Biomolecular Structure 2005

Take home exams allow a creativity that is not available with in-class exams. In particular, it allows you to exercise facility with computer and on-line resources that are key to structural analysis and genomics.

This exam is to be worked on independently. There is evidence from the previous exam that this was not adhered to by some of you. Failure to fully follow the rules may result in the scoring of an "F" for the *entire course*, with other disciplinary interactions likely.

You may not talk, or in *any other way* communicate with *anyone* other than Professor Martin or Professor Hardy about *any* aspect of this exam. You are not to share any material of any kind with each other (or with anyone else).

> I understand the above rules and hereby certify that the answers I provide are solely my own. I did not communicate with anyone other than Professors Martin and Hardy about any aspect of this exam. I did not help anyone nor did I receive any help. I did not share any resources.
> I understand the consequences of failure to fully follow these rules.

Signature: ______________________________________

1. (10 points) You have graduated from U. Mass as an expert in Biomolecular Structure. As a post-doctoral Fellow, your challenge is to create a protein of new function. We have discussed at least four ways to create a protein of new function. Describe what protein you will start with, what new function you will create and what methods you will use to create your new function.

   Methods: a) gene shuffling b) rational or computational design c) directed mutagenesis based on homology d) binary patterning and selection e) phage display f) random mutagenesis and selection.

   Should also discuss a screen or selection or test for the new function.

2. (10 points) Predicting protein structure and function from amino acid sequence is difficult. To predict structure from sequence, it is necessary to use a computer algorithm to calculate the energy of various conformations of the sequence of unknown structure. *In silico* the protein is allowed to adopt various conformations and the algorithm is used to calculate the energy of each conformation. If the algorithm is good, the calculated energies should allow you to distinguish the proper fold from all improper folds.

   a. To select the proper fold, would you select the highest or the lowest energy structure?

   b. Based on what you have learned about protein structure, what parameters would you include in an algorithm so that you could most accurately calculate the energy of the conformation you are sampling?

   c. Which of those parameters would be most important? Why?

Parameters: a) hydrophobic burial b) hydrophilicity on outside c) amino acid propensity for helix and sheet d) side-chain rotomer strain e) salt bridges f) H-bonding potential g) conformation of local sequence in known structures (like Rosetta). h) etc.

Most important should be hydrophobic burial.

3. (10 points) We have discussed several proteins that undergo huge conformational changes.

   a. For each case of T7 RNA polymerase, Influenza Hemagluttinin, and GroEL, what drives the huge conformational change?

   t7 RNA polymerase: Displacement by the elongating transcript, which is ultimately driven by phosphoryl transfer (addition of NTP).

   Influenza Hemagluttinin: pH

   GroEL: ATP hydrolysis

   b. What can your surmise about the energetics of the various conformations that have been observed in crystal structures.

   The two conformations have similar energies and the two different conformations reflect the global free energy minimum in that particular condition.

4. (10 points) You are studying the system fireflies use to regulate their fluorescence. Assume that fireflies have evolved to quench their florescence in response to a fly swatter and that quenching of fluorescence should be very rapid and reversible. In class we discussed over a dozen different mechanisms for regulation of protein function. List these mechanisms, indicate which you think is the most likely to be involved in regulating firefly fluorescence, and justify why you predict that one over the others.

   Phosphorylation: fast and reversible, what nature usually uses for this kind of mechanism so this is probably the best

   Effector ligands (allosteric or competitive): Fine, but must be controlled before the signal is sent so that is an extra level of complexity

   Nucleotide hydrolysis: May be useful, but because it is dependent on energy balance in the cell it may not be good as a survival adaptation.

   Proteolysis cascade: not perfect because it is not reversible, but proteolysis can work rapidly so that would be an advantage.

   inteins, glycosylation, lipid modification, methylation, n-acetylation, sumoylation, nitrosylation, localization, pH, redox environment, compartmentalization and degradation: too slow to be useful for a rapid response

5. (10 points) Shown below are the sequences for Protein Phosphatase 1 (PP1) and Protein Phosphatase 5 (PP5). Using an online alignment tool like those we discussed in class (e.g. http://ca.expasy.org/) please align the two sequences and ***turn in your alignment***.

```
PP1:
      1 msdseklnld siigrllevq gsrpgknvql teneirglcl ksreiflsqp illeleaplk
     61 icgdihgqyy dllrlfeygg fppesnylfl gdyvdrgkqs leticlllay kikypenffl
    121 lrgnhecasi nriygfydec krryniklwk tftdcfnclp iaaivdekif cchgglspdl
    181 qsmeqirrim rptdvpdqgl lcdllwsdpd kdvqgwgend rgvsftfgae vvakflhkhd
    241 ldlicrahqv vedgyeffak rqlvtlfsap nycgefdnag ammsvdetlm csfqilkpad
    301 knkgkygqfs glnpggrpit pprnsakakk

PP5:
      1 ertecaeppr deppadgalk raeelktqan dyfkakdyen aikfysqaie lnpsnaiyyg
     61 nrslaylrte cygyalgdat raieldkkyi kgyyrraasn malgkfraal rdyetvvkvk
    121 phdkdakmky qecnkivkqk aferaiagde hkrsvvdsld iesmtiedey sgpkledgkv
    181 tisfmkelmq wykdqkklhr kcayqilvqv kevlsklstl vettlketek itvcgdthgq
    241 fydllnifel nglpsetnpy ifngdfvdrg sfsveviltl fgfkllypdh fhllrgnhet
    301 dnmnqiygfe gevkakytaq myelfsevfe wlplaqcing kvlimhgglf sedgvtlddi
    361 rkiernrqpp dsgpmcdllw sdpqpqngrs iskrgvscqf gpdvtkafle ennldyiirs
    421 hevkaegyev ahggrcvtvf sapnycdqmg nkasyihlqg sdlrpqfhqf tavphpnvkp
    481 mayantllql gmm
```

a. What is the % identity of the two sequences?

b. Assuming there is a structure of PP1 and not one for PP5, would this be an ideal case to apply homology modeling?

c. On the sequence of PP1, indicate the sites of digestion by the Arg-C protease based on an on-line digestion prediction tool.

d. On the sequence of PP5 indicate which Ser, Thr and Tyr residues are likely to be phosphorylated. Explain how you made your predictions. If you used an on-line tool, please cite it.

```
a. LALIGN finds the best local alignments between two sequences version 2.0u66 September
1998 Please cite: X. Huang and W. Miller (1991) Adv. Appl. Math. 12:373-381

 Comparison of:
(A) ./wwwtmp/lalign/.14211.1.seq PP1                                    - 330 aa
(B) ./wwwtmp/lalign/.14211.2.seq PP5                                    - 493 aa
 using matrix file: BL50, gap penalties: -14/-4

  44.8% identity in 239 aa overlap; score:  699 E(10,000): 3.6e-55

            50        60        70        80         90       100
PP1       LSQPILLELEAPLKICGDIHGQYYDLLRLFEYGGFPPESN-YLFLGDYVDRGKQSLETIC
          : .  : : :  . .::: :::.:::: .:: .:.: :.: :.: ::.::::. :.:.:
PP5       LVETTLKETEK-ITVCGDTHGQFYDLLNIFELNGLPSETNPYIFNGDFVDRGSFSVEVIL
       220       230         240       250       260       270

            110       120       130       140       150       160
PP1       LLLAYKIKYPENFFLLRGNHECASINRIYGFYDECKRRYNIKLWKTFTDCFNCLPIAAIV
           :...:. ::..: :::::::  ..:.::::  : : .:. .... :.. :. ::.:  .
PP5       TLFGFKLLYPDHFHLLRGNHETDNMNQIYGFEGEVKAKYTAQMYELFSEVFEWLPLAQCI
         280       290       300       310       320       330

            170        180       190       200       210       220
PP1       DEKIFCCHGGL-SPDLQSMEQIRRIMRPTDVPDQGLLCDLLWSDPDKDVQGWGENDRGVS
          . :..  :::: : :  ....::.: :  . ::.: .::::::::. . .: . . ::::
PP5       NGKVLIMHGGLFSEDGVTLDDIRKIERNRQPPDSGPMCDLLWSDPQPQ-NGRSISKRGVS
         340       350       360       370       380        390
```

```
          230       240       250       260       270       280
 PP1    FTFGAEVVAKFLHKHDLDLICRAHQVVEDGYEFFAKRQLVTLFSAPNYCGEFDNAGAMM
          :: .:.  ::....:: : :.:.:  .:::     . ::.::::::: .. : ....
 PP5    CQFGPDVTKAFLEENNLDYIIRSHEVKAEGYEVAHGGRCVTVFSAPNYCDQMGNKASYI
```

b. The identity is 48% for the 239 a.a. region shown above, which is not the entire length of the protein. For this region of the protein the homology is high enough to allow reliable homology modeling. Note this model will only be meaningful within the domain where the sequence homology is high above 40%. In the other regions of the protein, the homology model would be quite meaningless.

c._Arg-C proteinase cleavage positions:

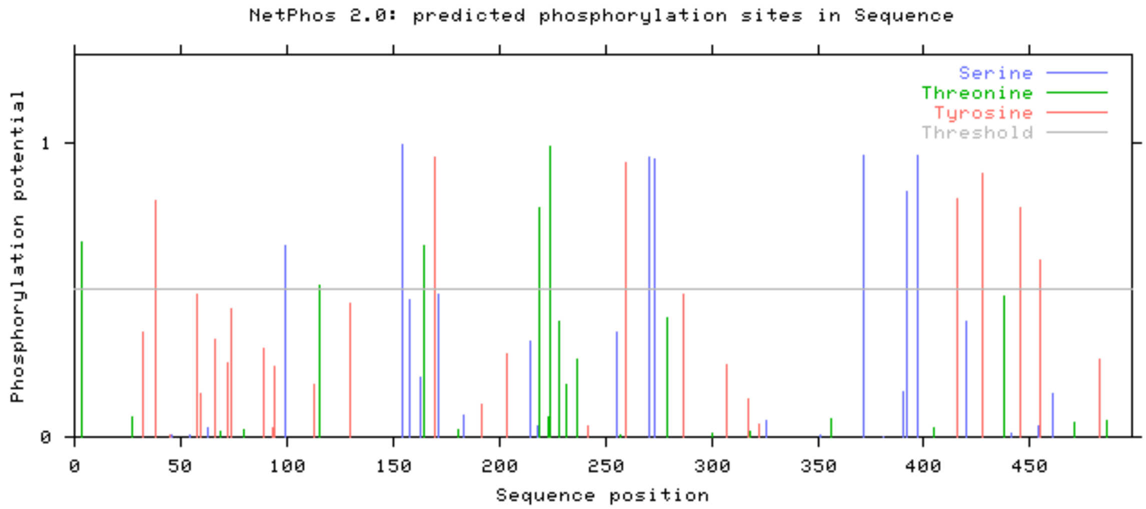
15 23 36 43 74 96 122 132 142 143 187 188 191 221 246 261 317 323

At the C-terminal side of Arg (R), ie., the red residues below:

```
  1 msdseklnld siigrllevq gsrpgknvql teneirglcl ksreiflsqp illeleaplk
 61 icgdihgqyy dllrlfeygg fppesnylfl gdyvdrgkqs leticlllay kikypenffl
121 lrgnhecasi nriygfydec krryniklwk tftdcfnclp iaaivdekif cchgglspdl
181 qsmeqirrim rptdvpdqgl lcdllwsdpd kdvqgwgend rgvsftfgae vvakflhkhd
241 ldlicrahqv vedgyeffak rqlvtlfsap nycgefdnag ammsvdetlm csfqilkpad
301 knkgkygqfs glnpggrpit pprnsakakk
```

d. Phosphorylation prediction by NetPhos 2.0 Server - prediction results

```
ERTECAEPPRDEPPADGALKRAEELKTQANDYFKAKDYENAIKFYSQAIELNPSNAIYYGNRSLAYLRTECYGYALGDAT      80
RAIELDKKYIKGYYRRAASNMALGKFRAALRDYETVVKVKPHDKDAKMKYQECNKIVKQKAFERAIAGDEHKRSVVDSLD     160
IESMTIEDEYSGPKLEDGKVTISFMKELMQWYKDQKKLHRKCAYQILVQVKEVLSKLSTLVETTLKETEKITVCGDTHGQ     240
FYDLLNIFELNGLPSETNPYIFNGDFVDRGSFSVEVILTLFGFKLLYPDHFHLLRGNHETDNMNQIYGFEGEVKAKYTAQ     320
MYELFSEVFEWLPLAQCINGKVLIMHGGLFSEDGVTLDDIRKIERNRQPPDSGPMCDLLWSDPQPQNGRSISKRGVSCQF     400
GPDVTKAFLEENNLDYIIRSHEVKAEGYEVAHGGRCVTVFSAPNYCDQMGNKASYIHLQGSDLRPQFHQFTAVPHPNVKP     480
MAYANTLLQLGMM                                                                        560
..T..................................Y..........................................      80
..................S...............T......................................S......     160
....T....Y................................................T....T................     240
...................Y..........S.S...............................................     320
...................................................S...................S....S...     400
...............Y...........Y................Y.........Y.........................     480
```



560

7 Serines: 99, 154, 271, 273, 372, 392, 397

5 Threonines: 3, 115, 165, 219, 224

7 Tyrosines: 38, 170, 260, 416, 428, 445, 455

6. (5 points) In protein sequence alignment algorithms, there are all sorts of parameters one can set (and get correspondingly different alignments!). Two of these are:

   Gap penalty – penalty to create a gap in the alignment

   Gap length penalty – penalty for extending the cap

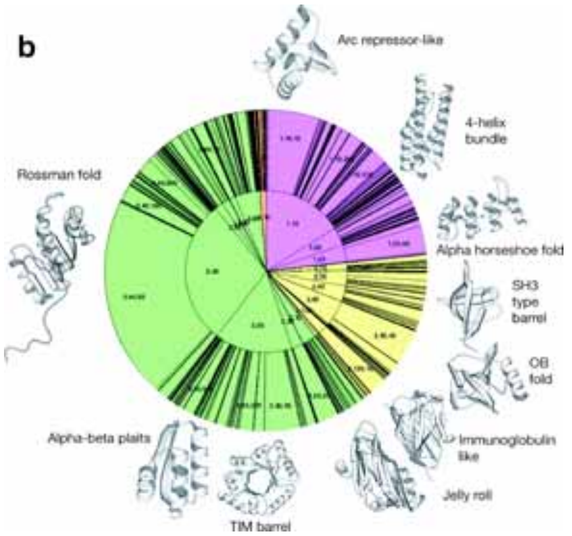
   Why are there two separate parameters for insertion of gaps into alignments? Explain this in terms of things you have learned in this class.

   When aligning two (or more) proteins, the need to introduce a gap in one protein's sequence presumably arises because the other protein contains an insertion. We need to invoke a "penalty" for this, however, or the algorithm would simply insert 1-2 aa insertions all over the place to make up for what really are sequence variations – hence the "Gap Penalty." Insertions are most likely to occur at loops on the protein surface. During evolution, if a surface loop will tolerate a single aa insertion, odds are good that it can accommodate more. Hence the "Gap Length Penalty" should be lower than the penalty for initiating a gap in the first place (remember that this is all done without knowledge of the protein structure, so we're hoping that the inserted gap really does have structural significance).

7. (5 points) Histidine is the amino acid most commonly found at catalytic active sites. Why?

   The His imidazole is the only sidechain that has a pKa close to neutral. Thus it can readily function as an acid-base catalyst. It can also function as a nucleophile, use its charged form to stabilize charged transition states, and can both accept and donate H-bonds.

8. (10 points) A popular classification scheme for family and superfamily folds is called CATH. An analysis of the distribution of different kinds of folds shows an unusually large percentage of proteins with an α/β architecture (Rossman fold, TIM barrel, α/β plaits). Think carefully about the common structural features of these folds and propose why they might show up so often in evolution. Talk both in structural and evolutionary terms.



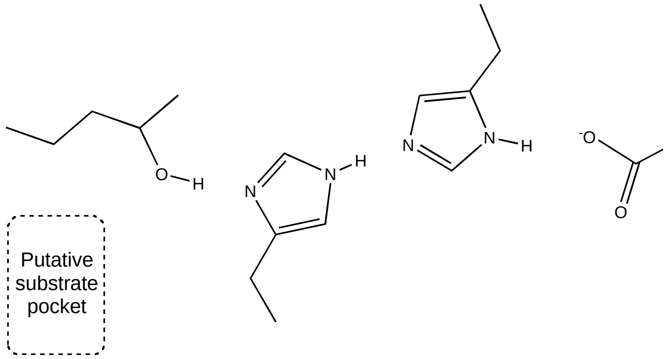
   A variety of answers are acceptable here, all based on the idea that the family structure allows energetically more facile substitutions than in some other families. The simplest is to note that the α/β architecture provides a nice hydrophobic packing, but with less severe constraints than in some other fold classes. The sheet can "slide around" a bit relative to the helices. The four helix bundle, in contrast, has rather rigid requirements for packing the ridges of the helices against one another. Sliding of the helix along its axis requires simultaneous rotation. Thus deviations needed to accommodate amino acid changes can be more difficult.

   This is a question that prompted you to "speculate wildly" (so long as you retain structural "reasonableness."

9. (15 points)You are trying to decide on the function of an ORF protein whose structure you have just determined. You find a conserved set of residues positioned as at right and suspect that a substrate molecule might bind in the pocket indicated.

   a. What general kind of reaction with the substrate (using simple organic nomenclature) might the Thr residue be involved in? Complete the structures of the amino acids, including double bonds and assignment of hydrogens.

Answer: the Thr could be positioned for nucleophilic attack. One His acts to deprotonate the Thr hydroxyl, while the second His insures proper protonation of the first (ie., stables the appropriate tautomer), and the carboxylate insures/stabilizes proper protonation of the second.



   b. What other considerations might have gone into the identification of the substrate binding pocket?

Answer: we might have noticed that the pocket is concave, with patches of conserved hydrophobicity.

10. (15 points) Go retrieve structure 1TUP from the PDB. This is the tumor suppressor protein p53 bound to DNA.

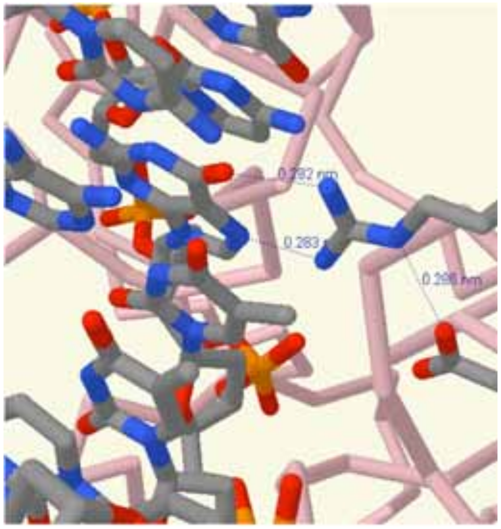
   a. identify an amino acid that is intimately involved in the *recognition* of a specific base in the DNA.

   b. very often, other residues in the protein are used to position the residue making primary contact with the DNA (a term called "buttressing). Identify an amino acid that is carrying out a buttressing function for your DNA contact residue identified in part (a) above (if you can't find one, then go back to (a) and choose another DNA contact residue).

   c. present a figure illustrating the interaction (using PyMOL or your favorite visualization program).

Answer: see picture at right. Arg 280 contacts G253, while Asp 281 interacts with a N on Arg 280 to orient the side chain for optimal contact.

Other answers are acceptable, but should involve two contacts (ideally) with base atoms that will provide direct sequence specificity (contacts with the phosphates will not provide DNA sequence specificity).


0.282 nm
0.283
0.295 nm