

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 105, but your grades would be out of 100 (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	/ 30	Question 9	/ 20	Total
Question 7	/ 10	Question 10	/ 15	
Question 8	/ 10	Question 11	/ 20	

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

1. **(3 points)** Which of the following is a *non-synonymous* mutation? *You can find the RNA translation code in Problem 7.*

a) AGA->AGC      b) GGU->GGA      c) CGG->AGG  
d) (a) and (b)      e) none of the above

2. **(10 points)** Consider the following DNA motif profile shown below. Which is the profile-most probable 3-mer in string ATTCAGGA? (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) ATT      b) TTC      c) CAG      d) AGG      e) GGA

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

a) 1000 genomes project      b) Pubmed  
c) Short Read Archive      d) (a) and (b)      e) (a), (b) and (c)

4. **(10 points)** I used k-means clustering with  $k=2$  and obtained the two cluster centers in Table 1. What are the cluster assignments for the three genes in Table 2? Show your work (drawing the genes and center helps).

**Table 1. Centers**

	Time 0	Time 1	Time 2
Center 1	1	0	1
Center 2	-1	0	-1

**Table 2. Genes**

	Time 0	Time 1	Time 2
Gene A	2	0	2
Gene B	-1	0	0
Gene C	2	0	-1

- a) A: 1, B: 1, C: 2
- b) A: 2, B: 2, C: 1
- c) A: 1, B: 2, C: 2
- d) A: 1, B: 2, C: 1
- e) A: 2, B: 1, C: 2

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. **(3 points)** 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. (10 points)** (You can refer to the genetic code figure below). Consider RNA sequence S=... CGCAUAUGAACAAGAC...

- Write down the aminoacid sequence resulting from translation of the open reading frame (ORF) starting in the second position of the string (i.e., first codon is GCA).
- Specify a *synonymous* nucleotide substitution in this ORF, i.e., does not change aminoacid sequence.
- Specify a *non-synonymous* nucleotide substitution in this ORF.
- Specify a substitution that closes this ORF; write down the resulting aminoacid sequence.

		Second Letter					
		U	C	A	G		
1st letter	U	UUU   Phe UUC   UUA   Leu UUG	UCU   UCC   Ser UCA   UCG	UAU   Tyr UAC   UAA   Stop UAG   Stop	UGU   Cys UGC   UGA   Stop UGG   Trp	U C A G	3rd letter
	C	CUU   CUC   Leu CUA   CUG	CCU   CCC   Pro CCA   CCG	CAU   His CAC   CAA   Gln CAG	CGU   CGC   Arg CGA   CGG	U C A G	
	A	AUU   AUC   Ile AUA   AUG   Met	ACU   ACC   Thr ACA   ACG	AAU   Asn AAC   AAA   Lys AAG	AGU   Ser AGC   AGA   Arg AGG	U C A G	
	G	GUU   GUC   Val GUA   GUG	GCU   GCC   Ala GCA   GCG	GAU   Asp GAC   GAA   Glu GAG	GGU   GGC   Gly GGA   GGG	U C A G	

**Problem 8. (10 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. (20 points)** Consider the evolution game you worked on in Project 2. Remember that the binding rule in that case was `T[C|G]GTNNNNT[A|G]NT`, (i.e., in position 2 either `C` or `G` allows binding, in positions with `N` any base allows binding, and in position 10, either `A` or `G` allows binding).

- (a) Write down a profile consistent with this binding rule, that is, a profile you would estimate from a population of 12-mers evolved according to the rules of our simulation in Project 2. Show and explain your work.

- (b) Calculate the entropy of the profile you wrote down as answer for (a).

- (c) Calculate the *relative* entropy of the profile for background frequencies  $b_A=1/3$ ,  $b_C=1/6$ ,  $b_G=1/6$ ,  $b_T=1/6$ .

**Problem 10. (15 points).** Write down an expression for the probability that the length  $k$  prefix of a randomly generated DNA string of length  $n$  is equal to the reverse complement of its length  $k$  suffix. E.g., ACGTATTAACGT is one such string for  $n=12$  and  $k=4$ .

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

(b) Solve assuming  $k \leq n < 2k$

**Problem 11 (20 points)** We've seen in class two algorithms that use probability estimates as part of an optimization problem: (a) in the Gibbs sampling algorithm for motif finding, we used the 'profile probability' of a k-mer to sample positions in DNA sequences containing a protein binding site, and (b) in the EM algorithm used in soft k-means, we used 'assignment probability' to calculate cluster centers using weighted averages. Design an EM algorithm to solve the motif finding problem.

1. In soft k-means, the parameters of interest were the  $k$  centers. What is the parameter of interest in motif finding?
2. In soft k-means, *HiddenMatrix* was a matrix with a row for each gene and a column for each cluster center. What does  $HiddenMatrix_{ij}$  correspond to in soft k-means?
3. How is each entry  $HiddenMatrix_{ij}$  calculated in soft k-means? Write the mathematical expression.
4. Now, let's define a similar *HiddenMatrix* for motif finding. It should have  $t$  rows (number of strings in Dna), how many columns should *HiddenMatrix* have in this case? What does  $HiddenMatrix_{ij}$  correspond to in this case?
5. How would you compute  $HiddenMatrix_{ij}$  for motif finding? Write a mathematical expression. Note: this is the *E-step* in your algorithm.
6. In soft k-means, *HiddenMatrix* was used to calculate weighted means. Write the mathematical expression to calculate the  $i$ -th cluster center. Note: this is the *M-step* in fuzzy k-means.
7. Given *HiddenMatrix* for motif finding as you've defined above, how would you use it to calculate a motif profile? The key here is how to calculate *weighted counts*. Write a mathematical expression for entry  $p_{cj}$  corresponding to nucleotide  $c$  and position  $j$  of the profile. Note: this is the *M-step* of your algorithm.



