20.320 Problem Set #2

Due on September 30rd, 2011 at 11:59am. No extensions will be granted.

General Instructions:

1. You are expected to state all of 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 must be printed and attached to the solutions, and the corresponding code should be submitted on Course website. For ease of grading (and in order to receive partial credit), your code must be well organized and thoroughly commented with meaningful variable names.
2. You will need to submit the solutions to each problem to a separate mail box, so please prepare your answers appropriately. Staple the pages for each question separately and make sure your name appears on each set of pages. (The problems will be sent to different graders, which should allow us to get graded problem sets 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. Python codes when relevant should be submitted on Course website.
4. The problem sets are due at noon on Friday the week after they were issued. 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 is available on the Course website and follow the guidelines carefully. Especially review the guidelines for collaboration on code. NO sharing of code is permitted.

Note: To unzip the files necessary for this homework from the command line, go to the directory containing ps2_files.zip and type

```
unzip ps2_files.zip
```

Problem 1: Alanine scanning -

A number of crystallographic studies have shown that the binding interfaces between proteins are generally large and include many intermolecular contacts. However, structural analysis alone cannot show whether all of these contacts are important for tight binding. For determining which residues are important for binding between two proteins, alanine scanning is often performed. In this technique, each amino acid is substituted individually by an alanine. The impact in the interaction can then be measured and the importance of the residue can be assessed. In this problem, you will determine the importance of certain residues of interleukin-4 (IL-4) for binding to its receptor IL4-BP. IL-4 binds to its receptor with a K_D of 150pM, corresponding to a binding free energy, ΔG of -13.38 kcal/mol (at 25C). Your alanine scanning experiment gives the result listed in the table below (data from http://nic.ucsf.edu/asedb/).

Mutation	ΔΔG (kcal/mol)
I5A	1.2
T6A	-0.1
Q8A	0
E9A	3.1
I11A	0.1
T13A	1
N15A	0
S16A	-0.17
E19A	-0.3
K77A	0.16
Q78A	0.13
R81A	0.48
F82A	-0.08
K84A	0.35
R85A	0.42
R88A	3.7
N89A	1.6
W91A	0.7

a) Given the ΔΔG for these mutations, which residues are the most important for binding? How much do they change the K_D (fold-wise)?

b) Why might you not be surprised that residues that seem to affect binding are spaced 4 apart or 1 apart? Pull up a crystal structure of IL-4 (try pdb 2INT) in PyMOL and highlight those residues. Include a picture in your assignment.

c) Suppose a different mutation on a residue that is buried inside the protein, F55A, yields a K_D of 120nM. Why is there such a dramatic effect on the K_D when this residue is known not to interact with the receptor?

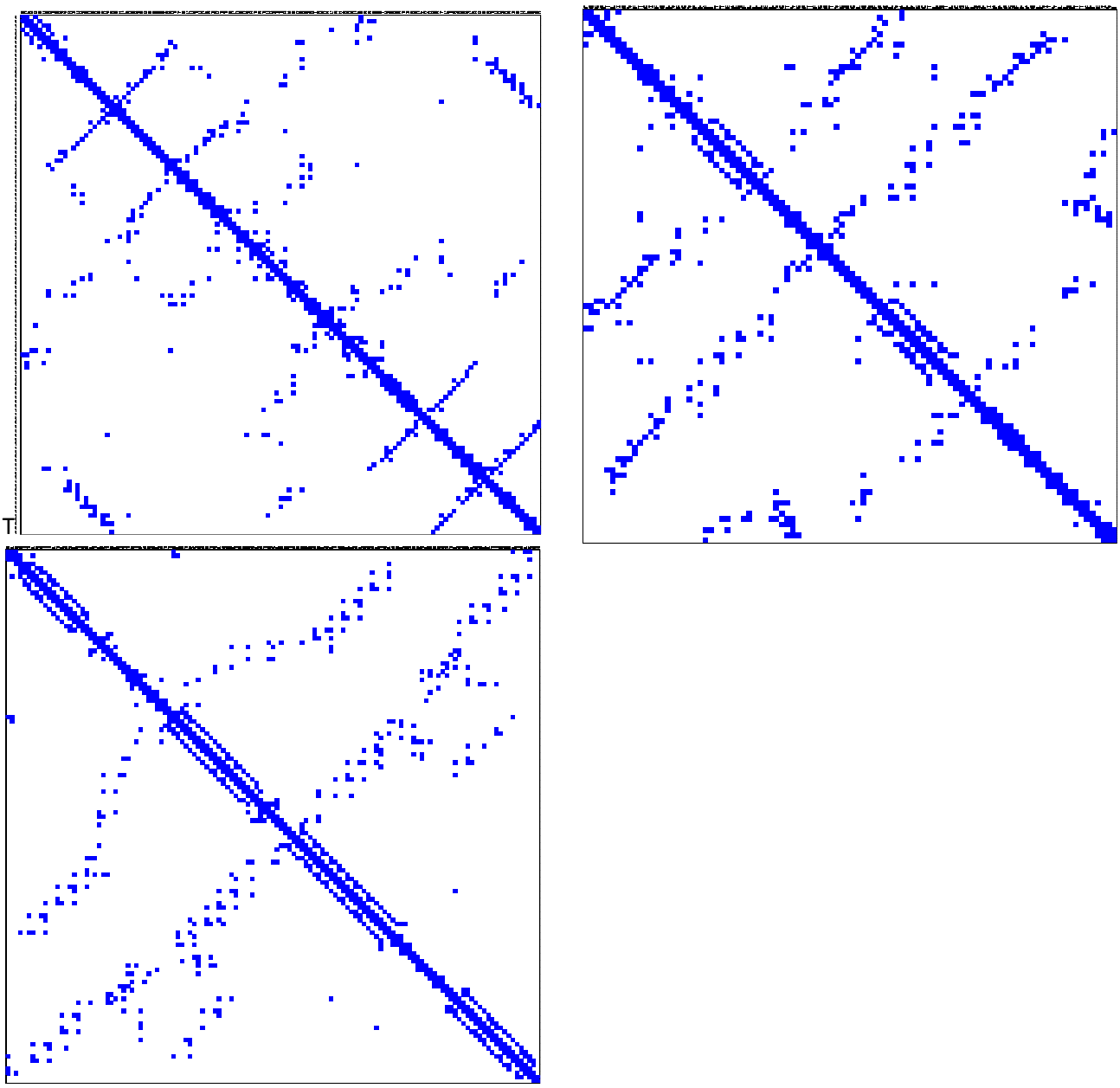
d) What does this tell you about the limitation of alanine scanning?

Problem 2: Contact maps

a) Given what you know about the structure of alpha helices, what might you expect the contact map to look like for a single alpha helix? Provide a sketch, highlighting the key features.

b) What might you expect it to look like for two parallel strands of a beta sheet? Again, provide a sketch.

c) You made a contact map for IL-4 but it got messed up in your stack of papers with two other contact maps. Which contact map is IL-4? How do you know?



d) On the IL-4 contact map, circle and identify two secondary structure elements by their local contacts (on-diagonal). Circle and identify the (non-local) contacts between the two strands of the beta sheet. Are the strands parallel or antiparallel?

e) Are the majority of the non-local contacts in this protein between parallel or antiparallel secondary structures?

Problem 3: Destabilizing mutations

In this problem you are going to look at alanine scanning in the context of a dimer, specifically one known as a leucine zipper. Leucine zippers are DNA-binding domains consisting of a coiled coil of alpha helices. For this problem specifically, you are going to look at the leucine zipper of GCN4, a yeast transcription factor that controls the amino acid starvation response. Take a look at the structure of pdb 2ZTA using PyMOL.

a) Is the protein a homo or hetero-dimer? Is it parallel or antiparallel?

b) Identify the leucine residues. Color the rest of the protein blue and the leucines red. Do you notice any patterns in the location (both 3D and sequentially)? Include a picture, in which you show the cartoon representation of the entire protein, and only show sticks for the leucine residues.

c) To computationally probe the importance of these residues, PyRosetta was used to make each possible single-residue mutation to alanine along chain A. These have been given to you as PDBs in the Amuts folder. Write a python script to calculate the change in score induced by each single-residue mutation. Use the standard scoring function. Plot the change in score (relative to the original protein) versus residue number (mut0 is a change in residue 1, etc).

d) To view the contributions from each term in the energy function, revise your script to output scores from several new scoring functions you will create. Each score will have weight 1 for only one energy term, and thus will output the (unweighted) contribution of that energy term. Output the energies for `fa_atr`, `fa_rep`, `fa_sol` for each mutation and plot them together.

Problem 4: Designing and Predicting Helical Peptides

In this problem, you will attempt to rationally redesign a peptide to adopt a desired structure based on your knowledge of the principles of secondary structure formation and aided by a computer algorithm. You have been provided with a Python implementation of the Chou-Fasman algorithm for detecting alpha helices. You should open and read the script to see what the code does. You will be editing the script minimally to choose which sequences you want to analyze with the Chou-Fasman algorithm, and then running the script from command line “python choufasman.py”.

a) Run the code with the peptide $(Ala_5Gly_4Ser)_2$. What do you predict its secondary structure to be? In solution, do you think that a large or a small fraction of the peptide will actually adopt the predicted structure? Justify your answer.

b) Find a variant that differs from the starting peptide by at most 3 amino acids and is not predicted to be helical.

c) Find 3 variants that each differ from the original peptide by at most 5 amino acids and are strongly predicted to be helical through their length. Make sure the 3 variants are all different and non-trivial.

d) Based on your knowledge of the C-F algorithm, explain why each of your variants in the above questions produced the changes in expected helicity.

e) Try out the Chou-Fasman predictions on the following sequence from IL-4:

HKCDITLQEIIKTLNSLTEQKTLCTELTVTDIFAASKNTTEKETFCRAATVLRQFYSHHEKDTRCLGATAQQFHRHKQLIRFL
KRLDRNLWGLAGLNSCPVKEANQSTLENFLERLKTIMREKYSKCSS

Where are the predicted helices?

f) Compare the results to the secondary structure from the PDB (2INT) (hint: the assigned secondary structure is actually explicitly noted in the pdb file - open the PDB file in a text editor and look through it to find this). What is the percentage of correctly classified residues (treat sheets in the pdb as simply non-helical)? Does the percentage of correctly classified residues fairly represent the accuracy of the algorithm? Justify your answer.

Problem 5: DNA Sequence Motif Prediction

Suppose you had performed a protein-binding microarray experiment to determine the specificity of a particular transcription factor (TF), and you found your transcription factor bound to the following 6-mers:

```
AATGGT
TATACT
AACGTT
CATGAT
AATGCA
```

a) In the position weight matrix (PWM), our model is

P(seq | TF binds) = P(position 1 | TF binds)*P(position 2 | TF binds)*P(position 3 | TF binds)...

Calculate the matrix of probabilities for each nucleotide by position.

b) Under this model, what are the probabilities of observing the following sequences given that the transcription factor binds them:

i) AATGCC

ii) AATACT

c) An interesting question one might ask about the probability of (ii) is how it compares to the probability of observing that sequence, given that the sequence is NOT bound by a transcription factor. Suppose each base is equally likely in the background DNA. What is the probability of observing (i) and (ii) above?

d) We'd often like to compare the probability of the data under two different models, so we take the ratio of the two. This is called a likelihood ratio, and is P(data|model 1) / P(data | model 2). In our case this is P(sequence | TF binds) / P(sequence | background DNA). Calculate the likelihood ratio for (i) and (ii).

e) Show algebraically that if each entry at *i*, *j* in the PWM is

log[P(nucleotide *i* at pos *j* | TF binds) / P(nucleotide *i* | background)]

then the log likelihood ratio for a sequence is simply the sum of the corresponding entries in the PWM.

f) PWMs make a particularly important assumption. Suppose you performed another experiment with an RNA-binding protein and had observed the following set of bound sequences:

```
AGACUGCCCAGUCU
AGAGUGCUCACUCU
UCCCUGAUUAGGGA
AGCCUUAACAGGCU
UGACUGCCCAGUCA
AAACUUAUAAGUUU
UCCCUUACUAGGGA
AGACUUUUAAGUCU
```

What is special about these sequences? Suppose you had observed a sequence UGCCUGCCCACUUU. Would you expect your protein to bind this sequence? (Hint, you do not need to calculate the PWM). Why or why not? Would a PWM be appropriate to estimate the likelihood ratio for this sequence being a binding site?

20.320 Analysis of Biomolecular and Cellular Systems
Fall 2012