WRITE ALL ANSWERS IN YOUR BLUE BOOKS. THIS PAPER WILL NOT BE GRADED.

20.320 Exam 1
Thursday October 13th
9:35-10:55

Instructions:

0. Write your name of the front cover of the blue book.

1. Answer all questions in the blue books. This exam paper will not be graded.

2. All questions can be answered in at most a few sentences. We will deduct points for excessively long replies, even if they contain the right answer.

3. State all assumptions for each problem.

4. In order to pace yourself please note that the maximum possible score on this exam is 100.

5. There are six problems on ten pages (total) in the exam.

Question 1 – Protein Folding and Synthesis (12 points)

Directions: Answer each of the following questions in your blue books with just a few sentences. Include equations where relevant. *Excessively long answers will not be graded.*

You are using a bacterial system to produce a protein to be used as a therapeutic. The bacterial product is chemically pure and identical to that produced in humans. However, the material you obtain is not enzymatically active. You presume that improper folding might be a problem.

a. In your blue book, draw the thermodynamic “pathway” for a simple two-state model of protein folding on a reaction coordinate diagram, and label important features of the curve. Please label your axes as well.

b. Consider an experiment where you rapidly dilute a concentrated, denatured protein and measure the rate at which it refolds (your protein starts in urea, which denatures proteins, and is diluted in aqueous solution). Write a differential equation for the rate of protein folding, determine the solution to the equation (it was solved for you in class!) and qualitatively sketch how the “faster” and “slower” proteins would look on a plot of unfolded protein vs. time. (Assume that you can ignore the reverse reaction in which some of the newly folded protein unfolds).

c. You attempt to denature and refold the bacterial protein *in vitro*. While you are able to obtain some active protein, the yield is low. When you increase the initial concentration of your protein, the yield from the system drops further. What might cause this?

d. You have identified mutations that increase your yield when the protein is refolded at high concentrations. However, these mutations have no effect on the yield when the protein is refolded at very dilute concentrations. Provide an explanation for these observations.

Question 2 – Sequence Motifs
(20 points)

Directions: Answer each of the following questions in your blue books with just a few sentences. Include equations and diagrams where relevant. *Excessively long answers will not be graded.*

ChIP-Seq is a common experiment that allows the experimenter to identify pieces of DNA to which a particular protein is bound and is often used for determining sequence patterns for binding transcription factors.

Suppose you decided to work with a transcription factor, *yfp2* and ran a ChIP-Seq experiment which showed that *yfp2* bound to a 8-mer. You calculate the probability matrix, given below:

	1	2	3	4	5	6	7	8
A	0.3	0.1	0.25	0.09	0.15	0.3	0.2	0.5
C	0.05	0.1	0.6	0.01	0.4	0.01	0.2	0.15
G	0.45	0.1	0.05	0.2	0.4	0.19	0.3	0.15
T	0.2	0.7	0.1	0.7	0.05	0.6	0.3	0.2

a. Under this model, what is the probability of observing the most likely sequence? Why is it so low?

b. Does the fact that the probability is so low suggest that the transcription factor won't bind this sequence? Briefly explain.

c. Write an equation for the log-likelihood ratio that you would use to determine whether a sequence was more likely to be a binding site for *yfp2* or a random region in this genome.

Continued on next page

Question 2 – continued

d. In class we discussed the advantages of using sequence motifs over consensus sequences. However, there are some features that sequence motifs do not capture. Carefully examine the sequences below, which are a representative part of a much larger data set.

Identify a sequence pattern that is not captured by a sequence motif built from these sequences (even if you include pseudocounts). In other words, what types of sequences are unlikely to ever occur in this full dataset but would be scored well by a motif built on these data. (You do not need to compute the motif to answer this question).

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
A	C	T	A	T	T	C	G	T	A	G	T
G	G	C	A	T	T	C	G	T	G	C	C
A	G	A	A	T	T	C	G	T	T	C	T
G	G	T	A	T	T	C	G	T	A	C	C
C	C	G	T	T	T	C	G	A	C	G	G
C	T	G	C	T	T	C	G	G	C	A	G
A	A	G	C	T	T	C	G	G	C	T	T
C	G	A	C	T	T	C	G	G	T	C	G
C	A	C	G	T	T	C	G	C	G	T	G
C	T	A	G	T	T	C	G	C	T	A	G
C	G	C	G	T	T	C	G	C	G	C	G
A	A	G	A	T	T	C	G	T	C	T	T
A	A	C	T	T	T	C	G	A	G	T	T
T	G	G	T	T	T	C	G	A	C	C	A
T	A	C	T	T	T	C	G	A	G	T	A

Question 3 - Protein Secondary Structure (18 points)

Directions: Answer each of the following questions in your blue books with just a few sentences. Include equations where relevant. *Excessively long answers will not be graded.*

Chou-Fasman:

a. Why does Chou-Fasman perform less well on predicting sheets than helices?

b. Chou-Fasman is a knowledge-based algorithm. Explain what this means and what the knowledge-based parameters are.

c. List two reasons why you don’t find Proline in an alpha helix. What about Glycine (one reason)?

Coiled coils:

a. Looking down the length of one helix in a coiled coil, there is often a periodic patterning of hydrophobic and hydrophilic residues. What is the period of this repeat?

b. We looked at a general type of coiled-coil domain called a leucine zipper. Explain the importance of the leucines in this domain.

c. Suppose you wanted to create a new algorithm, based on Chou-Fasman, that would search for coiled-coils in an amino acid sequence data. Describe briefly (no more than 3 sentences) how your algorithm would work, and what data you would need to be able to construct the algorithm.

Question 4 - Levinthal's Paradox and Combinatorial Search (20 points)

Directions: Answer each of the following questions in your blue books with just a few sentences. Include equations where relevant. *Excessively long answers will not be graded.*

a. What does Levinthal's paradox tell us about how proteins actually fold?

b. You're in luck! You have decided to work on your design project, and the Athena cluster is completely empty tonight. You anticipate you have time to explore 10^40 different protein conformations.

   You are searching for a novel sequence that will fold into a stable structure. You intend to try every amino acid at every position, and a reasonable number of backbone angles for each residue. Please write an equation for the length of protein you can design, and estimate the value of this equation. You do not need to give a precise answer. Make sure you state any of your assumptions.

c. You quickly realize that you will not be able to search all possible conformations for a reasonably sized protein. Instead you decide to search for the most stable conformation of each potential sequence using the type of Metropolis algorithm described in class. Describe the steps in this algorithm. Include simulated annealing in your answer.

Continued on next page

This course makes use of Athena, MIT's UNIX-based computing environment. OCW does not provide access to this environment.

Question 4 – continued

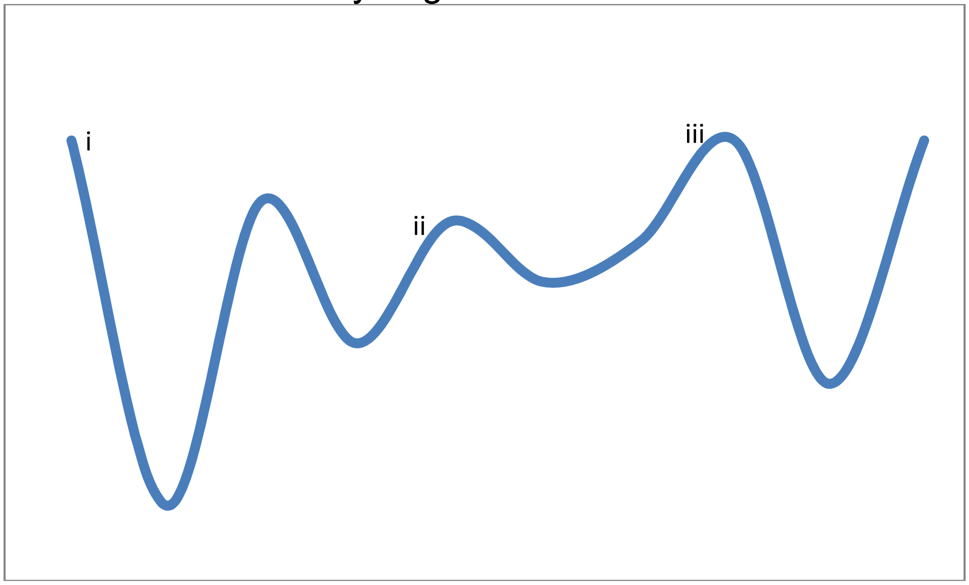
d. Copy the following diagram into your blue book and indicate the final state or states reached by the algorithm on the energy landscape below at

   i. high temperature,

   ii. low temperature, and

   iii. with simulated annealing

   Use the indicated start position for each trajectory and assume that the algorithm terminates either when it reaches a steady state or after a very large number of iterations.



Make your answer clear for each part. It may help to use separate diagrams for each section i, ii, and iii.

Be sure to number your steps for clarity, and include the answer in your blue book! (Sketch the landscape, DOES NOT HAVE TO BE EXACT OR A WORK OF ART)

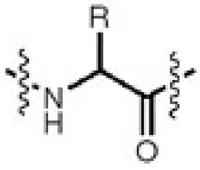
Question 5 (10 points)

Directions: Answer each of the following questions in your blue books with just a few sentences. Include equations where relevant. *Excessively long answers will not be graded.*

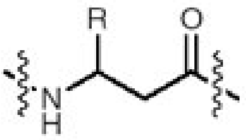
a. Name three levels of protein representation discussed in class.

b. Give one advantage or disadvantage of using each representation.

c. Below are structures of the naturally occurring alpha-peptides and of a beta-3 peptide. If you are trying to make a protein therapeutic, why might you prefer the beta-3 peptide over the alpha-peptide.



α-peptide



β_3-peptide

d. For your first attempt at designing a therapeutic molecule out of beta-3 peptides you synthesize a polymer of these peptides with the same sequence of side chains at the “R” positions as in a naturally occurring globular protein. Will this new molecular have the same function as the naturally occurring version? Briefly explain your answer.

e. What’s the difference between positive and negative design? Why should you do negative design? (Hint, think about Prof. Keating’s bZIP problems!)

Question 6 (20 points)

Directions: Answer each of the following questions in your blue books with just a few sentences. Include equations where relevant. *Excessively long answers will not be graded.*

A. You have discovered that there are three common alleles of a gene encoding a key signaling protein that regulates cell growth. These variants differ in the extent to which they are phosphorylated by a kinase called Exm320.

The kinase specificity is described by the following pattern: AXSP. All three protein variants contain this sequence. The highest affinity variant has a short amino-acid insertion approximately 100 amino acids away from the site of phosphorylation.

Very briefly explain how this sequence might increase the levels of phosphorylation.

Continued on next page

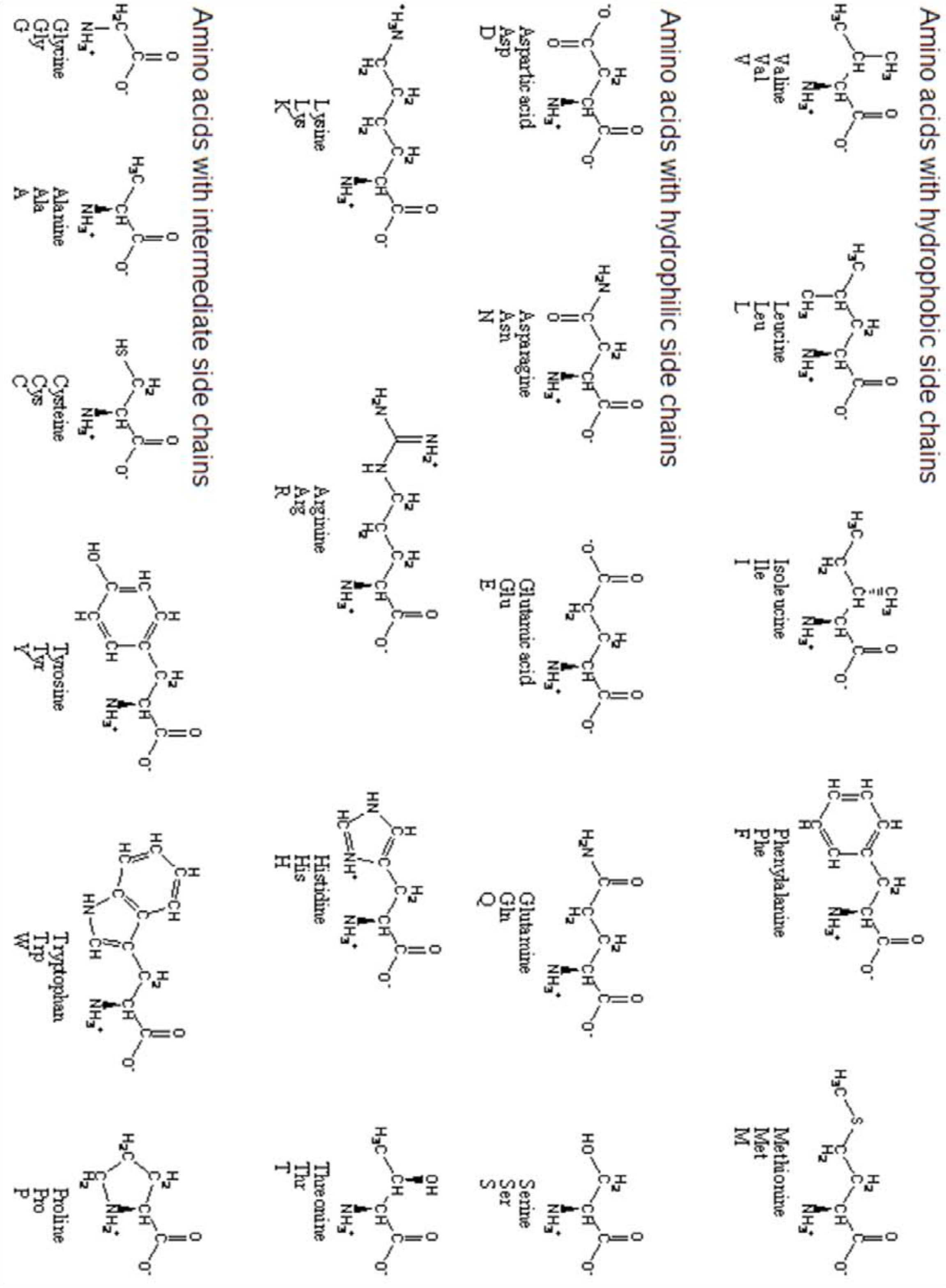
Question 6 – continued

B. The remaining two variants lack this region and differ in the sequences shown below. The site of the phosphorylation is indicated by an asterisk (*).

Best & Second Best Substrates:	A	A	L	A	G	L	A	V	I	I	A	V	S	P
Poor Substrate:	A	R	L	A	D	L	A	R	K	I	A	V	S	P
													*	

You presume that the difference in sequence must explain the difference in their strength as substrates. To prove your hypothesis, you synthesize two peptides containing the sequences shown above. To your surprise and disappointment, you find that they are both phosphorylated with equal efficiency.

Very briefly provide a **single plausible explanation** for the observations that (1) the bottom sequence causes the protein containing it to be a worse substrate than the top sequence and (2) the peptides containing these sequences are both equally good substrates. Your answer should not be the same one you used in part "A". Rationalize your answer based on the ***relative positions of amino acids*** that vary among these sequences.


Amino acids with hydrophobic side chains
Valine
Val
V
Leucine
Leu
L
Isoleucine
Ile
I
Phenylalanine
Phe
F
Methionine
Met
M
Amino acids with hydrophilic side chains
Aspartic acid
Asp
D
Asparagine
Asn
N
Glutamic acid
Glu
E
Glutamine
Gln
Q
Serine
Ser
S
Lysine
Lys
K
Arginine
Arg
R
Histidine
His
H
Threonine
Thr
T
Amino acids with intermediate side chains
Glycine
Gly
G
Alanine
Ala
A
Cysteine
Cys
C
Tyrosine
Tyr
Y
Tryptophan
Trp
W
Proline
Pro
P

20.320 Analysis of Biomolecular and Cellular Systems
Fall 2012