20.320 — Problem Set # 7

November 5^{th}, 2010

Due on November 12^{th}, 2010 at 11:59am. No extensions will be granted.

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. 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. Staples 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 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 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 was provided on the first day of class and follow the guidelines carefully.

67 points for problem set 7.

1 Designing 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.

a) Start 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.

Solution:

Ala_5Gly and $GlySerAla_5Gly$ (positions 1-6 and 9-16) are predicted to be alpha-helical, the flexible linker regions in between will be unstructured.
In solution, a small fraction of this peptide will be in exactly this conformation. As can be seen from printing the helical propensities of all positions, the helices are just barely predicted to be helical (average propensity near 1 for the helical stretches). In other words, the enthalpy of folding is small and entropy will distribute the molecule's residence time over many conformations. Many such nearby conformations exist with the flanking glycines unstructured (the flanks of the helical regions have the lowest individual propensities).
4 points total: 2 for correct prediction of helicity, 2 for cogent justification based on propensities and entropy that most of the time, the peptide will reside in similar but different conformations.

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

Solution:

E.g. $Ala_2ProAla_2Gly_4SerAlaProAlaProAlaGly_4Ser$.
1 point.

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; the graders will use common sense to judge whether a variant is a genuinely distinct peptide or a minor variation on annother of your solutions.*

Solution:

E.g.

- $Ala_5Gly_2AlaGlySerAla_5GlyAla_4$
- $Ala_5Gly_2AlaGlySerAla_5Trp_4Ser$
- $Ala_5Gly_2GluGlySerAla_5GlyIleMetAlaSer$

3 points.

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.

Solution:

Points to note include

- Proline must not occur.
- Glycine can occur, but no more than twice in a row.
- The algorithm is not perfect — sterically very impropable sequences such as the Trp_4 in the second example above are predicted to be alpha-helical if the numerical criteria are fulfilled.
- The nucleation-propagation model will be very sensitive to single mutations in sequences of borderline helicity.

6 points, 2 for each sequence. It is acceptable to discuss all sequences in the same paragraph. They should have been chosen to illustrate as many aspects of the algorithm as possible.

14 points for problem 1.

2 Multiple sequence alignment

Receptor tyrosine kinases of the Epidermal Growth Factor Receptor (EGFR) family are essential to numerous phyisological and pathological processes. In human, 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 containts EGF-like domains known to activate EGF receptors. Some gaps have been ommitted to make 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 alignement was found.

MSA 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) Deduce 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.

Solution:

`x(4)-C-x(3,7)-C-x(4,5)-C-x(4,13)-C-x(1)-C-x(2)-G-[E,G,N]-[F,Y]-x(4)-G-x(2)-[RD]-C-x(1)`

2 points

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

Solution:

- Cysteines (6): Crucial for protein folding, the 3 intramolecular disulfide bond is the base of EGF-like ligands.
- Glycine: Compact, small volume, good packing, high flexibility

6 points

c) What amino acids were somewhat preservered? (can be mutated to another amino acid with similar properties)

Solution:

- Tyrosine/Phenylalanine: With high contact surface generally abundant in protein-protein interfaces
- Arginine: Positively charged side chain could be important for hydrogen bonding and hence for protein function.

6 points

d) Compute the log-odds matrix for the first five positions of this alignment. Assume that all amino acids are equally probable in the background and add a pseudocount of 0.1%.

Solution:

to compute the log-odd matrix with pseudo count given a particualr probability of finding residue a in position i p(a,i) proceed as follows:

1. Calculate the frequencies p(a,i) for each residue at each position.
2. Add the pseudocount 0.001 (0.1%) to all 0 probabilities.
3. Normalize each p(a,i) by the sum of the probabilites for that given i position.
4. Find the odds by dividing by the background probability $p_b = 0.05$
5. Take the log of base 2 (any base is ok).

Amino acid	Position 1	Position 2	Position 3	Position 4	Position 5
A	-5.6682	0.8451	-5.6453	-5.7748	-5.6710
A	-5.6682	-5.6465	-5.6453	-5.7747	4.2948
C	-5.6681	0.8454	0.8466	1.7173	-5.6444
D	-5.6681	1.8456	0.8467	0.7380	-5.6444
E	-5.6681	2.4310	-5.6450	-5.7450	-5.6444
F	-5.6681	-5.6453	-5.6450	-5.7449	-5.6444
G	3.4353	-5.6453	-5.6450	-5.7448	-5.6444
H	-5.6550	0.8466	-5.6450	-5.7447	-5.6444
I	2.4218	0.8467	0.8468	2.3322	-5.6444
K	-5.6520	0.8468	-5.6449	-5.7199	-5.6444
L	-5.6520	-5.6449	-5.6449	-5.7199	-5.6444
M	-5.6520	-5.6449	1.8469	-5.7198	-5.6444
N	-5.6520	-5.6449	-5.6447	2.7721	-5.6444
P	-5.6520	-5.6449	0.8471	-5.6942	-5.6444
Q	-5.6520	0.8469	0.8472	1.7977	-5.6444
R	1.8399	-5.6448	1.8473	-5.6856	-5.6444
S	-5.6505	-5.6448	0.8474	-5.6855	-5.6444
T	-5.6505	-5.6448	0.8475	-5.6855	-5.6444
V	-5.6505	-5.6448	-5.6444	-5.6855	-5.6444
W	-5.6505	-5.6448	-5.6444	-5.6854	-5.6444

5 points

e) From a structural perspective, why would a particular amino acid be conserved while another would not at a specific position in a protein?

Solution:

The following positions in a protein sequence are preferentially preserved:

- secondary strutures (alpha helix, beta sheets)
- Amino acids that have important interaction with ions that mediate protein folding
- Amino acids that are responsible for protein function

4 points: need two different explanations for full credit

f) Why are proline and glycine likely to disrupt helical secondary structures?

Solution:

Similar sterical hindrance when rotating the phi angle, therefore there is a high entropic cost in constricting this amino acids in a alpha-helica structure. *4 points*

27 points for problem 2.

3 Bone morphogenic protein-2

Bone morphogenic protein-2 (BMP-2) is a 116 amino acid protein morphogen and part of a larger family of BMPs. Bone morphogenetic proteins regulate many developmental processes during embryogenesis as well as tissue homeostasis in the adult. Signaling of bone morphogenetic proteins is accomplished by binding to two types of serine/threonine kinase transmembrane receptors termed type I and type II. Because a large number of ligands signal through a limited number of receptors, ligand-receptor interaction in the BMP superfamily is highly promiscuous, with a ligand binding to various receptors and a receptor binding many different BMP ligands. You can obtain structural data for BMP2 from the Protein Databank (PDB) using accession 2QJB. The co-ordinates of the α-carbons of the amino acids can be extracted from the PDB file using PyRosetta or any other method of your choice. For the remainder of this problem, consider only the 116 amino acids of chain A of the BMP2 dimer bound to its receptor. Be mindful that there is not fully characterized structural information for all residues in this protein chain. Indeed only residues 12-114 have been resolved. PyRosetta will read those residues with index 1-103.

a) Determine the 103*103 'distance matrix' for BMP2 chain A by computing the distances between the alpha carbons of all amino acid pairs. Represent this distance matrix as a 'binary contact matrix' (BCM). The contents of the BCM must be either 0 (for no contact) or 1 (for contact). For computing the BCM, you may assume that any two amino acids with C_α atoms that are less than 6 angstroms apart make contact with each other. The use of MATLAB or Python to derive the BCM is permitted.

Solution:

See Python code.
6 points

b) Based on the BCM, determine the sum of sequence separations ΔS_{ij} in residues (between contacting residues i and j), for all the residues in BMP2.

$$S_{TOT,\,r} = \sum_{j=1}^{n} \Delta S_{rj}$$

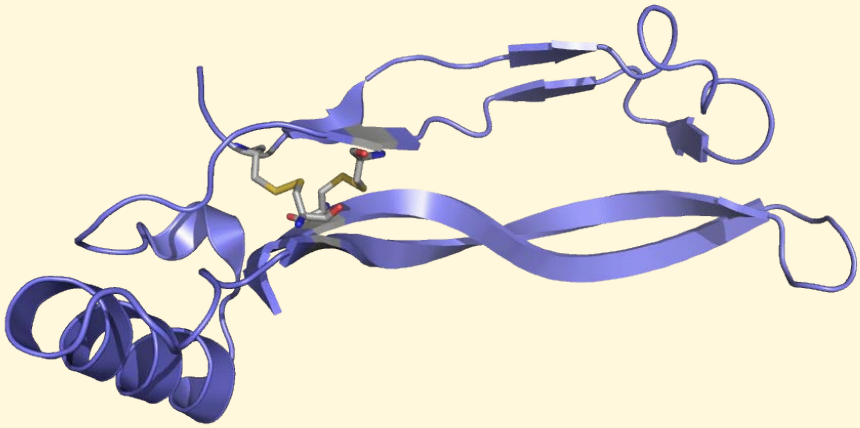
Solution:

See `-!4,!"` code.
6 points

c) For which amino acid is S_{TOT} the greatest (only one amino acid)? What types of interactions does this amino acid make with its contacts? *Hint: find which residue it is most likely to contact with given the BMP results. The PDB file can help you.* How will these interactions influence the stability and the kinetics of folding?

Solution:

- The index returned by the `-!4,!"` code indicates residue 100, which corresponds to Cystein 111 in BMP-2.
- There is 9 predicted residues with which this residue interacts. However, we know that typically cysteins form disulfide bonds. If we look at the PDB file, we see that Cys111 is involved in a disulfide bond with Cys43. We can use our `-!4,!"` code to see which residue it is predicted to interact with and we obtain (4, 5, 32, 69, 70, 71, 99, 100, 101). Position 32 corresponds indeed to Cys43.
- Cys14 (residue 3) — Cys79 (residue 68) is very close and also an acceptable answer.
- Disulfide bonds are very strong and will hold the protein in a stable structure.
- The disulfide-binded pairs Cys14–Cys79 and Cys43–Cys111 are highlighted in the PDB structure below:



6 points

d) Calculate the total number of amino acid contacts (N) within BMP (*i.e.* the number of pairwise interaction between residues of the protein) and use this to estimate the contact order (CO) for BMP-2 using:

$$\mathrm{CO} = \sum_{i=1}^{\mathrm{N}} \sum_{j=i+1}^{\mathrm{N}} \frac{\Delta S_{\mathrm{ij}}}{\mathrm{L} \cdot \mathrm{N}}$$

Solution:

CO = 0.093

4 points

e) Assuming that the logarithmic rate constant of protein folding $\ln(k_{eff})$ is proportional to the contact order (CO) as shown by Plaxco et al. J Mol Biol. 1998, estimate the rate constant of folding (k_{eff}) using a constant of proportionality of 300, i.e. using the expression:

$$\ln(k_{eff}) = -300 \cdot CO$$

Solution:

$k_{eff} = 6.7\ 10^{-13} s^{-1}$

2 points

f) Do you suppose that the relationship between folding rate and contact order provided here holds for all proteins - why or why not? Does this relationship hold for BMP2 - why or why not? If not, will the actual rate constant of folding for BMP2 be lesser/greater than the calculated rate constant.

Solution:

No, the relationship above between binding kinetics and contact order does not hold for all proteins because the biochemical interactions dictating the formation of the folded structure varies from protein to protein.
No, this relationship is not likely to hold true for BMP2 because of the presence of the rate-limiting disulfide bonds that slow the folding kinetics.
2 points

26 points for problem 3.

Codes for Problem 3

PS7Q3.py:

```
from rosetta import *
rosetta.init()
from matplotlib import pylab
import math

out = open('PS7Q3data.txt', 'w')

p = Pose("2QJB.pdb")

matrix = []

for i in range(1,p.total_residue()+1):
     matrix.append([])
     CA_index_i = p.residue(i).xyz("CA")
     out.write('\n')
     for j in range(1,p.total_residue()+1):
        CA_index_j = p.residue(j).xyz("CA")
        Vector_distance = CA_index_i - CA_index_j
        Norm = Vector_distance.norm
        matrix[i-1].append(Norm)
        out.write('%4.2f\t'%Norm)
```

PS7Q3.m

```
function matlabcodePS7Q3()
clc;
clear all;

load PS7Q3data.txt

A = PS7Q3data;
B = zeros(size(A));
Sum_DeltaS = 0;
max = 0;
posx = 0;
posy = 0;

for i = 1:length(A)
    Sum_DeltaS = 0;
    for j = 1:length(A)
        if A(i,j) < 6
            B(i,j) = 1;
            Sum_DeltaS = Sum_DeltaS + abs(j-i);
        end
    end
    if Sum_DeltaS > max
        max = Sum_DeltaS;
        aa_index = i;
    end
end

sprintf('Max Sum of DeltaS %4.0f', max)
sprintf('Interaction < 6.0 A with max Sum_Delta S is for residue %4.0f', aa_index)

N_contacts = 0;
Index = zeros(1,9);
for i = 1:length(A)
    if B(aa_index, i) == 1
        N_contacts = N_contacts + 1;
        Index(N_contacts) = i;
    end
end
sprintf('Cystein 111 has %4.0f predicted interactions', N_contacts)
Index

% Calculating the Contact Order
Sum_Sum_DeltaS = 0;
for i = 1:length(B)
    for j = i:length(B)
        if  B(i,j) == 1;
            Sum_Sum_DeltaS = Sum_Sum_DeltaS + abs(j-i);
        end
    end
end
L = length(B);
N = 1/2*sum(sum(B))

CO = Sum_Sum_DeltaS / (N*L)
sprintf('Contact order is %4.0f', CO)

```

```
57  keff = exp(-300*CO)
```

20.320 Analysis of Biomolecular and Cellular Systems
Fall 2012