20.320 Problem Set #3

Due on October 7th, 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 and the corresponding code must be printed and attached to the solutions. 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 stellar site and follow the guidelines carefully. Especially review the guidelines for collaboration on code. NO sharing of code is permitted.

Problem 1 – bZIP specificity
(20 points)

Amy Keating gave a guest presentation about optimizing binding for bZIP using protein binding arrays. Their group used protein binding assays to identify fragments that would optimally bind bZIP in a selective manner.

A. What is bZIP? Why was it important to design peptides that would bind this molecule and what biological goal were they approaching by optimizing for this binding event? (4 points)

B. What are the four types of natural specificity that Professor Keating mentioned in lecture? These were the four ways that the cell could control which protein binding interactions occur. (4 points)

C. Generally explain their peptide array approach – how does it find optimal binding partners? What do you measure in the experiment? What results came out of their approach? (4 points)

D. What was their rationale for using a computational approach to also predict protein specificity? What factors was the computational approach able to reveal about bZIP specificity? (4 points)

E. Explain the tradeoff between specificity and stability. Conceptually, how did their CLASSY algorithm deal with this trade-off? (4 points)

Problem 2 – Thermodynamic Cycles and Alanine Scanning (16 points)

In class we talked about re-designing GCSF as a potential therapeutic. Optimizing specificity of interaction is one part of that task, but before we can start designing, it's important to characterize how the natural molecule binds and how residues in the protein's amino acid sequence contribute to the molecule's natural binding. As such, we'll use alanine scanning and thermodynamic cycles to look at the interaction.

TABLE II
Binding of G-CSF mutants to Ba/F3 cells expressing WT-GR or (R288A)GR

G-CSF mutant	Receptor			
	WT-GR		(R288A)GR	
	K_d nM[a]	Mut/WT[b]	K_d nM[a]	Mut/WT[b]
WT	0.045 ± 0.008	1.0	0.37 ± 0.03[c]	1.0
E19A	0.050 ± 0.004	1.1	0.29 ± 0.03	0.78
K23A	0.077 ± 0.015	1.7	0.95 ± 0.11	2.5
E46A	0.076 ± 0.003	1.7	3.32 ± 0.86[c]	8.9
D112A	0.060 ± 0.003	1.3	4.06 ± 0.85	10.9

[a] Data are mean ± range of two assays, including data shown in Fig. 4.
[b] Ratio of K_d for mutant G-CSF/WT G-CSF.
[c] Data are mean ± S.D. of three assays, including data shown in Fig. 4.

A. Draw out the four thermodynamic cycles for different GCSF mutants binding to the wild-type receptor. Be sure to label the ligand and receptors along with each ΔG correctly. (4 points)

B. Compute the ΔΔG between all mutant pairs. Just calculate the free energy of mutation in the background of the wild-type receptor. (6 total ΔΔG's) at normal body conditions (37° C and 1 atm pressure). (4 points)

C. Given these ΔΔG's, which mutations destabilize binding to the wild-type receptor? Consider what these mutated residues may have been contributing to the protein before being switched to an alanine (4 points)

D. Suppose we want to look at the WT and E46A GCSF variants with WT-GR and R288A-GR. Draw out the double mutant cycle. Be sure to label the ligand and receptors along with the ΔG's and ΔΔG's correctly. (note: you can draw it as a cube, or simplify it, but it must contain all of the components). (4 points)

Problem 3 – Rotamer Packing
(55 points)

In the previous problem we looked at using thermodynamic cycles to analyze how changing multiple residues to alanine affected binding of the growth factor to the receptor. Now we will use an energy minimization algorithm to perform a rotamer search to repack the side-chains of the GCSF/GCSF complex into a new, relatively low-energy state after mutating. This time we will be mutating the aspartic acid at residue 110 to a histidine.

A. In order to make the calculations manageable we will only mutate single amino acids – such as mutating Asp110 -> His110. Briefly discuss the implications of mutating a single residue on:
    a. Overall protein structure
    b. Backbone conformation
    c. Protein packing/folding
    d. Protein Binding (take a look at 1CD9_AB.pdb for this one). (5 points)

B. Download pdb files 1CD9_A.pdb and 1CD9_AB.pdb from Course website. These files contain the structure of the unbound GCSF protein and the structure of GCSF and the extracellular portion of the GCSF receptor, respectively.

   Look at both structures in PyMol. Attach pictures of the structures with the aspartic acid high-lighted in a different color than the rest of the protein. Given the discussion in class what can you say about how Asp contributes to binding affinity vs binding specificity in this interaction? Why? *** Note: if you are unable to print in color or choose to submit photos online, please indicate on your problem set that you have submitted online (this helps our graders) (7 points)

C. The interaction energies of these particular side chains depend on their orientation. Different side-chain "packing" leads to the development of different rotamers – each that have different energies of folding. PyRosetta can help us look at how different iterations of folding/packing residues on the protein can change the energy. The program optimizes new folding through a Monte-Carlo algorithm (the details of the algorithm aren't important, just know that it will help you optimize rotamer packing). For this problem you are going to repack residue 110 and look at how the energy changes.

   Intro from PyRosetta's tutorials: (Just for reference)
   Rosetta has a side-chain repacking routine pre-packaged as a "mover", which carries out a computational search each time it is applied. The specific scope of the packing is specified in a PackerTask object, which we can specify via commands or from an input file.

Useful PyRosetta commands: (These you will need to know)
Create a PackerTask as follows. This will set the task to allow packing only of residue 49:

```
task_pack = standard_packer_task(pose)
task_pack.restrict_to_repacking()
task_pack.temporarily_fix_everything()
task_pack.temporarily_set_pack_residue(49,True)
```

Confirm your settings using:

```
print task_pack
```

We now can create a PackRotamersMover:

```
packmover = PackRotamersMover(scorefxn,task_pack)
```

Apply the packmover to your pose with:

```
packmover.apply(pose)
```

For this problem you are going to repack residue 110 and look at how the energy changes. Familiarize yourself with the new PyRosetta commands and then write a python script to use PyRosetta to repack residue 110 of the 1CD9_A.pdb file. Because repacking is a stochastic process, write your script such that it will repack the residue 10 times and take the average of all ten scores. What is the score before and after packing? Has it changed significantly? How can you explain the change/no-change in the two scores? (15 points)

D. Mutagenesis: We can now follow a similar analysis after mutating residue 110 from Asp to His. Again, PyRosetta can help us do this, this time using their Design capabilities.

Design operations are easiest to specify through a data file called a "resfile." You can create a resfile for a given pdb file or pose using:

```
generate_resfile_from_pdb("1CD9_A.clean.pdb","1CD9_A.resfile")
```
OR
```
generate_resfile_from_pose(pose,"1CD9_A.resfile")
```

Inside the resfile you will see a list of all residues and NATRO next to it, indicating that it is set to use the native rotamer. NATRO can be changed to the following:

NATRO	use native amino acid and native rotamer (does not repack)
NATAA	use native amino acid, but allow repacking to other rotamers
PIKAA ILV	use only the following amino acids and allow repacking between them
ALLAA	use all amino acids and all repacking

Edit the resfile to allow force residue 110 to be Histidine ("110 A PIKAA H") and save the file as "1CD9_A-D110H.resfile". Create a new task for design from the resfile:

```
task_design = TaskFactory.create_packer_task(pose)
```
***** note that this method has changed names recently and may be mis-documented on the PyRosetta site!*
```
task_design.read_resfile("1CD9_A-D110H.resfile ")
```

Create a new PackResiduesMover

```
packmover2 = PackRotamersMover(scorefxn, task_design)
```

with the design task and use it to mutate residue 110 to histidine. **What is the new score? (Again, write a script to repack 10 times and find the average score). Is the mutation more or less stable? Discuss why histidine may be more or less stable for the protein.** (15 points)

E. Hypothesize a side chain substitution that would be more favorable for the protein. State which residue you are selecting, and why you think it might be more favorable. (3 points)

F. Now change that residue using the same steps from part D. and report the new energy of the protein. Did the energy increase or decrease? Is this what you expected? Discuss why your residue selection may have increased or decreased the energy of the protein. (10 points)

Problem 4 – Multiple Sequence Alignment
(9 Points)

Receptor tyrosine kinases of the Epidermal Growth Factor (EGFR) family are essential to numerous physiological and pathological processes. In humans, 12 EGFR family ligands have been identified and a significantly conserved section of the multiple sequence alignment (MSA) of some members of this family is shown below. We have also included the extracellular matrix protein Tenascin-C which contains EGF-like domains known to activate EGF receptors. Some gaps have been omitted to simplify the problem. In the MSA, the amino acids are represented by their one-letter amino acid code. Capital letters indicate a significant alignment while lowercase letters indicate no significant alignment.

```
AREG_HUMAN/142-182      KKNPCNaefqNFCIH-GECKYIEH---LEAVTCKCQQEYFGERCG
BTC_HUMAN/65-105        HFSRCPkqykHYCIK-GRCRFVVA---EQTPSCVCDEGYIGARCE
EGF_HUMAN/972-1013      SDSECPlshdGYCLHDGVCMYIEA---LDKYACNCVVGYIGERCQ
EREG_HUMAN/64-104       SITKCSsdmnGYCLH-GQCIYLVD---MSQNYCRCEVGYTGVRCE
HBEGF_HUMAN/104-144     KRDPCLrkykDFCIH-GECKYVKE---LRAPSCICHPGYHGERCH
NRG1_HUMAN/178-222      HLVKCAekekTFCVNGGECFMVKDlsnPSRYLCKCQPGFTGARCT
NRG2_HUMAN/341-382      HARKCNetakSYCVNGGVCYYIEG---INQLSCKCPNGFFGQRCL
NRG3_HUMAN/286-329      HFKPCRdkdlAYCLNDGECFVIETl-tGSHKHCRCKEGYQGVRCD
NRG4_HUMAN/5-46         HEEPCGpshkSFCLNGGLCYVIPT---IPSPFCRCVENYTGARCE
TGFA_HUMAN/43-83        HFNDCPdshtQFCFH-GTCRFLVQ---EDKPACVCHSGYVGARCE
TENA_HUMAN/559-590      KEQRCP----SDCHGQGRCVDG---------QCICHEGFTGLDCG
```

A. Complete the PROSITE consensus pattern for the above alignment by manual pattern recognition of the MSA. If you are not familiar with PROSITE notation: http://en.wikipedia.org/wiki/Sequence_motif. (1 point)
x(4)- -x(3,7)- -x(4,5)-C-x(4,13)-C-x(1)-C-x(2)-[ ]-[F,Y]-x(4)- -x(1)

B. What amino acids were absolutely preserved throughout the evolution of this family? Give a rationale why each was preserved. (2 points)

C. What amino acids were somewhat preserved? Or which amino acids could be mutated to another amino acid with similar properties? (2 points)

D. Compute the log-likelihood matrix for the first five positions of this alignment (use base 2 this time). Assume that all amino acids are equally probable in the background and add a pseudocount of 0.1% (4 points)

20.320 Analysis of Biomolecular and Cellular Systems
Fall 2012