

DUE: Monday Feb 29 at the BEGINNING of class.

Hand In: Answer the questions on paper, number your answers, show all work (as necessary), identify key assumptions, and indicate all collaborations and 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 or R), 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.

Your work must be legible -- if your handwriting isn't great, type it up and print it.

Questions

1. You will need the Hmwk3.fa file on D2L for this question. This fasta file contains a segment of an *archaea* 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 and answer the following questions.

a) Let's first assume there that are NO closely-related archaeal species in the database. Which flavor of BLAST is most sensitive for comparing your sequence to species that are NOT closely related? Explain in one sentence.

(b) Given the computational demands of the most sensitive form of BLAST, you decide to instead use BLASTX. Search only the first 6000 nucleotides against the non-redundant (nr) database, limited to "Archaea" and "Bacteria" organisms only. (Note you can adjust all of these parameters from the standard NCBI BlastX interface). 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.) What species contains a sequence that is the most similar by overall score? What fraction of the query is included in this alignment? What coordinates of the query match and on which strand?

(c) Looking at the "Taxonomy Report" (find by first clicking on the "Taxonomy Reports" link), which species had the most hits to this sequence?

(d) Given the protein similarity of your top hit, if you repeated your search and wanted to have the most accurate scoring matrix for the top hit, which BLOSUM matrix would you choose? (explain)

(e) Do you think you know exactly what species this sequence is taken from? (any web search or bioinformatics tool is fair game) If yes, give your evidence.

2. Answer the following questions using PubMed, the NCBI biomedical literature index/search engine. Provide the accession number(s) (PMID) for all information you utilize to answer these questions.

a) Prof. David Haussler has been a prolific, leading scientist in genome research, although he did not start in biomedical research. In what year did Prof. Haussler score his first publication in the journal "Science"?

b) Prof. Haussler's brother, Mark R. Haussler, is also a noted scientist. How many publications were they co-authors on, and what molecule did they study together?

c) Using only papers published between 2004 and 2008, name two parts of the human body that have been found to contain archaea?

3. You consider using an HMM approach to model protein secondary structure prediction. The straight-forward approach uses three secondary structure confirmations: " α -helix", " β -strand", and "turn" as the hidden states emitting observable amino acids. It is assumed that the frequencies/probabilities of each of the twenty amino acids can be determined from experimental data for each of those confirmations.

a) Draw the state diagram (circles and arrows) of the HMM.

b) How many emission parameters are needed to describe this model?

c) How many transition parameters are needed to describe this model?

4. You are excited about being able to use the human genome browser 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 D2L, named Lupski2010.pdf.]

(a) 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) Use the genome.ucsc.edu human genome browser (version NCBI36/hg18) to answer the following questions. 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) Which DNA sequencing technology was used?

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

5. ****Extra Credit**** This question also refers to Charcot–Marie–Tooth paper discussed in question #4. It is extra credit for ALL students (graduate and undergraduate).

Use the genome.ucsc.edu human genome browser (version NCBI36/hg18)

(a) Using coordinates and/or protein sequence from Figure 2 from the paper you found in question 6 (you need the full text, not just the abstract), and its legend, 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).

(b) 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).

(c) 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.

6. (Advanced) Consider the two state HMM describing DNA sequence that was discussed in class. Namely where one state was GC–poor (we will call this state L) and one state is GC–rich (we will call this state H).

Consider the following parameters of the model:

$T(H,H) = 0.5$ $T(H,L) = 0.5$ $T(L,H) = 0.4$ $T(L,L) = 0.6$

Emissions:

	A	C	G	T
H	.2	.3	.3	.2
L	.3	.2	.2	.3

The probability of starting in H or L is 0.5 => $T(0,L) = 0.5$ $T(0,H) = 0.5$

a) What algorithm is used to calculate the most likely path for a sequence?

b) What is the most likely path for the sequence GGCACTGAA?