DUE: Friday March 9 at the BEGINNING of class.
Hand In: Answer the questions on paper, number your answers. Your work must be legible -- if your handwriting isn't great, type it up and print it.

For full credit you must identify key assumptions and provide reasoning (or show work) behind answers. Whenever possible, partial credit will be given if adequate work is shown. Remember I encourage working together, but you MUST indicate all collaborations and/or assistance received or given.

Note that showing work means that if you utilize software for assistance (programs you write or stock software such as Excel), you should indicate as such and provide sufficient details so that I can judge the work. That may mean sending me (by email) source code or associated Excel files.

Questions 1-5 are required of all sections. Questions 6 is required of all graduate students (5520 and 7000) but undergraduates may elect to do the advanced problem for extra credit.

All problems have a maximum value of 10 points. Sub-problem values are marked when appropriate.

Questions

1. (a) 1pt What is the advantage of a seeded method like BLAST compared to a Needleman-Wunsch alignment ?

(b) 3pt Which flavor of BLAST is most sensitive for comparing your sequence to a species that are NOT closely related? Explain in one sentence.

(c) 3pt In Karlin-Altshul statistics, what are four main assumptions?

(d) 3pt Describe a scenario where BLAST will be incapable of finding a high quality match, even when one exists in the database. To get full credit, you must specify the word hit length utilized by BLAST in your scenario AND state the actual percent identity of your query to the best sequence in the database (which BLAST is unable to find).

2. You will need the Hmwk3.fa file on Canvas for this question. This fasta file contains a segment of an *archaeal* genome that is not quite finished. The goal is to analyze this segment. Using NCBI Blast: http://blast.ncbi.nlm.nih.gov/Blast.cgi follow the directions below and answer the following questions. The documentation for NCBI Blast is available at:

ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_BLASTGuide.pdf

You decide to use BLASTX (select Blastx from the front page).

Search only the first 6000 nucleotides (query subrange 1-6000) against the non-redundant (nr) database. You are only interested in the top hits, so set the "Max target sequences" parameter to 50 and word size to 6. (Note that these are “Algorithm Parameters” and can be adjusted by expanding the form at the bottom.)

(a) 2 pt What species do you think this sequence originated from? Why?

(b) 2 pt What fraction of the query is included in the best alignment? 27% Is this hit a complete gene?

(c) 2 pt What coordinates of the query match and on which strand in the best alignment? (Note that you must view the actual alignment to answer this question.)

(d) 1pt Looking at the "Taxonomy Report" (find by first clicking on the "Taxonomy Reports" link at the top of the results), how many hits were observed to Thermoportei?

(e) 3pt In total, How many genes are in the first 6000 bases of this sequence? For full credit, give the names of each gene.

3. You are given the protein sequence for Yfg1 and asked to identify distinct domains within the protein. As a first step you use BLAST to search Yfp1 against the non-redundant database (nr) and the top hit is as follows:

Name	Score	Query	E-value	Ident	Accession
insulin precursor	320	100%	1e-95	100%	NP_001191615.1

Observing that your best hit was to a RefSeq entry, you re-run your search against the RefSeq database (all parameters exactly the same, only changing the database utilized) and obtain the following top hit:

Name	Score	Query	E-value	Ident	Accession
insulin precursor	320	100%	9e-111	100%	NP_001191615.1

(a) 3 pt How did you know the hit was to a RefSeq gene?

(b) 7pt Give what you know about how E-values are calculated, why has the E-value changed between the two searches?

4. 10 pt Explain the difference between an algorithm and a scoring scheme. Use an example from class to clarify your definition.

5. Suppose that you are worried that you might have a rare disease. You decide to get tested, and suppose that the testing methods for this disease are correct 99 percent of the time (in other words, if you have the disease, it shows that you do with 99 percent probability, and if you don't have the disease, it shows that you do not with 99 percent probability). Suppose this disease is actually quite rare, occurring randomly in the general population in only one of every 10,000 people.

10 pt You obtain a positive test result. What is the probability that you have the disease? [Hint: Use Bayes Rule]

6. (Advanced) You are excited about being able to use the human genome browser (at UCSC) to look more closely at the molecular basis of human genetic diseases in the news. To start with, you decide to investigate one of the genes mentioned in the NY Times article "Disease Cause is Pinpointed with Genome" by Nicholas Wade (http://www.nytimes.com/2010/03/11/health/research/11gene.html?_r=0).

One of the two papers described in this article has two authors "Lupski JR" and "Gibbs RA", and titled " **Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy"**. [The article is downloadable on Canvas.]

Use the NYT article, the paper and genome.ucsc.edu human genome browser (version NCBI36/hg18) to answer the following questions.

(a) 1pt What is the human gene name, and abbreviation (six letters) that was found to causative of Charcot-Marie-Tooth neuropathy in this subject's genome (you can use this gene name to find the gene quickly, use the "gene" window).

(b) 2pt How many exons does this gene have (use "RefSeq Genes" track, or dark blue "UCSC Genes" track)? What is the "genomic size" (full length, including exons and introns)?

(c) 1pt Which DNA sequencing technology was used?

(d) 1pt The paper describes following two independent mutations through the extended family, and showed only those who inherited both mutations had the disease. What are these mutations? (i.e. Q340R)

(e) 1pt For family members with only one bad allele (haploinsufficiency), what were two typical symptoms?

(f) 1pt Using coordinates and/or protein sequence from Figure 2 from the paper find the position in the UCSC genome browser of the mutation that normally codes for Tyrosine. Figure 2C gives this alignment, but it does NOT mention that there are two species that have the precise mutation variant responsible for this disease. What are these two species?
(Hint: in Multiz alignment of 44-vertebrates, click on the settings bar (grey vertical bar on left), and select "+" at the top to select and see all species in the Multiz alignment track. Another hint: note that the gene is on the reverse/minus strand, so to "turn it around" with 5' end on the left, click on the "reverse" button just below the browser window (between the "configure" and "refresh" buttons).

(g) 1pt You want to develop a genetic test for this mutation, so you need to find the closest "SNP" (single nucleotide polymorphism). You notice there is a SNP in the "Simple Nucleotide Polymorphisms" track right next to your mutation. What is the dbSNP id # (starts with "rs"), and the chromosome coordinate (i.e. chrX:12345443).

(h) 2pt You notice that this mutation is found in a relatively small exon. If you were to go looking in the largest exon for other genetic mutations, which exon would that be? Give the exon number and first five nucleotides (on the 5' end) of this exon.