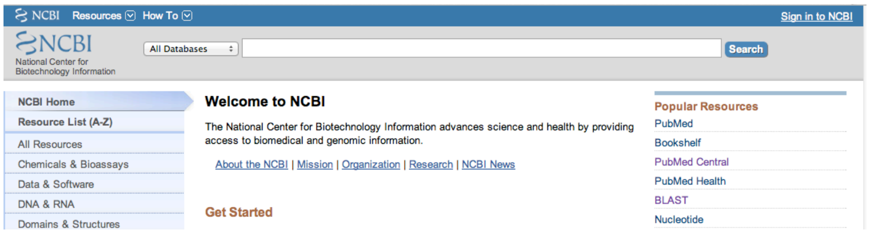
University of San Diego

Biochemistry Lab
Informatics and PyMOL Workshop

Introduction: *This handout along with the videos and links is designed to guide you through the basics of finding and examining amino acid and nucleotide sequences and their conserved structural and functional domains. Then you will learn how to align the protein sequences to find important amino acids. Finally you will learn how to render/model a structure to examine the structure and function of any known protein structure. Following completion of this handout, you will be able to finish the informatics and Pymol homework assignment linked on the class webpage.* **Note**: The movies that are linked on the class webpage were made by a TA at Univ MN Duluth. He created some of our lab and shared the movies with us.

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 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 "Tools" and "How to" there are a number of tutorials to advance your knowledge if you find something of interest or are looking for additional help.

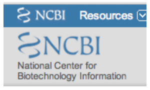


Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

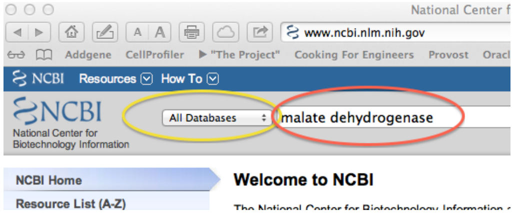
About the NCBI | Mission | Organization | Research | NCBI News

Get Started



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, the protein or nucleotide accession number or even part or all of a protein or nuclotide sequence, you can find information about the protein or gene and find the protein/nucleotide accession number. Each time a protein or nucleotide sequence is entered into one of the databases, the protein/gene is given an accession number. Protein accession numbers for the GenBank/NCBI database are 3 letters + 5 numerals (ex). Nucleotide accession numbers have 1 letter and 5 numerals OR 2 letters + 6 numerals (ex). Nucleotide sequences determined from mRNA (coding region) will be noted as the mRNA accession numbers and begin with two letters (NM_001282404). Gene sequences will also be listed with a Gene ID number. 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 protein or nucleotide 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. Protein and gene 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.

1. From the NCBI homepage, enter "malate dehydrogenase". Before clicking enter, make sure the menu selection is set for "All Databases". Look down the page. Take some time to click on a number of the features including gene, protein and nucleotide sequence. Look at the several helpful features within each header.



malate dehydrogenase

2. Return to 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 for tips searching Gene.

3. Locate the desired Gene record in the results and click the symbol to open the record.

4. 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.

5. Repeat using the Protein database in the pull-down menu.

6. Go to www.rcsb.org and type in "malate dehydrogenase" then under the molecule name, click on one of the possibilities. In one of the records, select "molecular description" and invest time looking at the features you can discover on the enzyme.

Step 2) Searching for protein vs DNA sequences (~2 hrs)

How does one find a DNA or protein sequence?

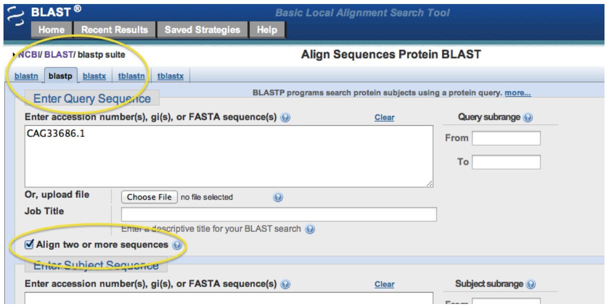
- Go to PubMed and search for malate dehydrogenase. Look at links – what kind are they? Think about how many papers you would have to read to find the sequence.
- Record the reaction catalyzed by the enzyme and one or two specific biochemical characteristics of the enzyme.
- Go to the pulldown window 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, but using the advanced search function may help. (Using a specific accession number found in a journal publication can help you narrow the search tremendously).
- Find the human 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 the **gi. Number** (this is important for BLAST searches)
- Scan down and find the **CDS** number – The CDS is the "coding DNA sequences". 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 aa translation and the nucleotide sequence.
- 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? Instead, search the protein database using NP_005909.2   What is the difference? Is this the same protein sequenced as before?

Step 3) Aligning aa sequences (~1 hr)

If you need to compare two sequences: say against a protein you have that doesn't have a structure, or you read a paper that indicates amino acid 212 was important for something. How would you directly line the sequences up? One way is to manually slug it out by hand. Not fun. Instead, you can take the amino acid (or nucleotide) sequences and align them using the BLAST search tool.



Align Sequences Protein BLAST

BLASTP programs search protein subjects using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s): CAG33686.1

Query subrange From To

Or, upload file: no file selected

Job Title: Enter a descriptive title for your BLAST search

☑ Align two or more sequences

Enter Subject Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Subject subrange

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.

BLAST 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 protein and nucleotide 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 aa 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. 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!

Step 4) Structural Data Base and Rendering (~3 hrs)

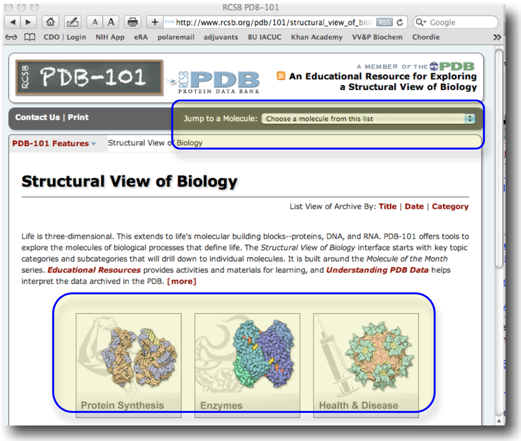
Next you will be working with PyMOL. This program is used to display structures (called rendering), focus in on the amino acid sequences in specific regions, and lay two structures on top of the other to compare the shape of the protein. You can even make mutations to see what would happen in the structure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

1) Watch the RCSB PDB Tutorial linked on the class webpage to get a simple overview of the RCSB website. Follow the link from the lab webpage. 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 class webpage. If you need additional help or are curious to learn more about using PyMOL, please review the links within the PyMOL Wiki.

4) Choose one of the MDH isoforms that is interesting to you. Ensure that the protein structure has a ligand, inhibitor, activator, substrate or some other molecule (protein, DNA, RNA, carbohydrate, lipid or small molecule) in the structure. This will be the protein you will work with for the following questions.



RCSB PDB-101

http://www.rcsb.org/pdb/101/structural_view_of_biol

RSS
Google
CDO | Login
NIH App
eRA
polaremail
adjuvants
BU IACUC
Khan Academy
VV&P Biochem
Chordie
RCSB PDB-101
RCSB PDB PROTEIN DATA BANK
A MEMBER OF THE PDB
An Educational Resource for Exploring a Structural View of Biology
Contact Us | Print
Jump to a Molecule: Choose a molecule from this list
PDB-101 Features
Structural View of Biology

Structural View of Biology

List View of Archive By: Title | Date | Category

Life is three-dimensional. This extends to life's molecular building blocks--proteins, DNA, and RNA. PDB-101 offers tools to explore the molecules of biological processes that define life. The *Structural View of Biology* interface starts with key topic categories and subcategories that will drill down to individual molecules. It is built around the *Molecule of the Month* series. ***Educational Resources*** provides activities and materials for learning, and ***Understanding PDB Data*** helps interpret the data archived in the PDB. **[more]**