20.320 Problem Set 1
September 10, 2009

#

This problem set consists of three problems designed to reinforce your knowledge of protein structure and energetics and to develop your skills at computationally analyzing protein sequences and structures:

Questions one and two relate to the structure of influenza hemagglutinin, which is the protein that allows the flu virus to enter human cells. (The swine flu is called H1N1 because it carries type 1 Hemagglutinin and type 1 Neuraminidase.) Question three examines the protein that causes the deadly genetic disease cystic fibrosis.

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 18th. 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 1
Question 1

#

Hemagglutinins are a general class of factors that increase the affinity of red blood cells for each other, causing clumps to form (the clumping of red blood cells is referred to as hemagglutination). Although some hemagglutinins are expressed under normal conditions (for instance, blood group antigens and the Rh factor), many pathogens express hemagglutinins and hemagglutinin-like proteins to help them adhere to and invade host cells more effectively. For example, the influenza viruses express hemagglutinin glycoproteins on their surfaces that play a key role in the initial binding between virus and host cell.

Influenza hemagglutinin is particularly interesting because it exploits several features of the cell's endocytic pathway to protect the virus from degradation and to facilitate its release into the cytoplasm. Once the virus attaches to the exterior of the cell it is internalized in a membrane-bound compartment called an endosome, which fuses with a lysosome to begin digesting what the cell internalized. Key to this digestive process is the acidification of the endosome, since the enzymes involved in digestion are only active when the pH is substantially lower than in the cytoplasm. The influenza virus is able to exploit this acidification process using hemagglutinin. The hemagglutinins on the viral surface undergo a pH-dependent conformational change, exposing a hydrophobic pocket that can insert into the membrane of the endosome and fuse the endosomal and viral membranes together. This allows the virus to escape degradation and transit into the cytoplasm.

Structural data for both native HA (http://www.rcsb.org/pdb/cgi/explore.cgi?pdbId=3EYJ) and HA at endosomal pH(http://www.rcsb.org/pdb/cgi/explore.cgi?pdbId=1HTM) can be obtained from the Protein Data Bank (PDB).

a) Write a Python program to parse the PDB files and extract the phi and psi angles for the HA_2 chain (chain 'B' in the PDB files) of Hemagglutinin in its native state and at endosomal pH. Use this to create a Ramachandran plot for both structures. (Note: Since chain 'B' in PDB file 1HTM only contains residues 40-153 of chain 'B' in PDB file 3EYJ, only consider those residues.) For this problem, use the Biopython package. Biopython is set of tools for biological computation written in Python and is free to download here: http://biopython.org/wiki/Download Source code for the PDB package can be found here: http://www.biopython.org/DIST/docs/api/Bio.PDB-module.html

   Use the following code segment as a model for parsing a PDB file:

```
for model in Bio.PDB.PDBParser().get_structure("HA_Native", "3EYJ.pdb") :
        polypeptides = Bio.PDB.PPBuilder().build_peptides(model["B"])
        for poly_index, poly in enumerate(polypeptides) :
```

   The following command is used to print the phi and psi angles of a polypeptide:

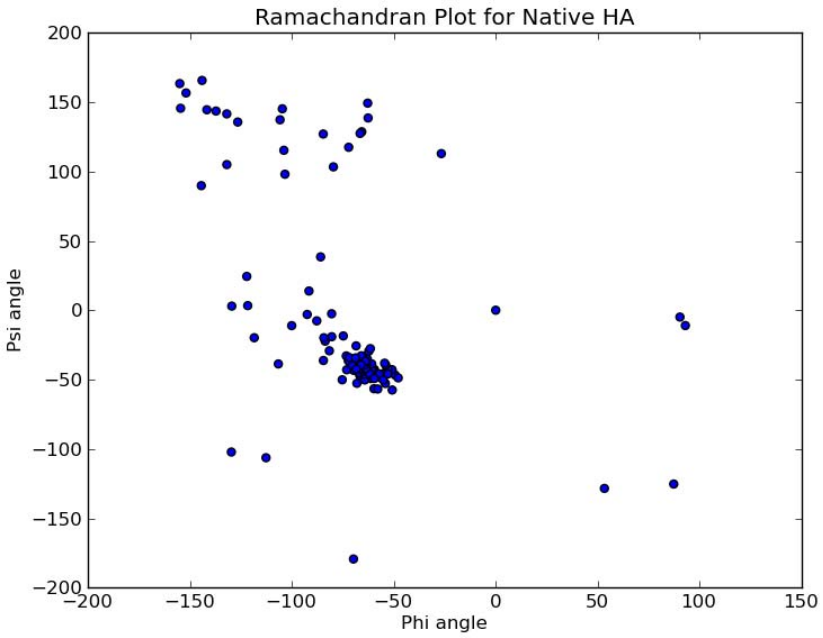
```
poly.get_phi_psi_list()
```

20.320 Problem Set 1
Question 1

```
#============================================================================#
# Problem Set 1 Question 1
# August 31, 2009


import Bio.PDB
import math
import numpy as np
import pylab

# Part A: Create Ramachandran plot for Native and Endosomal HA

# Parse PDB file for Native HA, extract angles, save them to array "native"
native = [[0,0]]
native_range = range(39, 153)
for model in Bio.PDB.PDBParser().get_structure("HA_Native", "3EYJ.pdb") :
        polypeptides = Bio.PDB.PPBuilder().build_peptides(model["B"])
        for poly_index, poly in enumerate(polypeptides) :
            phi_psi = np.float_(poly.get_phi_psi_list())
            phi_psi_deg = phi_psi * 180 / math.pi
            for res_index in native_range :
                 native = np.append(native, [phi_psi_deg[res_index,:]], axis=0)

# Parse PDB file for Endosomal HA, extract angles, save them to array "endo"
endo = [[0,0]]
endo_range = range(0, 114)
for model in Bio.PDB.PDBParser().get_structure("HA_Endo", "1HTM.pdb") :
        polypeptides = Bio.PDB.PPBuilder().build_peptides(model["B"])
        for poly_index, poly in enumerate(polypeptides) :
            phi_psi = np.float_(poly.get_phi_psi_list())
            phi_psi_deg = phi_psi * 180 / math.pi
            for res_index in endo_range :
                 endo = np.append(endo, [phi_psi_deg[res_index,:]], axis=0)

# Create Ramachandran plots for each conformation
pylab.figure(1)
pylab.scatter(native[:,0], native[:,1], c='b', marker='o')
pylab.xlabel('Phi angle')
pylab.ylabel('Psi angle')
pylab.title('Ramachandran Plot for Native HA')

pylab.figure(2)
pylab.scatter(endo[:,0], endo[:,1], c='b', marker='o')
pylab.xlabel('Phi angle')
pylab.ylabel('Psi angle')
pylab.title('Ramachandran Plot for Endosomal HA')

# Part C: Create scatter plots for Phi/Psi angles by residue
indices = range(0, 115)
pylab.figure(3)
pylab.scatter(indices, native[:,0], c='b', marker='o')
pylab.scatter(indices, endo[:,0], c='r', marker='o')
pylab.xlabel('Residue Position in Common Sequence')
pylab.ylabel('Phi Angle')
pylab.title('Phi Angles')

pylab.figure(4)
pylab.scatter(indices, native[:,1], c='b', marker='o')
pylab.scatter(indices, endo[:,1], c='r', marker='o')
pylab.xlabel('Residue Position in Common Sequence')
pylab.ylabel('Psi Angle')
pylab.title('Psi Angles')

# Part D: Helical in one conformation but not the other
```

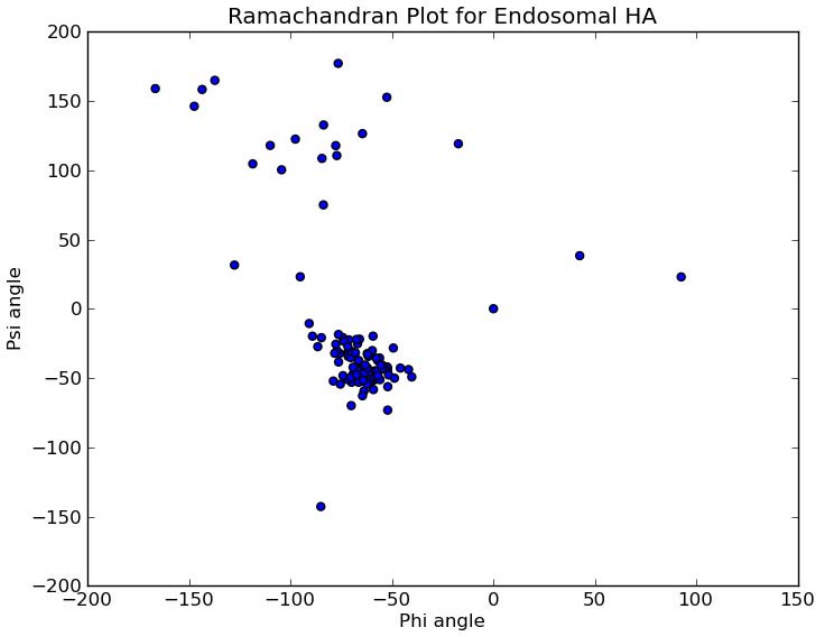
#

```
# Fill arrays with indices of helical residues
native_helices = []
for index in indices :
    if native[index, 0] < -57 :
        if native[index, 0] > -71 :
            if native[index, 1] > -48 :
                if native[index, 1] < -34 :
                    native_helices.append(index+39)

endosome_helices = []
for index in indices :
    if endo[index, 0] < -57 :
        if endo[index, 0] > -71 :
            if endo[index, 1] > -48 :
                if endo[index, 1] < -34 :
                    endosome_helices.append(index+39)

# Search for indices appearing only once
unique = 0
for index in native_helices :
    if endosome_helices.count(index) == 0 :
        unique += 1
for index in endosome_helices :
    if native_helices.count(index) == 0 :
        unique += 1

print "Helical Residues, Native: ", native_helices
print "Helical Residues, Endosomal: ", endosome_helices
print "Unique helical residues: %i" % unique

pylab.show()
```



.

#



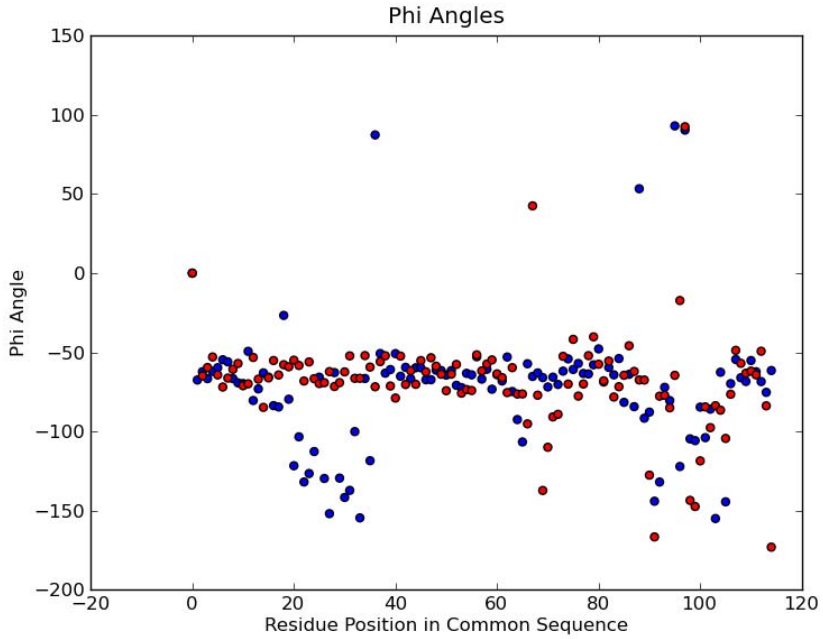
b) What do the Ramachandran plots tell you about the secondary structure of HA in these two conformations?

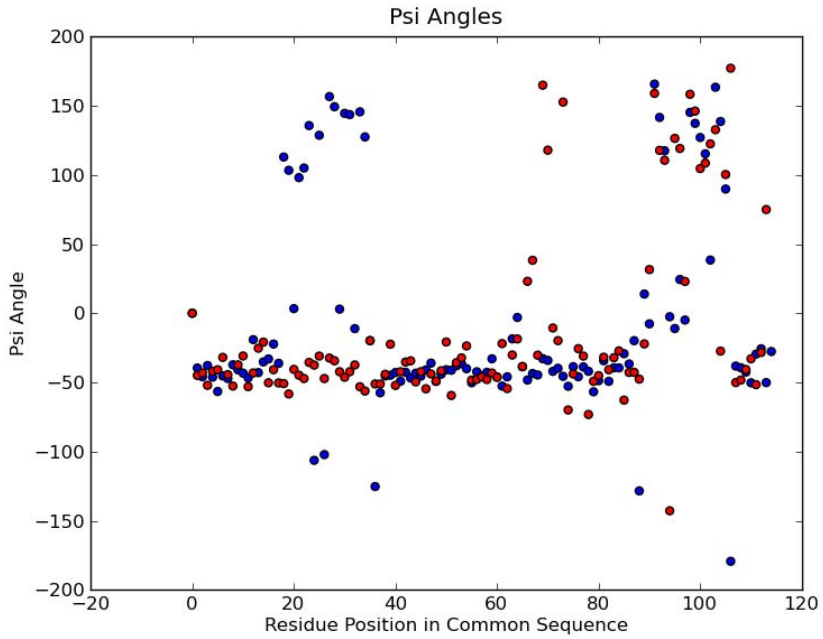
The Ramachandran plots indicate that both conformations have many residues in alpha-helical conformations, with more endosomal hemagglutinin having more helical residues and fewer in beta sheet conformations.

c) Plot phi angles vs. residue number for the two conformations on the same plot. (Each position on the x-axis should have two data points, representing the phi angle of that residue in the two structures). Make a similar plot for psi angles.

20.320 Problem Set 1
Question 1

#





Blue dots correspond to residues in native hemagglutinin, while red dots correspond to hemagglutinin residues at endosomal pH.

#

d) Based on the phi/psi angles, determine the number of residues that are alpha helical in one structure but not in the other. Define a helical residue as one where phi is between -57 and -71˚, and psi is between -34 and -48˚.

Based on the given criteria, the program calculates the following residues as helical for hemagglutinin at native and endosomal pH:

```
Helical Residues, Native:  [40, 41, 42, 43, 47, 48, 49, 53, 77, 78, 81, 82,
83, 84, 85, 86, 88, 89, 90, 91, 93, 96, 97, 99, 106, 107, 110, 111, 112, 114,
115, 116, 117, 120, 122, 125, 147, 148]

Helical Residues, Endosomal:  [41, 44, 46, 48, 60, 61, 63, 65, 68, 69, 71,
81, 82, 88, 91, 96, 97, 99, 119, 126, 127, 148]

Unique helical residues: 40
```

20.320 Problem Set 1
Question 2

#

Key to the function of influenza hemagglutinin is its pH-dependant conformational change in the endosome, fusing the viral membrane with the endosomal membrane and allowing release of the virus into the cytoplasm.

a) Of the principal forces responsible for maintaining the tertiary structure of a protein, which would be most strongly affected by the acidification of the surrounding environment?

Salt bridges (also charge-charge interactions or ionic bonds) would be most strongly affected by pH. Acidifying the environment could protonate amino acids with pKa values below physiological pH, either conferring a positive charge (e.g. histidine) or neutralizing a negative charge (e.g. glutamate, aspartate). This would affect which salt bridges could form and which protein conformation would be most energetically favorable. Of the other forces responsible for maintaining tertiary structure, neutralizing charge would not significantly alter the hydrophobicity of a protein nor would it adversely affect hydrogen bonding.

Structural studies have shown that several histidine residues play a key role in mediating this pH-dependent conformational change of influenza hemagglutinin.

b) What property of histidine makes it especially suited to this role? Which other amino acid residues could potentially serve the same function? Be sure to justify your choices.

Histidine is unique in that its side-chain pKa value is 6.1, which is close to physiological pH. At physiological pH (7.4), histidine is singly protonated, uncharged, and therefore incapable of forming salt bridges. Upon the acidification of the endosome, histidine becomes doubly protonated with a net positive charge. Presuming the pH of the endosome remained above the pKa values of glutamate and/or aspartate, the doubly protonated histidine could then interact with either of these residues.

Glutamate and aspartate could serve a similar function, but the endosome would have to be made much more acidic. Both have side-chain pKa values close to 4.0; therefore at higher pH values they can form salt bridges with positively charged residues.

One way to determine which residues are vital to the structure and function of a protein is to align the sequences of many variants of the protein and look for conserved residues (those that are present in the same position in each protein variant). We can do this by comparing the hemagglutinins across various serotypes of human influenza A. On the Course website, you will find a document containing the amino acid sequences of several influenza hemagglutinins (H1, H2, H3, H5, H7, and H9).

c) Use CLUSTALW to find histidine residues that are conserved across all six sequences. Attach the CLUSTALW alignment, highlighting the residues you find. CLUSTALW is available on Athena clusters, or you can find a web client here: *http://www.ebi.ac.uk/Tools/clustalw2/index.html*

```
gi|63054902|gb|AAY28987|        ----MAIIYLILLFTAVRG-------DQICIGYHANNSTEKVDTILERNV 39
gi|251757610|gb|ACT15357|       ---MEKIVLLLAIVSLVKS-------DQICIGYHANNSTEQVDTIMERNV 40
gi|256383631|gb|ACU78205|       --MKAILVVLLYTFATANA-------DTLCIGYHANNSTDTVDTVLEKNV 41
gi|81174796|gb|ABB58945|        -----------LMVTAINA-------DKICIGYQSTNSTETVDTLTKTNV 32
gi|254564370|gb|ACT67810|       MKTIIALSYILCLVFAQKLPGNDNSTATLCLGHHAVPNGTIVKTITNDQI 50
gi|115279133|gb|ABI85000|       MNTQI-LAFIACMLIGTKG-------DKICLGHHAVANGTKVNTLTERGI 42
                                             .   .         :*:*:::  .   *.*: :  :

gi|63054902|gb|AAY28987|        TVTHAKDILEKTHNGKLCKLNGIPPLELGDCSIAGWLLGNPECDRLLSVP 89
gi|251757610|gb|ACT15357|       TVTHAQDILEKTHNGKLCNLDGVKPLILRDCSVAGWLLGNPMCDEFLNVP 90
gi|256383631|gb|ACU78205|       TVTHSVNLLEDKHNGKLCKLRGVAPLHLGKCNIAGWILGNPECESLSTAS 91
gi|81174796|gb|ABB58945|        PVTQAKELLHTEHNGMLCATNLGHPLILDTCTIEGLIYGNPSCDLLLGGR 82
gi|254564370|gb|ACT67810|       EVTNATELVQSSSTGEICDS-PHQILDGKNCTLIDALLGDPQCDGFQN-K 98
gi|115279133|gb|ABI85000|       EVVNATETVETVNIKKICTQ-GKRPTDLGQCGLLGTLIGPPQCDQFLE-F 90
                                 *.:: : :.      :*           * : . : * * *: :

gi|63054902|gb|AAY28987|        EWSYIMEKENPRDGLCYPGSFNDYEELKHLLSSVKHFEKVKILPKDR-WT 138
gi|251757610|gb|ACT15357|       EWSYIVEKINPANDLCYPGNFNDYEELKHLLSRINHFEKIQIIPKNS-WS 139
gi|256383631|gb|ACU78205|       SWSYIVETSSSDNGTCYPGDFIDYEELREQLSSVSSFERFEIFPKTSSWP 141
gi|81174796|gb|ABB58945|        EWSYIVERPSAVNGMCYPGNVENLEELRLLFSSASSYQRVQIFPDTI-WN 131
gi|254564370|gb|ACT67810|       KWDLFVER-SKAYSNCYPYDVPDYASLRSLVASSGTLE---FNNESFNWT 144
gi|115279133|gb|ABI85000|       DANLIIER-REGTDVCYPGKFTNEESLRQILRGSGGID---KESMGFTYS 136
                                . . ::*      . *** .. :  .*:  .      :          :

gi|63054902|gb|AAY28987|        QHTTT-GGSRACAVSGNPSFFRNMVWLT--KKGSNYPVAQGSYNNTSGEQ 185
gi|251757610|gb|ACT15357|       DHEAS-GVSSACPYQGRSSFFRNVVWLT--QKDNAYPTIKRSYNNTNQED 186
gi|256383631|gb|ACU78205|       NHDSNKGVTAACPHAGAKSFYKNLIWLV--KKGNSYPKLSKSYINDKGKE 189
gi|81174796|gb|ABB58945|        VTYSG--TSSACSN----SFYRSMRWLT--QKDNTYPVQDAQYTNNRGKS 173
gi|254564370|gb|ACT67810|       GVTQN-GTSSACIRRSKNSFFSRLNWLT--HLNFKYPALNVTMPNNEQFD 191
gi|115279133|gb|ABI85000|       GIRTN-GATSACRR-SGSSFYAEMKWLLSNSDNAAFPQMTKSYRNPRNKP 184
                                       : **      **:  : **     .  :*       *

gi|63054902|gb|AAY28987|        MLIIWGVHHPNDETEQRTLYQNVGTYVSVGTSTLNKRSTPEIATRPKVNG 235
gi|251757610|gb|ACT15357|       LLVLWGIHHPNDAAEQTRLYQNPTTYISVGTSTLNQRLIPKIATRSKVNG 236
gi|256383631|gb|ACU78205|       VLVLWGIHHPSTSADQQSLYQNADAYVFVGSSRYSKKFKPEIAIRPKVRD 239
gi|81174796|gb|ABB58945|        ILFMWGINHPPTDTVQTNLYTRTDTTTSVTTEDINRAFKPVIGPRPLVNG 223
gi|254564370|gb|ACT67810|       KLYIWGVHHPGTDKDQIFXXAQASGRITVSTKRSQQTVIPNIGSRPRVRN 241
gi|115279133|gb|ABI85000|       ALIIWGVHHSGSATEQTKLYGSGNKLITVGSSKYQQSFTPSPGARPQVNG 234
                                 * :**::*.     *            * :.  .:   *  . *. *..

gi|63054902|gb|AAY28987|        QGGRMEFSWTLLDMWDTINFESTGNLIAPEYGFKISKRGSSGIMKTEGTL 285
gi|251757610|gb|ACT15357|       QSGRMEFFWTILKSNDAINFESNGNFIAPENAYKIVKKGDSTIMKSELEY 286
gi|256383631|gb|ACU78205|       QEGRMNYYWTLVEPGDKITFEATGNLVVPRYAFAMERNSGSGIIISDTPV 289
gi|81174796|gb|ABB58945|        LQGRIDYYWSVLKPGQTLRVRSNGNLIAPWYGHILSGESHGRILKSDLNS 273
gi|254564370|gb|ACT67810|       IPSRISIYWTIVKPGDILLINSTGNLIAPRGYFKIRS-GKSSIMRSDAPI 290
gi|115279133|gb|ABI85000|       QSGRIDFHWLLLDPNDTVTFTFNGAFIAPDRASFFR--GESLGVQSDVPL 282
                                  .*:.  * ::.  : : .  .* ::.*     :   . .  : ::

gi|63054902|gb|AAY28987|        E-NCETKCQTPLGAINTTLPFHNVHPLTIGECPKYVKSEKLVLATGLRNV 334
gi|251757610|gb|ACT15357|       G-NCNTKCQTPIGAINSSMPFHNIHPLTIGECPKYVKSNRLVLATGLRNS 335
gi|256383631|gb|ACU78205|       H-DCNTTCQTPKGAINTSLPFQNIHPITIGKCPKYVKSTKLRLATGLRNV 338
gi|81174796|gb|ABB58945|        G-NCVVQCQTERGGLNTTLPFHNVSKYAFGNCPKYVGVKSLKLAVGMRNV 322
gi|254564370|gb|ACT67810|       G-KCNSECITPNGSIPNDKPFQNVNRITYGACPRYVKQNTLKLATGMRNV 339
gi|115279133|gb|ABI85000|       DSGCEGDCFHSGGTIVSSLPFQNINPRTVGKCPRYVKQTSLLLATGMRNV 332
                                   *   *    * : .  **:*:   : * **:**    * **.*:**

gi|63054902|gb|AAY28987|        PQIE--------SRGLFGAIAGFIEGGWQGMVDGWYGYHHSNDQGSGYAA 376
gi|251757610|gb|ACT15357|       PQGER----RRKKRGLFGAIAGFIEGGWQGMVDGWYGYHHSNEQGSGYAA 381
gi|256383631|gb|ACU78205|       PSIQ--------SRGLFGAIAGFIEGGWTGMVDGWYGYHHQNEQGSGYAA 380
gi|81174796|gb|ABB58945|        PARS--------SRGLFGAIAGFIEGGWPGLVAGWYGFQHSNDQGVGMAA 364
gi|254564370|gb|ACT67810|       PE--------KQTRGIFGAIAGFIENGWEGMVDGWYGFRHQNSEGRGQAA 381
gi|115279133|gb|ABI85000|       PENPKQAYQKRMTRGLFGAIAGFIENGWEGLIDGWYGFRHQNAQGEGTAA 382
                                *          .**:********.** *:: ****::*.* :* * **
```

```
gi|63054902|gb|AAY28987|        DKESTQKAFDGITNKVNSVIEKMNTQFEAVGKEFSNLERRLENLNKKMED 426
gi|251757610|gb|ACT15357|       DKESTQKAIDGVTNKVNSIIDKMNTQFEAVGREFNNLERRIENLNKKMED 431
gi|256383631|gb|ACU78205|       DLKSTQNAIDEITNKVNSVIEKMNTQFTAVGKEFNHLEKRIENLNKKVDD 430
gi|81174796|gb|ABB58945|        DRDSTQKAIDKITSKVNNIVDKMNKQYEIIDHEFSEIETRLNMINNKIDD 414
gi|254564370|gb|ACT67810|       DLKSTQAAIDQINGKLNRLIGKTNEKFHQIEKEFSEVEGRIQDLEKYVED 431
gi|115279133|gb|ABI85000|       DYKSTQSAIDQITGKLNRLIDKTNQQFELIDNEFSEIEQQIGNVINWTRD 432
                                * .*** *:* :..*:* :: * * ::  : .**..:* ::  : :   *

gi|63054902|gb|AAY28987|        GFLDVWTYNAELLVLMENERTLDFHDSNVKNLYDKVRMQLRDNVKELGNG 476
gi|251757610|gb|ACT15357|       GFLDVWTYNAELLVLMENERTLDFHDSNVKNLYDKVRLQLRDNAKELGNG 481
gi|256383631|gb|ACU78205|       GFLDIWTYNAELLVLLENERTLDYHDSNVKNLYEKVRSQLKNNAKEIGNG 480
gi|81174796|gb|ABB58945|        QIQDIWAYNAELLVLLENQKTLDEHDANVNNLYNKVKRALGSNAMEDGKG 464
gi|254564370|gb|ACT67810|       TKIDLWSYNAELLVALENQHTIDLTDSEMNKLFEKTKKQLRENAEDMGNG 481
gi|115279133|gb|ABI85000|       SMTEVWSYNAELLVAMENQHTIDLADSEMNKLYERVRKQLRENAEEDGTG 482
                                   ::*:******* :**::*:*  *:::::*:::.:  * .*. : *.*

gi|63054902|gb|AAY28987|        CFEFYHKCDDECMNSVKNGTYDYPKYEEESKLNRNEIKGVKLSSMGVYQI 526
gi|251757610|gb|ACT15357|       CFEFYHRCDNECMESVRNGTYDYPQYSEEARLKREEISGVKLESIGTYQI 531
gi|256383631|gb|ACU78205|       CFEFYHKCDNTCMESVKNGTYDYPKYSEEAKLNREEIDGVKLESTRIYQI 530
gi|81174796|gb|ABB58945|        CFELYHKCDDRCMETIRNGTYNRGKYKEESRLERQKIEGVKLESEGTYKI 514
gi|254564370|gb|ACT67810|       CFKIYHKCDNACIGSIRNGTYDHDVYRDEALNNRFQIKGVELKS-GYKDW 530
gi|115279133|gb|ABI85000|       CFEIFHKCDDQCMESIRNNTYDHTQYRTESLQNRIQIDPVKLSS-GYKDI 531
                                **:::*:**: *: :::*.**:   *  *:  :* :*. *:*.*    .

gi|63054902|gb|AAY28987|        LAIYATVAGSLSLAIMMAGISFWMCSNGSLQCRICI 562
gi|251757610|gb|ACT15357|       LSIYSTVASSLALAIMVAGLFLWMCSNGSLQCRICI 567
gi|256383631|gb|ACU78205|       LAIYSTVASSLVLVVSLGAISFWMCSNGSLQCRICI 566
gi|81174796|gb|ABB58945|        LTIYSTVASSL------------------------- 525
gi|254564370|gb|ACT67810|       ILWISFAISCFLLCVALLGFIMWACQKGNIRCNICI 566
gi|115279133|gb|ABI85000|       ILWFSFGASCFLLLAIAMGLVFICIKNGNMRCTICI 567
                                :   :   ..:
```

CLUSTALW identifies 3 histidine residues (indicated in bold above) that are conserved across all six hemagglutinin sequences.

d) Assuming the pH of the acidified endosome is 4.5, which types of residues would you expect to see complexed with these key histidines? Based on your CLUSTALW analysis, identify the other residues that are likely involved with this pH-dependant transition.

If the pH of the acidified endosome is 4.5, a significant number of glutamate (E, pKa = 4.07) residues will be negatively charged, while the majority of aspartate (D, pKa = 3.86) residues will be negatively charged. These would form salt bridges with the positively charged histidine residues and stabilize the conformation of hemagglutinin. If these residues are required for this stabilization, they should be conserved as well. The CLUSTALW analysis indicates there are several aspartate and glutamate residues that are conserved across all six sequences (highlighted in green).

#

e) Use Biopython to compute the distance between the alpha carbons of the conserved histidine residues you identified in Part (c) and the other conserved residues you identified in Part (d). Report the minimum distance you find for each conserved histidine. Does any pair of residues seem especially close together? Some hints:
   1. For this exercise, as with Question 1, only consider residues in the "B" chain of hemagglutinin at endosomal pH. This sequence is posted on the Course website in FASTA format.
   2. It will help to repeat your CLUSTALW alignment from Part (c) with this new sequence – this will help you find the residues you are looking for.
   3. You can copy and paste the FASTA sequence directly into the list of hemagglutinin sequences you analyzed in Part (c).
   4. You should only be looking for residues that are conserved across all seven sequences in your new alignment.
   5. `residue["CA"].coord` returns the coordinates (*x, y, z*) of the alpha carbon of `residue`

When we consider chain "B" of influenza hemagglutinin at endosomal pH, only the last histidine in the sequence (H142) is absolutely conserved. We are therefore interested in the distances between H142 and the absolutely conserved acidic residues in chain "B", namely E57, E61, D86, E97, E103, D109, D112, and D145.

The following code will calculate and print the distance between H142 and the specified residues:

```
# Problem Set 1 Question 2
# August 31, 2009


import Bio.PDB
import numpy
res_index = [17, 21, 46, 57, 63, 69, 72, 85]
for model in Bio.PDB.PDBParser().get_structure("HA_Endosomal", "1HTM.pdb") :
    polypeptides = Bio.PDB.PPBuilder().build_peptides(model["B"])
    for poly_index, poly in enumerate(polypeptides) :
       key_his = poly[102]
       print "Distances from %s%i: (angstroms)" % (key_his.resname, key_his.id[1])
       for index in res_index :
          dist_vector = poly[index]["CA"].coord - key_his["CA"].coord
          distance = numpy.sqrt(numpy.sum(dist_vector * dist_vector))
          output = "%s%i %f" % (poly[index].resname, poly[index].id[1], distance)
          print output
```

This code produces the following output:

```
Distances from HIS142: (angstroms)
GLU57 44.210747
GLU61 37.950996
ASP86 5.515404
GLU97 17.523859
GLU103 27.225662
ASP109 30.871565
ASP112 32.381725
GLN125 14.174937
```

Based on this, His142 seems especially close to Asp86, potentially indicating a salt bridge between these two residues.

20.320 Problem Set 1
Question 3

#

Cystic fibrosis (CF) is a genetic disorder caused by a mutation(s) in the cystic fibrosis transmembrane conductance regulator (CTFR) gene. CTFR, the protein product, is a traffic ATPase that transports chloride ions across epithelial cell membranes. Mutations lead to improper folding of CTFR and prevent proper chloride ion transport across these cell membranes. The ΔF508 mutation, aka the deletion of the phenylalanine (F) at position 508, is the most common mutation associated with cystic fibrosis.

a) Explain why the deletion of the phenylalanine (F) at position 508 might lead to misfolding (discuss the amino acid & its impact on structure).

The deletion of phenylalanine at position 508 leading to a folding defect can be explained by two possibilities: 1. the loss of the spacing effect of the peptide backbone, or 2. the loss of the phenylalanine side chain. Phenylalanine is an aromatic amino acid and therefore contributes to a hydrophobic region of the nucleotide binding domain. Since the phenylalanine side chain is partially surface-exposed, deletion of this amino acid can introduce local structural changes to the amino acid residues surrounding F508. Deletion of the peptide backbone would bring together two amino acid side chains that originally were separated. This would change the conformational space of the original surface and could lead to misfolding. (Experiments have shown that the peptide backbone is critical for the folding efficiency. Side chain mutations have little effect in proper folding.) (Thibodeau et al. Nat Struct Mol Biol 12 2004 p10-16)

The ΔF508 mutation along with several other known mutations that cause CF, occur in a region of the CTFR known as a nucleotide binding domain (NBD1). In an experiment, (Qu & Thomas JBC 271:13 1996 p. 7261-7264) NBD1 and NBD1 with the ΔF508 mutation (NBD1ΔF) were tested for folding yield at different temperatures.

b) Calculate the ΔG_{fold} (kcal/mol) at 37°C and 25°C of NBD1 and NBD1Δ using the following data: From the paper, we know that "at 2 μM final NBD1 concentration and 37°C, 63% of the wild type polypeptide folds into the soluble conformation, while only 38% of the ΔF508 assumes the folded conformation. At 18 μM final polypeptide concentration and 25 °C, 29% of the wild type domain reaches the native state in contrast to 19% of the ΔF508 mutant." Are these values reasonable? Explain.

From lecture, we know that: $\Delta G_{\text{fold}} = -RT \ln K_{\text{fold}} = -RT \ln\left(F/U\right)$

Therefore, at 37˚C:

WT: $\Delta G_{\text{fold}} = -\left(1.987\,{}^{\text{cal}}/_{\text{mol–K}}\right)\left(310\text{ K}\right)\ln\left({}^{63\%}/_{37\%}\right) = -328\,{}^{\text{cal}}/_{\text{mol}} = -0.33\,{}^{\text{kcal}}/_{\text{mol}}$

ΔF508: $\Delta G_{\text{fold}} = -\left(1.987\,{}^{\text{cal}}/_{\text{mol–K}}\right)\left(310\text{ K}\right)\ln\left({}^{38\%}/_{62\%}\right) = 302\,{}^{\text{cal}}/_{\text{mol}} = 0.30\,{}^{\text{kcal}}/_{\text{mol}}$

And at 25˚C:

WT: $\Delta G_{\text{fold}} = -\left(1.987\,{}^{\text{cal}}/_{\text{mol–K}}\right)\left(298\text{ K}\right)\ln\left({}^{29\%}/_{71\%}\right) = 530\,{}^{\text{cal}}/_{\text{mol}} = 0.53\,{}^{\text{kcal}}/_{\text{mol}}$

ΔF508: $\Delta G_{\text{fold}} = -\left(1.987\,{}^{\text{cal}}/_{\text{mol–K}}\right)\left(298\text{ K}\right)\ln\left({}^{19\%}/_{81\%}\right) = 858\,{}^{\text{cal}}/_{\text{mol}} = 0.86\,{}^{\text{kcal}}/_{\text{mol}}$

This same group determined the free energy change of denaturation ΔG_D of wild-type NBD1 along with various other mutants from known CF cases at 37°C in a separate publication (Qu et al, JBC 272:25 1997 p 15739-15744), using somewhat different experimental conditions from the 1996 paper.

Protein	$\Delta G_{D,0}$ (kJ/mol)	$\Delta\Delta G_{D,0}$ (kJ/mol)
NBD1	15.5	
NBD1ΔF	14.4	-1.1
NBD1-R553M	16.6	1.1
NBD1ΔF-R553M	14.1	-1.4
NBD1-S549R	16.7	1.2
NBD1-G551D	16.6	1.1

c) Given the values of the ΔGs calculate the K_{fold} (ratio of folded to unfolded) of wild-type NBD1 and all of the mutants. ($\Delta G_{fold} = -\Delta G_D$)

Rearranging the above equations: $K_{\text{fold, NBD1}} = \exp\left(-\frac{\Delta G_{\text{fold}}}{RT}\right) = \exp\left(\frac{\Delta G_{\text{D}}}{RT}\right)$

Therefore,

$$K_{\text{fold, NBD1}} = \exp\left(\frac{15.5\,{}^{\text{kJ}}/_{\text{mol}}}{0.0083\,{}^{\text{kJ}}/_{\text{mol K}} \times 310\text{K}}\right) = 409$$

Similarly, the following K_{fold} values can be calculated:

Mutant	ΔG_D (kJ/mol)	K_{fold} (Unitless)
NBD1	15.5	409
NBD1ΔF	14.4	267
NBD1-R553M	16.6	627
NBD1ΔF-R553M	14.1	238
NBD1-S549R	16.7	651
NBD1-G551D	16.6	627

d) Is the ΔG_{fold} for the wild type in Part (c) the same as your answer to Part (b)? Explain.

No, the ΔG_{fold} values are quite different between experiments. The wild-type protein has a ΔG_{fold} of < 1 kcal/mol in the first set of experiments, but roughly 3 kcal/mol in the second set (recall that 1 kcal = 4.184 kJ). This could be due to differences between conditions under which the experiments were performed. Salt concentrations and buffer compositions could affect the reported ΔG values, as could crowding effects from differing concentrations of the protein.

20.320 Analysis of Biomolecular and Cellular Systems
Fall 2012