Problem Set 1: Review of Molecular Biology

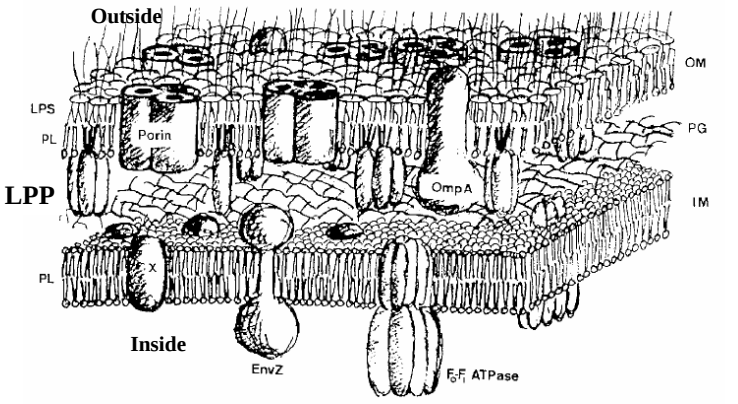
In this problem set, as in every problem set, as in everything you do for the rest of your life, accompany each answer with the evidence and reasoning that led you to your belief.

1. Estimate the length of a typical bacterial gene, a bacterial protein, and bacterial genome. How many genes in a typical bacterial genome? (Your experience with *What is a Gene* should come in handy here)

2. How big is the human genome (in nucleotides)? (BioBIKE will be of no help here. You'll need to go to an outside source). How many genes would you estimate it has? In fact, the human genome has around 20,000 protein-encoding genes. What gives?

3. 34.4% of the genome of the cyanobacterium Prochlorococcus marinus MIT9312 consists of A's. From this information, what would you predict is the frequency of the sequence CCCGGG? Give the answer in nucleotides per occurrence.

4. The cells of many bacteria are limited by two membranes: the outer membrane and the inner cytoplasmic membrane (see figure at right). The most abundant protein in the membranes is called lipoprotein (LPP), which stabilizes the structure. In order to understand its high level of expression, Nakamura and Inouye (1980) compared the DNA from *E. coli* with that of another enteric bacterium, *Serratia marcescens*, in the region of the gene encoding LPP.



LPS lipopolysaccharide; LPP lipoprotein; PL phospholipids; OM outer membrane; IM inner membranes; PG peptidoglycan

Lukas Buehler

Consider Figure 2 from their article:

Nakamura K, Inouye M (1980). DNA sequence of the *Serratia marcescens* lipoprotein gene. Proc Natl Acad Sci USA 77:1369-1373.

4a. From the DNA sequence of *E. coli* shown in that figure, write out the first 10 nucleotides of the double stranded DNA starting at the point where transcription of *lpp* mRNA begins, labeling the ends of the two strands as 5' or 3'.

4b. Write out the sequence of the first 10 nucleotides of *lpp* mRNA of *E. coli*.

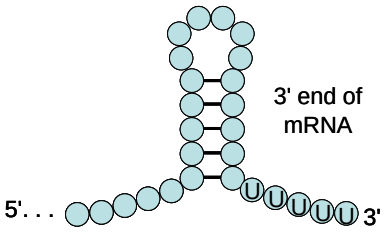
4c. Write out 12 nucleotides of the *lpp* mRNA of *E. coli*, starting with the first codon of the gene. Put a space between each codon. Then write under each codon the encoded amino acid of the *E. coli* lipoprotein.

4d. Between the *E. coli* and *S. marcescens* DNA sequences, Nakamura and Inouye have placed various geometric shapes: parallelograms, trapezoids, and triangles. What, specifically, does the triangle between positions +94 and +95 in the *S. marcescens* sequence signify? What is the biological consequence of the horizontal ***length*** of the triangle? What if it were slightly smaller?

4e. The density of geometric shapes is much higher in the second line of the figure than in the third and fourth. Why?

4f. The density of circled amino acids is much higher at the beginning of the gene (before the vertical arrow) than in the middle. Why?

The end points of RNA transcripts are often split palindromic sequences (the length varies from gene to gene), capable of forming hairpin loop structures followed by several U's (see figure at right). The region of basepairing need not be perfect, if there are sufficient pairs to maintain the hairpin structure.



4g. Does the mRNA of the *lpp* gene end with something like this structure? If so, draw it in the form shown at the right, but using actual nucleotides rather than circles.

5. Presuming that all mutations are equally likely, what is the probability that a mutation of an aspartate codon will lead to a codon encoding a charged amino acid? A hydrophobic amino acid?*

6. Go to PhAnToMe/BioBIKE and display the sequence of Synechococcus Phage Syn5 (nickname Syn5). Suppose you harbor some doubts that the given start codon for the gene csv5_gp02 (given as starting at coordinate 790) is correct. Argue for or against each of the propositions below:

    a. The gene may really start at coordinate 785 (the letter A)
    b. The gene may really start at coordinate 739 (the letter A)
    c. The gene may really start at coordinate 781 (the letter G)

7. Consider the same gene. Presuming that the given start codon ***is*** correct, predict the severity of the phenotype (i.e. how much the function of the protein may be affected) by the following mutations:

    a. The nucleotide at coordinate 785 is changed to a T
    b. The nucleotide at coordinate 790 is changed to a T
    c. The nucleotide at coordinate 795 is changed to a T
    d. The nucleotide at coordinate 796 is changed to a T
    e. The nucleotide at coordinate 792 is deleted
    f. The three nucleotides from coordinates 793 to 795 are deleted

8. Devise a way to produce the reverse complement of a DNA sequence. Specifically:

    a. Write an algorithm, either in plain English or in symbols of your choice, that will take a given single-stranded DNA sequence (5'->3') of arbitrary length and produce its complement (5'->3').

    b. Consider how to implement this algorithm in BioBIKE (without using its special functions). Note what functionality you need but don't exists.

* The following web site might be of some use: http://www.cgl.ucsf.edu/chimera/docs/UsersGuide/midas/hydrophob.html