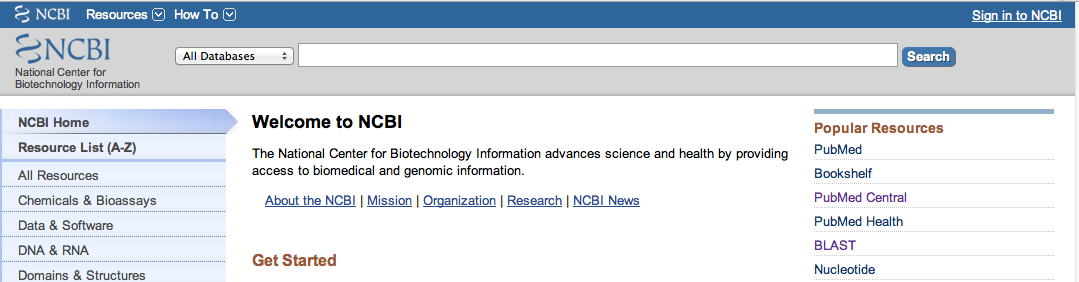
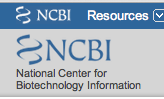
***Introduction****: This handout, along with the videos and links provided on the class website, is designed to guide you through the basics of finding and examining nucleotide and amino acid sequences and their conserved structural and functional domains. Then you will learn how to align the protein sequences to find important amino acids. You will learn how to render/model a structure to examine the structure, function and interaction interface of any known protein(s)****.*** ***Following completion of this handout, you will be able to finish the Bioinformatics and Bioinformatics homework (HW4) assignment linked on the class webpage****.*

**Step 1: Basic introduction to find the function of a gene or gene product using NCBI (~1 hr)**

First, a simple introduction is required. Go to the NCBI page at [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov) This is the National Center for Biotechnology Information hosted by the National Library of Medicine and the National Institutes of Health. Take a moment or two and click on different options, including PubMed, BLAST, Nucleotide, Protein, and Domains & Structures. Notice under the Resources List a “Training and Tutorials” tab. There are a number of tutorials to advance your knowledge if you find something of interest or are looking for additional help.

After you have seen some of the options and resources at the NCBI website, return to the NCBI home page by clicking on the NCBI icon. Now we will start to use NCBI to find the function of a gene or gene product. There are multiple ways to start. If you know the name of a gene, protein or nucleotide accession number or even part or all of a nucleotide or protein sequence, you can find information about the gene and/or protein and find the related accession numbers. Each time a nucleotide or protein sequence is entered into one of the databases, the gene/protein is given an accession number. Nucleotide accession numbers have 1 letter + 5 numerals OR 2 letters + 6 numerals. Nucleotide sequences determined from mRNA (coding region) will be noted as the mRNA accession numbers and begin with two letters (ie. NM\_001282404). Gene sequences will also be listed with a Gene ID number. Protein accession numbers for the GenBank/NCBI database are 3 letters + 5 numerals. For older records, you will find a both a “version” and “gi” as part of the accession number. This is an older version of the nucleotide or protein accession number. If any change in the record of the gene or protein occurs, the version number is increased by one decimal and a new gi number is assigned. Gene and protein records can have multiple identification numbers associated with them. It is always best to keep that in mind when searching. Click here if you want to learn more information about the [accession, version and gi number](http://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/Format/exercises/qa_accession_vs_gi.html).

Let’s begin searching for information on your protein of choice. **START WITH A GENE OR PROTEIN NAME!**

1. From the NCBI home page, click on the Search pull-down menu to select the Gene database, type the Gene Name (malate dehydrogenase) in the text box and click Go. See [Gene Help](http://www.ncbi.nlm.nih.gov/books/NBK3841/) for tips searching Gene.
2. Locate a desired Gene record in the results and click the symbol to open the record.
3. Functional information will be located in the Summary, Bibliography, and General gene info sections. Also, see the Links list for resources such as Conserved Domains and BioSystems.
4. Repeat using the Protein database in the pull-down menu.

**NOTE the difference between the identifier for a protein vs a DNA sequence.**

**Step 2: Searching for DNA vs protein sequences**

How does one find a DNA or protein sequence?

* Go to PubMed and search for malate dehydrogenase. Look at the links – what kind are they? Think about how many papers you would have to read to find the sequence.
* Go to the pulldown window next to the SEARCH box and find the Nucleotide database option. Enter malate dehydrogenase. Can you find it? Try searching for MDH, MDH1, MDH2 and MDH2 Homo sapiens. Sometimes it is hit and miss with an educated guess. Try the advanced search function. Does this limit your hits to more relevant entries? (Using a specific accession number found in a journal publication can help you narrow the search tremendously).
* Find the human MDH2 or the MDH2 gene and protein records. Are there different variants? If so, what are the differences*?*

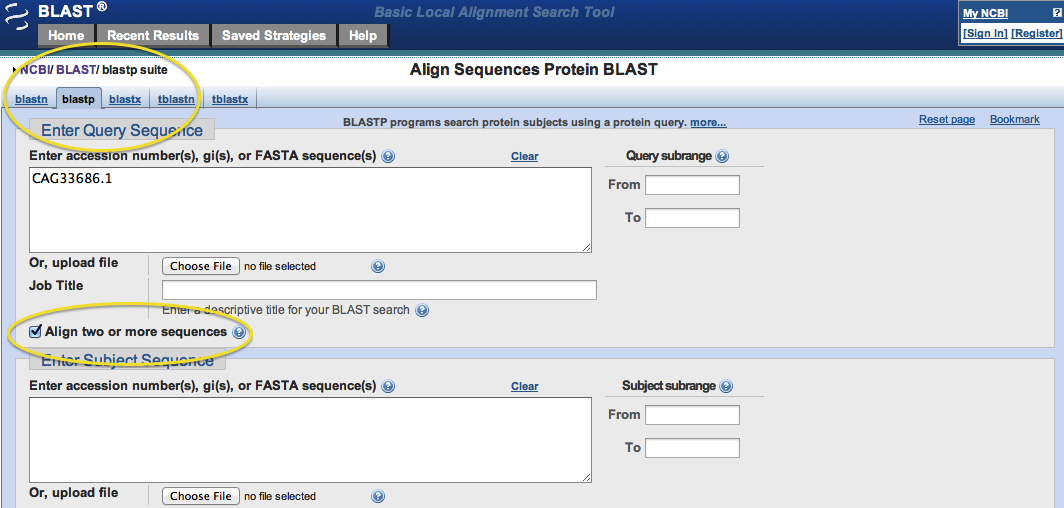
***Be careful when you click on the different links.***

*You may be taken to a different database such as “gene”, “nucleotide” or “protein”. The database for each page will be shown at the top and will have very different information.*

* In the nucleotide window, search for gi: NM\_005918.3
* Scan through the record. On the right of the window, you will find options to use the information with the record. Notice there are links to conserved domains, sequence features, and articles about the gene. Take a moment to look at those links. Continue to search down the record. You will see information about the size of the gene and/or mRNA, the location of the protein, if it is a variant or not, etc. Scanning further you will find several PubMed references. Read the titles and click on one or two of the links.
* Find and record the **gi. Number** (this is important for BLAST searches)
* Scan down to find and record the **CDS** number – The CDS is the “coding DNA sequence”. This is where the DNA sequence start site is. – A common mistake is to assume that the first nucleotide is the first amino acid codon.
* Continue down to find both the nucleotide sequence and the amino acid translation.
* At the top of the record, click on the FASTA link. This is a format used to compare sequences.
* Click on the Graphics link. Search for hyperlinked sections. Notice the NAD binding site, dimerization interface, and so on. Here you can find the graphical representation of a number of important domains of the protein/gene.
* Search using the nucleotide accession number under the protein database pulldown menu. What happens? Does the window stay in the protein database? Now try searching the protein database using NP\_005909.2 What is the difference? Is this the same protein sequence as before?

**Step 3: Aligning amino acid sequences**

This is useful when you need to compare two sequences: say against a protein you have that doesn’t have a solved structure, or you read a paper that indicates amino acid 212 is important for something. How would you directly line the sequences up? You could try to do it manually, but that would not be fun. Instead, you can use computational alignment tools to examine similarities among nucleotide or protein sequences.



The Basic Local Alignment Search Tool (BLAST) is a program on the NCBI website that will compare two sequences and match these sequences based on best matches. Very few if any sequences are the same, have the same length, or will start with the same sequence. Therefore BLAST helps you align two sequences to find where amino acids or nucleotides are common or unique. *You can also search* a known sequence (protein or nucleotide) against a short region – say a series of amino acids, by typing or copy/pasting the sequence into the search.

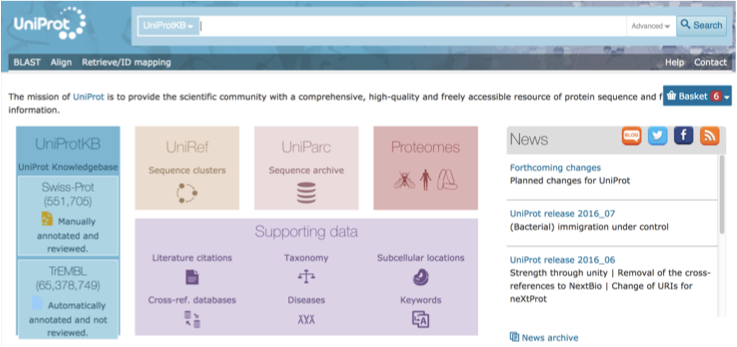
[BLAST](http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastHome) has different versions for nucleotides and proteins. You must use the correct sequence accession number – a common mistake is using a nucleotide accession number in a protein BLAST search. Click on the BLAST home page and familiarize yourself with the nucleotide and protein BLAST windows.

Click on the NCBI icon in the top left corner, and use the pull down menu to select the Protein: Sequence Database. Then type in human MDH1 and find the *Homo Sapiens* record (right hand side of the screen).

* You will find the “Run BLAST” link to the right. Click on it. Notice your protein record has already been loaded into the BLAST protein search. Convince yourself that you are in the protein BLAST by looking at the tab. It should read “blastp”.
* Click on the “Align two or more sequences” box and enter the gi, FASTA, or other protein accession number for MDH2.
* Find which amino acid in MDH1 corresponds to residue 58 in MDH2.
* Do the same thing using two nucleotide sequences. Pick any two MDH sequences. You will need to make certain you change the program settings from blastp to blastn (n = nucleotide, p = protein) in the pull-down window.
* Finally, go to the protein record number and find out which triplet bases code for the MDH1 amino acid you queried. Don’t forget that the first nucleotide is not the start site. See above for the CDS number! This will be important when comparing critical amino acids and domains between different isoforms of MDH.

***FYI -*** A second tool for examining sequence alignments of more than three sequences is Clustal Omega. Multiple sequence alignments (MSAs) are used to examine conservation of an amino acid or nucleotide at a position across several sequences of the same gene or protein. Correlation between conservation and function has been observed suggesting an evolutionary pressure for the organism to preserve a defined residue or sequence in order to maintain function.

**Now that you have a question that forms the basis for the sequences** you will align, where can you find these sequences? Several places – NCBI as outlined above but two other databases can serve as resources: [OrthoDB](https://www.orthodb.org/) and [UniProt](https://www.uniprot.org/). Choose UniProt to complete this exercise. Find the link on the class webpage for the UniProt Clustal Alignment turorial and watch before proceeding.



Search for human cytosolic malate dehydrogenase. There are more than two human cytosolic MDH results. What is the difference between the two? Click on the two human MDH genes and conduct an alignment (look for the alignment tab). Is there a difference? Next, select the filter by “Reviewed” on the top left. Can you now see the difference? If not, in NCBI search using the gene name for each of the two records… What is the ME1 protein that comes up after filtering “Reviewed”? Is this MDH too?

The point is to carefully review the records to make sure you are always using the correct record and one that has been carefully reviewed. Many a long, lost frustrating hour has been lost researching the wrong record.

Using UniProt, browse down the human MDH1 and the MDH2 UniProt records. Review the information available to the researcher.

* Can you find sites of post-translation modifications?
* There are a few identified interactions with MDH1 and MDH2 posted using STRING or other database engines; check some of these links out.
* Are there known mutations of either MDH isoform? What can you learn about the splice variants of MDH1 and MDH2?

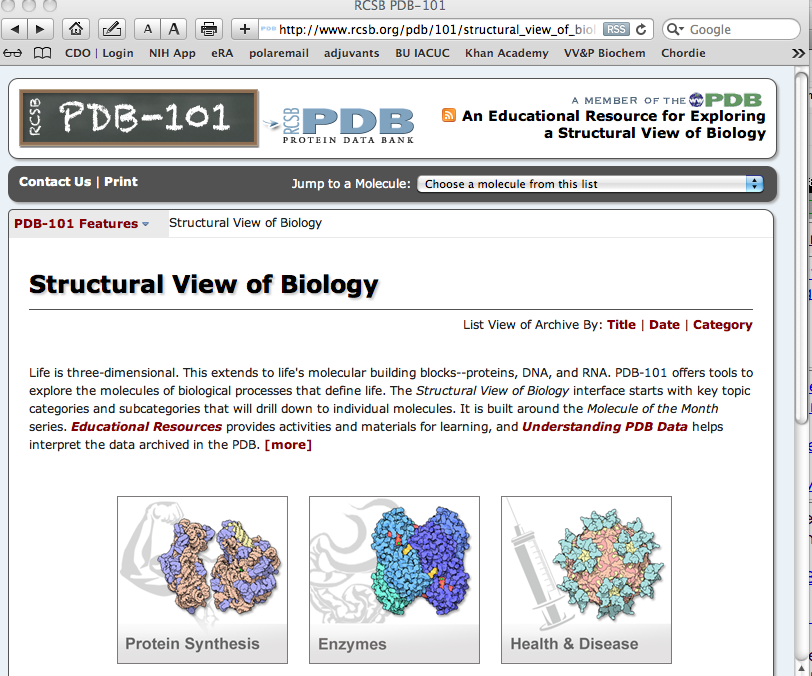
Use the alignment tool to align three of the MDH1 splice variants and research how these RNA splice varients differ from each other… If you aren’t sure, review (google) what an RNA splice varient is and how they are created.

* Align at least three of the human cytosolic MDH varients. How do they differ?
* Like other organellar proteins, mitochondrial proteins are synthesized with a “transit sequence”. The sequence is targeted by several proteins to shuttle the protein into the organell of choice where the sequence is usually removed. Thus it is important to understand if the mitochondrial MDH clone we will use has the transit sequence or not… Hint – it does not. Find the FASTA sequence in UniProt for human MDH2 with and without the transit sequence and conduct another clustal alignment. Notice how the numbering changes. FYI our sequence of human MDH is the canonical (find that info on the UniProt page) without the transit sequence BUT for synthesis in E.coli, the first amino in the hMDH2 sequence has the required MET codon… Remember this factoid – it will come back in a big way later.
* Finally align the longest splice varient of human cytosolic MDH with the cannonical MDH (without the transit sequence). Notice the regions of high and low homology. Use the options to identify the active site, sites of phosphorylation or substrate binding… See any interesting trends?
* Lastly – recall which is the amino (N-)terminus and the carboxy (C-) terminus in these sequences/alignments.

**Step 4: Structural Database, Rendering and Protein-Protein Interaction Analysis**

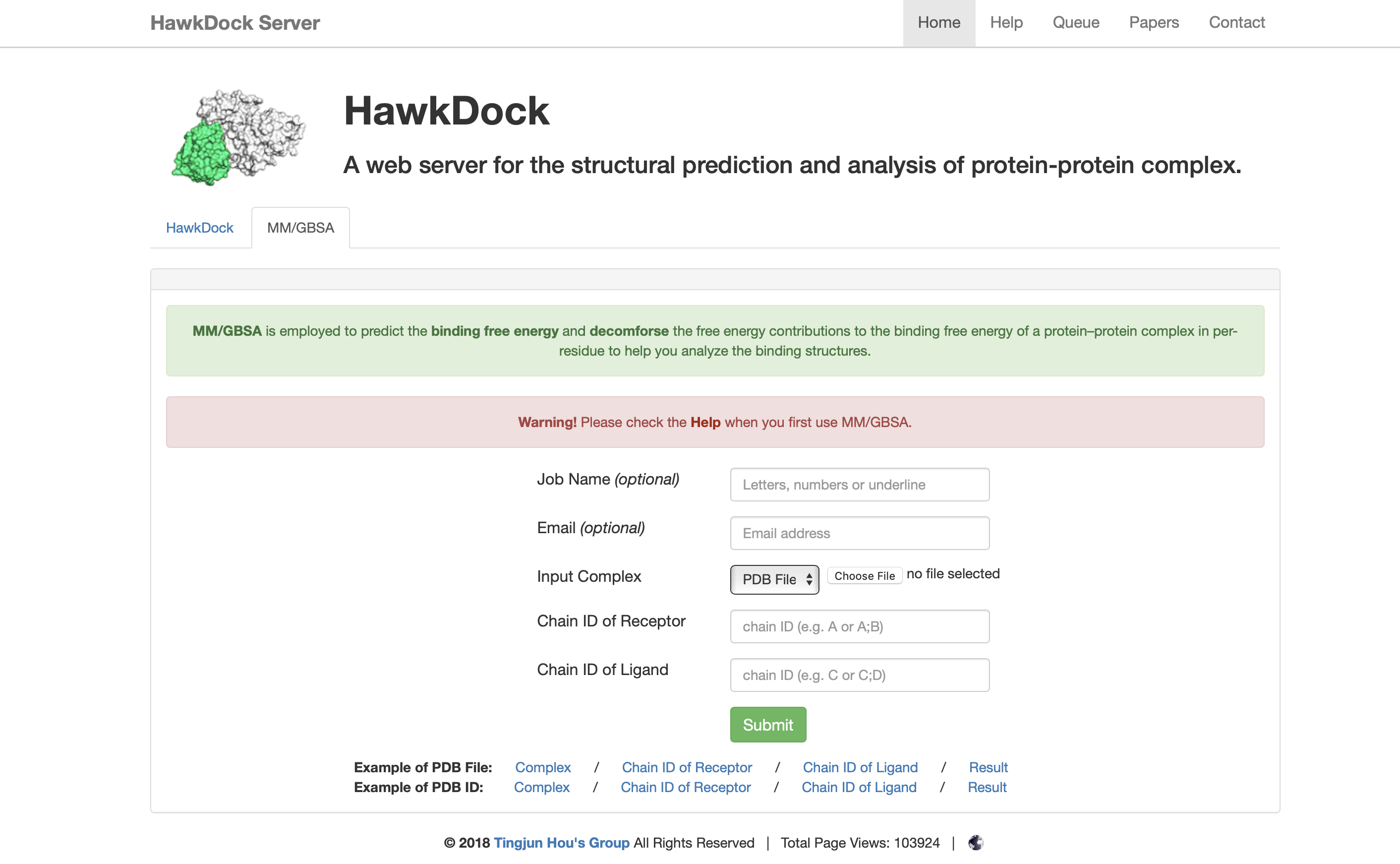
Next, you will be working with PyMOL. This program is used to display structures (called rendering), focus in on the amino acid sequences in specific regions, and align structures with each other to compare the shape of the protein. You can even make mutations to see what would happen in the structure and use it to create publication quality images of a protein structure or domain.

*To be prepared for lab, you should review the videos posted on the course website*. You do not need to prepare your notebook in advance, but feel free to record any questions you have and bring them with you to class. To install PyMOL on to your own computer to use, see the link provided on the class web page to complete your free registration for an educational use only copy.



1. Watch the RCSB PDB Tutorial linked on the class site to get a simple overview of the RCSB website. . Take a moment to see the kinds of information found in PDB 101 and RCSB.
2. Next, click on and review the PyMOL Wiki. Take a moment to find and bookmark the Practical PyMOL for Beginners link on the Tutorials page. You will want to come back to this page as you learn how to use PyMOL.
3. Watch each of the PyMOL movies linked on the course site (we will do this together in class). If you need additional help or are curious to learn more about using PyMOL, please review the links within the PyMOL Wiki.
4. The goals after watching all of these tutorials, are to:

* Determine the appropriate PDB file to use for rendering
* How to conduct the basic rendering tools of PyMOL (you can, if interested, look at other rendering programs freely availible – this is not required for this course).
* Overlay two similar sequences and highlight the important differences and similarities
* Zoom in and focus on key residues and ligands of a protein/nucleotide structure
* Appropriatly lable and create publiction quality images and movies of sequences.
* Show/hide key regions of a structure, find specific amino acid residues in the sequence and structure
* Measure distances between atoms of amino acids.
* Identify dimer interfaces and the interface of protein-protein interactions

**Hawk Dock** – in class, with your instructor, you will use one of the pre-docked MDH-CS PDB files to measure the free energy of the interaction and disect out the amino acids interacting in the interface by energy. Hawk Dock is one of several free software programs that will measure possible interactions between two proteins. This program predicts the “best fit” based on several fit requirements as described in the linked publication on Hawk Dock. Because these are predicted each program will give several docked models. Each slightly different. No one model is “right”. Each will have different binding energies. But the “lowest free energy” dock, does not translate to the “correct” dock between two proteins. These are only computational models, predictions, for a scientist to then test using mutational analysis. The more input into a docking program the more reliable the results will be. For instance, when hMDH2 and CS were crosslinked and they most likely Lys-Lys crosslinks are inputted into such programs, a “restricted” dock will be limited to include those known cross sites. Some times docking with and without these sites can help provide insight to the potenial likelyhood of the actual interface.

The MM/GBSA tab in the Hawk Dock website, uses docked protein PDB files to measure the binding free energy of a complex of two proteins and the individual residues from each chain. This can help you identify the potential importance of a amino acid interaction in the interface of two proteins. Click on the examples shown at the bottom of the page and see one of the online outputs. Notice it is important to idenify one protein as the ligand and the other as the receptor. For MDH-CS it doesn’t matter which is labled as receptor or ligand, just make sure you remember. Revew the publication on Hawk Dock to better understand the nature of the program. Then follow the instructions on the PPI link to enter your docked MDH-CS protein into the program.

**In lab:** Use the homework assignment to guide you through the steps to create your images!