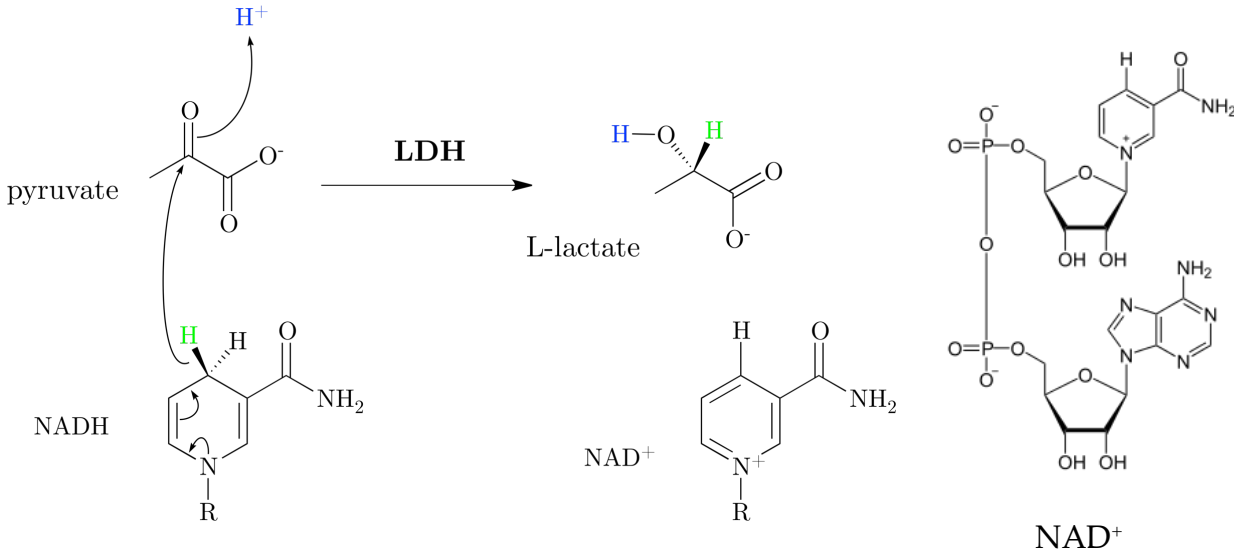
EXERCISE 7: SEARCHING FOR PDB FILES

In this lesson you will learn the following skill:
• How to locate the structure file for a given protein in the PDB.

Up till now, you have been given the PDB ID numbers of the structures to download for viewing in Jmol. What if you want to locate the PDB file for your favorite protein? How can you search for that file? If your search returns a long list of structure files, how will you know which one to use? In this exercise you will learn how to conduct a search and select the proper file for use with Jmol.

We will search for a structure file for the enzyme lactate dehydrogenase (LDH), which participates in the pathway of glycolysis. The enzyme catalyzes the reaction:

$$\text{pyruvate} + \text{NADH} + \text{H}^+ \rightarrow \text{L-lactate} + \text{NAD}^+$$



From: https://en.wikipedia.org/wiki/Lactate_dehydrogenase, https://en.wikipedia.org/wiki/Nicotinamide_adenine_dinucleotide

The keto group of pyruvate is reduced to a hydroxyl group. The hydrogen added during the reduction is removed from the nicotinamide ring of NADH, converting it to NAD^+. The complete structure of NAD^+ is shown at the right. The enzyme has binding sites for each of the substrates.

Go to the PDB web page. Instead of entering a PDB ID, type **lactate dehydrogenase**. Click on **Go**. As of January 25, 2021, this brings up a list of 170,093 structures where the words "lactate" and "dehydrogenase" appear somewhere in the PDB file.

To get a more restricted list, click on **Advanced Search** under the search window. Under **Attributed** click on the double down arrow. In the **Polymer Molecular Features** section click on **Macromolecular Name**. Make sure that the second pulldown menu is set to **contains phrase**, and type **lactate dehydrogenase** in the search window. Click on the search icon at the lower right. We now have a list of 196 structures. To see all the structures on one page, set the **Displaying___Results** pulldown menu (upper right) to **200**.

Let's refine the search and try to find the human enzyme. In the list of **Refinements** on the left, notice the category **SCIENTIFIC NAME OF SOURCE ORGANISM**. Click on **More...** to see an even longer list. Then click on **Homo sapiens (43)** and then the search icon to bring up the 43 human structures. To make things easier for now, on the horizontal bar above the results, click on **Compact**.

Of these 43 structures, two (1T2F and 1I0Z) are for the heart isozyme (H isozyme) of LDH, also called B LDH, and the other 41 are for the muscle isozyme (M isozyme), also called A LDH.

You can also use the **Score** pulldown menu to help find the file you are looking for. Selecting **Resolution: best to worst** will bring the structures with the best resolution to the top of the list.

You can use the **Experimental Method** column of the **Refinements** to restrict the list to structures determined by X-ray diffraction or those determined by NMR or by cryo-electron microscopy. Unfortunately you have to run the Advanced Search Query again to reveal all the choices. In this case one out of 196 structures was determined by cryo-electron microscopy rather than X-ray diffraction. Check the **ELECTRON MICROSCOPY** box and then the search icon to reveal this structure.

On the banner above the results, click on **Tabular Report** and then on **Ligand** in the pulldown menu. The files are listed with information about bound ligands, including the 3-letter abbreviation of each ligand (e.g. NAI is NADH).

Go back to the **Advanced Search Query Builder** at the top. Click **Add Field** and then the double down arrow. In the **Chemical Component** section click on **Chemical ID(s)**. Enter **NAI** in the search window and click on the search icon to see the structures that contain bound NADH.

Question 1: How many lactate dehydrogenase structures contain bound NAD^+? Bound NADH? Write a Jmol command to select the NADH. (Hint: The answer is not **select nadh**.)

Question 2: How many structures of mouse (*Mus musculus*) lactate dehydrogenase are available in the PDB?

Question 3: Find the highest-resolution structure for human enolase with bound 2-phosphoglyceric acid. State the PDB ID number, the resolution, the number of polypeptide chains, and the identity of the bound ligands.

Question 4: Locate a structure for rabbbit (*Oryctolagus cuniculus*) actin that comprises only actin rather than a complex of actin with one or more other polypeptides. Describe the structure (number of chains, ligands, experimental method, resolution).

EXERCISE 8: WHAT'S IN A FILE?

In this exercise you will examine the text of a PDB structure file and learn what is contained in it.

You will learn the following skill:
• How to display and read the header of a PDB file and how to get useful information that will help you to manipulate the image of the protein in Jmol and to understand what you are seeing.

We will examine a PDB file for the heart isozyme of human lactate dehydrogenase (LDH). The enzyme catalyzes the reaction

$$\text{pyruvate} + \text{NADH} + \text{H}^+ \rightarrow \square\text{lactate} + \text{NAD}^+$$

The reaction is shown on the first page of Exercise 7, including stuctures of all the substrates and products. In Exercise 7 we determined that there are two files for this enzyme, one of which is 1I0Z (the third character is a zero).

Search for the PDB page for 1I0Z. The information on the page is similar to what was noted in Exercises 1 and 2. There is an abstract of the journal article in which the structure was originally published, along with information about when the structure was deposited and later modified, as well as information about the X-ray crystallographic data and links to the original X-ray data. The **Macromolecules** section tells us that there are chains designated A and B.

The **Ligands** section tells us that we also have 1,4-dihydronicotinamide adenine dinucleotide (NADH) abbreviated NAI, and oxamic acid, an analog of pyruvate, abbreviated OXM. Oxamic acid is simply pyruvic acid with the methyl group replaced by an amino group. It is an inhibitor that occupies the site where pyruvate binds.

Let's now take a look at the actual PDB file. To the right of the PDB ID number click on **Display Files** and then **PDB Format**. We see a browser tab displaying the text of the file. This file contains all the information that Jmol and other programs need to locate every atom of a protein in space, to identify each amino acid residue and hetero group by name and number, to locate the various elements of secondary structure, and to distinguish the different polypeptide chains. It also contains information that may be important to you if you want to use this structure to explore the protein, including information about which are the more reliable and less reliable regions of the model. Let's take a closer look at the contents of the file:

The initial information provides the name and source of the enzyme, the submission history and the title of the paper where the structure was published. There is information about the crystal preparation and the X-ray diffraction results. There is a remark (REMARK 300) that "the biological assembly is a tetramer generated from the dimer in the asymmetric unit." We will explain what that means in Exercise 10.

REMARK 465 tells us that leucine 333 was not located in the experiment. This is the C-terminal residue. The C-terminal residue probably had some mobility within the crystals, causing it to lack a well-defined position. Occasionally residues from the middle of a polypeptide chain are also missing from the PDB model. It is important to know this when you are using the model.

There is information about atoms in close contact and dihedral angles that are outside the allowed regions of the Ramachandran plots. These may be areas where the protein is slightly strained or possibly areas where the model is less reliable.

Some abbreviations for the NADH and oxamic acid binding sites are defined.

The amino acid sequence is given.

Then the HET (hetero = non-amino acid) molecules are identified: note that there is one NADH and one OXM for each chain.

The locations of the helices andsheets are specified as well as the positions of three *cis* peptide bonds and the amino acids comprising the different named sites.

After this there is a list of every atom in the structure along with its x, y, and z coordinates, occupancy and B-factor (temperature factor). For example, atom number 26 is the gamma carbon atom of lysine 4 in the A chain. Its x, y, and z coordinates with respect to the origin are 11.808, 72.154, and 35.838. The occupancy is 1.00, as it is with all the atoms in this file, indicating that all the molecules in the crystal have exactly the same conformation. (In some structures the molecules are found in multiple conformations, and a given atom may have more than one position, each with a designated fraction of occupancy.)

The B-factor (temperature factor) provides information about how much the position of an atom varies about its specified position. A high value of the B-factor indicates either that the atom experiences a large degree of freedom of movement due to thermal motion, or that the protein is disordered at this position and that the atom occupies a range of different positions in different molecules within the crystal. The B-factor is equal to $8\pi^2$ (which is equal to 79) times the average value of the square of the displacement of the atom due to thermal vibration. So a B-factor of 79$Å^2$ means that the average displacement is 1 Å; a B-value of 30 indicates a displacement of 0.62 Å. In this

model the B-factors of the A chain are 30-56 $Å^2$ for the first 20 residues, the last 6 residues, and for a stretch of 17 residues (212-228) in the middle of the chain, and lower (as low as $7Å^2$) everywhere else.

The CONECT information at the end tells the program which atoms in the HET molecules are connected by bonds.

Question: Examine the PDB file 1TAQ. What protein is this? How many different polypeptide chains are present, and what is the length of each? Are any residues missing from the structure, and if so, which ones? Are there any *cis* peptide bonds, and if so, which one(s)?