Math 5233 Midterm, due Friday March 9th at 4 pm.

Groups:
Group 1: Brad, Luke, Amanda, Marissa.
Group 2: Shane, Doug, Bethany, Noland.
Group 3: Terrence, Eric, Alayna, Maria.

(1) Many Plasmodium proteins have long asparagine-rich inserts. For well-understood proteins, these inserts seem to be located away from the active sites of the protein and probably form soluble, unstructured (not helices or sheets) loops between necessary components of the protein. For sequence alignment of such a protein to a standard database (such as SwissProt or the nonredundant protein database 'nr' at NCBI), what choice of scoring matrix, gap opening penalty, and gap extension penalty might help relative to the NCBI Blastp defaults of: BLOSUM62 for the scoring matrix, -11 for the gap opening penalty, and -1 for gap extension? You may wish to experiment by looking at the results for proteins such as PFC0250c (an endonuclease), PFF0830w (alpha adaptin), PF11_0395 (guanylyl cyclase), PFL1880w (CoA ligase), or others. Explain your choices and their possible pros and cons.

(2) In the (assigned reading) article '*Quod erat demonstrandum?* The mystery of experimental validation of apparently erroneous computational analyses of protein sequences', the first case studied is that of a putative archaeal cysteine-tRNA synthetase. The candidate is a protein MJ1477 in the Archaeal thermophile Methanocaldococcus jannaschii.

(a) Find the relevant pathway for cysteine-tRNA synthesis in the KEGG Pathway database for M. jannaschii. What enzymes are linked to this reaction in KEGG?

(b) Get the protein sequence for MJ1477 and use Blastp to find similar sequences. Are there any good hits to cysteine-tRNA synthetases?

(c) Do you think that MJ1477 is a cysteine-tRNA synthetase? In other words, who is right? To make up your mind, you may wish to search other databases such as PubMed for more information.

(3) Suppose that P and Q are probability distributions on N items - that is, $P = \{p_1, \ldots, p_N\}$ and $Q = \{q_1, \ldots, q_N\}$ are both lists of N numbers between 0 and 1, which sum to 1. Show that the difference in entropy between

$$H_P - H_Q = -\sum_{i=1}^{N} p_i \log p_i + \sum_{i=1}^{N} q_i \log q_i$$

is equal to the relative entropy

$$H(P||Q) = -\sum_{i=1}^{N} p_i \log (p_i/q_i)$$

of P relative to Q, if Q is the uniform distribution (each item has a $1/N$ probability).

(4) Suppose we view each of the approximately 30,000 proteins in the human genome as a message from the genome to the cell, and suppose that their frequency of expression follows Zipf's law: the most frequent protein is twice as likely as the second-most frequent protein, three times as likely as the third-most frequent protein, and so on. What is the entropy of this distribution? How does it compare with the entropy of the uniform distribution for 30,000 items?