

**BNFO 301: Introduction to Bioinformatics**  
**Introduction to Molecular Biology - Problem Set**

Questions relating to **What is a cell? How does it work?**

A1. Which of the following are hydrophobic? Hydrophilic? Amphipathic?

- |                |               |
|----------------|---------------|
| A. vinegar     | D. sugar      |
| B. skin        | E. wax        |
| C. tooth paste | F. rabid dogs |

A2. In general, hydrophilic molecules have a difficult time passing cell membranes unless the cell makes accommodations for them. Presuming there are no such accommodations, which of the following molecules would not easily get into a cell?

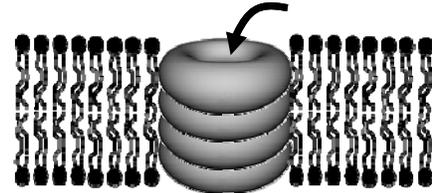
- |           |                |
|-----------|----------------|
| A. sodium | C. ethanol     |
| B. sugar  | D. amino acids |

A3. Consider that at an air-water interface, amphipathic molecules expose their hydrophobic surface to air. Draw a picture of what a soap bubble might look like at the molecular level, using a long-sticked popsicle to represent a molecule of soap.

Questions relating to **Proteins: What are they? What do they do?**

B1. Proteins that bind to DNA are typically rich in arginines and lysines? Why?

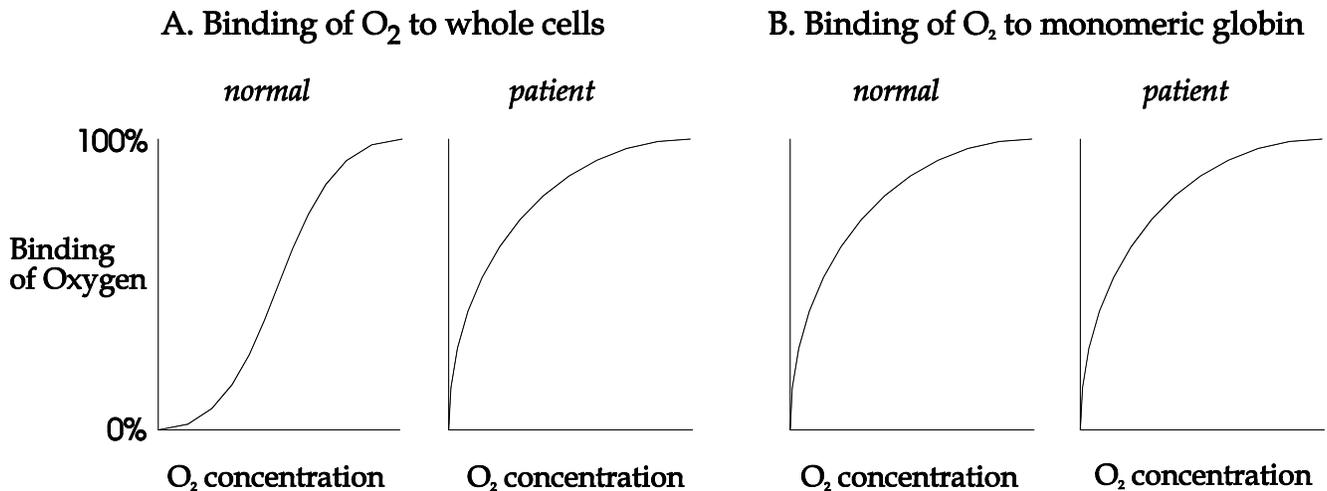
B2. Some antibiotics form rings that stack and create a pore through the membrane. Consider a cyclic polypeptide antibiotic in which each ring is composed of the four amino acids: serine, glycine, threonine, and alanine. If each atom of the backbone is about 2 angstroms in length, estimate the circumference of the pore (presume it to be a circle) and the diameter of a molecule that could fit through it. Approximate the circumference ( $\pi$ -diameter) to be 3-diameter. (Show work)



B3. Peptide hormones can be as small as five amino acids in length, but most protein have polypeptide chains ranging from 70 to 1000 amino acids. Suppose you have some reason to believe that a small protein, precisely 100 amino acids in length, can transform glucose into gold. You set out to synthesize every possible 100-amino acid protein until you find the one you want. How many such proteins may you have to go through? How long would it take?

B4. Before you cook an egg, the egg "white" is not at all white: it's clear. After you cook the egg, the "white" is white, because the large amount of globular protein has denatured (i.e., unfolded), and as a consequence, the protein has precipitated. Why should unfolding globular protein that are normally soluble in water cause them to stick to each other (which is what "precipitate" means)?

- B5. Lactate dehydrogenase (the last enzyme in human anaerobic glycolysis) is a soluble, multimeric protein. If you were to try to fold a single linear polypeptide chain of lactate dehydrogenase, you would find it impossible to do so without leaving a large number of hydrophobic amino acids exposed to water. Explain.
- B6. A patient exhibits signs of anemia. The red cell count is normal, as is the amount of hemoglobin, as judged by the binding of antibody directed against hemoglobin, but the binding of oxygen to whole cells is atypical (Fig. A below). You isolate hemoglobin from the patient and test binding of oxygen to the monomeric globin subunits. It is normal (Fig. B below). What mutation might account for these findings?

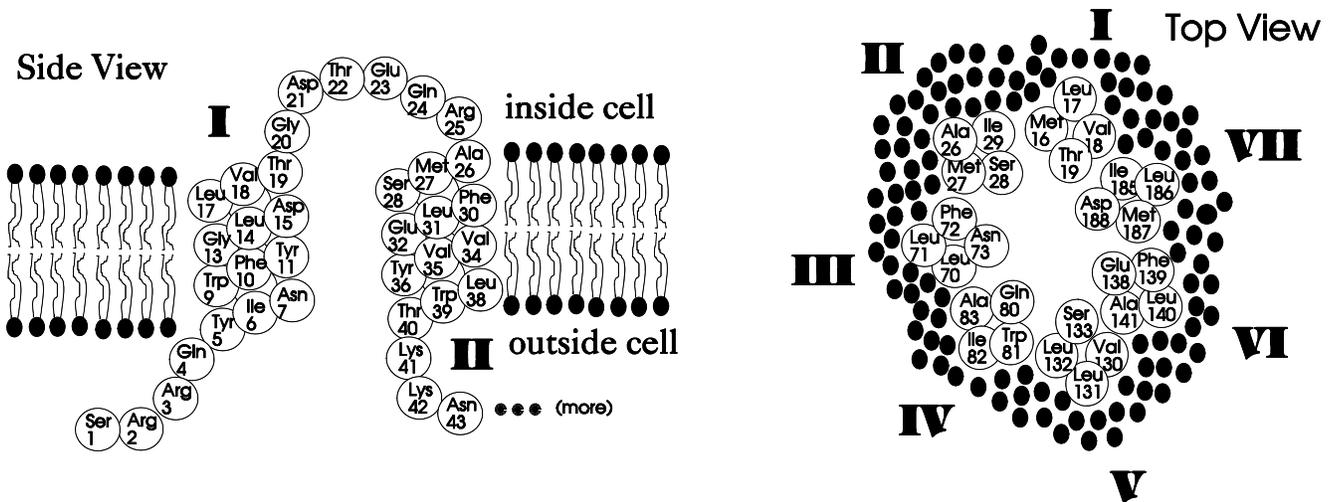


- B7. Of the 19 amino acids of glycophorin that lie within the membrane, some are hydrophilic (see Fig. 6 in notes entitled *Protein*). What do you think may be the significance of this?
- B8. A child presents to you, her pediatrician, with all the classical symptoms of diabetes. Upon testing, you find that antibody against insulin detects only very low levels of insulin in her blood, but she responds normally to administered insulin. You are surprised to find, however, that the same antibody detects levels of insulin in the pancreas that are grossly higher than normal. What mutation might account for these findings?
- B9. An enzyme has a molecular weight of 60,000 daltons. When it is exposed to detergent, the protein breaks up to identical inactive components with molecular weights of 20,000 daltons. If the detergent is removed by dialysis, the 60,000-dalton protein reforms and regains enzymatic activity. You have isolated two mutant proteins. Mutant 1 shows no enzymatic activity and has a molecular weight of 20,000 daltons whether or not detergent is present. Mutant 2 has a molecular weight of 60,000 without detergent and 20,000 with detergent but shows no enzymatic activity in either case.
- Suggest defects to explain the behavior of each of the mutant enzymes.
  - A person is heterozygous for Mutant 2 (i.e., has 50% Mutant 2 enzyme and 50% normal enzyme). How would you explain an observation that the person has 87.5% of the enzymatic activity of a normal person? How would you explain an observation of 12.5% activity?
  - Ascribe the terms "dominant" or "recessive" to the mutation leading to Mutant 2, according to the two situations presented in **b**.

B10. Many proteins that form channels through membranes pass through the membrane multiple times. For example, rhodopsin, the light receptor protein in the rod cells of the retina, passes through the membrane seven times as alpha-helical chains. Below is a cartoon showing the side view of part of a hypothetical channel-forming protein -- call it rhodopsin. The circles are amino acid residues, the number of each corresponding to the amino acid's position in the chain. The roman numerals refer to membrane-spanning alpha-helical segments of the protein (only the first two are shown here). The top view shows how the seven  $\alpha$ -helices participate in the formation of a pore through the membrane. The pore serves as the means by which protons can pass the membrane in response to light.

Congenital retinitis pigmentosa is a genetic disease leading to night-blindness. The disease exhibits a variety of symptoms of different severities, which, in many cases, have been linked to specific mutations in rhodopsin. For each given molecular outcome, choose one or more plausible amino acid mutations that could account for it. In each case, explain, briefly, why your choice(s) would lead to the outcome.

- |  |  |
|--|--|
| <p>a. Rhodopsin found in cytoplasm, fails to insert in membrane.</p> <p>b. Radical change in structure of rhodopsin. Channel doesn't form properly.</p> <p>c. Overall structure of rhodopsin normal, but channel does not conduct protons.</p> <p>d. Structure and function of rhodopsin normal.</p> | <p>A. Insertion of three glutamates between Thr<sub>22</sub> and Glu<sub>23</sub>.</p> <p>B. Insertion of three glutamates between Phe<sub>30</sub> and Leu<sub>31</sub>.</p> <p>C. Glu<sub>138</sub> mutated to arginine.</p> <p>D. Asp<sub>188</sub> mutated to leucine.</p> <p>E. Mutation in amino acid not found in mature rhodopsin.</p> |
|--|--|



Questions relating to **DNA and RNA**

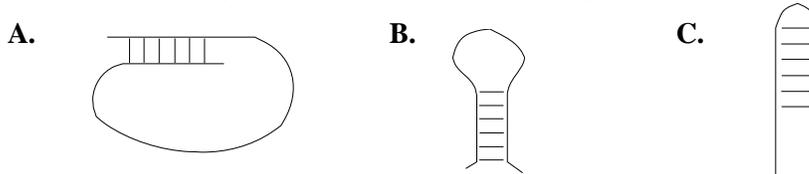
- C1. You are caring for a patient with an unknown viral disease. The virus is known to contain single-stranded DNA. To analyze the virus, you obtain DNA from unaffected cells and from purified virus. Unfortunately, you forget to label the two tubes. The two DNA samples are analyzed for base composition, giving the results shown below. Which tube contains viral DNA, and why?

	<b>G</b>	<b>A</b>	<b>T</b>	<b>C</b>
<b>Tube 1</b>	20.9	28.8	29.0	21.3
<b>Tube 2</b>	21.1	29.1	20.8	29.0

- C2. Consider the sequence below of one strand of a DNA fragment.

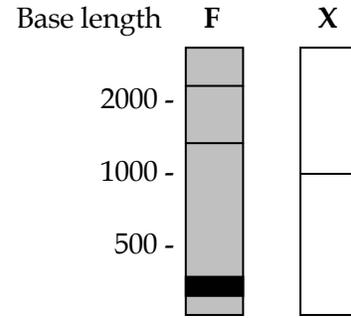
5'-AGAGAGAGCTAAGGTCTCTCC-3'

Which of the following is a likely structure for the single-stranded fragment to assume?



- C3. AIDS is caused by a virus (HIV) that contains single-stranded RNA as its genetic material. Upon infection, the virus first replicates from its RNA a second DNA strand, using the enzyme reverse transcriptase. The ends of the newly synthesized strand are connected, using the enzyme DNA ligase. Then the RNA strand is replaced by DNA, using the enzyme DNA polymerase. The anti-AIDS drug AzaT is an inhibitor of one of the three enzymes I named. Which one do you suppose it is and why?
- C4. You isolate a new restriction enzyme EcoRX (an enzyme that cuts DNA at a specific position). You find that the enzyme cuts the sequence ATGGTATACTGAACGAA once. Like many proteins that recognize DNA, restriction enzymes generally recognize palindromic sequences.
- Presuming that EcoRX does so as well, what is the sequence it recognizes and cuts?
  - About how often would you expect to find this sequence in a random piece of DNA? (form of answer should be “one site every X nucleotides”)
  - What assumptions did you make in determining your answer to Part b.?
  - The cyanobacterium *Anabaena* has a genome 7.2 megabases in size and about 21% of its bases are guanines. Predict the number of EcoRX sites in the *Anabaena* genome.
- C5. You isolate total RNA from *Clostridium botulinum* (the pathogenic agent of botulism) and run it out on a gel under conditions (somewhat different from those in your lab) such that the RNA is fully extended (i.e., the RNA migrates solely on the basis of its length). You stain the gel so that the RNA is visible under fluorescent light and see the gel shown below (F). Then you blot the gel, which means that you cause the RNA to be transferred to filter paper (see text for details) and probe it with a radioactive fragment of DNA containing the gene encoding botulinum toxin (therefore the probe will bind only to RNA carrying the sequence for the toxin). The blot is autoradiographed (exposed to X-ray film) shown as X.

- Why don't the two bands seen by fluorescence appear by autoradiography?
- What are the two bands seen by fluorescence?
- If the botulinum toxin gene encodes a protein 300 amino acids long, what fraction of its mRNA is devoted to that gene? (Show work)



C6. Belozersky and Spirin published a listing of nucleic acid base compositions from a large number of bacteria [(1958) Nature 182:111-112]. Part of the list is shown below. What striking feature is evident in the comparison of DNA and RNA compositions? Given our present knowledge of RNA, how do you account for their findings?

Species	DNA Base Composition					RNA Base Composition				
	Bases (moles per cent)				G+C	Bases (moles per cent)				G+C
	G	A	C	T	A+T	G	A	C	U	A+U
<i>Proteus vulgaris</i>	19.8	30.1	20.7	29.4	0.68	31.0	26.3	24.0	18.7	1.22
<i>Escherichia coli</i>	26.0	23.9	26.2	23.9	1.09	30.7	26.0	24.1	19.2	1.21
<i>Erwinia carotovora</i>	27.1	23.3	26.9	22.7	1.17	29.5	26.5	23.7	20.3	1.14
<i>Mycobacterium vadosum</i>	29.2	20.7	28.5	21.6	1.37	31.7	23.8	23.5	21.0	1.23
<i>Pseudomonas aeruginosa</i>	33.0	16.8	34.0	16.2	2.03	31.6	25.1	23.8	19.5	1.24

Questions relating to **Translation** and the **genetic code**

D1. The table to the right shows some of the original data used to assign amino acids to triplet codons. 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. From the results shown, deduce as much as you can about which triplet codons encode which amino acids.

**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

D2. List the changes that can be produced by a single basepair mutation in the AGA codon encoding arginine and label each silent (no effect on protein structure), conservative (mild effect on protein structure), hydrophobic-to-hydrophilic, hydrophilic-to-hydrophobic, or other.

\*The differences in incorporation of radioactivity between the two experiments is not significant. Each experiment should be considered separately. data is from Speyer et al (1966) Cold Spring Harbor Symposium of Quantitative Biology 31:559-567

D3. It is unfortunately quite common for humans or computers to make errors in determining the sequence of nucleic acids. It is particularly common for a string of G's or C's (e.g., CCCCC) to be read erroneously as one too few or one too many. Suppose that you have

sequenced some DNA and you are certain that the sequence listed below contains the translational start of a large protein (greater than 300 amino acids), beginning with the normal start codon.

GGGGAGGATAGCCATGCCAGCCCCCTAATTAGGGGGAGTTTCTCTGCAAAA

- What should convince you that there is an error in your sequence?
- Presuming that there is only one error, a deletion or insertion of a single base, which do you suppose it is and where?

D4. Hemophilia A is an X-linked disease associated with the absence of an essential blood clotting factor, factor VIII (if you don't have any idea what an X-linked trait is, don't worry about it). Factor VIII is encoded by the gene called *FACTOR8*. This gene was cloned from several individuals -- some affected, some not -- and sequenced. A portion of each sequence that you're sure contains the beginning of the gene (i.e., the start codon) was compared with the same portion of the wild-type sequence, as shown below. Each sequence contains only one mutation, shown emphasized.

Wild-type	5'-GGAGTTGAGTCATGGACTCTAAGCAGCGATCCACAAAG...
Individual a	5'-GGAGTT <b>T</b> AGTCATGGACTCTAAGCAGCGATCCACAAAG...
Individual b	5'-GGAGTTGAGTCAT <b>T</b> GACTCTAAGCAGCGATCCACAAAG...
Individual c	5'-GGAGTTGAGTCATGGACTCT <b>T</b> AGCAGCGATCCACAAAG...
Individual d	5'-GGAGTTGAGTCATGGACTCTAAGCAG <b>C</b> TATCCACAAAG...
Individual e	5'-GGAGTTGAGTCATGGACTCTAAGCAGCGATCCACT <b>T</b> AAG...

For each individual, choose from the list below to describe what you predict would be the severity of the phenotype, and give the reason for your choice.

- Severe hemophilia
- Mild hemophilia
- No hemophilia

D5. Complete the following table:

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										