HST.950/6.872 Problem Set 2

Due 11/03/2005

1. Klenk H.-P. and colleagues published the complete genome sequence of the organism Archaeoglobus fulgidus in *Nature*. See:
   Klenk, H. P., et al. "The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon Archaeoglobus fulgidus." *Nature*. 390, no. 6658 (November 27, 1997): 364-70.
   You will have to use additional resources to find the sequence.
   1. Find the protein sequence of the hypothetical protein AF1226 precursor. What is the sequence of amino acids (in single letter representation) from positions 141-147?
   2. Write a Matlab function that calculates how many nucleotide sequences can give rise to an arbitrary amino acid sequence. The input to the function should be a string (amino acid sequence) and the output should be the number of potential nucleotide sequences. This link will be helpful: http://www.bio.davidson.edu/courses/Molbio/aatable.html

      As will this Table:

AMINO ACID	RNA CODON
ALANINE (A)	**GCC, GCA, GCG, GCU**
ARGININE (R)	**AGA, AGG, CGU, CGA, CGC, CGG**
ASPARAGINE (N)	**AAC, AAU**
ASPARTIC ACID (D)	**GAC, GAU**
CYSTEINE (C)	**UGC, UGU**
GLUTAMIC ACID (E)	**GAA, GAG**
GLUTAMINE (Q)	**CAA, CAG**
GLYCINE (G)	**GGA, GGC, GGG, GGU**
HISTIDINE (H)	**CAC, CAU**
ISOLEUCINE (I)	**AUA, AUC, AUU**
LEUCINE (L)	**UUA, UUG, CUA, CUC, CUG, CUU**
LYCINE (K)	**AAA, AAG**
METHIONINE (M) (Start)	**AUG**

PHENYLALANINE (F)	**UUC, UUU**
PROLINE (P)	**CCA, CCC, CCG, CCU**
SERINE (S)	**UCA, UCC, UCG, UCU, AGC, AGU**
THREONINE (T)	**ACA, ACC, ACG, ACU**
TRYPTOPHAN (W)	**UGG**
TYROSINE (Y)	**UAC, UAU**
VALINE (V)	**GUA, GUC, GUG, GUU**
Stop	**UAA, UAG, UGA**

3. Using the Matlab function from part 2., calculate the number of sequences that could give rise to the 7 amino acid sequence found in part 1.

2. Neurofibromatosis-1 is an inherited disorder characterized by formation of neurofibromas (tumors involving nerve tissue) in the skin, subcutaneous tissue, cranial nerves, and spinal root nerves. NF1 is an autosomal dominant trait, caused by a mutation in the NF1 gene.
    1. The Affymetrix microarray probe (from human genome array: HG-U133+ 2.0) for the NF1 is as follows:

       AAGTGCCATGTTCCTCAGATTTATC

       If one starts with a sequence of length n, derive a formula for the probability that it will match is random 25 base sequence exactly assuming 1) base types are uniformly distributed within each position and 2) each base position is independent of the other.
    2. Using the same assumptions as above, what is the expected number of 25-base sequences that will match the sequnce given in part 1, within the entire human genome (3x10^9 base pairs) going in the 5' to 3' direction (i.e. only looking in one direction)?
    3. Using the same assumptions as above, what is the theoretical ratio of the number of serine to tryptophan amino acids in the human genome?
    4. Are the numbered assumptions made in question parts 1, 2, and 3, always correct? Explain why or why not.
    5. Diagnosis of the disease in this question is typically done via clinical evaluation rather than by microarray analysis. Describe a use for having this gene as a probe in the HG-U133+ 2.0 Genome Array. How might it be useful in other settings, in conjunction with other technologies, or in medicine in the future?

3. You have a sample of human DNA. However, in order to analyze it, you need to 'amplify the signal' by creating multiple copies of the sequence. This can be done via a process called PCR (Polymerase Chain Reaction). In one step of the process, a DNA polymerase is used. Such polymerases typically come from organisms like: *Thermus aquaticus*, *Pyrococcus furiosus*, and *Thermococcus litoralis*. What would happen if you used human DNA polymerase? Why? (What do the aforementioned organisms have in common?)

4. The following is a protein expression circuit perturbation-based profile. It lists which proteins are present in a particular system under different trial circumstances. In each trial (except t=0) one protein is either added or removed.

Trial	**Protein 1**	**Protein 2**	**Protein 3**	**Protein 4**
t=0	0	1	1	1
t=1	1	0	1	removed
t=2	0	0	removed	1
t=3	1	removed	1	1
t=4	added	1	1	1

    1. Use Boolean algebra and DeMorgan's Law (as necessary) to find a simple expression (using only primitives AND, NOT, OR) that is consistent with this protein expression circuit perturbation-based profile.
    2. Your collaborators' experiments suggest that all proteins except protein #3 are involved in the p53 pathway (involved in many cancers). Based on Biocarta (www.biocarta.com) pathways, what are possibilities for proteins 1, 2, and 4 given the protein circuit you derived in part 1?

5. The Human Massome (http://chocolate.chip.org/~protcoop/protind.php) is a research project that seeks to view proteins and their interactions from a mass spec perspective. It contains the largest collection of human protein interactions(>100K protein interactions). It can be searched by inputting the masses of suspected protein interactions in the form. The first line is for the minimum and maximum expected mass of the first protein (or protein product) respectively. The second line is for the minimum and maximum expected mass of the second protein (or protein product) respectively.

    1. You do a SELDI-based mass spec experiment to analyze certain binding proteins in cancer. After some analysis, you see that the following peaks (circled) that are of interest. The mass spec laser/instrument has been optimized for the range of 20000-35000 in this experiment. Assume the instrument accuracy is 400 ppm (parts per million). What are the possible protein masses for the three circled peaks? List a few potential identities for these three.

2. Your collaborator verifies that the proteins represented by the peaks, do, in fact, bind. However, there are only two proteins involved (though the collaborator is unable to identify them in time for your publication deadline). Using the "Human Massome," what might the identity of these proteins be? Why could the collaborator only find complexes with two different types of proteins- is he/she missing something?
3. The collaborator feels bad after having missed the deadline for the publication. So, he offers to help identify some other interactions. Via gel electrophoresis techniques, he finds that a protein with a mass of around 81,372 Da interacts with the 18,012 Da protein above. Why is SELDI mass spec not really suitable to find the new protein (why is it not suitable- in this experiment as well as in general?)?
4. The collaborator is on a roll- and has found out that another protein with a mass of around 55,344 interacts with the new protein found in the last part. What could this protein be?
5. (Optional) Looking at the Biocarta (www.biocarta.com) pathway for CD95, what new pathway (not listed) does this new finding suggest? Are interactions from previous parts of this question in the Biocarta pathway diagram?