Biol 591 Introduction to Bioinformatics (Fall 2003)

Scenario 1: Problem Set 1M – Genome Comparison

*Some questions refer to articles on the sequencing of E. coli O157:H7 [Perna et al (2001), Nature **409**:529-33; Hayashi et al (2001), DNA Res **8**:11-22]. See Scenario 1 main page for links to these articles.*

PS1M-1. The table below was prepared from the data in the codon tables for *Borrelia burgdorferi* and *Mycobacterium tuberculosis* included in the notes. The second and third columns indicate the amount of each amino expressed as a percentage of the total potential protein content of each species. The last column reflects the number of codons available for each amino acid. (For example, Ala is encoded by 4 codons; 4 ÷ 64 = 6.3%.)

a. Do the number of codons available for each amino acid appear to have a bearing on the frequency with which individual amino acids are used? Put another way, is there a correlation between the percentage of all codons devoted to coding for a particular amino acid and the abundance with which that amino acid is used?

b. Comparing the two bacteria, *B. burgdorferi* uses much more Lys in its protein than does *M. tuberculosis*. Why might that be? (Hint: look at a codon table.)

c. We might imagine that both organisms are likely to encode similar proteins and need amino acids with similar properties. What amino acids have similar properties to Lys? If *B. burgdorferi* uses more Lys than does *M. tuberculosis*, does *M. tuberculosis* use more of these other amino acids than *B. burgdorferi*?

Amino Acid	% of all amino acids in		% of all codons
	B. burgd.	***M. tuberc.***	
Ala	4.8	13.2	6.3
Arg	3.1	7.5	9.4
Asn	7.5	2.5	3.1
Asp	5.3	5.8	3.1
Cys	0.7	0.9	3.1
Gln	2.7	3.1	3.1
Glu	7.2	4.7	3.1
Gly	4.9	9.7	6.3
His	1.2	2.3	3.1
Ile	9.8	4.2	4.7
Leu	10.2	9.7	9.4
Lys	11.0	2.1	3.1
Met	1.8	1.9	1.6
Phe	5.5	2.9	3.1
Pro	2.4	5.9	6.3
Ser	7.4	5.6	9.4
Thr	4.5	5.9	6.3
Trp	0.4	1.5	1.6
Tyr	4.1	2.1	3.1
Val	5.1	8.6	6.3

PS1M-2. One error in the process of DNA sequence analysis that can be particularly problematic is the omission of a single nucleotide. This will cause a frame shift in a protein coding sequence, resulting in an incorrect translation of the protein by our computers. One use of codon frequency tables is to tell us when this is likely to have occurred. Assume the following DNA sequence codes for a protein in a close relative of *M. tuberculosis*. It has been decoded below in each of the three possible reading frames. Frame 1 begins with the first nucleotide, frame 2 the second, and frame 3 the third. From the *M. tuberculosis* codon use table provided in the notes, which translation is most likely correct?

a. 1 **b.** 2 **c.** 3

Frame	C A C C T T G C A C G A G T A C A T C G G C
1	His Leu Ala Arg Val His Arg -
2	- Thr Leu His Glu Tyr Ile Gly
3	- - Pro Cys Thr Ser Thr Ser Ala

PS1M-3. Sickle cell anemia results from a single nucleotide change in the human beta globin gene. The mutation affects one codon:

Normal	**Sickle cell**
GAG	GTG
Glu	___

What does the new codon encode? Choose all the terms from the following list that describe this mutation:

a. substitution
b. insertion
c. deletion
d. silent
e. nonsense
f. missense
g. transition
h. transversion

PS1M-4. In Fig. 5 of the notes for this scenario, gene C1 is an ortholog of gene B1. If we were attempting to derive the pedigree in the figure by comparing the sequences of the genes in organisms B and C, it might be difficult to determine whether gene C1, C2, or C3 is the true ortholog of gene B1 because all of the genes shown will have similar DNA sequences. This is one of the challenges facing us in our BLAST search and the authors of the two papers. Assume that organism B is *E. coli* K-12 and organism C is an O157:H7 strain. If the positions of genes B1, C1, C2, and C3 on the outer ring of Fig. 1 of the *Nature* paper (p.530) were provided to us, how might that help us determine whether C1, C2, or C3 is an ortholog of B1?

PS1M-5. Table 1 of the *Nature* paper (p. 532) summarizes the nucleotide differences (also known as "single nucleotide polymorphisms" or "SNPs") between suspected orthologous genes in the two strains. In this table, position 1 of codon 1 from one protein is compared to position 1 of codon 1 from the ortholog in the other strain. This is repeated for positions 2 and 3 of the codon. This process is then repeated for every codon of every ortholog. Thus, if we were to add an additional codon for comparison whose sequence was GAT in EDL933 and ACG in MG1655, the tallies in the table would change as follows. Top section, under column "A": 865→866; middle section, under column "C": 166→167; bottom section, under column "G": 1619→1620. (The totals would also change.) Make sure that you see this.

a. In which position (first, second, or third) did most of the differences occur? Why do you think this was the case?

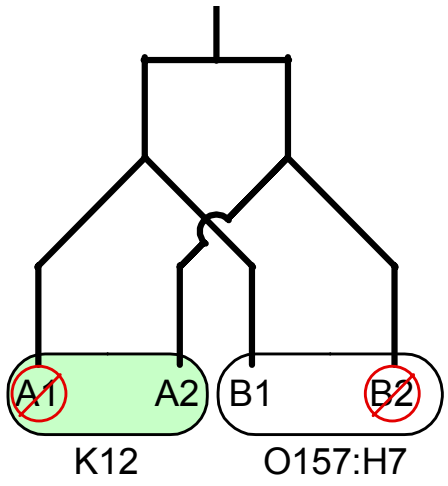
b. It is assumed that nucleotide differences between orthologous genes result from mutations in one or both of the genes. Looking only at the third position, were transition mutations or transversion mutations more frequent? Why might this be?

c. Looking at the first and second positions, were transition mutations or transversion mutations more frequent? Does your explanation in question **b** apply to these results as well? Why or why not?

PS1M-6. Just above Table 1 on p. 532 of the *Nature* paper, the authors discuss "hypervariable" genes, which are defined as orthologous genes that are much more divergent than the majority of orthologous genes. One explanation provided for this observation is that the cells experienced positive selection for divergence of these proteins. This is sometimes observed in pathogen genes encoding proteins that interact with the host's immune system. Variation in surface proteins allows the pathogen to avoid the immune response, so amino acid changes in such proteins tend to be advantageous rather than disadvantageous.

a. If this explanation is correct, would you expect the pattern of nucleotide variation observed in Table 1 and discussed in question **PS1M-5a** to be observed in the hypervariable genes? Why or why not?

b. A second explanation provided for the existence of apparent hypervariable orthologs is "differential paralogue retention from an ancient random duplication." This possibility is illustrated below. Genes A1 and B2 are lost, leaving A2 and B1 as the sole remaining homologs. In this case, are genes A2 and B1 orthologs? If not, what is their relationship?



PS1M-7. On page 14 of the *DNA Research* article, the authors report that the amount of DNA in the O157:H7 strain not present in the K-12 strain is equivalent to the entire genome of the bacterium *Borrelia burgdorferi* (which causes Lyme disease). In other words, our hope of quickly finding a few genes that are responsible for O157:H7's virulence are in doubt. The task could yet be accomplished if we understood the function of the proteins encoded by these genes. This might seem within reach since we arguably have a more complete understanding of the protein of *E. coli* than any other organism. Looking at Table 3 of the same article, what percentage of the 1632 genes that are found only in O157:H7 have functions assigned to them? What does this tell you about our knowledge of this simple organism?

PS1M-8. From your reading of one or both papers, which of the following differences were discovered between the two strains? Select all that apply.

a. Genes were found in O157:H7 that were not found in K-12
b. Genes were found in K-12 that were not found in O157:H7
c. Genes were found that were present in both strains, but whose sequences were not identical.

PS1M-9. From the results of your program (modified from *BlastParser.pl*) and the results of your colleagues, what genes may be candidates to determine the pathogenicity of *E. coli* O157:H7? Which ones seem the most likely?