For this homework, each of you will be assigned a particular human protein to model. I have picked out proteins of interest in cancer biology. I have colleagues at Fox Chase Cancer Center who work on some of these, and others of them are sequenced in gene panels given to our patients. We are trying to develop predictors of the phenotypes of missense mutations in some of these genes, and having better models would be really helpful. So this homework is like a group project to improve the interpretation of gene sequencing data for real patients. I hope it will be fun and interesting. If the results are publishable, we may consider submitting one or more of these to PLOS ONE for publication.

The goal of the homework is to use structure prediction servers to obtain models of the assigned protein and to interpret as much biology on these proteins as possible in light of the models. As I showed during the lectures and demonstration, each available server has strengths and weaknesses and greater or lesser flexibility in choosing templates and visualizing the results. In this assignment, I want you to use several different servers for structure prediction and to gather biological information from as many sources as possible.

Perform the following tasks and provide a write-up in the form of a short paper that might appear in a journal with an Abstract, Introduction, Methods, Results, and Discussion. As a guide, the lengths of each section should be at least: Abstract – 150 words, Introduction – 500 words, Methods – 250 words, Results – 1000 words, Discussion – 500 words. They can be longer or shorter as needed, but I want you to think about the modeling results, not just run some web servers and make some images. You should include figures to show interesting features of the models. Such features might include active sites, the positions of known disease mutations or even variants of unknown significance (VUS), the structures of homo- or heterooligomeric complexes or complexes with ligands, inhibitors, or nucleic acids.

Here are the steps you should go through and the things you should look at in the structure predictions.

1. Get the protein sequence from Uniprot. Learn something about the function of the protein, what it interacts with, is it an enzyme, are there disease-related mutations, have people made mutations experimentally and do these have a phenotype, are there different splice forms (isoforms), etc. There are links to protein-protein interaction databases that you can use to find additional protein interactors. You will need all this information for the introduction, results, and discussion sections.

2. Run at least three different disorder predictors on the sequence. There is a list here: http://www.disprot.org/predictors.php. IUpred, PONDR-FIT, Spritz, and DISOPRED are all good servers, but some of the others might be interesting. Are the predictions consistent? Are there long regions of disorder (>30 amino acids) and where are they? You can download the images of the disorder prediction that the servers provide and line them up by scaling the size of each image horizontally. Or (even better) you can download the raw data and put them into a graphing program (R, Matlab, even Excel), and put them into one graph. Make sure you understand what is the cutoff for something to be predicted as “disordered” (most of them are 0.5 on a 0 to 1.0 scale, but DISOPRED has a 0.05 cutoff. Here’s an example of human STK6.
3. Run the Pfam server to find the Pfam domains in your protein sequence: [http://pfam.xfam.org/search](http://pfam.xfam.org/search). Use the Pfam site to find out about each Pfam domain – what it does, what proteins the domain is found in. Present the results as a bar plot of the domains (like you see in most papers on proteins). Powerpoint would work. You should mark the residue numbers where each domain starts and ends (which are often missing in papers). You can add as many annotations as you can find on where the domains are, where the disorder is (from step 1), where any post-translational modification sites are, where disease-related mutations are, and where other molecules are known to interact. Something like this:

![Diagram of protein domains](image.png)

4. Look up the Pfams on ProtCID: [http://dunbrack2.fccc.edu/protcid](http://dunbrack2.fccc.edu/protcid) (under the browse button, look for the first letter of the Pfam and then under that link, look for the Pfam). It’s possible some Pfams from the Pfam website will not be on ProtCID, either because there really are no structures in that Pfam, or because ProtCID is based on Pfam version 27, while the website is on version 29. We are updating the site but it’s not done yet. When you look up a Pfam, you get a page of hits that describes what is in each structure that contains the Pfams – the Pfam architecture of the chain contains that has the Pfam (under “Chain Arch”), other chains in the structure including peptides (under “Entry Arch”), the biological assemblies of all the entries that contain the Pfam (e.g. PDBBA = A2; PISABA = A4 for homodimer and homotetramers; A2B2 is a heterotetramer of two different kinds of proteins). Then look at the ligands in the last column. If you click on each one, you will get a description of the ligand. Or you can click on the link at the top that says “Click for description of ligands.” Here is a snapshot:

![Snapshot of Pfam domains](image.png)

| PDB Entries for P53  
(HMM Length = 196) |
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Click here for the interaction network</td>
</tr>
<tr>
<td>Click here for descriptions of ligands</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PDBID</th>
<th>Entity ID</th>
<th>Chain Arch</th>
<th>E-Value</th>
<th>SeqBeg</th>
<th>SeqEnd</th>
<th>HmmBeg</th>
<th>HmmEnd</th>
<th>Entry Arch</th>
<th>UniP Codes</th>
<th>PDBBA</th>
<th>PISABA</th>
<th>Ligands</th>
</tr>
</thead>
<tbody>
<tr>
<td>3p01</td>
<td>1</td>
<td>(P53)</td>
<td>1.26E-109</td>
<td>1</td>
<td>196</td>
<td>1</td>
<td>196</td>
<td>(P53)</td>
<td>(P53, HUMAN)</td>
<td>A2</td>
<td>A2</td>
<td>(ZN)</td>
</tr>
<tr>
<td>3p05</td>
<td>1</td>
<td>(P53)</td>
<td>1.26E-109</td>
<td>1</td>
<td>196</td>
<td>1</td>
<td>195</td>
<td>(P53)</td>
<td>(P53, HUMAN)</td>
<td>A4</td>
<td>A</td>
<td>(DNA,ZN)</td>
</tr>
<tr>
<td>3d00</td>
<td>1</td>
<td>(P53)</td>
<td>1.26E-109</td>
<td>1</td>
<td>194</td>
<td>1</td>
<td>195</td>
<td>(P53)</td>
<td>(P53, HUMAN)</td>
<td>A4</td>
<td>A</td>
<td>(DNA,ZN)</td>
</tr>
<tr>
<td>3dd8</td>
<td>1</td>
<td>(P53)</td>
<td>1.26E-109</td>
<td>1</td>
<td>194</td>
<td>1</td>
<td>195</td>
<td>(P53)</td>
<td>(P53, HUMAN)</td>
<td>A4</td>
<td>A</td>
<td>(DNA,ZN)</td>
</tr>
<tr>
<td>4mar</td>
<td>1</td>
<td>(P53)</td>
<td>1.26E-109</td>
<td>1</td>
<td>194</td>
<td>1</td>
<td>195</td>
<td>(P53)</td>
<td>(P53, HUMAN)</td>
<td>A4</td>
<td>A</td>
<td>(DNA,ZN)</td>
</tr>
<tr>
<td>1gzh</td>
<td>1</td>
<td>(P53)</td>
<td>1.26E-109</td>
<td>1</td>
<td>194</td>
<td>1</td>
<td>195</td>
<td>(P53)</td>
<td>(P53, HUMAN)</td>
<td>AC,B,C2</td>
<td>A</td>
<td>(TI)</td>
</tr>
<tr>
<td>1gzh</td>
<td>1</td>
<td>(P53)</td>
<td>1.26E-109</td>
<td>1</td>
<td>194</td>
<td>1</td>
<td>195</td>
<td>(P53)</td>
<td>(P53, HUMAN)</td>
<td>AC,B,C2</td>
<td>A</td>
<td>(TI)</td>
</tr>
<tr>
<td>1b8h</td>
<td>1</td>
<td>(P53)</td>
<td>1.26E-109</td>
<td>1</td>
<td>194</td>
<td>1</td>
<td>195</td>
<td>(P53)</td>
<td>(P53, HUMAN)</td>
<td>AC,B,C2</td>
<td>A</td>
<td>(TI)</td>
</tr>
<tr>
<td>1b8h</td>
<td>1</td>
<td>(P53)</td>
<td>1.26E-109</td>
<td>1</td>
<td>194</td>
<td>1</td>
<td>195</td>
<td>(P53)</td>
<td>(P53, HUMAN)</td>
<td>AC,B,C2</td>
<td>A</td>
<td>(TI)</td>
</tr>
</tbody>
</table>
The E-value column will tell you whether the Pfam is a strong hit for that PDB sequence. Generally values smaller than 1.0e-5 are quite safe. You can also tell if the PDB sequence has the whole Pfam by looking at the HmmBeg and HmmEnd columns. The length of the HMM is given at the top of the page. From this analysis, you should write up some comments about what kinds of biological assemblies the Pfam makes, what type(s) of ligands it binds, and observe whether more than one of the Pfams in your query sequence is present in a single chain in the PDB, or potentially in heterodimers in the PDB. It occurs occasionally that multi-domain proteins can be modeled off of protein complexes in the PDB, where the domains of the target protein are actually in different chains in the template. What you learn from this page might be relevant when picking templates on the websites that allow you to choose templates, since it will be easy to superimpose the template onto the model so that you can add the ligands/peptides/nucleic acids/other proteins from the template to your models.

5. Submit your sequence to as many servers as you can, but at least 5. Give them your email address so you can access the results later. If they provide a web link to the results, include that in your report so I can look at it as well. Some of these will take a few days, so submit early so you have the results.

http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index
http://toolkit.tuebingen.mpg.de/hhpred (remember: to use Modeller, the password is MODELIRANJE).
http://zhanglab.ccmb.med.umich.edu/l-TASSER/
http://robetta.bakerlab.org/submit.jsp
http://www.reading.ac.uk/bioinf/IntFOLD/
http://swissmodel.expasy.org/
http://raptorx.uchicago.edu/StructurePrediction/predict/
http://sparks-lab.org/yueyang/server/SPARKS-X/
http://manaslu.aecom.yu.edu/M4T/

When you can choose your templates on a webserver, think about what to select. Templates that cover more of the sequence are good, as are ones that have the right ligands in them (e.g. DNA). Picking multiple templates (where the server gives you that option) that are different from each other is helpful. You can check their source organisms and sequences on the PDB’s website, if the prediction server does not provide that information.

6. Download the models and accompanying information provided by the servers, which may include alignments, secondary structure predictions, sequence identity, significance scores for the hits (e.g., E-values), and model quality assessment scores, if they are provided. MQA scores are predicted accuracies of the models – what would be expected at that sequence identity or sequence alignment score. Summarize the models you downloaded in a table – templates (PDB codes and protein names and species, resolution of the structures, etc.), sequence identity, significance scores, what regions of the target are covered, etc.

7. Align the model structures either within VMD or Pymol, or even better use a multiple structure alignment server like one of these:

POSA http://posa.godziklab.org/
Mustang http://pxgrid.med.monash.edu.au:8080/mustangserver/
Matt http://matt.cs.tufts.edu/
MultiProt http://bioinfo3d.cs.tau.ac.il/MultiProt/

After aligning the programs, visualize them in VMD or Chimera or Pymol. Where are there differences in the models, especially in the backbone coordinates? Are the models complete (with no internal gaps in the structure)? If the models are different in loops or the placement of secondary structure, is this due to the template that was used? (you can download the templates from the PDB and superimpose them on the models to see where the placement of loops or secondary structures might have come from).

8. If possible, try to model the protein in complex with some other molecule. Looking at the available templates, do they have ions, small molecules (whether inhibitors or ligands or substrates), peptides, nucleic acids, or other proteins? If the target interacts with other proteins (from the Uniprot page), is there a structure in the PDB
with Pfams from your target and one of the interactors (this can be a lot of work to determine, since you have to identify the Pfam of the interactors and then determine if there is a structure or template for that protein, and for the complex (which you can determine with ProtCID). But if possible, try to do this, since protein-protein and protein-nucleic acid interactions are the most likely way that the targets effect their functions. To model the protein in complex with a ligand or with DNA, find a template that has the ligand or DNA. Superimpose your model onto the template with the ligand or DNA. You may need to superpose the residues immediately adjacent to the ligand, rather then the whole structure. To make a model that accounts for the position of the ligand, you can remodel the side chains of the model with SCWRL4 as I showed in class.

9. Look for variants that are disease associated for the target. You can look on the Uniprot page but also in COSMIC and other cancer mutation databases. Some are locus specific databases (some of these pages list these):

- [http://www.hgmd.cf.ac.uk/docs/oth_mut.html](http://www.hgmd.cf.ac.uk/docs/oth_mut.html)
- [http://www.hgmd.cf.ac.uk/ac/index.php](http://www.hgmd.cf.ac.uk/ac/index.php)
- [http://grenada.lumc.nl/LSDB_list/lsdbs](http://grenada.lumc.nl/LSDB_list/lsdbs)
- [http://cancer.sanger.ac.uk/cosmic](http://cancer.sanger.ac.uk/cosmic)
- [http://docm.genome.wustl.edu/](http://docm.genome.wustl.edu/)
- [https://civic.genome.wustl.edu/#/home](https://civic.genome.wustl.edu/#/home)

If it is not necessarily the case that every mutation in a cancer is actually deleterious to the function of the gene. Some can be activating of the protein function, and some can have no effect at all; they are passenger mutations in a tumor that has mutations in other tumor-driving genes. If you can find mutations that are documented to be causative, that’s better. For a number of mutations (at most 10), find where they are in the model and how they might disrupt the structure. Are they buried? Are they in a ligand or nucleic-acid binding site? Are they in an interface with a likely protein partner (check ProtCID for these)?

10. If you’re brave, you can try docking the model of the protein model you built with some protein binding partner listed in the protein-protein interaction databases or Uniprot. You may need to model that protein as well, although an experimental structure would be better. Try the ClusPro server for this step. There are others.

- [https://cluspro.bu.edu/login.php](https://cluspro.bu.edu/login.php)

11. In the Discussion, try to summarize what can be learned from the structure in terms of function (usually binding other molecules) and disease mutations, and any other biological aspects you can discern. Submit a Word document (or pdf) of your results. Please also send the coordinates of the models (or Pymol or Chimera sessions). You can zip or tar up all the files and send to roland.dunbrack@gmail.com.
1. Target: ATM_HUMAN

Hint: the template is mTOR. There is an EM structure of mTOR, PDB entry 5FLC, that may appear as a possible template. But it is very low resolution; use one of the X-ray structures. Once you model ATM, the EM structure might be informative as to how ATM might interact with other domains within itself (the ARM/HEAT repeats that occur before where the template aligns). There are other proteins in 5FLC and parts of mTOR, but for much of the structure the sequence of the protein in the crystal is not aligned to the coordinates (they are all labeled “UNK” for “unknown”).

It would be good to model an inhibitor of mTOR into the ATM active site (perhaps some molecules inhibit both but this would be hard to determine).

The sequence may be too long for some servers and therefore it would be a good idea to break the protein into domains (according to the Pfams that are in the target and in each template). The first 700 residues or so are probably Armadillo/HEAT repeats, and it would be useful to see if I-TASSER can fold these up.

https://en.wikipedia.org/wiki/Armadillo_repeat
2. Target: BRCA2_HUMAN
There is a structure of rat BRCA2 Ob domains, and you should make models based on the available templates, which should be straightforward. The templates include the protein DSS1 (1MJE and 1IYJ) and DNA (1MJE); retrieve the human sequence of this protein and make a model of DSS1 from the same template and a model of the complex of BRCA2, DSS1, and DNA.

Another interesting structure is PDB entry 1N0W, which contains a fragment of BRCA2 bound to RAD51. This is one of the BRC repeats of BRCA2. It would be worthwhile to model the other BRC repeats bound to RAD51 (they all bind RAD51). Here is the alignment:

<table>
<thead>
<tr>
<th>Segment</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRC1_1002_1036</td>
<td>NHSFGGSFRT ASNKEIKLSE HNIKKSMFFK DIEE</td>
</tr>
<tr>
<td>BRC2_1212_1246</td>
<td>NEVGFRFYS AHGTKLNVST EALQKAVKLFS DIEN</td>
</tr>
<tr>
<td>BRC3_1421_1455</td>
<td>FETSDFPPQT ASGKNISVAK ESFNKVNFDD QKPE</td>
</tr>
<tr>
<td>BRC4_1517_1551</td>
<td>KEPTLLGFHT ASGKKVIAK ESLDVKNLFD EKEQ PDB 1N0W</td>
</tr>
<tr>
<td>BRC5_1664_1698</td>
<td>IENASALAFYT SCSRTSVSQ TSLLEAKWLR EGIS</td>
</tr>
<tr>
<td>BRC6_1837_1871</td>
<td>FEVGPPAFRI ASGGICVSH ETIKKVDFIT DSFS</td>
</tr>
<tr>
<td>BRC7_1971_2005</td>
<td>SANTCGIFST ASGKVQVSD ASLQNARQVFS EIED</td>
</tr>
<tr>
<td>BRC8_2051_2085</td>
<td>NSSAFSGFST ASGKVQVSILE SSLHKVKGVE EFDL</td>
</tr>
</tbody>
</table>

If you can model one of these segments based on 1N0W, check to see if there are mutations in any of these residues in the mutation databases.
3. Target: PMS2_HUMAN and MLH1 heterodimer C-terminal domains
C-terminal domain of PMS2_HUMAN is the main target. There is an experimental structure of the N-terminal domain (residues 1-364) and predicted disorder from ~380 to ~580. Templates align to the C-terminal domain from residue ~630 to the end. Ab initio servers like I-TASSER and Robetta might be able to fold some of this.

PMS2 heterodimerizes with MLH1. The C-terminal domain of PMS2 is (distantly) homologous to the C-terminal domain of MLH1. There are structures of a homodimer of MLH1’s C-terminal domain (PDB 3RBN) and a heterodimer of yeast PMS1 and yeast MLH1 (PDB: 4E4W).

The experimental structure of the C-terminal domain of MLH1 (PDB 3RBN residues 486-751) is missing some coordinates. Submit it to some of the servers, especially Robetta and I-TASSER to see if they will complete the coordinates (residues 486-495, 503-505, 718-728 are missing) of MLH1 by filling in the missing residues.

Make a heterodimer model of human MLH1 and human PMS2 by superimposing the experimental structure of human MLH1 C-terminal domain (PDB 3RBN) onto yeast MLH1 and the model of human PMS2 onto the experimental structure of yeast MLH1.
4. Target: MLH1_HUMAN and PMS2 heterodimer N-terminal domains

There are structures of both of these proteins (residues 3-336 of MLH1 and 1-365 of PMS2), but both are missing coordinates due to disorder in the proteins. Model both domains with the servers.

The biological assembly of the MLH1 structures is a homodimer. The PMS2 N-terminal domain is homologous to the N-terminal domain of MLH1. So the MLH1 homodimer can be used to build a model of the heterodimer of the N-terminal domains of MLH1 and PMS2. Make the heterodimer by superimposing the complete models. Are there clashes of the modeled loops? Remodel the side chains with SCWRL4.

Here is a schematic of the complexes.
5. Target: PMS1_HUMAN and MLH1 heterodimers
PMS1 and MLH1 form heterodimers of their N-terminal and C-terminal domains. There is no structure of the N-terminal domain of PMS1 (residues 1-331), nor of the C-terminal domain (residues 723-932) and so these are the targets for modeling by the servers.

There is a structure of the N-terminal domain of MLH1 (PDB 4P7A) and the C-terminal domain (PDB 3RBN) but they are missing some coordinates, so they can also be submitted to the servers to make complete models.

A heterodimer of the N-terminal domains of PMS1 and MLH1 can be made by superimposing the two models onto the MLH1 homodimer in PDB entry 4P7A. Similarly, a heterodimer of the C-terminal domains can be made by aligning the experimental structure of human MLH1 C-terminal domain (PDB 3RBN) onto the yeast MLH1 in PDB 4FMN, and the model of the human C-terminal domain of PMS1 onto yeast PMS1 in PDB 4FMN.

Here is a schematic of the complexes.

![Diagram of complexes](image)
6. Target: MLH3_HUMAN and MLH1 heterodimers

The N and C terminal domains of MLH3 forms heterodimers with the N and C-terminal domains of MLH1. MLH3 is a homologue of PMS2 and PMS1 which also form heterodimers with MLH1. There is no structure of MLH3 so the goal is to model both the N and C-terminal domains of MLH3 (residues 1-351 for the N-terminal domain; 1175-1410 for the C-terminal domain).

There is a homodimer structure of MLH1 (PDB 4P7A) that can be used to model the heterodimer of MLH3 and MLH1, since they are (distantly) homologous to each other. There is a heterodimer structure of the C-terminal domains of MLH1 and yeast PMS1 (PDB 4FMN). A model of the C-terminal domain heterodimer complex can be made by superimposing the experimental structure of the C-terminal domain of human MLH1 (PDB 3RBN) onto yeast MLH1 in PDB 4FMN, and the C-terminal domain model of human MLH3 onto the yeast PMS1 in PDB 4FMN.

There are coordinates missing in the experimental structures of MLH1 N and C terminal domains. The servers can be used to build models that fill in these loops pretty easily, prior to making the heterodimer models.

Here is a schematic of the complexes.
7. Target: RAD50_HUMAN

This protein is part of the MRN complex along with MRE11 and nibrin (also called NBS1). There are at least two domains that can be modeled (possibly more if HHpred can find more remote templates than PSI-BLAST) – at the N-terminal 150 amino acids and from 1094-1307 (at least). The template in PDB 5DA9 appears to have both parts but strangely in reverse order from RAD50_HUMAN. It is also a homodimer, so a homodimer of RAD50 can be modeled. Other templates are also available.

To model an interacting protein, get the sequence of human MRE11. There is a small portion of MRE11 (residues ~426-530) that is related to chains C and D in PDB 5DA9. It should be possible to model a heterotetramer of MRE11 and RAD50.
8. Target: ATR_HUMAN

Hint: the template is mTOR. There is an EM structure of mTOR, PDB entry 5FLC, that may appear as a possible template. But it is very low resolution; use one of the X-ray structures. Once you model ATR, the EM structure might be informative as to how ATR might interact with other domains within itself (the ARM/HEAT repeats that occur before where the template aligns). There are other proteins in 5FLC and parts of mTOR, but for much of the structure the sequence of the protein in the crystal is not aligned to the coordinates (they are all labeled “UNK” for “unknown”).

It would be good to model an inhibitor of mTOR into the ATR active site (perhaps some molecules inhibit both but this would be hard to determine).

The sequence may be too long for some servers and therefore it would be a good idea to break the protein into domains (according to the Pfams that are in the target and in each template). The first 1200 residues or so are probably Armadillo/HEAT repeats, and it would be useful to see if I-TASSER can fold these up. https://en.wikipedia.org/wiki/Armadillo_repeat
9. Target: ERCC2_HUMAN

We modeled ERCC3 in class. But ERCC2 is also important in nucleotide excision repair. The first template that will come up is the EM structure, 5FMF, which is too low resolution for good modeling. Try some of the other templates. One of the available templates is bound to DNA, so make a model of ERCC2 bound to DNA.