BIOL591: Introduction to Bioinformatics
Comparative genomes to look for genes responsible for pathogenesis

Reading:

(1) Scenario 2: (Course web site) ***Read this first !***

(2) Perna, N. T., G. Plunkett, 3rd, et al. (2001). "Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7." Nature **409**(6819): 529-33.

Site for viewing or downloading the above article:

http://www.nature.com/cgi-taf/DynaPage.taf?file=/nature/journal/v409/n6819/full/409529a0_fs.html&content_filetype=pdf

(3) Hayashi, T., K. Makino, et al. (2001). "Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12." DNA Res **8**(1): 11-22.

Site for viewing or downloading the above article:

http://www.people.vcu.edu/~elhaij/IntroBioinf/Scenarios/Hayashi2001.pdf

Note: it is helpful to view the figures in color.

(4) Bioinformatics p. 53-56.

The articles above offer two views of our topic. The authors have done the same thing you are going to do—compare the genes in two strains of *E. coli* by using the BLAST program to find genes that could be responsible for the virulence of the O157:H7 strain. We will discuss the BLAST program in more detail later. For now, it is enough to know that it is a program that can be used to identify gene or protein sequences present in a database that are similar to a given sequence. Together, the two papers provide a wealth of information about pathogenic bacteria, microbial genomics, and molecular evolution. The purpose of these notes is to provide some guidance for understanding the papers and, at the same time, to use the papers to introduce several concepts that will appear repeatedly throughout the course. Read these notes before beginning to read the papers.

The authors of each of these papers compared a pathogenic *E. coli* O157:H7 strain to the non-pathogenic K-12 laboratory strain, MG1655. (The Nature paper compares the K-12 strain to the O157:H7 strain EDL933 whereas the DNA Research paper uses O157:H7 strain RIMD 0509952, which is referred to as "O157 Sakai.") Begin by reading the abstracts of both articles. (The abstract of the Nature article is the first paragraph of the paper and is in bold type.) Most of the important points of the papers are summarized here. You will see that the DNA from the two strains was found to be very similar in each case, as expected when comparing two different strains of the same species. Differences were noted, however, which fell into three categories. First, some genes were found in O157:H7 that were not present in K-12. This group of genes is quite large, which is not surprising given that the O157:H7 strain possesses a much larger genome than the K-12 strain. Second, some genes were found in K-12 that were not present in the O157:H7 strain. Third, some genes were found that were very similar in both strains, but were not identical. Each of these categories will be examined in turn.

Let's begin with genes found in O157:H7 that were absent in K-12. How did this come about? Two possibilities come to mind. First, the ancestor of both strains may have had all the genes present in either, but after the two strains diverged, they may have each lost some genes, with the K-12 strain losing more genes than O157:H7. Alternatively, the O157:H7 strain may have acquired new genes from somewhere else. Is this reasonable? How could this happen?

We have learned that bacteria have several methods for acquiring DNA. In fact, many of these methods were first observed in *E. coli*. You will see in the DNA Research paper that the O157 Sakai strain possesses two plasmids, named pO157 and pOSAKI. Plasmids are pieces of DNA much like the chromosome, though usually smaller. These two plasmids are 92,721 and 3,306 base pairs (abbreviated "bp") whereas the chromosome is 5,498,450 bp. One feature of plasmids that is important here is that they can often encode the ability to transfer themselves from one cell to another. (An animation showing the transfer of a plasmid called "F" from one *E . coli* cell to another by a process called conjugation may be found at:

http://www.microbelibrary.org/images/Tterry/anim/Fmating.gif

Thus, the transfer of plasmids is one way in which bacteria can acquire new DNA.

Both papers also emphasize the transfer of DNA by viruses. Though this may be surprising, bacteria can be infected by viruses. Jonathon Swift seems to have imagined this more than 200 years ago when he wrote:

Great fleas have little fleas upon their backs to bite 'em,
And little fleas have lesser fleas, and so *ad infinitum.*

The viruses that infect bacteria are called "**bacteriophage**" or just "**phage**." Certain phage can follow either of two pathways, lytic growth or lysogeny. Both are depicted in Fig. 1.

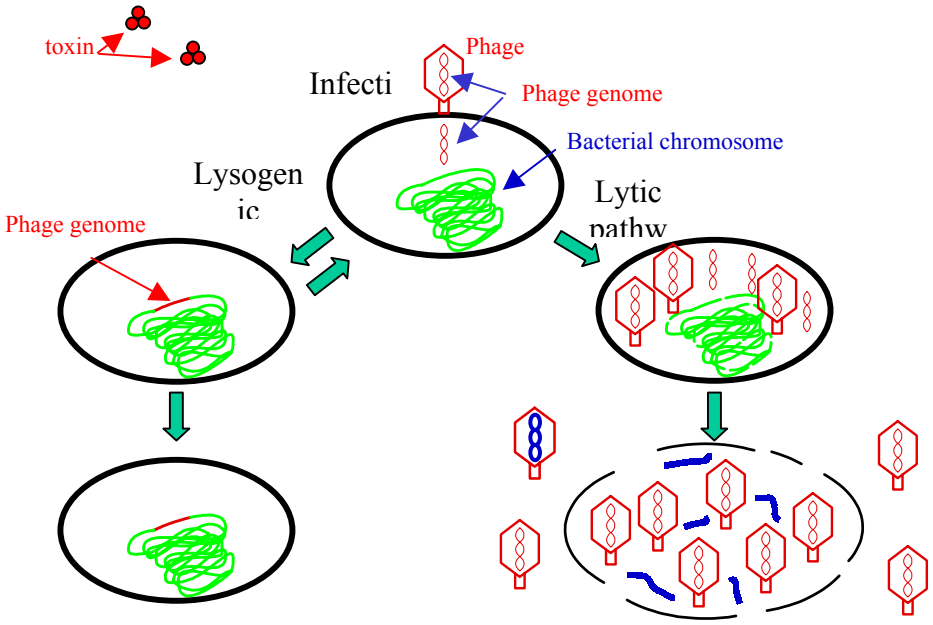

Fig. 1. Bacteriophage infection. The phage particle attaches to the bacterium and injects its DNA into the cell. The lytic pathway, shown at the right, results in cell lysis and production of phage particles, including one containing bacterial (blue) DNA. The lysogenic pathway (left) occurs when the phage genome integrates into the chromosome, becoming a prophage. Prophage genes may be expressed, producing proteins such as toxins. The cell can continue to replicate the phage DNA along with its own indefinitely, or the prophage genome can excise, resuming the lytic cycle.

In the lytic pathway, the virus hijacks the cell, causing it to replicate the virus DNA, package the DNA into protein shells, then break open, releasing progeny virus. In the lysogenic pathway, the phage integrates its DNA into the chromosome (at which point, it is called a **prophage**), but may later excise from the chromosome to resume the lytic pathway. Apart from killing bacteria, phage can genetically alter cells in two ways. First, genes contained on the phage can be expressed (transcribed and translated) just like cellular genes. These gene products can affect their hosts, in some instances changing a non-pathogenic bacterium into a pathogen (depicted above as toxin production encoded by the integrated prophage genome). Second, when virus shells are filled with DNA during lytic growth, it is possible for cellular DNA to be packaged into the shell along with or in place of virus DNA, as depicted in the figure above by the lone virus particle containing blue DNA. If this virus then infects another cell, it will transfer the DNA from the previous cell into the new host. This process, called transduction, is another means by which cells can exchange DNA. It appears that both of these processes have occurred in *E. coli*. Look for references to phage and prophage genes in the texts. (In the DNA Research paper, read section 3.2.)

So acquisition of foreign genes by shuttling of plasmids and viruses (and other mechanisms not mentioned here) does occur. This process, termed **horizontal transfer** or **lateral transfer** could theoretically explain the presence of unique genes in each strain. But there is also the possibility mentioned above that the presence of unique genes in O157:H7 is due not to their acquisition by this strain, but to their loss by K-12. Either explanation would produce the same result—the presence of genes in O157:H7 that are not found in K-12. Is there any way to distinguish between these possibilities? What we need is some way to identify foreign genes. Fortunately, examination of GC content and codon use can provide us with this information. Let's look at these concepts further.

We have mentioned previously that genomic sequences are often composed of unequal proportions of the four nucleotides, dATP, dTTP, dCTP, and dGTP. (We'll use the shorthand hereafter of "A," "T," "C," and "G.") Some organisms have far more C and G than A and T, some have far less.

SQ1: Chargaff's rule states that the amounts of A and T in any given double-stranded piece of DNA are equal, as are the amounts of G and C; that is, A=T and G=C. From what you know of the structure of DNA, why must this be true?

The proportion of Gs and Cs in DNA is usually referred to as "G+C content" or just "GC content." Why do different organisms have different GC contents? That is a matter of some debate. Certainly, there are many properties of a cell that change along with GC content, including codon usage, tRNA copy number, and melting temperature of the DNA and of RNA molecules such as tRNA and ribosomal RNA. Whether these other properties drive GC content change or merely reflect it is not entirely clear.

At any rate, if you look at the standard genetic code (at the "Resources and links" page of the course website) or the tables below, you will see that most amino acids are encoded by multiple codons. The frequency with which each of these "synonymous" codons is actually used is reflective of the GC content of the organism in which the gene is found. Below are tables indicating the frequency with which different codons are used for *Borrelia burgdorferi*, with 29% GC content in its coding DNA and *Mycobacterium tuberculosis*, with 65% GC.

***Borrelia burgdorferi*: 2294 CDS's (612759 codons)**

fields: [**triplet**] [amino acid] [fraction] [frequency: per thousand]

UUU	Phe	0.88	48.3	**UCU**	Ser	0.32	24.1	**UAU**	Tyr	0.77	31.6	**UGU**	Cys	0.68	4.9
UUC	Phe	0.12	6.3	**UCC**	Ser	0.05	3.4	**UAC**	Tyr	0.23	9.2	**UGC**	Cys	0.32	2.3
UUA	Leu	0.41	41.5	**UCA**	Ser	0.24	17.6	**UAA**	*	0.65	2.4	**UGA**	*	0.16	0.6
UUG	Leu	0.16	16.3	**UCG**	Ser	0.03	2.3	**UAG**	*	0.19	0.7	**UGG**	Trp	1.00	4.4
CUU	Leu	0.28	29.0	**CCU**	Pro	0.42	10.0	**CAU**	His	0.73	8.6	**CGU**	Arg	0.07	2.1
CUC	Leu	0.02	2.3	**CCC**	Pro	0.15	3.7	**CAC**	His	0.27	3.2	**CGC**	Arg	0.04	1.1
CUA	Leu	0.10	10.6	**CCA**	Pro	0.37	8.9	**CAA**	Gln	0.84	22.8	**CGA**	Arg	0.06	1.8
CUG	Leu	0.03	2.7	**CCG**	Pro	0.06	1.3	**CAG**	Gln	0.16	4.2	**CGG**	Arg	0.02	0.5
AUU	Ile	0.54	53.1	**ACU**	Thr	0.39	17.4	**AAU**	Asn	0.80	60.0	**AGU**	Ser	0.22	16.5
AUC	Ile	0.07	7.2	**ACC**	Thr	0.12	5.6	**AAC**	Asn	0.20	15.1	**AGC**	Ser	0.14	10.4
AUA	Ile	0.39	38.0	**ACA**	Thr	0.44	19.9	**AAA**	Lys	0.80	87.8	**AGA**	Arg	0.65	20.1
AUG	Met	1.00	18.1	**ACG**	Thr	0.05	2.2	**AAG**	Lys	0.20	22.2	**AGG**	Arg	0.18	5.5
GUU	Val	0.55	27.9	**GCU**	Ala	0.44	21.2	**GAU**	Asp	0.79	42.0	**GGU**	Gly	0.28	13.7
GUC	Val	0.05	2.4	**GCC**	Ala	0.11	5.1	**GAC**	Asp	0.21	11.3	**GGC**	Gly	0.16	7.7
GUA	Val	0.30	15.1	**GCA**	Ala	0.39	18.9	**GAA**	Glu	0.75	53.9	**GGA**	Gly	0.41	20.0
GUG	Val	0.11	5.4	**GCG**	Ala	0.06	2.7	**GAG**	Glu	0.25	17.8	**GGG**	Gly	0.15	7.4

Coding GC 29.27% 1st letter GC 38.52% 2nd letter GC 28.30% 3rd letter GC 21.01%

***Mycobacterium tuberculosis CDC1551*: 4187 CDS's (1329826 codons)**

fields: [**triplet**] [amino acid] [fraction] [frequency: per thousand]

UUU	Phe	0.21	6.2	**UCU**	Ser	0.04	2.3	**UAU**	Tyr	0.30	6.1	**UGU**	Cys	0.26	2.4
UUC	Phe	0.79	22.9	**UCC**	Ser	0.21	11.6	**UAC**	Tyr	0.70	14.5	**UGC**	Cys	0.74	6.9
UUA	Leu	0.02	1.7	**UCA**	Ser	0.07	3.8	**UAA**	*	0.15	0.5	**UGA**	*	0.55	1.7
UUG	Leu	0.19	18.1	**UCG**	Ser	0.35	19.5	**UAG**	*	0.30	1.0	**UGG**	Trp	1.00	14.8
CUU	Leu	0.06	5.6	**CCU**	Pro	0.06	3.6	**CAU**	His	0.29	6.6	**CGU**	Arg	0.12	8.7
CUC	Leu	0.18	17.2	**CCC**	Pro	0.29	17.0	**CAC**	His	0.71	16.0	**CGC**	Arg	0.38	28.7
CUA	Leu	0.05	4.8	**CCA**	Pro	0.11	6.4	**CAA**	Gln	0.26	8.2	**CGA**	Arg	0.10	7.6
CUG	Leu	0.51	49.7	**CCG**	Pro	0.54	31.7	**CAG**	Gln	0.74	22.9	**CGG**	Arg	0.33	24.9
AUU	Ile	0.15	6.5	**ACU**	Thr	0.07	3.8	**AAU**	Asn	0.21	5.2	**AGU**	Ser	0.07	3.7
AUC	Ile	0.79	33.4	**ACC**	Thr	0.59	34.6	**AAC**	Asn	0.79	19.4	**AGC**	Ser	0.26	14.6
AUA	Ile	0.05	2.3	**ACA**	Thr	0.08	4.8	**AAA**	Lys	0.26	5.4	**AGA**	Arg	0.02	1.4
AUG	Met	1.00	18.6	**ACG**	Thr	0.27	15.7	**AAG**	Lys	0.74	15.1	**AGG**	Arg	0.05	3.4
GUU	Val	0.10	8.2	**GCU**	Ala	0.08	11.2	**GAU**	Asp	0.28	15.9	**GGU**	Gly	0.19	18.6
GUC	Val	0.38	32.4	**GCC**	Ala	0.45	59.0	**GAC**	Asp	0.72	41.9	**GGC**	Gly	0.51	49.3
GUA	Val	0.06	4.9	**GCA**	Ala	0.10	13.0	**GAA**	Glu	0.35	16.2	**GGA**	Gly	0.10	10.0
GUG	Val	0.47	40.1	**GCG**	Ala	0.37	48.4	**GAG**	Glu	0.65	30.4	**GGG**	Gly	0.20	18.9

Coding GC 65.77% 1st letter GC 67.82% 2nd letter GC 50.22% 3rd letter GC 79.27%

Tables 1 and 2. Codon usage in *Borrelia burgdorferi* and *Mycobacterium tuberculosis* derived from sequence analysis of each genome. The number of coding sequences "CDS's" and codons from which these frequencies were derived are indicated above each table. [fraction]-proportion of occurrences of a particular amino acid encoded by a particular codon. For each amino acid, the fractions associated with each codon sum to 1. [frequency: per thousand]-the number of times each codon was used per genome ÷ 1000. * -stop codon.

Note that codons are translated into protein from messenger RNA, so the sequence of the RNA codons is shown. Replacing U with T results in the sequence of the non-coding strand of the gene's DNA, which is the sequence that is by convention reported as a gene's DNA sequence. For example, we would decode a reported DNA sequence of "ATG CTG TTT…" as "Met-Leu-Phe…"

SQ2 List the two triplets that code for Lys. What proportion of each is used in *Borrelia burgdorferi* compared to *Mycobacterium tuberculosis*? Is this finding surprising? Why or why not?

SQ3 There are 6 triplets that code for Arg. Find the two most commonly used in each of the two bacteria. What is the GC content of each of these codons?

SQ4 The GC content of the coding sequences available from *Bacillus anthracis* (the cause of anthrax) is 33.97%. Suppose we wanted to examine the DNA sequence of an isolate of *B. anthracis* that we suspect may have been altered by a genetic engineer to contain one or more foreign genes. By analysis of codon use, would it likely be easier to detect a foreign gene originating from *Borrelia burgdorferi* or from *Mycobacterium tuberculosis*?

So we see that we can in principle distinguish foreign DNA from native DNA by its GC content and codon use. This can allow us to distinguish acquisition of a new gene by O157:H7 from loss of a gene by K-12. The same holds true for genes present in K-12 that are absent in O157:H7, the second category of differences found. GC content and codon use are found in several of the figures and tables in the two papers. The third group of differences noted in the papers is genes that are recognizably similar in both strains, but which are not identical. Before discussing these genes, it would be helpful to discuss the terms used to describe related genes or proteins. (See also the assigned reading in the Bioinformatics text.)

If we were to use the BLAST program to identify proteins similar to the alcohol dehydrogenase gamma subunit (ADH3) from humans, we would find a good match to the amino acid sequence of the ADH from baboons (89% identity over a large region of the two proteins). There are three conceivable explanations for this finding. (1) The two proteins share this amount of identity by chance. The probability of this being true is provide by the BLAST program and turns out to be virtually nil. (2) These two proteins in these two organisms were originally unrelated, but evolved to have the same sequence in order to perform the same function. While unrelated proteins may evolve to perform the same function, their sequences remain largely dissimilar. (3) The two proteins are evolutionarily related. In this case, we hypothesize that a common ancestor of baboons and humans contained an alcohol dehydrogenase gene. When the human and baboon lineages separated, the alcohol dehydrogenase genes evolved separately, with each incorporating rare, independent mutations. In this case, the two genes (or proteins) would be described as **orthologs** (or orthologues). That is, they are derived from a common ancestor by vertical descent, as in Fig. 2. In this case, the degree of divergence between the two genes can be used as a measure of their relatedness and of the relatedness of the organisms that possess them. Thus, gene comparisons provide a window into evolutionary relationships that does not rely on archeological or anatomical comparisons.

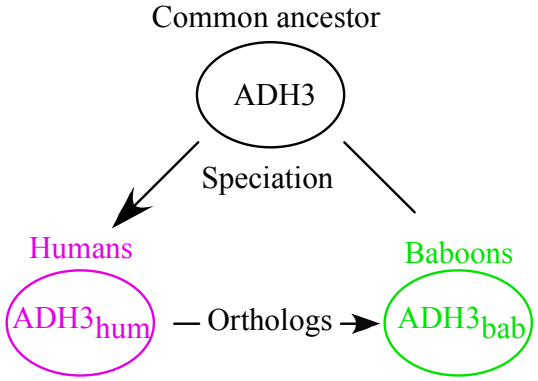

Fig. 2. Derivation of orthologs.

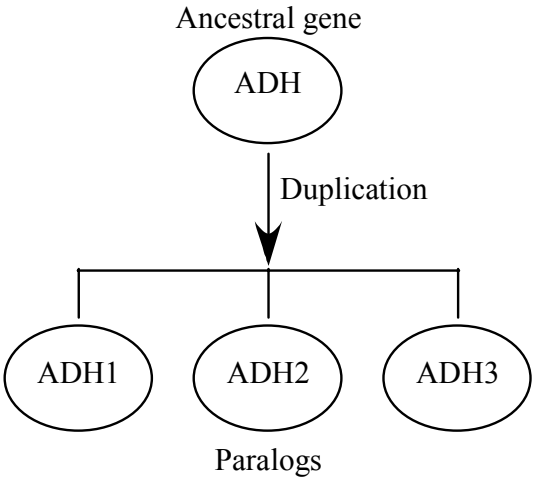

Fig. 3. Derivation of paralogs.

If you were to examine the results of the BLAST search, you would discover that a closer match to the human ADH3 protein is obtained with the human ADH1 and ADH2 proteins, which compose the alpha and beta chains of human ADH. How did three similar proteins arise in the same organism? The most likely mechanism is gene duplication, as depicted in Fig. 3. Genes that arise by duplication are called **paralogs** (or paralogues). More complicated relationships than those illustrated above can also be examined. Fig. 4 (next page) illustrates gene duplication as a solid horizontal line and a speciation event as a fork that looks like an upside down "Y." To distinguish orthologs from paralogs in this figure, determine whether the common ancestor of any two genes resides at a horizontal line or a fork. Any two genes whose common ancestor resides at a horizontal line are paralogs and any two genes whose common ancestor lies at a fork are orthologs.

SQ5: Are genes B1 and C2 orthologs or paralogs? People sometimes define paralogs as related genes within the same species that perform different functions and orthologs as related genes in different species performing the same function. Is your answer to the above question consistent with this definition?

Transfer of DNA from one species to another (by processes described above, including conjugation and virus-mediated transduction) is referred to as **horizontal transfer** or **lateral transfer** and is indicated by a dashed line in Fig. 4. A gene so transferred is sometimes referred to as a **xenolog** (or xenologue). This term is much less common than ortholog or paralog and is not used in either of the assigned articles, though the concept appears throughout each. Note that lateral transfer results in genes with evolutionary histories that do not accurately reflect that of the organism in which they are found.

Another term often encountered is homology. **Homology** refers to any two genes or proteins that are evolutionarily related, that is, descended from a common ancestor. Both orthologs and paralogs are therefore homologs (or homologues). It is also worth noting that it is incorrect to say that the human and baboon ADH3 proteins are "89% homologous". Proteins are

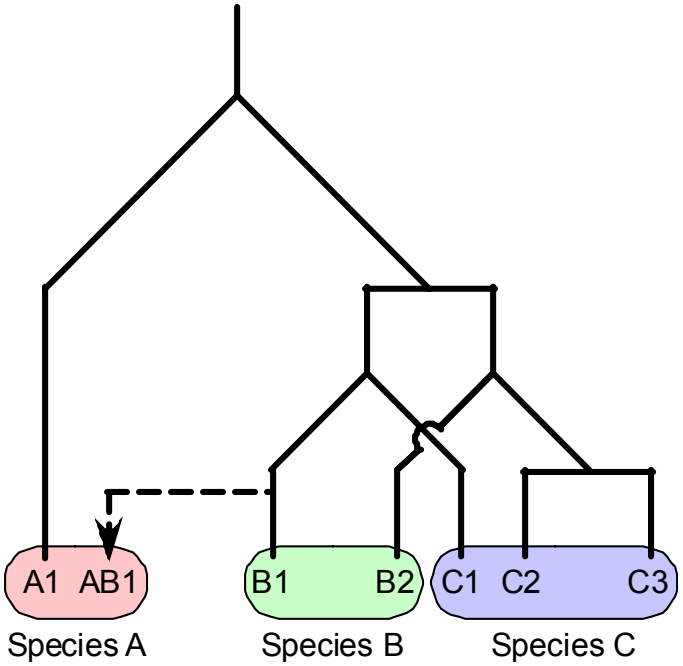

Fig. 4. Orthologs, paralogs, and xenologs. Related genes (A1—C3) from three species (A, B, and C) are shown. Gene duplication events are illustrated by solid horizontal lines. Speciation events are indicated by forks. Gene transfer from one species to another is indicated by a dashed line. Genes whose common ancestor resides at a horizontal bar are paralogs whereas those common ancestor resides at a fork are orthologs. For example, gene C2 is a paralog of gene C3 and an ortholog of gene B2. (Adapted from Jensen, R. A. 2001. Orthologs and paralogs - we need to get it right. Genome Biol **2:** Interactions1002; and Fitch, W. M. 2000. Homology: a personal view on some of the problems. Trends in Genetics **16:**227-231.)

either homologous (derived from a common ancestor) or they are not. Quantitation of similarity between proteins is best expressed in terms of "percent identity" or, if "similar" amino acids are counted, "percent similarity."

We now have the vocabulary to discuss the genes that are similar but not identical in the two strains. We are referring to orthologs. By definition, these genes were once identical. How did they develop differences in their sequences? DNA sequences undergo changes with time, principally due to errors that occur during DNA replication. These changes in DNA sequence are called **mutations**. Let's add a bit more vocabulary. A change of a single nucleotide is called a point mutation. Point mutations include deletions or insertions of single bases and substitutions of one base for another. Substitution mutations come in two varieties—transitions and transversions. A **transition mutation** is a change from one pyrimidine nucleotide to another or one purine nucleotide to another. (The pyrimidine nucleotides are C, T, and, in RNA, U. The purines are A and G.) A **transversion mutation** is a substitution of a purine with a pyrimidine or vice versa. Due to the chemistry of DNA, transition mutations occur more frequently than transversion mutations.

Substitution mutations within coding sequences fall into three categories, depending on the effect of the mutation on the encoded protein. A **silent mutation** has no effect on the amino acid sequence because it produces a synonymous codon. As an example, a mutation in the codon

GGA that changes its sequence to GG**G** is a silent mutation because both the old and new codons code for glycine. A **missense mutation** produces a codon that codes for a different amino acid. A mutation from GGA to G**A**A will result in production of the amino acid glutamate in place of glycine. A **nonsense mutation** produces a stop codon. An example would be a mutation from GGA to the stop codon **T**GA.

SQ6 There are 5 amino acids that are each encoded by exactly four codons. Choose one codon for any of these amino acids. Change the nucleotide in the first position to the three other possible nucleotides. How many of these mutations are silent mutations? Repeat for the second and third positions of the codon. Compare answers with someone else in the class who chose a different codon. Are your answers the same? Which codon position allows the most change without changing the amino acid sequence?

SQ7 There are two codons each for 9 of the amino acids. Choose any one of these 18 codons. Create a transition mutation in the third position of the codon. What is the result? Create a transversion mutation in the third position. What is the result? Choose another one of the 14 codons and repeat. In the third position, are transition mutations or transversion mutations more likely to result in a change in the amino acid encoded?

Table 1 of the Nature paper summarizes the changes in nucleotide sequence between orthologs in both strains. This table will be explored more fully in the problem set. We have now discussed the three categories of genes that differ between the two strains. Have we identified the gene differences responsible for the virulence of the O157:H7 strains? Read the papers and find out. Keep in mind the following additional points while reading:

- Regions of DNA unique to the K-12 and O157:H7 strains are called "K-Islands" and "O-Islands," respectively, in the Nature paper and "K-loops" and "S-loops" in the DNA Research paper.

- If you don't understand parts of the paper even after looking through these notes, skip to the next section. The areas of the papers that are most important for our purposes will be emphasized in the problem set.

- Remember that research articles are hardly ever easy to read. The secret is to focus on the parts of the article you DO understand (and to gradually expand that fraction) without getting overly concerned about the parts you DON'T understand. However, if you judge that some incomprehensible part is essential for your appreciation of a key point, then make a note of the term or concept that is throwing you. Bring these questions to class, or send them ahead by e-mail or questionnaire.