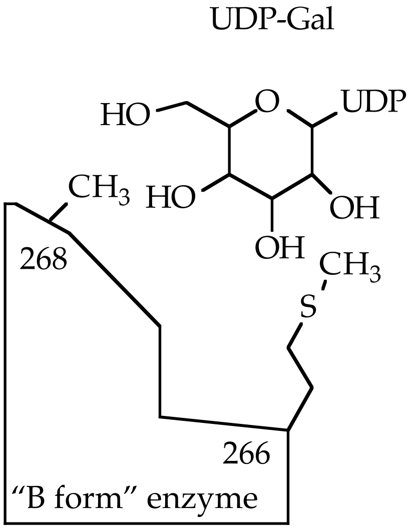
The DNA sequence of the coding strand in the region which encodes amino acids 266 and 268 of the I^A and I^B alleles is shown below (differences shown in **bold underlined** type):

Sequence of I^A allele: ...ACTACCTGGGGGGTTCTT...
Sequence of I^B allele: ...ACTAC**A**TGGGGG**C**GTTCTT...

e) Based on the mutation data, indicate the reading frame of the DNA sequences above.

f) Although the A and B glycosyltransferases differ at four places, only two of these contribute to their substrate specificity (the other two contribute only slightly to the substrate specificity). The structures of the active sites of the A and B forms of the blood type glycosyltransferase are shown below with their respective sugar substrates.





Based on these figures, explain how the different forms of the glycosyltransferase have different substrate specificities.

g) There are many rare i alleles known in the human population. For each of these mutations, provide a plausible explanation for why it would encode an inactive glycosyltransferase.

i) A mutation that changes amino acid 268 from Gly to Arg.

ii) A mutation that changes amino acid 309 from Tyr to a stop codon.

h) There is a rare allele at the blood type locus called *cis*-AB. The DNA sequence of this allele is intermediate between the I^A and I^B alleles; it contains some of the features of both. As a result, the enzyme catalyzes the reaction of the "O antigen" with **either** UDP-Gal or UDP-GalNAc. This produces the AB blood type.

i) Explain in terms of enzyme structure and function, how the changes in protein sequence in the protein encoded by the *cis*-AB allele described above could lead to the biochemical phenotype described above.

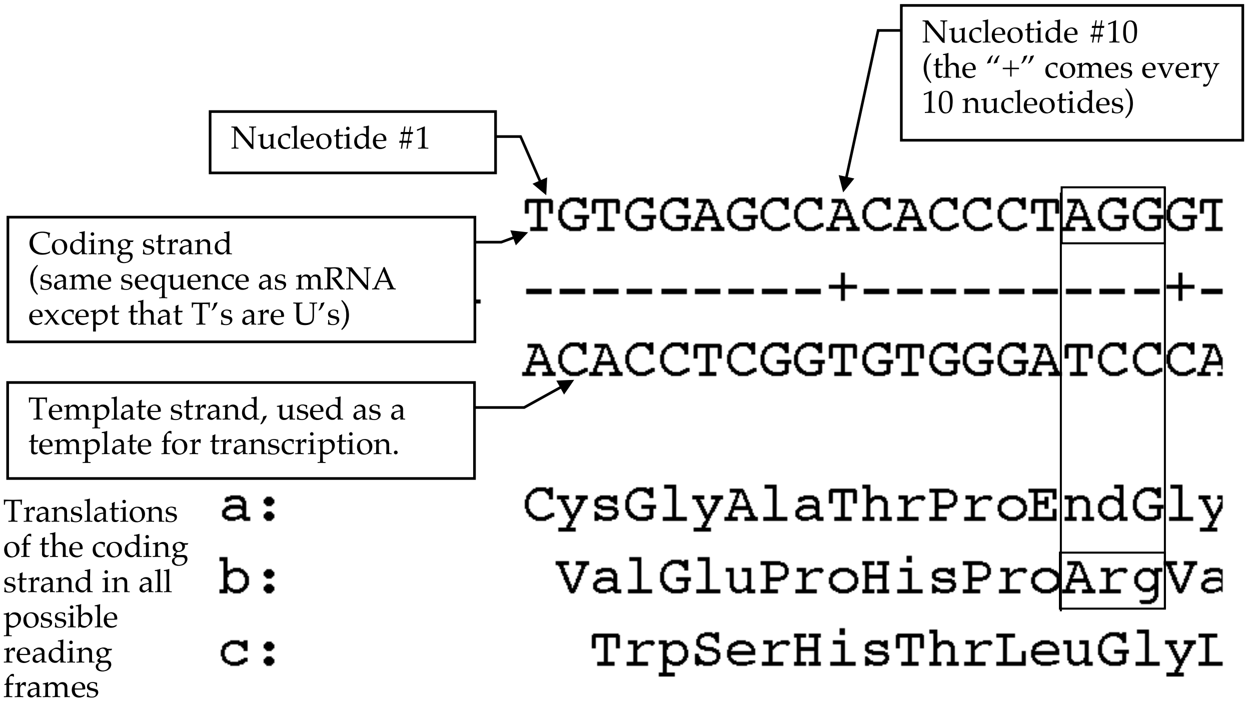
ii) Suppose that you are a researcher studying blood type. How would you tell if someone had the *cis*-AB allele (as opposed to having just the usual AB genotype I^AI^B)? Remember that, since you're dealing with people, you can't do crosses; you can look only at family histories (pedigrees). In other words, what blood type pedigree would indicate the presence of a *cis*-AB allele? Explain your reasoning.

(2) This problem applies genetics, biochemistry, and molecular biology to the protein hemoglobin, the protein that carries oxygen in the blood of humans. You will be given the DNA sequence of the β-globin gene from humans. This gene is located on chromosome 11. The DNA sequence includes the promoter, coding region, introns, exons, terminator, and so forth. First, you will use the protein and DNA sequences to draw a map of the major structural features of the β-globin gene. You will then be given specific mutations that have been found in this gene and will be asked to explain the effects of these mutations based on your knowledge of genetics, biochemistry, and molecular biology.

You will use the following tools as you see fit:

- Table of the genetic code. This can be found in your textbook.
- Table of amino acid structures and properties. This can also be found in your textbook.
- The program "Molecules in 3-dimensions" which you used in the Biochemistry chapter of this book to look at molecular structures in three dimensions. Access "Molecules in 3-d" at this site http://intro.bio.umb.edu/MOOC/jsMol/ and click on the link for this problem "Molecular Bio C1", and click on the "Load Hemoglobin and show 4 chains and heme" button. The remaining buttons help you to see the amino acids relevant to the Group B mutations. In each of these views, the indicated amino acid is shown as spheres; the rest of the protein is shown as yellow dots. Consult the Biochemistry chapter of this book to find out more about how to use "Molecules in 3-dimensions."

Here is how to interpret the DNA sequence on the following pages:



Nucleotide #10 (the "+" comes every 10 nucleotides)

Nucleotide #1

Coding strand (same sequence as mRNA except that T's are U's)

Template strand, used as a template for transcription.

Translations of the coding strand in all possible reading frames

```
    TGTGGAGCCACACCCTAGGGT
    ---------+----------+-
    ACACCTCGGTGTGGGATCCCA

a:  CysGlyAlaThrProEndGly
b:   ValGluProHisProArgVa
c:    TrpSerHisThrLeuGlyI
```

Under each line of double-stranded DNA sequence is a translation of the coding strand in all three possible reading frames. This is a convenience to save you the trouble of looking up the codons in the genetic code table. The three possible frames are:

- Frame (a) starts reading at the **<u>first</u>** nucleotide and is therefore

```
read as:          TGT, GGA, GCC, ACA, CCC, TAG….
or in mRNA:       UGU, GGA, GCC, ACA, CCC, UAG….
translated as:    Cys, Gly, Ala, Thr, Pro, End….
                                               ("End" = "STOP")
```

- Frame (b) starts reading at the **<u>second</u>** nucleotide and is therefore

```
read as:          GTG, GAG, CCA, …
or in mRNA:       GUG, GAG, CCA…
translated as:    Val, Glu, Pro, …
```

- Frame (c) starts reading at the **<u>third</u>** nucleotide and is therefore

```
read as:          TGG, AGC, CAC…
or in mRNA:       UGG, AGC, CAC…
translated as:    Trp, Ser, His…
```

Note that the sequences can be lined up in vertical columns. Therefore, if you want to find (for example) the codon and reading frame that correspond to the Arg in frame (b), just draw straight vertical lines up from each side of "Arg" to the codon as shown. This shows that the Arg was encoded by AGG preceded by a CCT in the same frame.

(I) <u>Make a map of the β-globin gene</u>

Using the amino acid sequence of the β-globin protein listed on the next page, make a map of the introns and exons in the β-globin gene. Here are some hints.

- β-globin is first made with a Met at the amino terminus (it starts from an AUG codon); that Met is removed before the protein is put into the red blood cells. The amino acids are numbered from the amino terminus of the mature protein – #1 is Val. (#0 is the starting Met.)
- The mRNA starts with nucleotide 101 and is synthesized 5′ to 3′ from left to right.
- The gene has three exons and two introns. Remember that introns usually start with GU and end with AG.
- The locations of the introns are marked in the protein sequence. Also, the protein sequence in the correct reading frame is underlined and the amino acids are numbered. Note that an intron can, and often does, splice in the middle of a codon.

β-globin protein sequence:

```
0   1   2   3   4   5   6   7   8   9   10  11  12  13  14  15  16  17
Met Val His Leu Thr Pro Glu Glu Lys Ser Ala Val Thr Ala Leu Trp Gly Lys
```

Intron 1 is inserted here.

```
 18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35
Val Asn Val Asp Glu Val Gly Gly Glu Ala Leu Gly Arg Leu Leu Val Val Tyr

 36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53
Pro Trp Thr Gln Arg Phe Phe Glu Ser Phe Gly Asp Leu Ser Thr Pro Asp Ala

 54  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71
Val Met Gly Asn Pro Lys Val Lys Ala His Gly Lys Lys Val Leu Gly Ala Phe

 72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89
Ser Asp Gly Leu Ala His Leu Asp Asn Leu Lys Gly Thr Phe Ala Thr Leu Ser
```

Intron 2 is inserted here.

```
 90  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106 107
Glu Leu His Cys Asp Lys Leu His Val Asp Pro Glu Asn Phe Arg Leu Leu Gly

108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125
Asn Val Leu Val Cys Val Leu Ala His His Phe Gly Lys Glu Phe Thr Pro Pro

126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143
Val Gln Ala Ala Tyr Gln Lys Val Val Ala Gly Val Ala Asn Ala Leu Ala His

144 145 146
Lys Tyr His
```

a) Using the line below, draw the map of the gene encoding this portion of β-globin. Be sure to include:

- Start of mRNA
- Start and stop codons
- Introns and exons

You can use problem (4.2.4) in the Molecular Biology chapter of this book as a guide to drawing gene maps.

```
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
0                  500                  1000                 1500
```

(II) Looking at mutations

There are mutations that result in the production of abnormal β-globin. Technically, the resulting disease phenotype is "β^0 thalassemia." The "β^0" refers to the complete absence of any β-globin protein. The precursors to red blood cells continue to make α-globin molecules. Unfortunately, in the absence of β-globin, the α-globin molecules stick together in large aggregates that destroy the red blood cells. Individuals with β^0thalassemia thus have no functional red blood cells and must receive frequent blood transfusions to live. β^0 thalassemia is inherited in an autosomal recessive manner.

A list of mutations that result in β^0 thalassemia is given below.

Mutation #	**Location**	**DNA change**	**Context of change**
A1	197	G ⇒ A	GT**G**GG ⇒ GT**A**GG
A2	202	A ⇒ T	GC**A**AG ⇒ GC**T**AG
A3	398	C ⇒ T	CC**C**AG ⇒ CC**T**AG
A4	170	delete A	TG**A**GG ⇒ TGGG
A5	175,176	delete AA	G**AA**GT ⇒ GGT
A6	176,177	insert G	GAAGT ⇒ GAA**G**GT

b) For each mutant,

- Give the changes in the amino acid sequence that would result from the mutation listed.

- Explain why the alteration in amino acid sequence would cause the resulting β-globin protein to be inactive.

- Explain in molecular terms why the phenotype of β^0 thalassemia is recessive.

There are other mutations that result in hemolytic anemia. Hemolytic anemia translates as "lack of red blood cells (anemia) due to red blood cells breaking (hemolysis)." The red blood cells break open because the abnormal β-globin sticks together in large aggregates that damage the red blood cells. The phenotype of hemolytic anemia is dominant and hemolytic anemia is inherited in an autosomal manner.

A list of these mutations is given below. These change only one amino acid and have varying effects on the function of hemoglobin.

Mutation #	**Location**	**Change**	**Context of change**	**Effect**
B1	1452	G ⇒ C	TG**G**GC ⇒ TG**C**GC	hemolytic anemia
B2	233	C ⇒ A	GG**C**CC ⇒ GG**A**CC	hemolytic anemia
B3	202	A ⇒ G	GC**A**AG ⇒ GC**G**AG	NORMAL HEMOGLOBIN
B4	1464	G ⇒ T	TG**G**TC ⇒ TG**T**TC	hemolytic anemia
B5	471	A ⇒ G	TC**A**TG ⇒ TC**G**TG	hemolytic anemia
B6	479	A ⇒ G	AG**A**AA ⇒ AG**G**AA	hemolytic anemia

c) For each mutant,

- Give the changes in the amino acid sequence that would result from the mutation listed.

- Explain why the alteration in amino acid sequence would cause the resulting β-globin protein to be inactive.

- Explain in molecular terms why the phenotype of hemolytic anemia is dominant.

The sequence of the gene encoding β-globin follows.

DNA sequence of the β-globin gene

```
     5′  TGTGGAGCCACACCCTAGGGTTGGCCAATCTACTCCCAGGAGCAGGGAGG
       1 ---------+---------+---------+---------+---------+ 50
     3′  ACACCTCGGTGTGGGATCCCAACCGGTTAGATGAGGGTCCTCGTCCCTCC

a:       CysGlyAlaThrProEndGlyTrpProIleTyrSerGlnGluGlnGlyGly -
b:        ValGluProHisProArgValGlyGlnSerThrProArgSerArgGluGly-
c:         TrpSerHisThrLeuGlyLeuAlaAsnLeuLeuProGlyAlaGlyArg  -



         GCAGGAGCCAGGGCTGGGCATAAAAGTCAGGGCAGAGCCATCTATTGCTT
      51 ---------+---------+---------+---------+---------+100
         CGTCCTCGGTCCCGACCCGTATTTTCAGTCCCGTCTCGGTAGATAACGAA

a:        GlnGluProGlyLeuGlyIleLysValArgAlaGluProSerIleAlaTyr-
b:         ArgSerGlnGlyTrpAlaEndLysSerGlyGlnSerHisLeuLeuLeu  -
c:       AlaGlyAlaArgAlaGlyHisLysSerGlnGlyArgAlaIleTyrCysLeu -


          →     start of mRNA
         ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACC
     101 ---------+---------+---------+---------+---------+150
         TGTAAACGAAGACTGTGTTGACACAAGTGATCGTTGGAGTTTGTCTGTGG

a:         IleCysPheEndHisAsnCysValHisEndGlnProGlnThrAspThr  -
b:       ThrPheAlaSerAspThrThrValPheThrSerAsnLeuLysGlnThrPro -
c:        HisLeuLeuLeuThrGlnLeuCysSerLeuAlaThrSerAsnArgHisHis-



         ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGG
     151 ---------+---------+---------+---------+---------+200
         TACCACGTGGACTGAGGACTCCTCTTCAGACGGCAATGACGGGACACCCC
          0                             10
a:       METValHisLeuThrProGluGluLysSerALAValThrAlaLeuTrpGly -
b:        TrpCysThrEndLeuLeuArgArgSerLeuProLeuLeuProCysGlyAla-
c:         GlyAlaProAspSerEndGlyGluValCysArgTyrCysProValGly  -



         CAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTAT
     201 ---------+---------+---------+---------+---------+250
         GTTCCACTTGCACCTACTTCAACCACCACTCCGGGACCCGTCCAACCATA
                    20                            30
a:        LysValAsnVALAspGluValGlyGlyGluAlaLeuGlyARGLeuValSer-
b:         ArgEndThrTrpMetLysLeuValValArgProTrpAlaGlyTrpTyr  -
c:       GlnGlyGluArgGlyEndSerTrpTrpEndGlyProGlyGlnValGlyIle -
```

```
         CAAGGTTACAAGACAGGTTTAAGGAGACCAATAGAAACTGGGCATGTGGA
     251 ---------+---------+---------+---------+---------+300
         GTTCCAATGTTCTGTCCAAATTCCTCTGGTTATCTTTGACCCGTACACCT

a:         ArgLeuGlnAspArgPheLysGluThrAsnArgAsnTrpAlaCysGly  -
b:       GlnGlyTyrLysThrGlyLeuArgArgProIleGluThrGlyHisValGlu -
c:        LysValThrArgGlnValEndGlyAspGlnEndLysLeuGlyMetTrpArg-


         GACAGAGAAGACTCTTGGGTTTCTGATAGGCACTGACTCTCTCTGCCTAT
     301 ---------+---------+---------+---------+---------+350
         CTGTCTCTTCTGAGAACCCAAAGACTATCCGTGACTGAGAGAGACGGATA

a:       AspArgGluAspSerTrpValSerAspArgHisEndLeuSerLeuProIle -
b:        ThrGluLysThrLeuGlyPheLeuIleGlyThrAspSerLeuCysLeuLeu-
c:         GlnArgArgLeuLeuGlyPheEndEndAlaLeuThrLeuSerAlaTyr  -


         TGGTCTATTTTCCCACCCTTAGGCTGCTGGTGGTCTACCTTTGGACCCAG
     351 ---------+---------+---------+---------+---------+400
         ACCAGATAAAAGGGTGGGAATCCGACGACCACCAGATGGAAACCTGGGTC
                                 31
a:        GlyLeuPheSerHisProEndAlaAlaGlyGlyLeuProLeuAspProGlu-
b:         ValTyrPheProThrLeuArgLEULeuValValTyrProTrpThrGln  -
c:       TrpSerIlePheProProLeuGlyCysTrpTrpSerThrProGlyProArg -


         AGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGG
     401 ---------+---------+---------+---------+---------+450
         TCCAAGAAACTCAGGAAACCCCTAGACAGGTGAGGACTACGACAATACCC
          40                            50
a:         ValLeuEndValLeuTrpGlySerValHisSerEndCysCysTyrGly  -
b:       ARGPhePheGluSerPheGlyAspLeuSerTHRProAspAlaValMetGly -
c:        GlySerLeuSerProLeuGlyIleCysProLeuLeuMetLeuLeuTrpAla-


         CAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTG
     451 ---------+---------+---------+---------+---------+500
         GTTGGGATTCCACTTCCGAGTACCGTTCTTTCACGAGCCACGGAAATCAC
                    60                             70
a:       GlnProEndGlyGluGlySerTrpGlnGluSerAlaArgCysLeuEndEnd -
b:        AsnProLysVALLysAlaHisGlyLysLysValLeuGlyALAPheSerAsp-
c:         ThrLeuArgEndArgLeuMetAlaArgLysCysSerValProLeuVal  -
```

```
            ATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGT
        501 ---------+---------+---------+---------+---------+550
            TACCGGACCGAGTGGACCTGTTGGAGTTCCCGTGGAAACGGTGTGACTCA
                                  80
a:           TrpProGlySerProGlyGlnProGlnGlyHisLeuCysHisThrGluEnd-
b:            GlyLeuAlaHisLeuAspASNLeuLysGlyThrPheAlaThrLeuSer  -
c:          MetAlaTrpLeuThrTrpThrThrSerArgAlaProLeuProHisEndVal -


            GAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGGTGAG
        551 ---------+---------+---------+---------+---------+600
            CTCGACGTGACACTGTTCGACGTGCACCTAGGACTCTTGAAGTCCCACTC
             90                            100         104
a:            AlaAlaLeuEndGlnAlaAlaArgGlySerEndGluLeuGlnGlyGlu  -
b:          GLULeuHisCysAspLysLeuHisValAspPROGluAsnPheARGValSer -
c:           SerCysThrValThrSerCysThrTrpIleLeuArgThrSerGlyEndVal-


            TCTATGGGACCCTTGATGTTTTCTTTCCCCTTCTTTTCTATGGTTAAGTT
        601 ---------+---------+---------+---------+---------+650
            AGATACCCTGGGAACTACAAAAGAAAGGGGAAGAAAAGATACCAATTCAA

a:          SerMetGlyProLeuMetPheSerPheProPhePheSerMetValLysPhe -
b:           LeuTrpAspProEndCysPheLeuSerProSerPheLeuTrpLeuSerSer-
c:            TyrGlyThrLeuAspValPhePheProLeuLeuPheTyrGlyEndVal  -


            CATGTCATAGGAAGGGGAGAAGTAACAGGGTACAGTTTAGAATGGGAAAC
        651 ---------+---------+---------+---------+---------+700
            GTACAGTATCCTTCCCCTCTTCATTGTCCCATGTCAAATCTTACCCTTTG

a:           MetSerEndGluGlyGluLysEndGlnGlyThrValEndAsnGlyLysGln-
b:            CysHisArgLysGlyArgSerAsnArgValGlnPheArgMetGlyAsn  -
c:          HisValIleGlyArgGlyGluValThrGlyTyrSerLeuGluTrpGluThr -


            AGAUGAATGATTGCATCAGTGTGGAAGTCTCAGGATCGTTTTAGTTTCTT
        701 ---------+---------+---------+---------+---------+750
            TCTACTTACTAACGTAGTCACACCTTCAGAGTCCTAGCAAAATCAAAGAA

a:            MetAsnAspCysIleSerValGluValSerGlySerPheEndPheLeu  -
b:          ArgEndMetIleAlaSerValTrpLysSerGlnAspArgPheSerPhePhe -
c:           AspGluEndLeuHisGlnCysGlySerLeuArgIleValLeuValSerPhe-
```

```
          TTATTTGCTGTTCATAACAATTGTTTTCTTTTGTTTAATTCTTGCTTTCT
      751 ---------+---------+---------+---------+---------+800
          AATAAACGACAAGTATTGTTAACAAAAGAAAACAAATTAAGAACGAAAGA

a:        LeuPheAlaValHisAsnAsnCysPheLeuLeuPheAsnSerCysPheLeu -
b:         TyrLeuLeuPheIleThrIleValPhePheCysLeuIleLeuAlaPhePhe-
c:          IleCysCysSerEndGlnLeuPheSerPheValEndPheLeuLeuSer  -


          TTTTTTTTCTTCTCCGCAATTTTTACTATTATACTTAATGCCTTAACATT
      801 ---------+---------+---------+---------+---------+850
          AAAAAAAAGAAGAGGCGTTAAAAATGATAATATGAATTACGGAATTGTAA

a:         PhePheSerSerProGlnPheLeuLeuLeuTyrLeuMetProEndHisCys-
b:          PhePheLeuLeuArgAsnPheTyrTyrTyrThrEndCysLeuAsnIle  -
c:        PhePhePhePheSerAlaIlePheThrIleIleLeuAsnAlaLeuThrLeu -


          GTGTATAACAAAAGGAAATATCTCTGAGATACATTAAGTAACTTAAAAAA
      851 ---------+---------+---------+---------+---------+900
          CACATATTGTTTTCCTTTATAGAGACTCTATGTAATTCATTGAATTTTTT

a:          ValEndGlnLysGluIleSerLeuArgTyrIleLysEndLeuLysLys  -
b:        ValTyrAsnLysArgLysTyrLeuEndAspThrLeuSerAsnLeuLysLys -
c:         CysIleThrLysGlyAsnIleSerGluIleHisEndValThrEndLysLys-


          AAACTTTACACAGTCTGCCTAGTACATTACTATTTGGAATATGTGTGTGC
      901 ---------+---------+---------+---------+---------+950
          TTTGAAATGTGTCAGACGGATCATGTAATGATAAACCTTATACACACACG

a:        LysLeuTyrThrValCysLeuValHisTyrTyrLeuGluTyrValCysAla -
b:         AsnPheThrGlnSerAlaEndTyrIleThrIleTrpAsnMetCysValLeu-
c:          ThrLeuHisSerLeuProSerThrLeuLeuPheGlyIleCysValCys  -


          TTATTTGCATATTCATAATCTCCCTACTTTATTTTCTTTTATTTTTAATT
      951 ---------+---------+---------+---------+---------1000
          AATAAACGTATAAGTATTAGAGGGATGAAATAAAAGAAAATAAAAATTAA

a:         TyrLeuHisIleHisAsnLeuProThrLeuPheSerPheIlePheAsnEnd-
b:          IleCysIlePheIleIleSerLeuLeuTyrPheLeuLeuPheLeuIle  -
c:        LeuPheAlaTyrSerEndSerProTyrPheIlePhePheTyrPheEndLeu -
```

```
         GATACATAATCATTATACATATTTATGGGTTAAAGTGTAATGTTTTAATA
    1001 ---------+---------+---------+---------+---------1050
         CTATGTATTAGTAATATGTATAAATACCCAATTTCACATTACAAAATTAT

a:         TyrIleIleIleIleHisIleTyrGlyLeuLysCysAsnValLeuIle  -
b:       AspThrEndSerLeuTyrIlePheMetGlyEndSerValMetPheEndTyr -
c:        IleHisAsnHisTyrThrTyrLeuTrpValLysValEndCysPheAsnMet-


         TGTGTACACATATTGACCAAATCAGGGTAATTTTGCATTTGTAATTTTAA
    1051 ---------+---------+---------+---------+---------1100
         ACACATGTGTATAACTGGTTTAGTCCCATTAAAACGTAAACATTAAAATT

a:       CysValHisIleLeuThrLysSerGlyEndPheCysIleCysAsnPheLys -
b:        ValTyrThrTyrEndProAsnGlnGlyAsnPheAlaPheValIleLeuLys-
c:         CysThrHisIleAspGlnIleArgValIleLeuHisLeuEndPheEnd  -


         AAAATGCTTTCTTCTTTTAATATACTTTTTTGTTTATCTTATTTCTAATA
    1101 ---------+---------+---------+---------+---------1150
         TTTTACGAAAGAAGAAAATTATATGAAAAAACAAATAGAATAAAGATTAT

a:        LysCysPheLeuLeuLeuIleTyrPhePheValTyrLeuIleSerAsnThr-
b:         AsnAlaPhePhePheEndTyrThrPheLeuPheIleLeuPheLeuIle  -
c:       LysMetLeuSerSerPheAsnIleLeuPheCysLeuSerTyrPheEndTyr -


         CTTTCCCTAATCTCTTTCTTTCAGGGCAATAATGATACAATGTATCATGC
    1151 ---------+---------+---------+---------+---------1200
         GAAAGGGATTAGAGAAAGAAAGTCCCGTTATTACTATGTTACATAGTACG

a:         PheProAsnLeuPheLeuSerGlyGlnEndEndTyrAsnValSerCys  -
b:       LeuSerLeuIleSerPhePheGlnGlyAsnAsnAspThrMetTyrHisAla -
c:        PheProEndSerLeuSerPheArgAlaIleMetIleGlnCysIleMetPro-


         CTCTTTGCACCATTCTAAAGAATAACAGTGATAATTTCTGGGTTAAGGCA
    1201 ---------+---------+---------+---------+---------1250
         GAGAAACGTGGTAAGATTTCTTATTGTCACTATTAAAGACCCAATTCCGT

a:       LeuPheAlaProPheEndArgIleThrValIleIleSerGlyLeuArgGln -
b:        SerLeuHisHisSerLysGluEndGlnEndEndPheLeuGlyEndGlySer-
c:         LeuCysThrIleLeuLysAsnAsnSerAspAsnPheTrpValLysAla  -
```

```
         GTAGCAATATTTCTGCATATAAATATTTCTGCATATAAATTGTAACTGAT
    1251 ---------+---------+---------+---------+---------1300
         CATCGTTATAAAGACGTATATTTATAAAGACGTATATTTAACATTGACTA

a:        EndGlnTyrPheCysIleEndIlePheLeuHisIleAsnCysAsnEndCys-
b:         SerAsnIleSerAlaTyrLysTyrPheCysIleEndIleValThrAsp  -
c:       ValAlaIlePheLeuHisIleAsnIleSerAlaTyrLysLeuEndLeuMet -


         GTAAGAGGTTTCATATTGCTAATAGUAGCTACAATCCAGCTACCATTCTG
    1301 ---------+---------+---------+---------+---------1350
         CATTCTCCAAAGTATAACGATTATCATCGATGTTAGGTCGATGGTAAGAC

a:         LysArgPheHisIleAlaAsnSerSerTyrAsnProAlaThrIleLeu  -
b:       ValArgGlyPheIleLeuLeuIleValAlaThrIleGlnLeuProPheCys -
c:        EndGluValSerTyrCysEndEndEndLeuGlnSerSerTyrHisSerAla-


         CTTTTATTTTATGGTTGGGATAAGGCTGGATTATTCTGAGTCCAAGCTAG
    1351 ---------+---------+---------+---------+---------1400
         GAAAATAAAATACCAACCCTATTCCGACCTAATAAGACTCAGGTTCGATC

a:       LeuLeuPheTyrGlyTrpAspLysAlaGlyLeuPheEndValGlnAlaArg -
b:        PheTyrPheMetValGlyIleArgLeuAspTyrSerGluSerLysLeuGly-
c:         PheIleLeuTrpLeuGlyEndGlyTrpIleIleLeuSerProSerEnd  -


         GCCCTTTTGCTAATCATGTTCATACCTCTTATCTTCCTCCCACAGCTCCT
    1401 ---------+---------+---------+---------+---------1450
         CGGGAAAACGATTAGTACAAGTATGGAGAATAGAAGGAGGGTGTCGAGGA
                                                      105
a:        ProPheCysEndSerCysSerTyrLeuLeuSerSerSerHisSerSerTrp-
b:         ProPheAlaAsnHisValHisThrSerTyrLeuProProThrAlaPro  -
c:       AlaLeuLeuLeuIleMetPheIleProLeuIlePheLeuProGlnLEULeu -


         GGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCA
    1451 ---------+---------+---------+---------+---------1500
         CCCGTTGCACGACCAGACACACGACCGGGTAGTGAAACCGTTTCTTAAGT
                   110                            120
a:         AlaThrCysTrpSerValCysTrpProIleThrLeuAlaLysAsnSer  -
b:       GlyGlnArgAlaGlyLeuCysAlaGlyProSerLeuTrpGlnArgIleHis -
c:        GlyAsnValLEUValCysValLeuAlaHisHisPheGlyLYSGluPheThr-
```

```
         CCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAAT
    1501 ---------+---------+---------+---------+---------1550
         GGGGTGGTCACGTCCGACGGATAGTCTTTCACCACCGACCACACCGATTA
                              130
a:       ProHisGlnCysArgLeuProIleArgLysTrpTrpLeuValTrpLeuMet -
b:        ProThrSerAlaGlyCysLeuSerGluSerGlyGlyTrpCysGlyEndCys-
c:         ProProValGlnAlaAlaTYRGlnLysValValAlaGlyValAlaAsn  -


         GCCCTGGCCCACAAGTATCACTAAGCTCGCTTTCTTGCTGTCCAATTTCT
    1551 ---------+---------+---------+---------+---------1600
         CGGGACCGGGTGTTCATAGTGATTCGAGCGAAAGAACGACAGGTTAAAGA
         140               146
a:        ProTrpProThrSerIleThrLysLeuAlaPheLeuLeuSerAsnPheTyr-
b:         ProGlyProGlnValSerLeuSerSerLeuSerCysCysProIleSer  -
c:       ALALeuAlaHisLysTyrHISEndAlaArgPheLeuAlaValGlnPheLeu -


         ATTAAAGGTTCCTTTGTTCCCTAAGTCCAACTACTAAACTGGGGGATATT
    1601 ---------+---------+---------+---------+---------1650
         TAATTTCCAAGGAAACAAGGGATTCAGGTTGATGATTTGACCCCCTATAA

a:         EndArgPheLeuCysSerLeuSerProThrThrLysLeuGlyAspIle  -
b:       IleLysGlySerPheValProEndValGlnLeuLeuAsnTrpGlyIleLeu -
c:        LeuLysValProLeuPheProLysSerAsnTyrEndThrGlyGlyTyrTyr-


         ATGAAGGGCCTTGAGCATCTGGATTCTGCCTAATAAAAAACATTTATTTT
    1651 ---------+---------+---------+---------+---------1700
         TACTTCCCGGAACTCGTAGACCTAAGACGGATTATTTTTTGTAAATAAAA

a:       MetLysGlyLeuGluHisLeuAspSerAlaEndEndLysThrPheIlePhe -
b:        EndArgAlaLeuSerIleTrpIleLeuProAsnLysLysHisLeuPheSer-
c:         GluGlyProEndAlaSerGlyPheCysLeuIleLysAsnIleTyrPhe  -


         CATTGCAATGATGTATTTAAATTATTTCTGAATATTTTACTAAAAAGGGA
    1701 ---------+---------+---------+---------+---------1750
         GTAACGTTACTACATAAATTTAATAAAGACTTATAAAATGATTTTTCCCT

a:        IleAlaMetMetTyrLeuAsnTyrPheEndIlePheTyrEndLysGlyAsn-
b:         LeuGlnEndCysIleEndIleIleSerGluTyrPheThrLysLysGly  -
c:       HisCysAsnAspValPheLysLeuPheLeuAsnIleLeuLeuLysArgGlu -
```

```
         ATGTGGGAGGTCAGTGCATTTAAAACATAAAGAAATGATGAGCTGTTCAA
    1751 ---------+---------+---------+---------+---------1800
         TACACCCTCCAGTCACGTAAATTTTGTATTTCTTTACTACTCGACAAGTT

a:         ValGlyGlyGlnCysIleEndAsnIleLysLysEndEndAlaValGln  -
b:       MetTrpGluValSerAlaPheLysThrEndArgAsnAspGluLeuPheLys -
c:        CysGlyArgSerValHisLeuLysHisLysGluMetMetSerCysSerAsn-



         ACCTTGGGAAAATACACTAT   3′
    1801 ---------+---------+ 1820
         TGGAACCCTTTTATGTGATA   5′

a:       ThrLeuGlyLysTyrThr     -
b:        ProTrpGluAsnThrLeu    -
c:         LeuGlyLysIleHisTyr   -
```

(3) In a fascinating and comprehensive study, Steward et al. (*Trends in Genetics* **19**[9]**:** 505-513 [2003]) looked at 5,686 different missense mutations, each of which led to an inheritable disease found in humans. They classified each mutation by the change in the amino acid sequence that resulted from the mutation, for example, Lys to Arg. Since there are 20 possible starting amino acids and, for each of them, there are 19 possible amino acids that they could be mutated to, there are 20 x 19 or 380 different types of missense mutations possible. They then determined how many of the 5,686 mutations fell into each category. Some types of mutation were relatively common, while others were relatively rare.

The frequency with which a given type of mutation leads to disease depends on two factors:

- How likely it is that a mutation could lead to that change
- How damaging that mutational change would be to the protein

The chance that a random mutation could lead to a particular change depends on the genetic code; for example, changing Gly (GGG) to Arg (AGG) requires only one base to be mutated, while Phe (UUU) to Asn (AAU) requires two bases to be mutated. Since changing two bases is much less likely than changing one, Gly to Arg mutations will occur more often than Phe to Asn mutations.

In fact, mutations are not completely random. In humans, it turns out that certain mutations occur more frequently than others. In humans, the C bases in CG sequences are sometimes modified by the addition of a methyl group. At a low frequency, these methyl-C's undergo spontaneous deamination and become T's. If this is not properly repaired, the GC sequence can become a CA or a TG sequence depending on which strand the methyl-C was in. Other mutations are known to occur at slightly lower frequencies.

The degree of damage that a particular amino acid change would do to a given protein depends on the properties of different amino acid side chains and their interactions as they influence protein structure and function. Remember that for a mutation to result in a genetic disease, it must have a substantial (usually negative) effect on the protein's function.

Using these factors, explain the following observations.

a) The most common type of mutation (229 of the 5,686 total) is Arg to Cys. Why would you expect this to be so frequent?

b) Another frequent type is Arg to Trp (197 of 5,686). Why would you expect this to be so frequent?

c) Another frequent type is Arg to His (217 of 5,686). Why is it surprising that this is so frequent?

d) The mutation Val to Pro was never observed in their set of 5,686 mutations. Why would you expect that it would be very infrequent?

e) The mutation Leu to Ile was very infrequent (3 of 5,686). Why would you expect that this would be rare?

f) The mutation Gly to Phe was never observed in their set of 5,686 mutations. Why would you expect that this mutation would be very infrequent?

(4) Below is the DNA sequence of the first part of a hypothetical gene. The promoter is underlined and transcription begins at and includes the bold G/C base pair.

```
5′ TACAC GCTTA GCTGA CTATA AGGAC GAATC GCTAC AACGA TGCGA-
   ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||
3′ ATGTG CGAAT CGACT GATAT TCCTG CTTAG CGATG TTGCT ACGCT-


  -TGCCA TCCGA TTGGT GTTCC TTCCA TGAAG GATGC ACAAC GCAAA 3′
   ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||
  -ACGGT AGGCT AACCA CAAGG AAGGT ACTTC CTACG TGTTG CGTTT 5′
```

a) What are the first 12 nucleotides of the transcript encoded by this gene? Label the 5′ and 3′ ends.

b) On the DNA sequence above, **circle** the DNA bases that encode the first amino acid of the protein.

c) What are the first four amino acids encoded by this gene? Label the N and C termini.

d) You want to create a system to translate a specific mRNA in a test tube. To an appropriate water and salt solution you add many copies of this mRNA and ATP. What other key components must you add?

You succeed in translating the mRNA in your test tube. You repeat the experiment with two identical test tubes. You add trace amounts of the antibiotic puromycin to test tube 2 only. Puromycin is a molecule that has structural similarities to the 3′ end of a charged tRNA. It can enter the ribosome and be incorporated into the growing protein. When puromycin is incorporated into the polypeptide, it stalls the ribosome and the polypeptide is released.

e) What effect would puromycin have on transcription?

f) What effect would puromycin have on translation?

g) In principle, there are two possible modes through which puromycin could bind to the ribosome:

A) Puromycin binds to a specific codon in the mRNA and stops translation there.
B) Puromycin does not bind to a specific codon and stops translation wherever the ribosome happens to be when the puromycin binds.

To distinguish between these hypotheses, you set up two test tubes:

- Test tube 1: The same translation mixture you described above.
- Test tube 2: The translation mixture with puromycin added.

You allow translation to occur for a little while and then examine the length of the polypeptide produced in both test tubes.

i) In test tube 1 you get a polypeptide that is 100 amino acids long. At least how many bases long was the complete mRNA that you added?

ii) Suppose that model (A) is correct. What would you expect to find in test tube 2?

- Only a single type of polypeptide.
- Only two types of polypeptides that are each different lengths.
- Only three types of polypeptides that are each different lengths.
- Only four types of polypeptides that are each different lengths.
- Polypeptides of all sizes, that is, dipeptides, tripeptides, … a polypeptide that is 100 amino acids long.

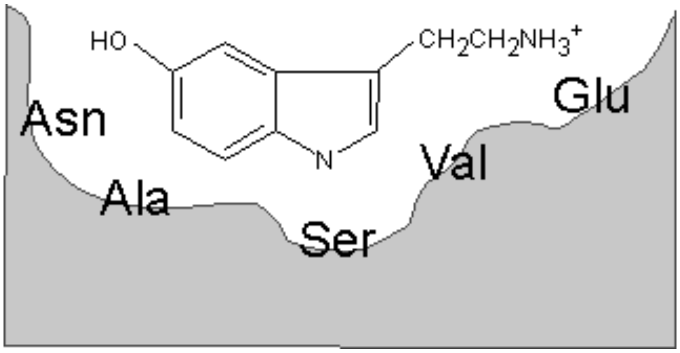
Explain your reasoning.

iii) Suppose that model (B) is correct. What would you expect to find in test tube 2?

- Only a single type of polypeptide.
- Only two types of polypeptides that are each different lengths.
- Only three types of polypeptides that are each different lengths.
- Only four types of polypeptides that are each different lengths.
- Polypeptides of all sizes, that is, dipeptides, tripeptides, … a polypeptide that is 100 amino acids long.

Explain your reasoning.

This gene encodes a protein that binds to the neurotransmitter serotonin, as shown below. The five amino acids involved in binding serotonin are shown.



Below is an internal part of the wild-type DNA sequence and the protein it encodes. The amino acids depicted in the picture above are underlined.

```
DNA         5′...ACC AAT GGA CCA GCA GGA AGC GGG GTA GCT GAG TAC...3′
                 ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
            3′...TGG TTA CCT GGT CGT CCT TCG CCC CAT CGA CTC ATG...5′

Protein     N-...Thr Asn Gly Pro Ala Gly Ser Gly Val Ala Glu Tyr...—C
```

h) You find the following alternative DNA sequence for this protein:

```
5′...ACC AAT GGA CCA GCA GGA TAG CGG GGT AGC TGA GTAC...3′
     ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||||
3′...TGG TTA CCT GGT CGT CCT ATC GCC CCA TCG ACT CATG...5′
```

i) Indicate (circle/underline) the site of the mutation on the sequence directly above.

ii) Does the alternative sequence have an insertion, deletion, or substitution mutation?

iii) Would you expect this DNA sequence to encode a protein that binds serotonin? Why or why not?

i) You find a third DNA sequence for this protein

```
5′...ACC AAT GGA CCA GCA GGA AGC GGG GTA GCT GAT TAC...3′
     ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
3′...TGG TTA CCT GGT CGT CCT TCG CCC CAT CGA CTA ATG...5′
```

i) Indicate (circle/underline) the site of the mutation on the above sequence.

ii) Does this third sequence have an insertion, deletion, or substitution mutation?

iii) Would you expect this DNA sequence to encode a protein that binds serotonin? Why or why not?