Fall 2014 **CMSC 423: MidTerm1** H. Corrada Bravo

Time: 1 Hour, 15 Minutes

WAIT FOR INSTRUCTIONS BEFORE BEGINNING

HONOR PLEDGE: "I pledge on my honor that I have not given or received any unauthorized assistance on this examination."

Signature and UID: ______________________________

Print name: ______________________________

○ ***Write your answers with enough detail about your approach and concepts used, so that the grader will be able to understand it easily.***

○ ***The sum of the grades is 75, but your grades would be out of 70 (thus you get 5 bonus points by solving all the problems).***

○ ***Select the best choice for the first 6 problems and mark it by X in the table below.***

Problem	**1**	**2**	**3**	**4**	**5**	**6**
A						
B						
C						
D						
E						

DO NOT WRITE BELOW THIS LINE

Questions 1-6	Question 9	Question 12
Question 7	Question 10	Total
Question 8	Question 11	

Multiple-choice Problems (Answer THE BEST CHOICE in the Table of the First Page and NOT HERE):

1. **(3 points**) Which of these best represents the relationship between genotype and phenotype?

   a) there is **no** relationship between genotype and phenotype

   b) an individual's phenotype **completely** determines their genotype

   c) an individual's genotype **completely** determines their phenotype

   d) an individual's phenotype **partially** determines their genotype

   e) none of the above

2. **(5 points**) Consider the following Profile for a DNA motif shown below. In which position of string `ATTCAGGA` is the highest probability 3-mer found? (Note: assume 1-indexing so the first character is in position 1. Show your work.)

	Pos 1	Pos 2	Pos 3
A	.8	.6	.4
C	.2	.3	.5
G	0	.3	.1
T	0	0	0

a) 1 b) 2 c) 4 d) 5 e) 6

3. **(2 points**) Which of the following resources *does not* contain high-throughput sequencing data from population experiments:

   a) KEGG Database b) 1000 genomes project

   c) Short Read Archive d) (a) and (b) e) all of the above

4. **(5 points**) Consider cyclic peptide N-I-C-E. I claim that its cyclospectrum is the following:

0-103-113-114-129-216-227-232-330-345-459

I am wrong. Select **the best** explanation why. You can find the integer mass table below.

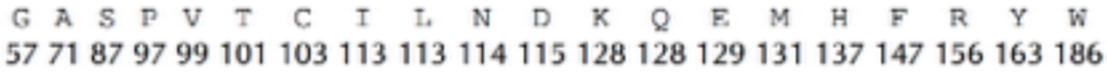
a) This is not a spectrum

b) It is missing the mass of peptide C-E-I

c) It is missing the mass of peptide C-E

d) This is its linear (not cyclic) spectrum

e) None of the above



G	A	S	P	V	T	C	I	L	N	D	K	Q	E	M	H	F	R	Y	W
57	71	87	97	99	101	103	113	113	114	115	128	128	129	131	137	147	156	163	186

5. **(2 points**) What is an *open reading frame* (ORF)?

a) any translatable sequence of nucleotides

b) any sequence of codons

c) a long enough sequence of aminoacids

d) a long enough sequence of codons without an intervening stop codon

e) None of the above

6. **(2 point**) Which of the following are examples of sequence mutations?

a) Single Nucleotide Polymorphism (SNP) b) insertion c) complementation

d) (a) and (b) e) None of the above

Questions (show all derivations as appropriate for full credit):

Problem 7. (6 points) (You can refer to the genetic code figure below). Consider nucleotide sequence S=…CGCATATGAACAAGA...

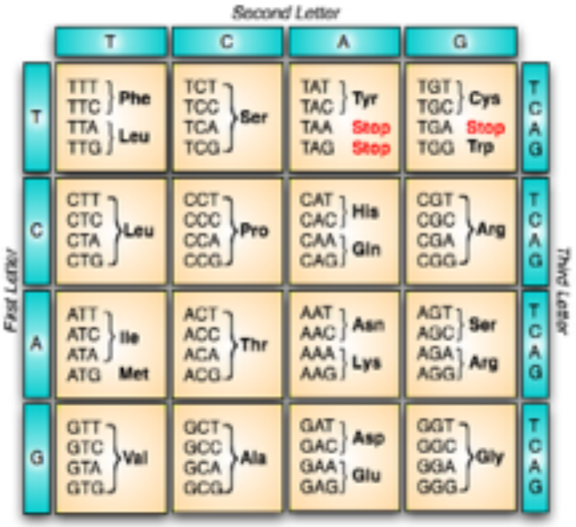
(a) How many *open reading frames* (ORFs) are there in this sequence (only considering forward strand)

(b) Write down the aminoacid sequence resulting from translation of one ORF (specify which).

(c) Specify a *synonymous* nucleotide substitution in this ORF, i.e., does not change aminoacid sequence.

(d) Specify a *non-synonymous* nucleotide substitution in this ORF.

(e) Specify a substitution that closes this ORF; write down the resulting aminoacid sequence.



	Second Letter				
First Letter	T	C	A	G	Third Letter
T	TTT TTC } Phe TTA TTG } Leu	TCT TCC TCA TCG } Ser	TAT TAC } Tyr TAA Stop TAG Stop	TGT TGC } Cys TGA Stop TGG Trp	T C A G
C	CTT CTC CTA CTG } Leu	CCT CCC CCA CCG } Pro	CAT CAC } His CAA CAG } Gln	CGT CGC CGA CGG } Arg	T C A G
A	ATT ATC ATA } Ile ATG Met	ACT ACC ACA ACG } Thr	AAT AAC } Asn AAA AAG } Lys	AGT AGC } Ser AGA AGG } Arg	T C A G
G	GTT GTC GTA GTG } Val	GCT GCC GCA GCG } Ala	GAT GAC } Asp GAA GAG } Glu	GGT GGC GGA GGG } Gly	T C A G

Problem 8. (6 points) Provide a definition of *reproducible data analysis.* Discuss its importance in experimental computational biology. Mention computational tools that can help ensure data analyses are reproducible.

Problem 9. (6 points) Suppose we have a set of four amino acids with the following masses: A=10, B=30, C=50, and D=90. Given the spectrum {0, 10, 30, 40, 50, 50, 60, 70, 90, 100} for a cyclic peptide, we can use the leaderboard cyclo-peptide sequencing algorithm to discover the correct sequence of amino acids. Assume the leaderboard limits the candidate peptides to the two highest scores in any given iteration of the algorithm.

Suppose we have the following candidates on the leaderboard: {AB, AC, AD}. Show the next iteration of the algorithm including expansion, trimming and the calculation of the next leaderboard.

Problem 10. (6 points) Consider the motif finding problem. Give an example of a profile that has low entropy, but high *relative* entropy relative to a set of background nucleotide frequencies. Your answer must include: (a) the profile, and (b) the set of background frequencies you are using to calculate relative entropy.

Problem 11. (12 points). Write down an expression for the probability that the size *k* prefix of a randomly generated DNA string of length *n* is equal to the reverse complement of its size *k* suffix. E.g., ACGTATTAACGT is one such string for n=12 and k=4.

(a) Solve assuming $2k \le n$

(b) Solve assuming $k \le n < 2\mathrm{k}$

Problem 12. (20 points) Consider the motif finding problem. In this question you will extend the algorithms you learned about to look for *maximum relative entropy* profiles.

(a) Argue why we should be looking for *maximum* relative entropy profiles, instead of *minimum* relative entropy profiles. Two or three sentences are sufficient.

(b) In the greedy algorithm and randomized motif search algorithms, we used the concept of *most probable* k-mer in a DNA string to determine next states to explore in the search. Given a profile *Profile,* and a DNA string, how is the *most probable* k-mer in a DNA string defined?

(c) Given a profile *Profile* and a set of *background frequencies,* define a score that you would use in these search algorithms in order to find *maximum relative entropy* profiles. You should write a mathematical expression that computes this score given: *Profile (defined by frequencies* p_{rj} *(the probability nucleotide r occurs in position j of the motif), background frequencies (*b_r*, the probability of observing nucleotide r anywhere in the genome) and a k-mer Pattern.* Write a sentence or two describing this scoring function, make sure to mention if this score is a probability (or not).

(d) In the Gibbs sampler a move was made by randomly choosing a k-mer within a DNA string with probability depending on the *Profile-probability* of each k-mer given the current *Profile.* How would you modify the Gibbs sampler to use your new score? Write out the pseudo-code for your modified Gibbs sampler.