7.36/7.91/20.390/20.490/6.802/6.874
PROBLEM SET 1. Sequence search, global alignment, BLAST statistics (19 Points)

Due: Thursday, February 20th at noon.

Problem 1. Sequence search (6 points)
To better understand inborn disorders of metabolism, you isolate a strain of mice that becomes ill unless fed a diet lacking phenylalanine. You sequence the genome of this mouse and find several differences from wildtype including a change to a region that encodes a highly expressed 68 nucleotide RNA which has sequence
5'-UGUACAUGAUGAAGUCAUAGCGAACGGAGAAGGGCCGGCUGAGGAA
ACUGCACGUCACCCUCCUGAAA-3'
in your strain and
5'-UGUACAUGAUGAAAACAGUCUCCCUCUUCUGAAUCUCGCUGAGGAA
ACUGCACGUCACCCUCCUGAAA-3'
in wildtype mice.
Search the sequence in your strain against the mouse genome and transcriptome using NCBI's BLASTn: from the BLAST homepage, click on "nucleotide blast" (not "Mouse") and use the "Mouse genomic + transcript" (G+T) Database, optimized for "Somewhat similar sequences". By expanding the "Algorithm Parameters" box at the bottom, set the Match/Mismatch scores to +1/-3.

(A) (1 pt.) How many statistically significant hits are there at an E-value of 0.05? In one sentence, what does an E-value of 0.05 mean? For transcript hits, what are the maximum reported scores, and are they raw scores or bit scores? (Click on the hyperlink to view individual hits.) To what parts of your RNA do these hits correspond, and what is the % match?

There are two transcript and two genome hits at an E-value of 0.05. The E-value is the expected number of hits with score at least as high as the hit's reported score when searching a query of length 68 nt against the Mouse G+T database. The maximum scores for the two transcript hits are 54 and 50.1 bits. The hit with score 54 bits corresponds to positions 38-68 of the query and has 97% identity to its match (matches 30 of 31 positions), while the hit with score 50.1 bits corresponds to positions 14-38 of the query and has 100% identity.

(B) (1 pt.) Using the E-value and reported score from the result with the highest % identity match from part (A), calculate the approximate length of the Mouse (G+T) Database.

Using the score S = 50.1 bits and E-value = $2* 10^{-4}$ along with m = 68nt in the formula $\mathrm{E-value} = mn2^{-S}$ yields a mouse G+T Database length of $n = 3.55 * 10^{9}$. Note that the mouse haploid genome assembly is about 2.7 billion base pairs, so after adding in transcript sequences, the estimate from the formula is around what we would expect (various corrections to the simple formula are made for base content, repetitive regions, and other parameters for the reported BLAST values).

(C) (1 pt.) Consider a query sequence Q of length L that matches perfectly to a sequence in the database, yielding a BLAST E-value E_1. How would the E-value change if only the first half of Q were searched against the database? In particular, would it stay the same, go up, go down, and how (linearly, exponentially, etc.)?

Intuitively, decreasing the length of query (and therefore match) should make the match more likely simply by chance and therefore less significant, so we should expect the E-value to increase. Quantitatively, if the sequence query length were halved ($m \rightarrow m/2$), the score S would decrease by a factor of 2 ($S \rightarrow S/2$) since there are half as many positions at which to accumulate positive match scores. Plugging these into equations for the original query sequence (with score E_1) and the half-length query sequence (with score E_2) yields:

$E_1 = mn2^{-S}$ and $E_2 = \frac{m}{2} n2^{-S/2} \Rightarrow E_1 2^{S} = 2E_2 2^{S/2} \Rightarrow E_2 = E_1 2^{(S/2 - 1)}$.

Thus, the E-value increases essentially exponentially, with an additional decreasing linear factor of 2 due to halving m. But this latter effect is much smaller than the exponential increase resulting from the decreased score.

(D) (1 pt.) Returning to the BLAST results from part (A), to what genes and RNA classes do the transcript hits with E-values below 0.05 belong? Does your RNA match the sense or antisense direction of these hits? (Click on the hyperlink of the hit and look at the "Strand" section, which tells you the DNA strand of the Hit/Query.)

Of the 2 statistically transcript significant hits at an E-value of 0.05, one matches nucleotides 14-38 of your RNA complementary to (matching the antisense direction of) an mRNA that encodes the phenylalanine hydroxylase (PAH) enzyme. Nucleotides 38-68 of your RNA match the sense direction of Snord100, a C/D Box snoRNA (a type of noncoding RNA that directs posttranscriptional modifications of other RNAs).

(E) (2 pts.) After performing an RNA-protein affinity purification (pull-down) from mouse cell lysates followed by mass spectrometry, you determine that your RNA interacts with the product of the *ADAR1* gene. What does this enzyme do, and what type of RNA does this enzyme act on? Looking back at the function and strand of the gene hit to the second part of your RNA, state a hypothesis as to how your RNA might function to cause your mouse's metabolic disorder. (Hint: on the BLAST hit entry corresponding to the mRNA, click on the "Graphics" link to see the hit in red and how your query at the bottom overlaps with it. If ADAR1 acts at the UAU codon, what is the resulting change during translation?)

The ADAR1 enzyme catalyzes A-to-I editing, post-transcriptionally deaminating adenosine in double-stranded RNA duplexes, yielding inosine. Since I is interpreted as G during translation, A-to-I changes in protein-coding sequences may lead to codon changes and altered functional properties of the proteins. In addition, A-to-I editing can play important roles in regulating gene expression, such as by altering alternative splicing, miRNA sequences, or miRNA target sites in the mRNA.

The *PAH* gene product is a critical enzyme in phenylalanine metabolism and catalyzes the rate-limiting step in its complete catabolism. Nucleotides 14-38 of your RNA overlap a region of the

PAH ORF antisense to the mRNA, including Tyrosine 414 encoded by the codon UAU. Deamination of this adenosine by ADAR would result in the ribosome interpreting a UGU codon, which encodes for the much smaller Cysteine. Thus, your mutant snoRNA provides an RNA duplex for ADAR1 to cause a missense mutation, which could resulting in reduced activity of the PAH enzyme and contribute to your mouse's metabolic disorder. Indeed, genetic Y414C mutations have been observed in human Phenylketonuria patients, and the mutation has been shown to induce global PAH conformational changes (Gersting *et al.* Am. Journ. Human Genetics *83* 2008 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2443833/pdf/main.pdf). Note that the RNA found in the wildtype mouse is very similar to the normal Snord100 snoRNA, which directs 2'O-ribose methylation of rRNA and does not affect PAH.

The example in this problem was inspired by SNORD115 (HBII-52), a human brain-specific C/D box snoRNA that exhibits sequence complementarity to an alternatively spliced transcript of the serotonin receptor. For more details of how SNORD115 regulates serotonin processing through A-to-I editing and alternative splice products, see Kishore and Stamm *Science* 2006 (http://www.sciencemag.org/content/311/5758/230.full.pdf).

Problem 2. Gapped sequence alignment (6 points)
In this problem, you will use the algorithms discussed in class to find the optimal alignment for a pair of short peptides.

(A) (1 pt.) In order to perform this alignment, you must first choose a scoring matrix. For example, you could use a constant match and mismatch penalty of 1 and -1, respectively, so that $S_{ij} = 1$ if $i = j$ and $S_{ij} = -1$ otherwise. Is this a good idea? Why or why not? In one sentence, briefly describe how you might obtain a better scoring matrix for protein comparison.

No - not all amino acid substitutions are equally (dis)favored. Some changes will more heavily impact protein structure and function than others, and will therefore evolve less frequently, and so they should be scored differently. For example, changing from one medium-sized hydrophobic residue to another (e.g., Val to Ile or Leu) within a signal peptide or transmembrane helix is often tolerated, but changing a hydrophobic to a charged residue could disrupt function in these contexts, and changing a buried medium-sized hydrophobic residue like Val to a much larger residue (e.g., Trp) could disrupt packing. Instead, commonly used scoring matrices are created by comparing related protein sequences and seeing how often evolution has allowed particular substitutions occur - these matrices better capture proteins' functional constraints than this simple +1/-1 scoring scheme.

(B) (1 pt.) You decide to explore more commonly used protein alignment scoring matrices instead. Compare the score for aligning two tryptophans (W) to the score for aligning two alanines (A) in the PAM250 scoring matrix. Both of these alignments are "matches", so why are these scores so different?

W-W pairings have a large positive score, while A-A pairings have a small positive score. This means that tryptophan residues are generally highly conserved, and changes from tryptophan to another amino acid are rare (and therefore generally evolutionarily unfavorable). Conversely, alanine is not as strongly conserved and changes relatively frequently. From a biochemical perspective, this makes sense since alanine is very small and won't generally have a big impact on protein structure (and is similar to many other nonpolar amino acids), while tryptophan is very big and changing it to almost anything else could dramatically alter protein structure.

(C) (2 pts.) Perform a **global** alignment of the two peptides ATWES and TCAET, using the Needleman-Wunsch algorithm to fill out the alignment matrix below. Use the **BLOSUM62** scoring matrix and a linear gap penalty of 2.
After filling out the matrix, circle the traceback path and write the final alignment. If there are multiple traceback paths, write out all top-scoring alignments.

Using the BLOSUM62 matrix in the textbook or commonly found online:

	Gap	A	T	W	E	S
Gap	0	**-2**	-4	-6	-8	-10
T	-2	0	**3**	1	-1	-3
C	-4	-2	1	**1**	-1	-2
A	-6	0	-1	**-1**	0	0
E	-8	-2	-1	-3	**4**	2
T	-10	-4	3	1	2	5

The traceback is highlighted in gray above. The final alignment is:

A	T	W	-	E	S
-	T	C	A	E	T

Note: There was a slightly different version of the BLOSUM62 matrix on the lecture slides (the scoring matrix was created from a different set of aligned sequences). This does not change the traceback or final alignment, only a few scores as shown below. Full credit was given for either answer. Using the BLOSUM62 matrix in the lecture slides:

	Gap	A	T	W	E	S
Gap	0	**-2**	-4	-6	-8	-10
T	-2	0	**3**	1	-1	-3
C	-4	-2	1	**3**	1	0
A	-6	1	-1	**1**	3	1
E	-8	-1	1	-1	**6**	4
T	-10	-3	4	2	4	8

(D) (2 pts.) Different scoring matrices and gap penalties can give very different alignment results. Below is the alignment of the peptides from part (C) using the **PAM250** scoring matrix (same gap penalty). The traceback path is shaded.

	Gap	A	T	W	E	S
Gap	0	-2	-4	-6	-8	-10
T	-2	**1**	1	-1	-3	-5
C	-4	**-1**	-1	-3	-5	-3
A	-6	-2	**0**	**-2**	-3	-4
E	-8	-4	-2	-4	**2**	0
T	-10	-6	-1	-3	0	3

What is the resulting alignment?

A - T W E S
T C A - E T

Compare the optimal alignments obtained using the BLOSUM62 and PAM250 scoring matrices. Why are they different?

The main reason the alignments are different is because of how strongly the C-W mismatch is penalized under the PAM250 matrix (score = -8), compared to in the BLOSUM62 matrix (score = -2). This means that under BLOSUM62 the C-W mismatch is tolerated without producing a gap, whereas under PAM250 a gap is preferred over the strong -8 penalty. Additionally, under PAM250, A-T pairings are more favorable (score = +1 vs. 0 for BLOSUM62).

Problem 3. Sequence similarity search statistics (7 points)
You are conducting local nucleotide sequence alignments with your favorite local alignment tool (e.g. BLAST) with match and mismatch scores of +1 and -1 respectively. You align a 100bp query sequence to a 1Mbp genome and find that a 20-nt subsequence from your query is a perfect match.

For each of the following cases, calculate the significance of a 20-nt perfect match (assume $K = 1$ in each case):
Note: The Gumbel distribution is continuous, so the P-value for a score x, $P(S \geq x)$, is equal to the formula $P(S > x)$ on the lecture slides for continuous x since a single point $P(S = x)$ has no probability mass. However, we are applying this continuous distribution to a scoring system that only takes on discrete values, so the $P(S = x)$ values in our scoring system have nonzero mass (a reasonable value for $P(S = x)$ would be $CDF(x+1) - CDF(x)$, where CDF is the cumulative distribution function given on the lecture slides). Thus, our intention was that the P-value is $P(S \geq 20) = P(S > 19)$, so 19 would be plugged into the Gumbel CDF formula; however, since the lecture slides and the textbook have different wording regarding $P(S \geq x)$ vs. $P(S > x)$, we will accept P-values with either 19 or 20 used in the Gumbel formula.
(A) (2 pts.) Query sequence and genome both have approximately balanced base composition A=C=G=T=25%).

Since every pair of nucleotides occurs with equal probability, the probability of a match (A/A, T/T C/C or G/G) is ¼, and the probability of a mismatch is therefore ¾. So to find λ, we need to solve $\frac{1}{4}\, e^{\lambda} + \frac{3}{4}\, e^{-\lambda} = 1$, which has solutions $\lambda = 0$ or $ln(3)$ (by substituting in $y = e^{\lambda}$). Since λ must be positive, we use $\lambda = ln(3)$. The score for the perfect 20nt match is x=20, so using the distribution of the scores $P(S > x) = 1 - \exp[-KMNe^{-\lambda x}]$, we obtain the P-value:

$$P(S \geq 20) = P(S > 19) = 1 - \exp[-(100)(1000000)e^{-19ln(3)}] = 0.0824.$$
(0.0283 for $x = 20$)

(B) (1 pt.) Query sequence and genome are both highly A-T rich (A=T=40%, C=G=10%).

A/A and T/T matches occur with probability 16/100 while C/C and G/G matches occur with probability 1/100. There are also two mismatches each with probability 16/100 (A/T and T/A) and two with probability 1/100 (C/G and G/C). The remaining 8 pairs are all mismatches with probability 4/100. Overall, the total probability of a match is 34/100 and probability of a mismatch is 66/100. We need to solve $(0.34)\, e^{\lambda} + (0.66)e^{-\lambda} = 1$, which has nonzero solution $\lambda = 0.6633$. The corresponding P-value is:

$$P(S \geq 20) = P(S > 19) = 1 - \exp[-(100)(1000000)e^{-19(0.6633)}] \approx 1.$$
(also ≈ 1 for $x = 20$)

(C) (1 pt.) Query is moderately **A+T**-rich (A = T = 30%, C = G = 20%) but genome is moderately **C+G**-rich (A = T = 20%, C = G = 30%).

In this case, all matches are equiprobable with probability (0.3)(0.2) = 0.06. Therefore the probability of a match is 4(0.06) = 0.24, and the probability of a mismatch is 1-0.24 = 0.76. Solving $(0.24)e^{\lambda} + (0.76)e^{-\lambda} = 1$, we obtain nonzero solution $\lambda = 1.153$, and the P-value is:

$$P(S \geq 20) = P(S > 19) = 1 - \exp[-(100)(1000000)e^{-19(1.153)}] = 0.0301.$$

(0.0096 for $x = 20$)

(D) (1 pt.) Briefly explain why the ordering of the P-values from (A) - (C) makes sense.

Since in (B) we are searching a highly A-T rich query against a highly A-T rich genome, we expect to see more similarity between the query and the genome by chance than in (A). Therefore, the match becomes much less significant than in (A). When the query is A-T rich and the genome is G-C rich as in (C), however, a match becomes less likely than if both query and genome had equiprobable base compositions as in (A), and so the P-value in (C) is smaller than in (A).

(E) (2 pts.) Design a new scoring system for application to searching a 20 nt query of unbiased composition against a highly A+T-rich genome (as in (B) above) that will increase the sensitivity for detection of matches to that genome by drawing lines from each box on the left to its new score in the right box (+1, 0, or -1 for different types of matches/mismatches). What would the P-value of a perfect match to this query (with 5 A's, 5 C's, 5 G's, 5 T's) be using your new scoring system?

Since C/C and G/G matches are unlikely by chance due to their low genome content, observing these matches provides the most evidence of a true alignment; they should therefore be given a score of +1. In contrast, because A/A and T/T matches will occur fairly often simply by chance due to their high genome content, these matches provide less evidence of a true alignment and should be given a score of 0. Mismatches generally provide evidence against a true alignment, so they should be given a score of -1.

With a query of unbiased content (A=C=G=T=25%) against the biased genome (A=T=40%, C=G=10%), there is 0.05 total probability of C/C or G/G match (score = +1), 0.2 probability of A/A or T/T match (score = 0), and 0.75 probability of a mismatch (score = -1). The equation $\frac{5}{100}\,e^{\lambda} + \frac{20}{100} + \frac{75}{100}\,e^{-\lambda} = 1$ leads to $\lambda = 2.7081$.

For a perfectly matched 20 nt query of unbiased content, there will be 10 matches of score +1 (C/C and G/G) and 10 matches of score 0 (A/A and T/T), for an overall score of +10. The P-value is therefore:

$P(S \geq 10) = P(S > 9) = 1 - \exp[-(20)(1000000)e^{-9(2.7081)}] = 5.1988 * 10^{-4}$

($3.4665 * 10^{-5}$ for $x = 10$)

7.91J / 20.490J / 20.390J / 7.36J / 6.802J / 6.874J / HST.506J Foundations of Computational and Systems Biology

Spring 2014