The answers below are taken from among the best answers submitted. Authors are noted before each answer.

1. In the paper *Science*. 2008 319(5868):1387-91, Baker and coworkers report the computational design of Retro-Aldol enzymes. Based on this paper answer the following questions.
   a. Why is the ability to computationally design enzymes desirable? Please provide concrete examples of how designed enzymes might be useful. (2 points)

   **MAX:** Designed enzymes could be used to catalyze reactions that are not currently performed in biology. For example, an enzyme could be designed to effectively break down cholesterol buildup in arteries. The protein could be designed to have a specific activity and stability, so the protein could be administered without side effects and it could be naturally degraded over time.
   
   Alternatively, enzymes could be designed that are more stable or active than their biological counterparts, enhancing the possibility of using enzymes in bio-matterial hybrid applications. Most material applications require that the system is stable and reliable for days or even years, so enhanced stability is extremely important. Membrane-bound enzymes require a lipid membrane to stabilize their structure; an enzyme could be designed that performs the same reaction in solution, allowing mass quantities of a chemical to be catalytically produced without a cellular environment. Since cell growth would not need to be supported, the products could be synthesized more efficiently in an industrial setting.

   b. Why has computational design been so challenging in the past? (3 points)

   **DAVID:** The authors of the paper state that previous attempts have been limited by computation power of computers in the past and the cost/time associated with synthesizing large ensembles of designed proteins. Also, the group mentioned that their outline of a general methodology for designing enzymes capable of multi-step reactions helps to overcome previous limitation stemming from the lack of such a methodology. I think that the limitations of computational models and accurate algorithms to predict allowable and “best” putative protein designs have also been a challenge to computational design. Additionally as pointed out in Lippow *et al* (2007), limitation of knowledge of functions of known enzymes has been a challenge to computation protein design, as has been the use of programs that rely heavily upon human intervention/input.

   c. Describe the process the Baker lab used to generate their panel of retro-aldolases. Be certain to comment on the number of designs assessed, the number of scaffolds in which design was successful and on experiments preformed that validated their success. (5 points).

   **DAVID:** In order to generate their retro-aldolase proteins, the Baker lab first chose the mechanism that they wanted their designed proteins to utilize to carry out the reaction (enamine catalysis by lysine). Having chosen the mechanism, the group next identified the intermediates and transition steps that would be required for this reaction and
gathered known models of these events. The group specifically allowed for the design of a site that would accommodate several different transition states and reaction intermediates because their model allowed for reactions using any of four different catalytic motifs. The next step was to superimpose these models to determine a putative (composite) active site for a protein catalyzing this reaction. Having designed a set of 181,555 putative active sites, the group identified scaffold proteins with catalytic pockets that could accommodate this site and then used the program RosettaMatch to generate a list of “matches” of positions in the scaffold protein where possible rotamers of the residues in the composite active site could be located to facilitate the intermediates and transition states required for the reaction. Having this list of matches, the group then optimized the active site by examining the combinations of positions of the side catalytic side chains to find orientations that gave reduced steric clashes but also retained the shape/geometry necessary for the reaction. The non-catalytic residues (present in the active site but not participating directly in the reaction) were then optimized to maximize the affinity of the transition state binding event. The binding energy of each of the resulting ensemble of putative protein designs was then determined and this, as well as assessment of the geometry of the active site-catalytic residues, was then used to predict the 72 designs with the highest enzymatic potential. The group synthesized what they predicted as the best designs (72 designs based on 10 different scaffold proteins), heterologously expressed these in *E. coli* and used a fluorescence-based assay to assess the efficiency of the 70 proteins that were produced in a soluble form.

d. Which of the four catalytic site geometries that they designed was the most successful in terms of catalytic rate enhancement? Why do you believe this was better than the other geometries used? (5 points)

**MAX:** Motif (IV) was most successful, with rate enhancements greater than four orders of magnitude than the uncatalyzed reaction. These designs employed a water molecule rather than an amino acid side chain to mediate proton shuffling in the reaction. The authors state that the water molecule “offers considerable versatility, because they can readily reorient to switch between acting as hydrogen bond acceptors and donors and involve neither delicate free-energy tradeoffs nor intricate interaction networks.” From the illustrations in figure 2.C., it is easy to see that this active site is the most flexible, an asset for any reaction with multiple transition states; therefore, it had a high chance to be successful. Motif IV may be the best design overall to catalyze the reaction, or other factors may have caused this conclusion:

They state that it is difficult to model the energetic tradeoffs that occur when polar side chains are solvated/desolvated or undergo favorable hydrophilic interactions. For this reason, it was more difficult to computationally generate plausible models for an extended hydrophilic network. The highest percentage of RosettaMatch-recommended scaffolds came from Motif IV (38/72 or 53%), so there were more Motif IV scaffolds that were given an experimental chance to be active. It is possible that RosettaMatch incorrectly ruled out some of the models based on an extended hydrophilic network.

2. Imagine you are living in the days before Roger Tsien’s Nobel Prize-winning work to make fluorescent proteins with many different absorbance and emission values (circa 1998). You want to generate some new variants of GFP that have alternative properties
using directed evolution, genetic engineering, computational design or a combination of all three. What kind of approach would you use to generate the colored FPs? Please be specific. Feel free to use examples from the existing literature to strengthen your case. (10 points)

DAVID: Living in the late 90s, I would likely shy away from computational design because of the reasons stated in question 1b above. Both the protein structure and sequence of GFP were known at this time, making directed evolution and genetic engineering both viable options. Because it is more easily adaptable to high-throughput screens, I would choose to utilize directed evolution. The coding sequence of wildtype GFP will be cloned into an expression vector in frame with a secretion signal. The resulting construct will then be used as a template for amplification with Taq polymerase or amplification under alternative error-prone conditions. The resulting PCR product will be a mix of copies of the gene with a wealth of different mutations/misincorporations. This product will be cloned into the same expression vector described above for wildtype protein. The resulting plasmid mixture (composed of many different gfp mutants) will be transformed into E. coli which, after a brief recovery period, will be aliquoted into the wells of 384-well plates containing media appropriate for expression and secretion of the GFP variants. At several timepoints after induction for GFP expression, the fluorescence will be determined using a fluorescence microplate reader and varying the excitation/emission filter sets. Cells expressing protein that exhibits excitation/emission spectra differing from wildtype GFP will be isolated from a likely heterogenous population by limiting dilution. Clones that exhibit the witnessed shift in GFP excitation/emission will then be cultured for plasmid isolation. Sequencing of the gfp insert will then be used to identify the mutation(s) giving rise to the GFP variant isolated.

LUIS:

3. For the protein you have selected and have had approved for your in class presentation use Consurf (http://consurf.tau.ac.il/index.html) to analyze the structure.
   a. Explain how Consurf works and the significance of the results. (2 points)

   ConSurf works by identifying the functional regions in a protein based on the evolutionary conservation of the amino acid in a particular chain. This approach takes the phylogenetic relationship of the proteins stored in the ConSurf
b. How many sequences were aligned during the analysis of your protein? Provide evidence for this number. (2 points)

YASH: During the multiple sequence alignment, 50 sequences with the lowest E-value were aligned. There were 501 PSI-BLAST hits, 443 of which are unique sequences, but only the ones with the lowest E-value were aligned.

The MSA for Hsp90 (2cg9, using the A chain), which was performed using CLUSTALW, is attached in the text file pdb2cg9.

c. Look at the results using the link to FirstGlance in Jmol. You can take screen shots of the results using command-shift-4 (Mac) or (PC). Turn in some annotated pictures of your conserved surface that depict the functionally important regions of your protein. (3 points)

YASH:
Figure 1 - Side view of Hsp90-nucleotide-p23/Sba1 monomer
Figure 2 – Side view of Hsp90-nucleotide-p23/Sba1 monomer
Figure 3 – Inside view of Hsp90-nucleotide-p23/Sba1 monomer
d. Which regions of the protein are most conserved? Which are least? Does this make sense based on the function of the protein? Justify why or why not. (3 points)

DAVID: The most conserved region of the protein is the helix-turn-helix region. This is not surprising in that this is a transcription factor and if it is unable to bind DNA, then this function has been severely inhibited. What was surprising, however, is that the reactive cysteine (C22) of the protein and the residues it interacted with (C127) were not highly conserved. The conservation score for C22 was 7 (out of 9) and the C127 was 2 out of 9. Such a low rating for 1 of the 2 available cysteines in the entire protein suggests that disulfide bond formation may not be a conserved mechanism in the homologues included in this query. The regions of the protein that are in contact in the dimer are also conserved, with conservation scores most often between 6 and 7 (out of 10). This is logical in that this interaction interface must be conserved to allow these regions of the
protein to facilitate dimerization. The least conserved portions of the protein are those that do not interact with either the DNA opposite unit of the dimer. This is logical in that these regions do not need to conserve a very strict chemistry of geometry in order to serve as a “linker” between the two DNA binding and interaction domain of a individual chain.

4. You have the opportunity to design a novel synthetic signaling pathway that would transmit the signal that a cell should swell in response to an increase in temperature.
   a. Which type of signaling pathway would you use and why would you use that pathway rather than any of the other signaling pathways/cascades we have discussed in class or read about in the book? (5 points)

   DAVID: We have discussed several pathways/cascades that lead to rapid activation (proteolytic cascade, kinase-kinase cascades) as well as rapid inactivation (ubiquitination, proteolysis). In designing this synthetic pathway, though, I want to incorporate both rapid activation and inactivation.

   Because the stimulus is an increase in temperature, my pathway would incorporate a protein that undergoes a conformational change at elevated temperatures. This protein, perhaps derived from a screen of temperature sensitive mutants, would bind to a second, catalytically-active protein to effectively block the active site and inhibit the catalytic protein from either activating downstream proteins that facilitate swelling. A recent Nature publication by Horsefield et al. (2006) discusses regulation of spinach aquaporin (transmembrane water channel). The study showed that this particular aquaporin is only open when two conserved serine residues are phosphorylated. Identification of the kinase responsible for this phosphorylation event and the subsequent identification/engineering of a temperature sensitive competitive inhibitor of this kinase would permit engineering of the following pathway. Once the temperature is reduced, the competitive inhibitor would be able to bind the kinase and prevent additional phosphorylation and opening of water channels.

   TsInhibitor:Kinase → heat → TsInHIBitor + kinase → phosphorylation of serines → opening of aquaporin.

   b. Contrast your pathway with the three next most probable pathways that you could select. (5 points)

   MAX: One possibility would be to design/alter a receptor that changes conformation at an elevated temperature, so that the intracellular domain acts to phosphorylate the proteins that activate the appropriate active transport channels and close passive transport channels. The disadvantage here is that it would be difficult to design such a conformation change, and that the receptor must be modified additionally so that the activity ceases when the membrane reaches a tension limit where it is about to burst. This pathway would act quickly by amplifying the signal from one receptor to a larger number of target proteins, and it would allow the cell to swell to a certain limit without bursting. However, it would be difficult to design such a complex receptor.
Another alternative would be to design a receptor that, when activated by an elevated
temperature, activates proteins that mark the appropriate channels for degradation. The
designed surface receptor could changes its intracellular domain conformation in response to an
elevated temperature. This pathway would be fast, but it would be wasteful because the channels
would have to be re-made, and it may interfere with other processes by degrading the wrong
proteins.

A third possibility would be to alter the existing protein channel transcription regulators
so that their conformations are temperature dependent. This technique could involve altering
multiple regulators (for passive and active transport), so it could become a complex problem.
Furthermore, the regulators would not be able to cause swelling to occur very quickly, since the
signal would be communicated at the transcriptional level only.

5. Kinetic data for a protease at neutral pH at 25°C are shown below. The enzyme
concentration is 50 nM. The conversion factor from relative units to μM/ second is . The
half-life of the uncatalyzed reaction at neutral pH at 25°C is 500 years.

a. Calculate $K_m$ and $k_{cat}$ for this enzyme. Show your work and please be sure that
your method of calculating these kinetic parameters is clear. (10 points)
b. What parameter do we use to determine how “good” an enzyme is? (2 points)
c. How good of an enzyme is this? (3 points)

<table>
<thead>
<tr>
<th>Substrate Concentration (micromolar, μM)</th>
<th>Initial Velocity (relative units/second)</th>
</tr>
</thead>
<tbody>
<tr>
<td>180</td>
<td>4.0</td>
</tr>
<tr>
<td>140</td>
<td>4.0</td>
</tr>
<tr>
<td>100</td>
<td>3.8</td>
</tr>
<tr>
<td>60</td>
<td>3.3</td>
</tr>
<tr>
<td>40</td>
<td>2.6</td>
</tr>
<tr>
<td>20</td>
<td>1.8</td>
</tr>
<tr>
<td>10</td>
<td>1.3</td>
</tr>
<tr>
<td>5</td>
<td>0.9</td>
</tr>
<tr>
<td>0</td>
<td>0.1</td>
</tr>
</tbody>
</table>

MICHAEL:

<table>
<thead>
<tr>
<th>[substr] (μM)</th>
<th>Int vel (rel unit/s)</th>
<th>int rate (μMs⁻¹)</th>
<th>1/[substr] (μM⁻¹)</th>
<th>1/(int vel) (s/rel unit)</th>
</tr>
</thead>
<tbody>
<tr>
<td>180</td>
<td>4.0</td>
<td>17.400</td>
<td>0.0056</td>
<td>0.0575</td>
</tr>
<tr>
<td>140</td>
<td>4.0</td>
<td>17.400</td>
<td>0.0071</td>
<td>0.0575</td>
</tr>
<tr>
<td>100</td>
<td>3.8</td>
<td>16.530</td>
<td>0.0100</td>
<td>0.0605</td>
</tr>
<tr>
<td>60</td>
<td>3.3</td>
<td>14.355</td>
<td>0.0167</td>
<td>0.0697</td>
</tr>
<tr>
<td>40</td>
<td>2.6</td>
<td>11.310</td>
<td>0.0200</td>
<td>0.0884</td>
</tr>
<tr>
<td>20</td>
<td>1.8</td>
<td>7.830</td>
<td>0.0500</td>
<td>0.1277</td>
</tr>
<tr>
<td>10</td>
<td>1.3</td>
<td>5.655</td>
<td>0.1000</td>
<td>0.1768</td>
</tr>
<tr>
<td>5</td>
<td>0.9</td>
<td>3.915</td>
<td>0.2000</td>
<td>0.2554</td>
</tr>
<tr>
<td>0</td>
<td>0.1</td>
<td>0.435</td>
<td>∞</td>
<td>2.2989</td>
</tr>
</tbody>
</table>
Using the Lineweaver-Burk double reciprocal equation,

\[ \frac{1}{v_o} = \left( \frac{K_M}{V_{\max}} \right) \frac{1}{[S]} + \frac{1}{V_{\max}} \]

a plot of \(1/v_o\) vs \(1/[S]\) will give a slope of \((K_M/V_{\max})\), a y-intercept of \((1/V_{\max})\) and an x-intercept of \(-(1/K_M)\)

From the equation of the linear fit of the graph, \(1/V_{\max} = 0.0575\)
Therefore \(V_{\max} = 17.4 \mu\text{Ms}^{-1}\) which is in agreement with the data provided.
Slope = 1.0454 = 0.0575 \(\times K_M\)
Therefore \(K_M = 18.18 \mu\text{M}\)
To calculate the \(k_{\text{cat}}\), the equation below is used

\[ k_{\text{cat}} = \frac{V_{\max}}{[E]_T} \]

\([E]_T = 50 \text{nM}, \text{total enzyme concentration}\)

\[ k_{\text{cat}} = \frac{17.4 \mu\text{Ms}^{-1}}{0.05\mu\text{M}} = 348\text{s}^{-1} \]

a. The efficiency of an enzyme is determined by \(k_{\text{cat}}/K_M\) and \(k_{\text{cat}}/k_{\text{uncat}}\)

b. The efficiency of the protease in question as calculated from \(k_{\text{cat}}/K_M\) is \(348\text{ s}^{-1}\)

\[ 1/18.18 \mu\text{M} = 19.14 \mu\text{M}^{-1}\text{s}^{-1} = 1.914 \times 10^7 \text{M}^{-1}\text{s}^{-1} \]

To use the \(k_{\text{cat}}\)-to-\(k_{\text{uncat}}\) ratio, the \(k_{\text{uncat}}\) has to be calculated from the half-life. Assuming a pseudo first order reaction,
\[ \ln 2 = k_{\text{uncat}} \times t_{1/2}; \]

\[ t_{1/2} = 500 \times 365.25 \times 24 \times 60 \times 60 \text{ s} = 1.5779 \times 10^{10} \text{ s} \]

\[ k_{\text{uncat}} = 0.693 \div 1.5779 \times 10^{10} \text{ s} = 4.3929 \times 10^{10} \text{ s}^{-1} \]

\[ k_{\text{cat}}/k_{\text{uncat}} = 348 \text{ s}^{-1} / 348 \text{ s}^{-1} = 7.9219 \times 10^{12} \]

c. From the value of the \( k_{\text{cat}}/K_M \), the enzyme can be said to be a very good one. Enzyme-substrate collision is in the range of \( 10^8 \text{ - } 10^9 \text{ M}^{-1}\text{s}^{-1} \) and perfect enzymes have efficiencies in this range where every collision leads to a reaction. Thus, an efficiency of just one or two orders of magnitude less is good. Carbonic anhydrase and fumarase are two examples of enzymes that have this efficiency. Chymotrypsin and Urease have lower values. Based on the \( k_{\text{cat}}/k_{\text{uncat}} \), the enzyme is a good one as well, better than fumarase and chymotrypsin (Science 1995, 267: 90-93)

6. During Craig Martin’s lectures he gave us a non-protein centric view of the world, explaining how nucleic acids can perform many binding and catalytic functions. Why do you think are there no living organisms that entirely lack proteins? (5 points)

YASH: Ribozymes are RNA enzymes that catalyze specific reactions in the cell. If all the enzymes were to be ribozymes, the enzymes would have a short half life due to the instability of RNA as it is prone to hydrolysis in the cell. Moreover, even if we consider RNA to assume 3D various structures and provide structural stability to the cell, it will have a shorter half life again. This in turn would lead the cell to produce more and more RNA, which is not desirable as RNA production comes at a cost of ATP. However, this cost will be much lesser if RNA is translated into proteins and the proteins can provide structural stability as their turnover rate is low. A similar argument can be made for RNA acting as a transport, receptor or signaling molecule.
LAWRENCE:

Proteins perform different variety of include catalysis, binding, signaling and structural. This makes proteins very important components of cells. Nucleic acids were thought to store genetic information only. However, later discoveries indicated that RNAs performed some catalytic function (ribozymes) as well signaling and regulatory functions (riboswitches). Even though RNAs perform some of the functions of proteins, and so could probably replace them, there has been as yet no living organism without proteins known. These could be because of the following

a. The number and types of reactions that proteins catalyze are so enormous and diverse that nucleic acids alone cannot perform this function.

b. There are 20 natural amino acids and only four ‘traditional’ nucleic acid bases. This means that the number of proteins that can be made from various permutations and combinations far outnumber that possible for nucleic acids. With the increasing number of proteins comes increasing structural diversity which allows proteins for every specific function possible.

c. The 20 amino acids come with various chemical properties. This ensures that proteins are well suited to perform their functions.

d. Another reason might be the fact that a combination of 2 different macromolecules is much better in functioning than only one of the two parts. Thus all processes in the cell which require a great deal of expertise are performed by a combination of the two. E.g. translation is carried out by the ribosome which has both rRNA and proteins. The ribosome itself is a riboprotein. In the same way transcription and replication requires the two macromolecules to work hand in hand.

7. You have graduated with honors from UMass and have taken a job with the US Drug Enforcement Agency. Your agency is looking for a way to treat people who have overdosed on cocaine. Given your expertise from Chem 791A you proposed developing catalytic antibodies that could be administered to overdose victims for digestion of cocaine in the bloodstream.

   a. What will you use as the hapten for the reaction? What factors did you consider in selection of this hapten? (5 points)

   b. Do you think these antibodies will be successful? What kinds of catalytic rates do you think you will need in order for them to be successful and is that magnitude of rate enhancement feasible with catalytic antibodies? (5 points).

LAWRENCE:

a. Cocaine, an alkaloid, is extensively metabolized in the body into benzoylecgonine and ecgonine methyl ester. Other minor metabolites include norcocaine, p- and m-hydroxycocaine and p- and m-hydroxybenzylecgonine. *(Journal of Analytical Toxicology 30(8) 2006 501-510).* Most of the cocaine metabolites are non toxic. The toxic metabolites include norcocaine, N-hydroxy norcocaine, cocaethylene which is formed when consumed with alcohol and cocaine itself. *(J.B.S. 2(1), 2006 pp53-58)*

The designed hapten should mimic cocaine but not the non-toxic metabolites. Cocaine will be the most appropriate hapten. However cocaine is rapidly hydrolyzed with a short half-life in the body and so cannot be
used. The hapten to be used for the study should be stable and long-lived so as to induce the formation of an antibody. To this end a hapten in which the two esters present in cocaine (responsible for its rapid hydrolysis) are replaced by a phosphonate ester and a carboximido group has been designed.

The hapten is able to mimic the negatively charged, tetrahedral transition state of the ester hydrolysis of cocaine.

In addition to this, its stability is enhanced by the presence of both the phosphonate group and the carboximido group. (*The Journal of Pharmacology and Experimental Therapeutics: 293* (3) 2000 952-962)

A further modification of this hapten has yielded the best antibody (Mab 15A10) so far. A carrier protein which helps in the absorption of the hapten is attached. This will be the hapten that I will use.

\[
R^1 = (\text{CH}_2)\text{N}^{14}\text{CO(\text{CH}_2)}\text{N}^{14}\text{CONH-carrier protein; } R^2 = R^3 = \text{H}
\]

For these antibodies to be successful, they need to have very high catalytic rates to complement those of the cocaine esterases in reducing the lifespan of cocaine in the body. The $k_{\text{cat}}$ of a cocaine esterase isolated from the bacterium *Rhodococcus* sp. MB1 is 7.8s⁻¹ and with a $K_M$ of 640nM. *(Nat. Struct. Biol. 9(1) 2002:17-21.)* Its $\frac{k_{\text{cat}}}{k_M}$ is about 1.22 * $10^8$. An antibody with similar or better kinetics will therefore be ideal. However, the best antibody for cocaine isolated to has a $k_{\text{cat}} = 2.3$ min⁻¹ and $K_M = 220\mu$M. *(J. Am. Chem. Soc., 118 1996, 5881-5890).* The $\frac{k_{\text{cat}}}{k_M}$ of the best available catalytic antibodies are in the ranges of $10^2$ and $10^6$ M⁻¹s⁻¹, *(Annual Review of Biochemistry 69: 2000 751-793,)* about 4 to 6 orders of magnitude less than what will be sufficient and so this is not feasible, as yet.

8. James Chambers spoke to us about two- and three-photon excitation experiments. Please give an example (in general terms) of when a two-photon experiment would be beneficial when compared to a single photon experiment. (5 points)

MUSLUM:

9. Draw the following on a clean sheet of paper. You can use structures from books as trace-through guides, if that helps you to draw clean structures.

a. A very common base pair in RNA structures is formed using the Watson-Crick faces of G and U. Draw the most stable GU base pair. (5 points)

b. In the ribosome (and other riboproteins), amino acids from the protein interact with base structures in the RNA. Show, in your drawing, how the amino acid Gln can interact with your GU pair, interacting with functional groups from both bases. (5 points)

See below – the preferred answer (interacting with both bases) is shown in black; the secondary answer is shown in green.

![GU base pair](image)

![Gln interaction](image)

c) As you showed above, the GU pair has two good hydrogen bonds. An AU pair also has two good hydrogen bonds. Explain how an RNA polymerase can allow U to be incorporated opposite A, but not opposite G. (5 points)

The distance and angles between the bonds to the sugars (glycosidic) are very different in a GU pair versus an AU pair. RNA polymerase has a set distance between the sugar-binding sites.
10. What do you see as the most valuable asset to have come out of the structural genomics initiative? (5 points)

DAVID: I think that the most valuable asset to derive form the structural genomic initiative has not been a massive number of structures; as we discussed in class, the number of protein structures produced has not been as high as anticipated. Rather, I think that the greatest effect has been the technological developments that have resulted. These technologies include - but are certainly not limited to- reducing the volumes/costs of protein required for crystallization, production and increasing availability of increasing automated and robotic technology, and development and distribution of more user friendly computational programs to make personal data comprehension and community sharing more feasible. An additional asset to come out of these efforts and the availability of increased funding opportunities has been the assembly of teams (working together or at distant labs) simultaneously pursuing different aspects/portions of common projects.