(16 Jan 2013)

Problem Set 1: Review of Molecular Biology

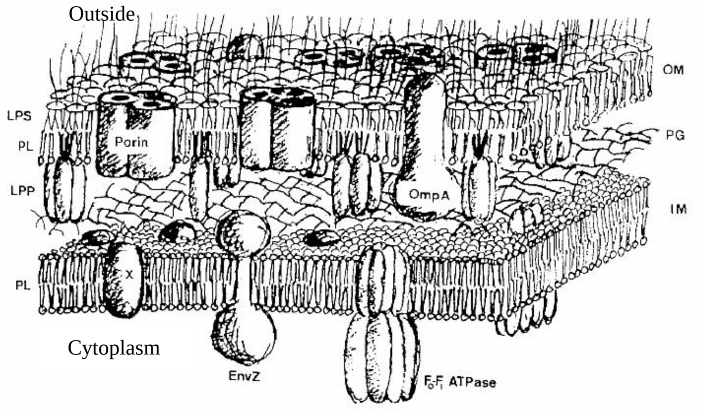
In this problem set, as in every problem set, as in everything you do for the rest of your life, accompany each answer with the evidence and reasoning that led you to your belief.

1. Estimate the length of a typical bacterial gene. A typical bacterial protein. A typical bacterial genome.

2. 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.

3. The cells of many bacteria are limited by two membranes: the outer membrane and the inner cytoplasmic membrane (see figure at right).

   For many years the laboratory of Masayori Inouye studied all matters related to the outer membrane, the proteins it contains, and the expression of those proteins. In one study, Nakamura and Inouye compared the DNA from *E. coli* and another enteric bacterium, *Serratia marcescens*, in the region of the gene encoding an outer membrane lipoprotein (*lpp*).



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

Lukas Buehler

   Consider Figure 2 from that article:

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

   3a. 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.

   3b. Write out the sequence of the first 10 nucleotides of *lpp* mRNA of *E. coli*.

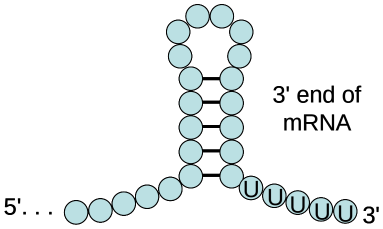
   3c. Write out 12 nucleotides of the *lpp* mRNA of *E. coli*, starting with the first codon of the gene. Put a space between each codon. Then write under each codon the encoded amino acid of the *E. coli* lipoprotein.

   3d. 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?

3e. The density of geometric shapes is much higher in the second line of the figure than in the third and fourth. Why?

3f. 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, 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.



3g. Does the mRNA of the *lpp* gene end with such a structure? If so, draw it, using actual nucleotides rather than circles.

4. 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?[*]

5. 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)

6. 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

7. Write an algorithm to amplify an entire fragment of DNA, given the sequence of that fragment, provided as a single string of nucleotides. Use PCR primers that are 20-nucleotides in length. You have at your disposal, any enzyme and chemicals you think you need, including a buffer sufficient to allow enzyme activity. You may obtain any DNA 20-nucleotide sequence so long as you specify the exact sequence. If you're unfamiliar with PCR, see the Introduction to Molecular Biology web page.

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

8. Devise a way to determine whether a given DNA sequence is a palindrome.

   8a. Write an algorithm, either in plain English or in symbols of your choice, that will do the job with a sequence exactly 6-nucleotides in length.

   8b. See how far you can get implementing this algorithm in BioBIKE. Note what functionality you need but don't know how to get.

   8c. Same as 8a, but modify the algorithm so that it works with a sequence of any length.

   8d. Modify it again so that it finds gapped palindromes, such as:

```
5'-GGAGCTTATGCTCC-3'
3'-CCTCGAATACGAGG-5'
```