20.320 Problem Set 2
September 18, 2009

#

General Instructions:

1. You are expected to state all your assumptions and provide step-by-step solutions to the numerical problems. Unless indicated otherwise, the computational problems may be solved using Python/MATLAB or hand-solved showing all calculations. Both the results of any calculations and the corresponding code must be printed and attached to the solutions.

2. You will need to submit the solutions to each problem to a separate mail box, so please prepare your answers appropriately. Staples the pages for each question separately and make sure your name appears on each set of pages. (The problems will get sent to different graders, which should allow us to get the graded problem set back to you more quickly.)

3. Submit your completed problem set to the marked box mounted on the wall of the fourth floor hallway between buildings 8 and 16.

4. The problem sets are due at noon on Friday September 25th. There will be no extensions of deadlines for any problem sets in 20.320. Late submissions will not be accepted.

5. Please review the information about acceptable forms of collaboration, which was provided on the first day of class and follow the guidelines carefully.

20.320 Problem Set 2
Question 1

#

In the last problem set, we discovered that Histidine 142 is conserved across many variants of influenza hemagglutinins and thus may play a key role in mediating their pH-dependent conformational change. Now, we will look at the effects the charge of His142 has on the electrostatics of both hemagglutinin conformations.

a) Write Biopython code to identify the charged residues in both native HA and HA at endosomal pH. As with PS1, consider only residues 40-153 of HA chain B.

b) Write an electrostatic function that computes the electrostatic potential for a protein structure. Your code should take two lists as input. The first list will contain the coordinates of a set of atoms and the second list will contain the charges. Make your code compatible with the Biopython package.
Use this code to compute the electrostatic potential of each of the following:

I. Native HA with uncharged His142
II. Native HA with charged His142 (all other histidines uncharged)
III. Endosomal HA with uncharged His142 (all other histidines charged)
IV. Endosomal HA with charged His142 (all other histidines charged)

For each case, sum the electrostatic potential of all pairs of charged amino acid using the distance between the charged atoms in your calculations. (Use atom 'OE2' for Glu, 'OD2' for Asp, 'NZ' for Lys, 'NH2' for Arg, and 'NE2' for His) Assume a dielectric constant value of 3.

c) Do the calculated electrostatic potentials make sense? If not, why not? How could we improve our energy calculation?

20.320 Problem Set 2
Question 2

#

Once again we will be looking at the HA_2 chain (chain 'B' in the PDB files) of influenza hemagglutinin in its native state and at endosomal pH using the same PDB files from PS1.

The conformation potentials used by Chou and Fasman[1] appear in the chart below. These were derived by examining 24 protein structures. The $P_{\alpha/\beta}$ value for each amino acid is proportional to the frequency of that amino acid in alpha helices/beta sheets and has been normalized so that they take on values between zero and two. The amino acids with $P_\alpha > 1$ are assumed to have a propensity for α-helices and similarly those with $P_\beta > 1$ are assumed to have a propensity for β-sheets. Thus Chou and Fasman classified amino acids as strong helix/sheet formers ($H_{\alpha/\beta}$), helix/sheet formers ($h_{\alpha/\beta}$), helix/sheet indifferent (I_α, $i_{\alpha/\beta}$), helix/sheet breakers ($b_{\alpha/\beta}$), and strong helix/sheet breakers ($B_{\alpha/\beta}$). These are also marked in the chart below. In order to understand the Chou-Fasman algorithm, we will use the algorithm to predict alpha-helix propensity in an input protein according to the following rules:

Criteria 1. *Helix Nucleation*. Locate clusters of four helical residues (h_α or H_α) out of six residues along the polypeptide chain. Weak helical residues (I_α) count as $0.5h_\alpha$ (i.e. three h_α and two I_α residues out of six could also nucleate a helix). Helix formation is unfavorable if the segment contains 1/3 or more helix breakers (b_α or B_α).

Criteria 2. *Helix Termination*. Extend the helical segment in *both* directions until terminated by tetrapeptides with $P_{\alpha,\ \text{average}} < 1.00$.

Criteria 3. Proline cannot occur in the alpha helix.

For this problem, you will need to download the following files and put them in one folder, including the pdb files for native HA (3EYJ.pdb) and endosomal pH HA (1HTM.pdb):

CFalphaPredict.py
ChouFasman.py

P_α			P_β		
Glu	1.51	H_α	Val	1.70	H_β
Met	1.45		Ile	1.60	
Ala	1.42		Tyr	1.47	
Leu	1.21		Phe	1.38	h_β
Lys	1.16	h_α	Trp	1.37	
Phe	1.13		Leu	1.30	
Gln	1.11		Cys	1.19	
Trp	1.08		Thr	1.19	
Ile	1.08		Gln	1.10	
Val	1.06		Met	1.05	
Asp	1.01	I_α	Arg	0.93	i_β
His	1.00		Asn	0.89	
Arg	0.98	i_α	His	0.87	
Thr	0.83		Ala	0.83	
Ser	0.77		Ser	0.75	b_β
Cys	0.70		Gly	0.75	
Tyr	0.69	b_α	Lys	0.74	
Asn	0.67		Pro	0.55	B_β
Pro	0.57	B_α	Asp	0.54	
Gly	0.57		Glu	0.37	

a) Write the `parsePDB` function in ChouFasman.py to identify alpha helical residues based on their phi-psi angles. Use the same criteria as in Problem Set 1 (phi is between -57 and -71°, and psi is between -34 and -48°). Your code from PS1 should require little, if any, modification. Use this code to load chain 'B' of the proteins (3EYJ and 1HTM) and return a list of helical residues. Attach a copy of your code and the output.

b) Complete the code `findAlpha()`, in the Python program CFalphaPredict.py to predict helical regions according to the Chou-Fasman rules above. Attach a copy of your code and output.

[1] Chou, P; Fasman, G.; "Empirical predictions of protein conformation," *Ann. Rev. Biochem.* **47** (1978) 251-276.

#

c) Modify the helix prediction criteria in your `findAlpha()` function from part (b) such that:
   - Helices nucleate with clusters of *five* helical residues (h_α or H_α) out of six residues along the polypeptide chain. As before, weak helical residues (l_α) count as $0.5h_\alpha$, (i.e. *four* h_α and two l_α residues out of six could also nucleate a helix).
   - Glycine (in addition to proline) cannot occur in the α-helix.

d) Compare the results from parts (a), (b), and (c) as well as the residues annotated as helical in the PDB file itself. [Hint: Look at the sequence details for both proteins on the PDB website for a nice graphical representation of each protein's secondary structure]. Which method yields results closer to those found experimentally? Explain the reasons behind the occasional failure of Chou-Fasman alpha-helix predictions.

e) Using the PyMOL PDB viewer (available for download at *http://delsci.com/rel/099*), look at the structures of these two proteins. Attach print outs of the structures from the viewer. Focusing on the HIS142 residue, explain the differences in the structures.

20.320 Problem Set 2
Question 3

#

In bacteria, the lactose repressor (*lacI*) is involved with regulating the transcription of genes involved in lactose metabolism. When lactose levels are low, *lacI* is bound to the *lac* operator, preventing the expression of β-galactosidase, which cleaves lactose into its galactose and glucose components. The following sequences were investigated for *lacI* binding in a 1987 paper mapping the recognition helix of *lacI* with the lac operator:

ACTTGTGAGC
ATTTGTGAGC
AAATGTGAGC
AATTGTGAGC
AACTGTGAGC
AATTGTGAGT
AATGGTGAGC
AAGTGTGAGC
AGTTGTGAGC

a) Calculate the $\log_2$(odds) matrix for these sequences. Use pseudocounts of 0.0025 for zero frequencies.

You are interested in a sequence of DNA from a newly discovered organism with an apparently functional lac regulation system. Based on your sequencing results, you predict that *lacI* binds somewhere in the following sequence.

ATCTCATATAATTGTGAGCTCTAATAGAGTTCATGAGCAATG

b) Calculate the $\log_2$(odds) score for each hypothetical binding site in your sequence of interest. Use these values to plot the $\log_2$(odds) score for each 10-base window as a function of the window starting point. For instance, the first value in your plot should be the $\log_2$(odds) score of the sequence ATCTCATATA.

c) Based on your results in Part B, determine the most likely *lacI* binding site in the given sequence. Report the $\log_2$(odds) score of your choice.

MIT OpenCourseWare
http://ocw.mit.edu

20.320 Analysis of Biomolecular and Cellular Systems
Fall 2012

For information about citing these materials or our Terms of Use, visit: http://ocw.mit.edu/terms.