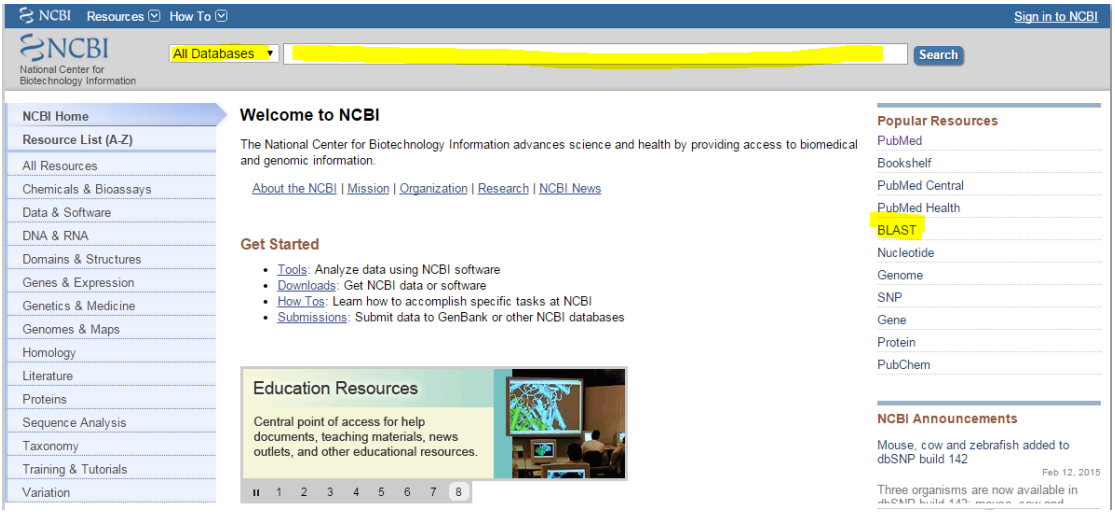
**Bioinformatics Introduction Worksheet** – The first part of this exercise is aimed at walking you through some of the key tools used by scientists to explore the relationship between genes and proteins throughout the Kingdoms of biology. The second part is an exercise that is aimed at helping you understand the DNA and protein that we're working with in lab this term. You'll need to prepare a document that addresses all questions found at the end of this document. It may help to use screenshots liberally.

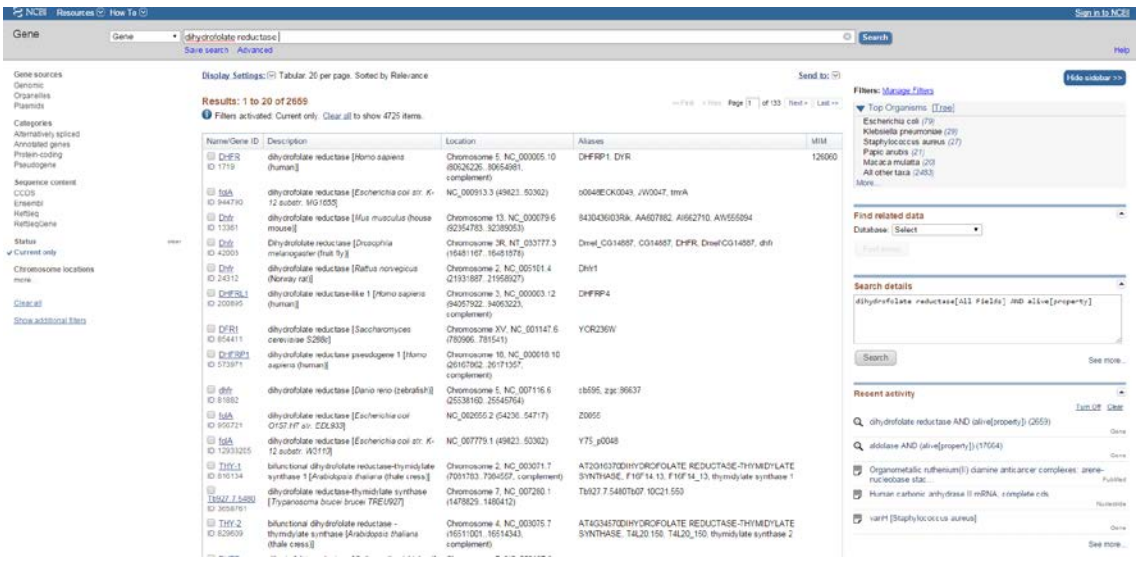
In this introductory exercise, image that you are a biochemist looking for information about the enzyme dihydrofolate reductase (DHFR). Your goal is to compare the enzymes from two different organisms: *Streptococcus pneumoniae* (a prokaryote) and *Saccharomyces cerve* (yeast).

Using the NCBI Interface:

1. Navigate to ncbi.nlm.nih.gov. This is an excellent tool that presents a very wide variety of other bioinformatics tools and databases in one convenient location. The tools that we will be using in this assignment are highlighted in the image below.




2. The dropdown box has a number of databases that are searchable. Select the "*gene"* option and search for dihydrofolate reductase.

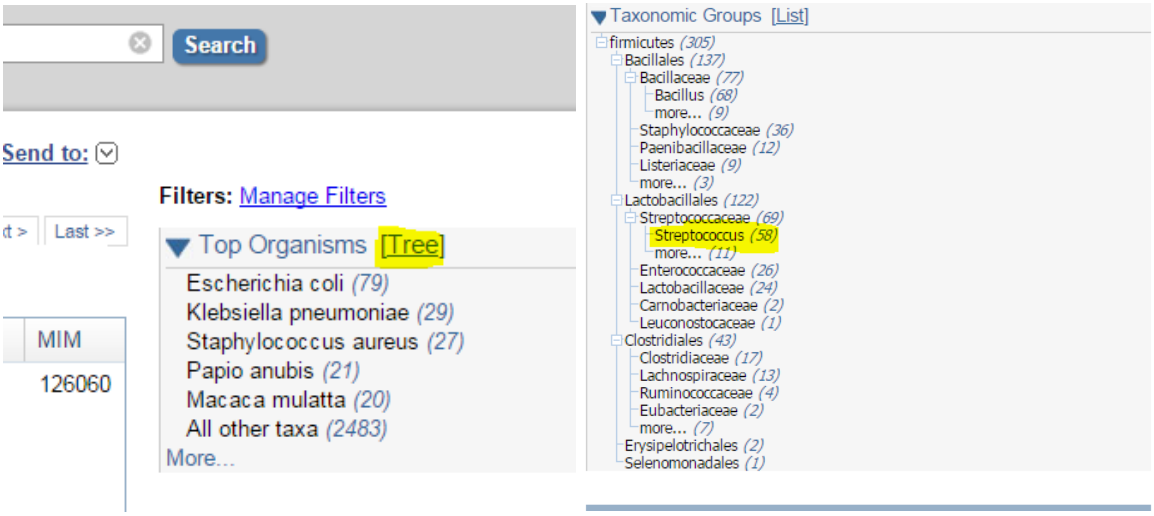



3. This search results in a huge number of genes that are described as dihydrofolate reductases – 2659 to be exact. Each line is a separate gene that has been added to this database – most from different organisms. Sometimes, you'll be lucky and find the organism of interest right away. In this case,

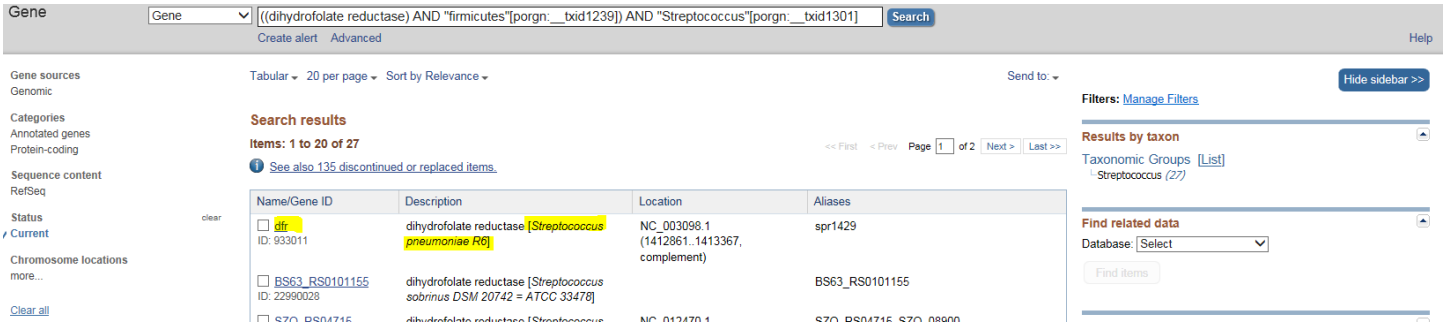
neither *Streptococcus pneumoniae* nor *Saccharomyces cerevisiae* are in the list. It's a huge waste of time to scroll through this list and find the organism, but fortunately NCBI provides a couple easy ways to filter out unwanted entries.

a. Use Boolean characters in the search bar to distinguish an organism:
    i. Search for: dihydrofolate reductase AND saccharomyces. You should end up with a list that has *Saccharomyces cerevisiae S288c* as the top hit. Go ahead and select the GeneID "DFR1". Open this in a new tab. We will come back to this one below.

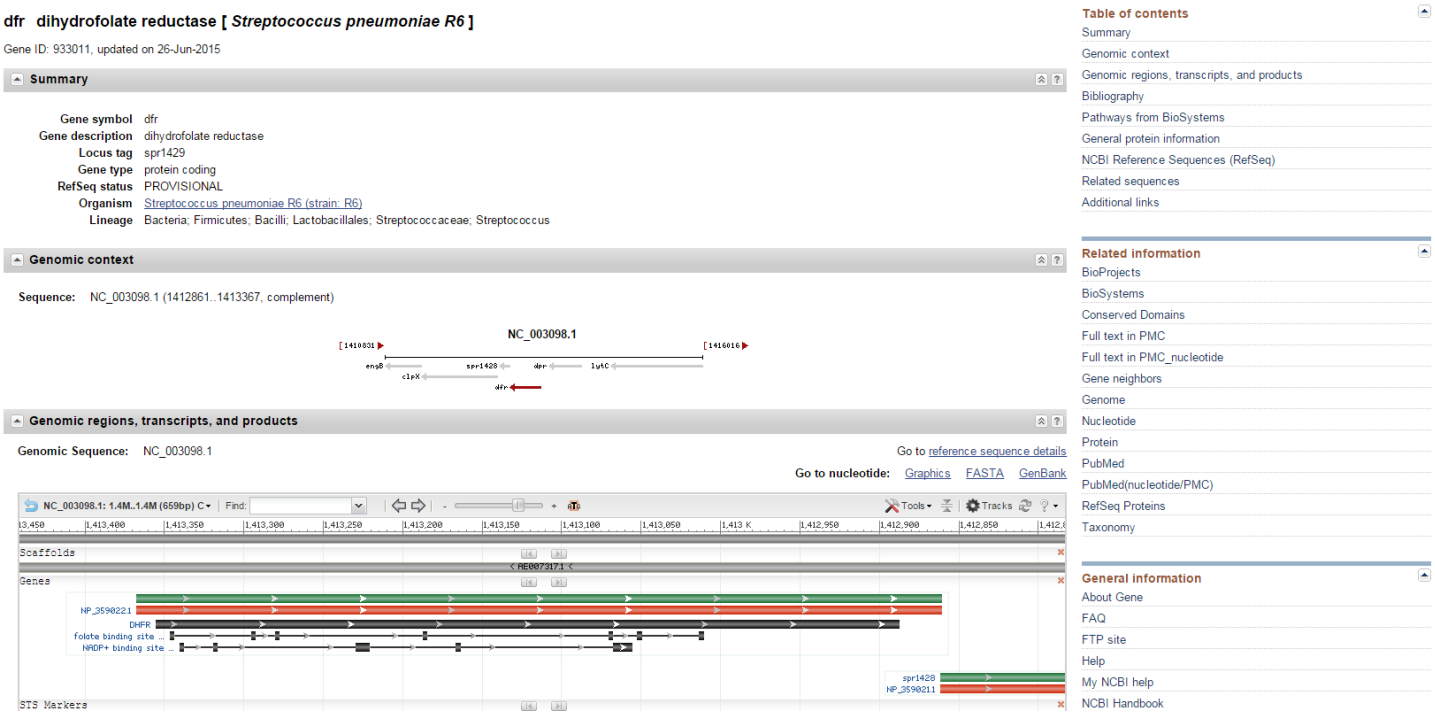
b. Use the taxonomy tool on the right side of your screen.
    i. Click on the Tree link. This brings up a larger list of organism based on the taxonomy. *Streptococcus pneumoniae* is a firmicutes (under bacteria) (click on the bacteria. You should now find Streptococcus in the list under Lactobacillales – click on it.




    ii. You should now find a Streptococcus pneumoniae entry within the first few lines. Specifically, you'll want to find the M1 GAS entry to follow along with the rest of this tutorial. Open this link in a new tab.

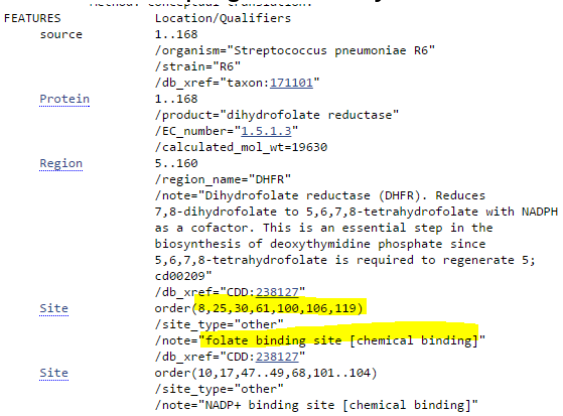


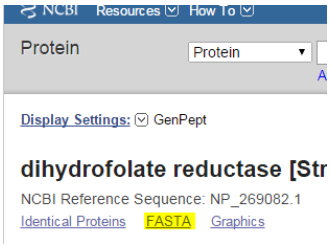
iii. You should now see a screen that looks like the image below.




The "Summary" box gives you information about the organism and gene ID. The "Genomic context" shows you what other genes are in the genomic "neighborhood" and can allow inferences about what role a gene plays. The "Genomic regions, …." box provides quick information about the transcription and translation products of this gene. In this case, the green bar gives you information about the gene, the red bar tells you about the translation product (…so the protein that gets made), and the gray bars describe "regions". Each region is a well characterized part of the protein that has a clear and known function. The first bar (labelled DHFR) covers most of the region of the translation product. This is the part of the protein that has a known function – hovering over the bar tells you what that function is ("Dihydrofolate reductase (DHFR). Reduces 7,8-dihydrofolate to 5,6,7,8-tetrahydrofolate with NADPH as a cofactor. This is an essential step in the biosynthesis of deoxythymidine phosphate since 5,6,7,8-tetrahydrofolate is required to regenerate 5, ..." ). The next one down is the folate binding site, which shows the specific locations that are known to interact with one of the substrates, folate (more on this when you cover protein structure in class). The next bar is the NADP+ binding site, which shows the regions that bind to the redox cofactor NADP+.

Go ahead and right click on the red box and select Views & Tools → GenBank View: NP_.. The NP means you're heading off to a site that gives you information about the protein. If you selected GenBank View: NC_.., you'll be heading to a site that tells you about the nucleotide sequence. In the top right corner, you'll see "Protein", which confirms that you're at the correct site. On this page, you have access to a lot of useful information including more details about the NADP+ and folate binding sites. The numbers shown here correspond to amino acid numbers that are known to interact with that substrate. For example, we can see that the folate binding site is made up of amino acids 8, 25, 30, 61, 100, 106, and 119. Looking all the way at the bottom of the page, you can see the actual amino acid sequence (using the single letter abbreviations). Together, you can tell that folate interacts with I (isoleucine-8), W (tryptophan-25), E (glutamic acid-30), etc.






A very useful format for bioinformatics is the FASTA format. Pretty much all tools that we'll see this term can function using this format – in some cases, it's the only format that is accepted! You can access the fasta sequence for the protein by selecting the FASTA link on the Protein page. This format has a ">" symbol on the first line – anything that follows this is tagged as information about the sequence. Once a new line is created (you know, by hitting "enter"), the rest of the information is read as the sequence. Go ahead and follow click on FASTA link on the Protein page and find the sequence. Do the same for the Nucleotide page. Open up a text document and paste both of these sequences into it.

```
>gi|15903472|ref|NP_359022.1| dihydrofolate reductase [Streptococcus pneumoniae R6]
MTKKIVAIWAQDEEGLIGKENRLPWHLPAELQHFKETTLNHAILMGRVTFDGMGRRLLPKRETLILTRNPEEKIDGVATFQDVQSVLDW
YQDQEKNLYIIGGKQIFQAFEPYLDEVIVTHIHARVEGDTYFPEELDLSLFETVSSKFYAKDEKNPYDFTIQYRKRKEV

>gi|15902044:1412861-1413367 Streptococcus pneumoniae R6 chromosome, complete genome
TTAGACTTCCTTTCTCTTGCGGTATTGGATGGTAAAATCATAAGGATTCTTCTCATCTTTGGCGTAAAATTTGCTTGAAACTGTCTCAA
AAAGAGACAAGTCAAGCTCTTCAGGGAAATAGGTATCTCCTTCCACCCGAGCATGAATGTGAGTGACAATCACTTCATCAAGGTAAGGT
TCAAAAGCCTGAAAAATTTGCTTCCCACCGATAATGTAGAGATTCTTTTCTTGATCCTGATACCAGTCAAGAACAGACTGGACGTCCTG
AAAAGTAGCAACCCCATCTATCTTTTCTTCCGGATTACGCGTCAAAATCAGGGTTTCCCGTTTTGGAAGCAAGCGACGCCCCA
TCCCATCAAAGGTCACACGCCCCATCAAGATAGCATGATTCAGAGTTGTTTCTTTAAAGTGCTGCAATTCTGCTGGCAAATGCCAAGGC
AGACGATTTTCCTTACCAATCAAACCCTCTTCATCCTGGGCCCAAATAGCTACGATTTTCTTAGTCAT
```

Now go back to the Yeast tab that you opened above (in 3a). You'll note that page looks very different. This is because *Saccharomyces cerevisiae* are eukaryotes; you should recall that the genomes of eukaryotes are much more complex than prokaryotes. For starters, the genome is organized into multiple chromosomes. In the Genomic context box, you can see that this gene in positioned on chromosome 15 and has a single exon (expressed regions of the gene).

The "Genomic regions, transcripts, and products", in this case, looks very different; there is only a green line. From this, you can access all information about the gene and protein, but it's not as simple as we saw in the prokaryotic version. This emphasizes the point that all the data you can collect from this database is dependent on someone adding the information correctly. Being versatile and willing to explore for a few minutes will go a long way to help you find the information that you need.

Open up the GenBank View: NC (right click on the green line and go to Views and Tools). As before, go ahead and access the FASTA sequence and paste it into your document.

Question: How can we find the protein sequence?

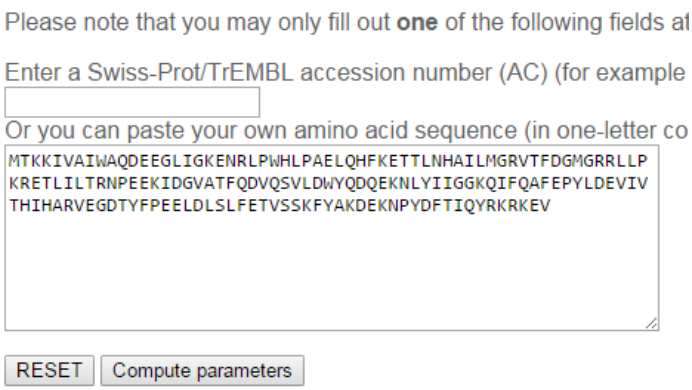
I'm glad you asked! In the FEATURES section (this is back on the main GenBank View: NC page), there is are several menus: gene, mRNA, and CDS are all of use. When you are dealing with a eukaryotic gene with several introns and exons, you can use the mRNA menu (transcript_id) to get the GenBank page for the mRNA. Similarly, we can find information about the protein in the CDS (this stands for conserved domains) section. Find the Protein_ID link – click on it and it will bring you to the protein page for the yeast DHFR. You'll note on this page that there is a folate and NADP binding site in this protein as well. Now go ahead and grab the FASTA sequence.

Ok, so we're done with the NCBI site for now – you may want to keep it open because we'll be coming back to use one more tool – BLAST. For now, navigate to www.expasy.org. This is a site that collects a very wide

variety of bioinformatics tools – some are similar to NCBI, some are not. For now, the two useful tools we'll use are ProtParam and ClustalW2

**ProtParam** – this is a great tool that gives you a ton of information about a protein based on its sequence. You can find it by clicking on Proteomics → Protein characterization and function → ProtParam (also, Googling ProtParam goes right to it).


Once you get to the ProtParam Tool, paste in the sequence of the *S. pneumoniae* DHFR (from your FASTA sequence). Make sure to put only the sequence, NOT the >info header.

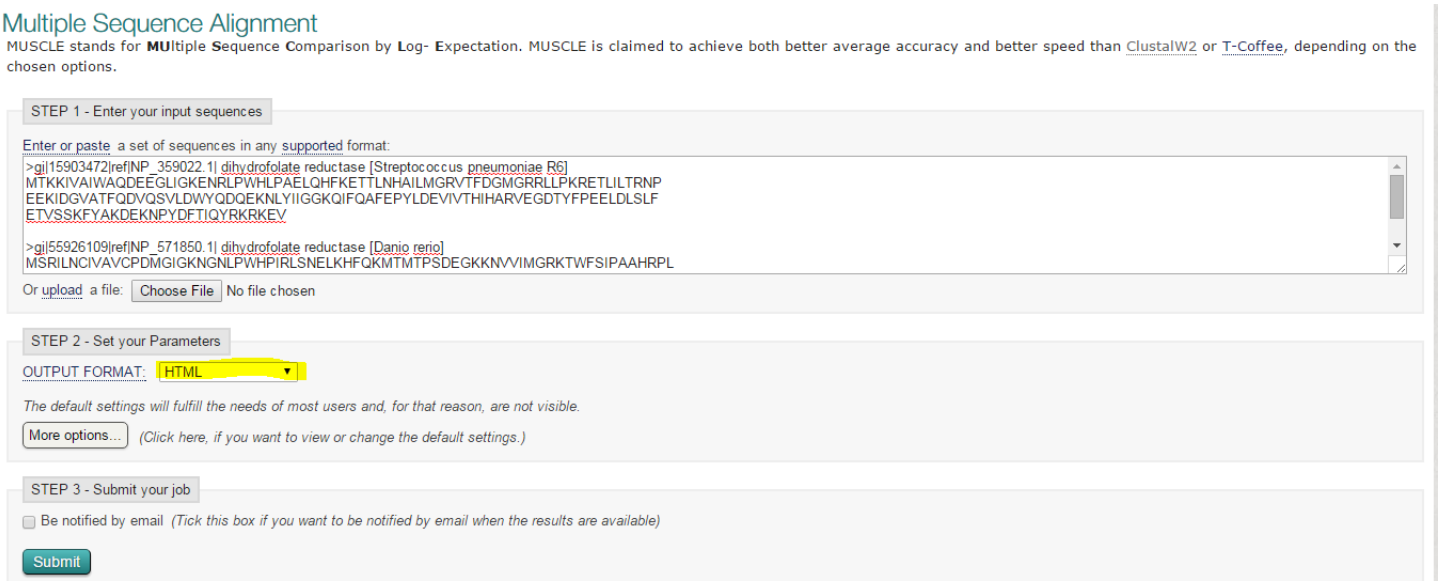



Computing the parameters gives you a ton of useful information about this protein including the molecular weight (19,761.5 Da), theoretical pI (5.42), and extinction coefficient (aka molar absorptivity - 26,930 $M^{-1}cm^{-1}$). This information is very useful when purifying and characterizing proteins. You'll absolutely need to use these numbers for OUR protein throughout the term.

If you do the same thing for the yeast DHFR, you'll see that all three of these values are similar: MW = 21,698 Da, pI = 8.40, and Ext. Coeff = 24,075. This suggests that the proteins are pretty darn similar. We'll confirm this in a little bit.

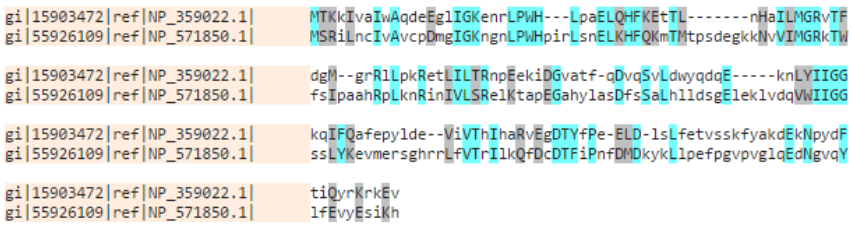
**MUSCLE** – this tool allows you to directly compare 2 or more sequences and look for similarities and differences. You can find it under Proteomics → similarity search/alignment → MUSCLE.

When you reach the MUSCLE page, you can paste two or more FASTA sequences into the text box (the MUST be FASTA). Let's first compare the two protein sequences.




In the "Visualization" activity that we'll do later in the term, you will need to output format to be Pearson/FASTA – keep this in mind!

Submit the job and wait for the alignment to appear. When it does, you should see an image that looks like this:



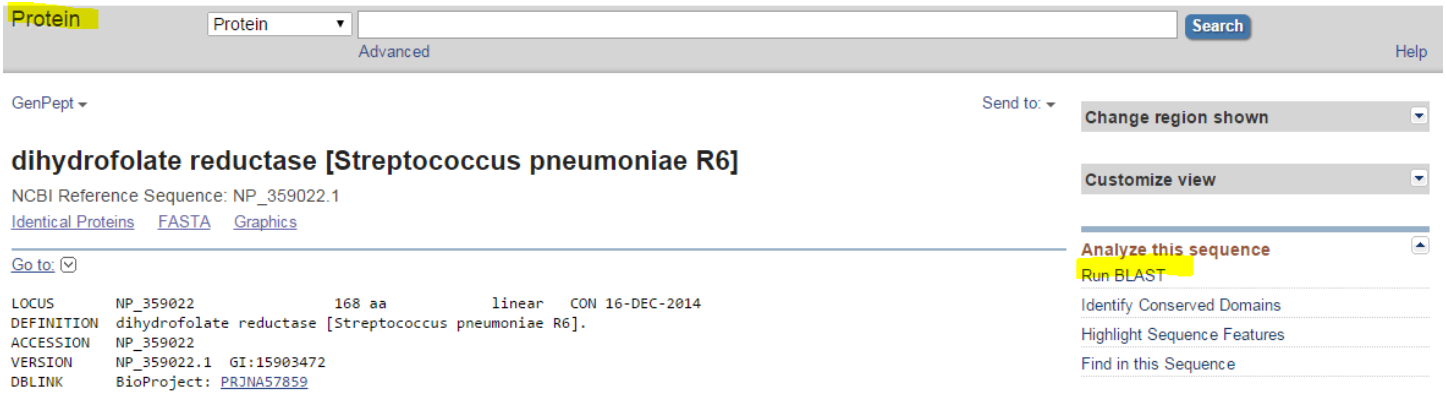

The image summarizes the relationship between the two proteins. The blue boxes identify amino acid residues that are identical between the two organisms, while the gray boxes are amino acids that have similar characteristics (e.g. hydrophobic or acidic). The more blue boxes that appear, the more identical the proteins are. If there is a lot of gray but not blue, then we say that the proteins have a high degree of similarity but not identity. Later in this exercise, we'll see how to calculate the exact % identity and % similarity. The dashes indicate regions where there is no corresponding amino acid to align: this is common in "loop" regions of protein structure (more on this in a few weeks).

We'll think more about these sequence alignments and how they are useful as the term progresses, but for now, you should be comfortable making a sequence alignment.

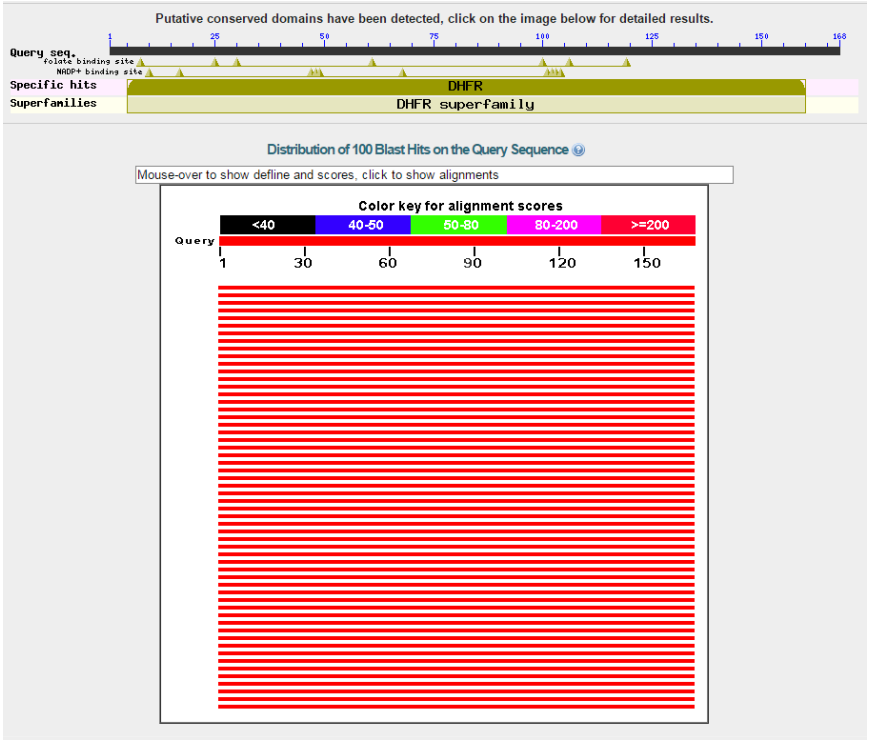
**BLAST**

The last tool we're going to learn about it the BLAST Search tool – which is pretty much amazing. This lets you take a known amino acid or DNA sequence and determine if there is a similar gene/protein in different organism. To exemplify the power of this tool, let's use the amino acid sequence of DHFR from S. pneumoniae to find the DHFR protein in *S. cerevisiae*. You can get to the BLAST page several ways – the

most convenient might be to Google “blast search”. For this exercise, the most convenient way to get there is from the *S. pneumoniae* DHRF protein page.




The great thing about this approach is that the accession number (NP_359022.1) is already put into the Sequence box. If you got here from some other avenue, you can paste the FASTA sequence right into the box and you’ll have the same results. At this point, you can click on the BLAST button – you should get lots of hits.



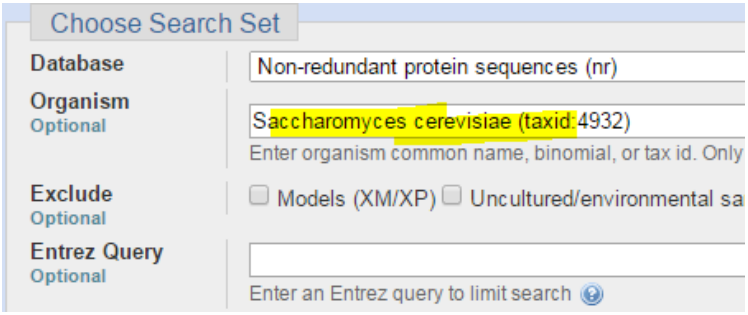

The data is presented in a very logical way. Each line corresponds to a protein from a different organism. The length of the bar indicates how much of the query (the *S. pneumoniae* protein) is represented in the match – a long bar means that there is a match for the entire protein – the shorter bars mean that these matched proteins don’t match up with the full *S. pneumoniae* sequence. The color code is how good the match is. Red means REALLY good match while black means a really bad match. Placing your mouse cursor over the bars, or scrolling down allows you to see that the best hits are from *S. pneumoniae* or other *Streptococci*. This is not surprising since these are all very similar bacteria.

Other important things to note:

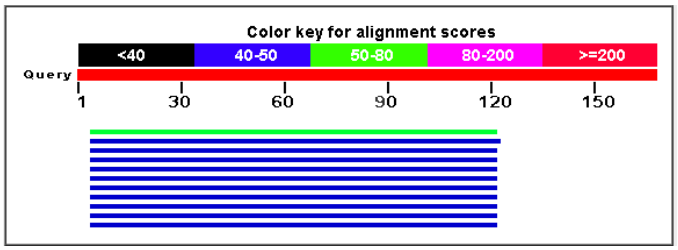
- Each hit has a corresponding score (red lines are >200, black lines are <40). These scores tell you how perfect the match is.

- Each hit show the Identity and Similarity (Positives). These correspond to identical amino acids and amino that have the same chemical properties (e.g. Asp and Glu), respectively. So in the example shown below (which is the very last one on the BLAST output page), the *S. pneumoniae* protein is 95% identical and 98% similar to the query.
- The gaps are listed – this corresponds to regions where there are no amino acids that match up – so perhaps one protein has an extended loop and the other has a β turn. In this case, there are not any gaps because the proteins match up so well.

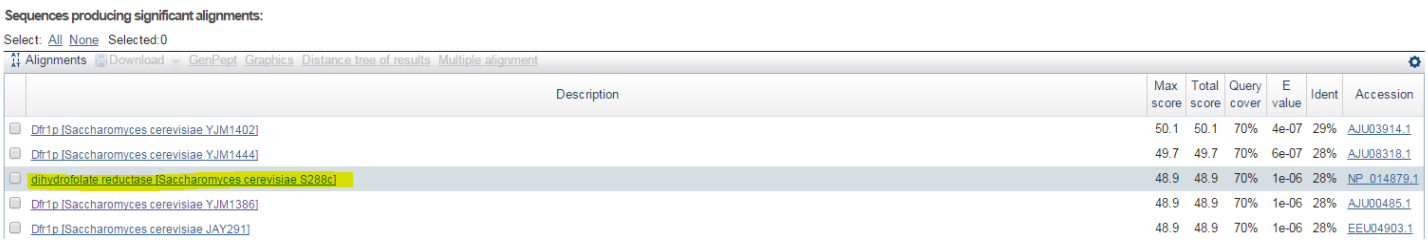

You'll note that *S. cerevisiae* is not in this list, so we still haven't found the desired protein using BLAST. In fact, you'll never find it in this list because it's not similar enough to appear. Never fear, we can still find it! Back on the BLAST homepage, we can use the Organism filter. Type Saccharomyces cerevisiae into the Organism box (make sure NOT to select exclude) and BLAST this.



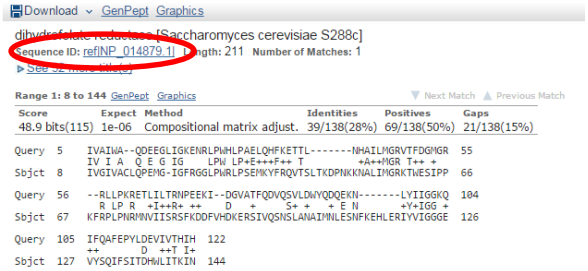

Now you have a MUCH smaller hit list and the hits have much lower scores (green and blue lines), indicating that the hits are not as similar as we saw before. None-the-less, hovering over the boxes shows that these are still DHFR enzymes!




Scroll down a little bit and you'll see that the S288c strain is in the hit list. Click on it.



This takes you down the page to the entry for the S288c protein. As you can see, DHFR in *S. cerevisiae* is 28% Identical and 50% similar (positives) to the *S. pneumoniae* protein; quite a few gaps are observed. If you compare this with the alignment file from MUSCLE, you'll see that the same matches are observed.




How can you know for sure that the protein you found here is exactly the same one that you found through NCBI? Easy! Look at the sequence ID – the link tells you the protein reference number, NP_014879.1. If you go back to the FASTA sequence that you found through NCBI, you'll note that this is exactly the same number; this unambiguously tells you that you have the same protein. Click on the Sequence ID link (circled above) and it will take you to a familiar looking page (NCBI protein page). Now click on the Gene link (or the Nucleotide link will take you to the mRNA page) to verify that this protein is coded for the same gene that we saw before.

**The assignment:**

1. Determine the gene and protein sequence for the lactate dehydrogenase gene from *Bacillus subtilis str 168* (a firmicutes). Report these in FASTA form.
2. From what you learn on the the NCBI GenBank View_NP page, please comment on each of the following:
    a. What redox cofactor does LDH use? (NAD, FAD, NADP, or Heme).
    b. This enzyme converst (S)-lactate and NAD+ to what two products?
    c. How many amino acids are present in LDH from *B. subtilis*?
3. Use the BLAST tool to determine if there is an LDH gene present in *Escherichia coli*. In lab this term, we will be using *E. coli* to make our protein. Note: do NOT select the MULTISPECIES or Partial enzyme for this question.
    a. Does Lactate dehydrogenase exist in E. coli? If so, what is the Sequence ID for the most similar to the *B. subtilis* enzyme?
    b. For the most similar protein, what is the % identity and % similarity?
    c. Does the *E. coli* enzyme use the same cofactor?
    d. Using MUSCLE, prepare a sequence alignment of these two proteins.
4. Using the ProtParam tool, determine the MW, pI, and extinction coefficient of LDH from *B. subtilis* and *E. coli*.

Note: Make sure to save a file that contains all of this information: it will prove to be very valuable over the next several months.