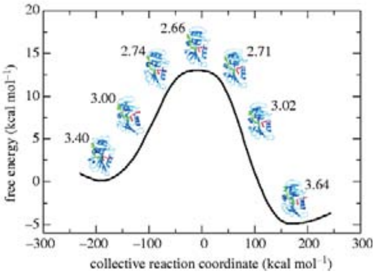
Question 1 – Protein Folding and Synthesis
(12 points)

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. (2 points)

I was looking just for an unfolded state, a (relatively) higher activation energy and then a lower folded state energy (the proteins in the picture are irrelevant).

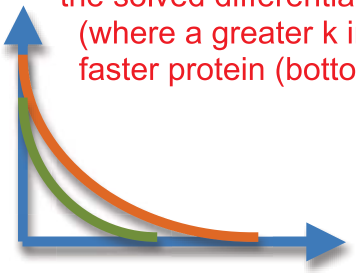


b. Consider an experiment where you rapidly dilute a concentrated, denatured protein and measure the rate at which it refolds (your protein starts in urea, which denatures proteins, and is diluted in aqueous solution). Write a differential equation for the rate of protein folding, determine the solution 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). (3 points)

the rate equation: d U(t)/dt = -kU(t)

the solved differential: $U(t) = U_0e^{-kt}$ (where a greater k indicates a faster protein (bottom curve)

“faster” protein)
and slower protein (top curve)



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? (3 points)
It's likely that the protein is forming aggregates.

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. (4 points)
The mutations must be preventing aggregate formation. These mutations wouldn't affect yield at low concentrations since aggregation is only really a problem at high concentration.

(If they explained that aggregation was only a problem at high concentrations in part c, I still gave them full credit here, but I was looking to see that they connected the mutations to the aggregates and that they recognized aggregation is less of a problem at dilute concentrations).

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? (4 points)
(2 points for explanation and 2 for the calculation)
The most likely sequence can be constructed simply by choosing the most likely sequence at each position. There are four sequences of equal maximum probability:

GTCT[C/G]T[G/T]A. The probability is simply the product of their entries: 0.45*0.7*0.6*0.7*0.4*0.6*0.3*0.5 = 0.0047628.

It is very low because there are many possible observable sequences under our model (every possible sequence is also possible in our model with nonzero probability!)

b. Does the fact that the probability is so low suggest that the transcription factor won't bind this sequence? Briefly explain. (4 points)

This is the probability of seeing a sequence, assuming that the transcription factor binds that sequence. This tells us nothing about the validity of our assumption. Put differently, P(A|B) says nothing of P(B). To assess the validity of our assumption, we compare the likelihood of our data under this assumption to the likelihood assuming it's not a transcription factor (this comparison is the likelihood ratio).

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.
(4 points)
log(P(sequence|motif model) / P (sequence | background/genome model)) =
log(P(sequence|motif model)) – log(P(sequence|background/genome)) =
sum over all positions i [log(p_i|motif)] - 8 log 0.25

Note that log(a/b) does not equal log(a)/log(b)

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**	**1**	**1**	**1**
									0	**1**	**2**
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

(8 points)
In all cases in this matrix, positions 1-4 are perfectly complementary to positions 12-9. This is an example of dependence (very very strong dependence) between positions, which a motif is incapable of

capturing. For example, a motif might score the following sequence very well:

CGGT TTCG TGGT

but since the data seems to suggest a palindrome is necessary, this sequence is unlikely to perform the same function as the above sequences.

A palindrome may be conserved to allow the RNA to form a stem-loop structure (hairpin).

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?

   (3 points)

   alpha helices rely more on local contacts whereas beta sheets tend to form interactions with non-local contacts. It's difficult to predict regions of beta sheets from sequential amino acids.

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

   (3 points)

   Knowledge-based means that the algorithm is based off of empirically derived data. I accepted answers that said the algorithm was based on experiments or a database of structure. The parameters are the helical "propensities".

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

   (3 points)

   Proline can't hydrogen bond and will have steric hinderance due to it's bulky ring structure. Glycine is too flexible and has a high entrophic cost to be included in a helical structure.

Coiled coils:

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

   (2 points) The repeat for a coiled-coil is once every 7 mers. I gave 1 point if they said the repeat was i, i+3 or i+ 4 for a helix.

b. We looked at a general type of coiled-coil domain called a leucine zipper. Explain the importance of the leucines in this domain.
(2 points)
I was looking for the fact that leucines for stable hydrophobic interactions between the two coils.

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.
(5 points)
Since Chou-Fasman works with helical 'propensities' – some empirical measure of how likely a particular amino acid is to be in a helix - we might consider using a similar set of 'propensities' for each amino acid to be in a coiled coil. However, this might run into issues, since we've seen that the position in the coiled coil matters substantially. Hydrophobic and hydrophilic patterns in the 7-residue repeat [heptad] should probably not be ignored. To account for this, incorporate position-specific propensities, and scan along a sequence, looking for a heptad (or longer) with high average propensity – then extend outwards till the propensity drops below some threshold.

It is true that helical propensity is necessary – however it is completely insufficient to predict coiled coils.

2 points deducted for missing the importance of position-specificity
1-5 points deducted for unclear / insufficient description of 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? (5 points)
Proteins do not fold by randomly sampling every possible conformation (this would take the age of the universe for every protein to fold!) Instead, there must be some kind of directed folding

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. (5 points)
   Assume (and many valid assumptions can be made here) that you are allowing 2 backbone conformations for each residue (say, helix and sheet), and all 20 amino acids with one rotamer conformation at each position. The expected length you're able to design would then be

$$(20 * 2)^n = 10^{40}$$

   n ~ 15 residues (a very short peptide, probably to short to fold anyway). It's hard to calculate n exactly, but you could estimate that $40^n = 10^{40}$ will mean that n has to be less than 40 (by about 2 fold)

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

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.

1) Random start state S_i, with Energy E_i
2) Make a random perturbation to a new neighbor state S_{test} with energy E_{test}
    a. If $E_{test} < E_i$
        i. $S_{i+1} = S_{test}$
        ii. $E_{i+1} = E_{test}$
    b. Else, set $S_{n+1} = S_{test}$ with Probability $P = e^{-(Etest-Ei)/kT}$,determined by a random number, otherwise $S_{i+1} = S_i$
3) Return to Step 2, repeat a set number of times

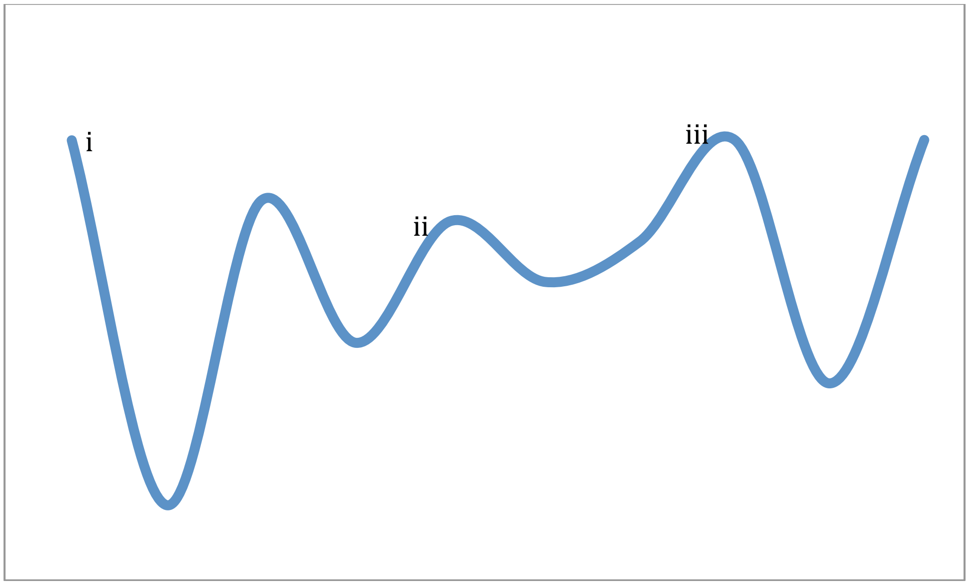
With simulated annealing: Decrease temperature as you proceed through this process. Essentially, this allows a more likely uphill walk at earlier trials (higher temperatures) to allow exploration of the landscape, with a decrease in temperature causing higher energies to be rejected, allowing the algorithm to find a minima.

Continued on next page

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) (3 points)

i) could end up virtually anywhere – can walk uphill or downhill freely, and may not ever settle into a minima

ii) Would likely end up in one of the two local minima closest to it (cannot make it uphill often)

iii) Would optimally find the global minimum, but may also end up in the near-global minimum close to it or another local mininum

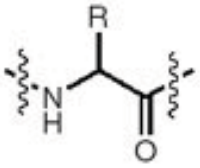
A wide variety of answers were accepted here that either showed or stated understanding of the above concepts.

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.*

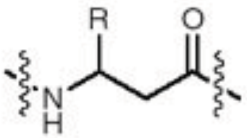
a. Name three levels of protein representation discussed in class.
(2 points)
atomic coordinates, amino acid sequence, dihedral angles, secondary structure, tertiary structure (only 3 were required)

b. Give one advantage or disadvantage of using each representation.
(2 points)
Atomic coordinates require much more detail about the protein, whereas secondary and tertiary descriptions are much more succinct, but lack the "resolution" of atomic coordinates.

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

(2 points)
The beta peptide is less likely to be cleaved by proteases and so your protein won't be degraded as easily.

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.
(2 points)
Since the residues themselves would have a different structure, it's likely that the structure of the new peptide would have an altered secondary structure and different function.

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

(2 points)

In positive design you are selecting for an advantageous property (such as binding to a particular fragment of DNA) and in negative design you are trying to remove unfavorable qualities (such as reducing non-specific binding).

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. (10 points for each part)*

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.

The Basic concept we were looking for here was ***allostery*** – this extra site away from the site of phosphorylation interact with the kinase favorably to increase binding.

Most correct specific examples of allostery or explanations showing understanding of how this site could impact phosphorylation were acceptable.

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.

Essentially, these amino acids likely are part of a helical sequence (hydrophilic residues on one side, hydrophobic on the other). Recall that helices are too large to fit into kinase active sites, and must unwind in order to do so. This would likely cause the helical part of the bottom protein to be phosphorylated less frequently as unwinding takes energy. However, short peptides are often unstructured in solution, so these two peptides would be both unstructured and therefore phosphorylated with equal efficiency.