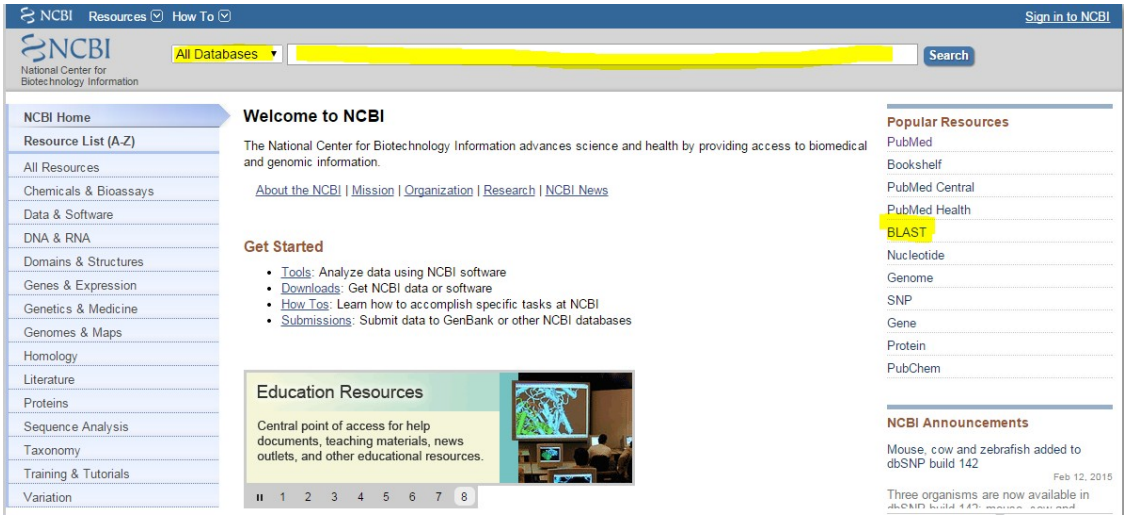
Bioinformatics Introduction Worksheet – The first part of this exercise is aimed at walking you through some of the key tools used by scientists to explore the relationship between genes and proteins throughout the Kingdoms of biology. The second part is an exercise that is aimed at helping you understand 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.

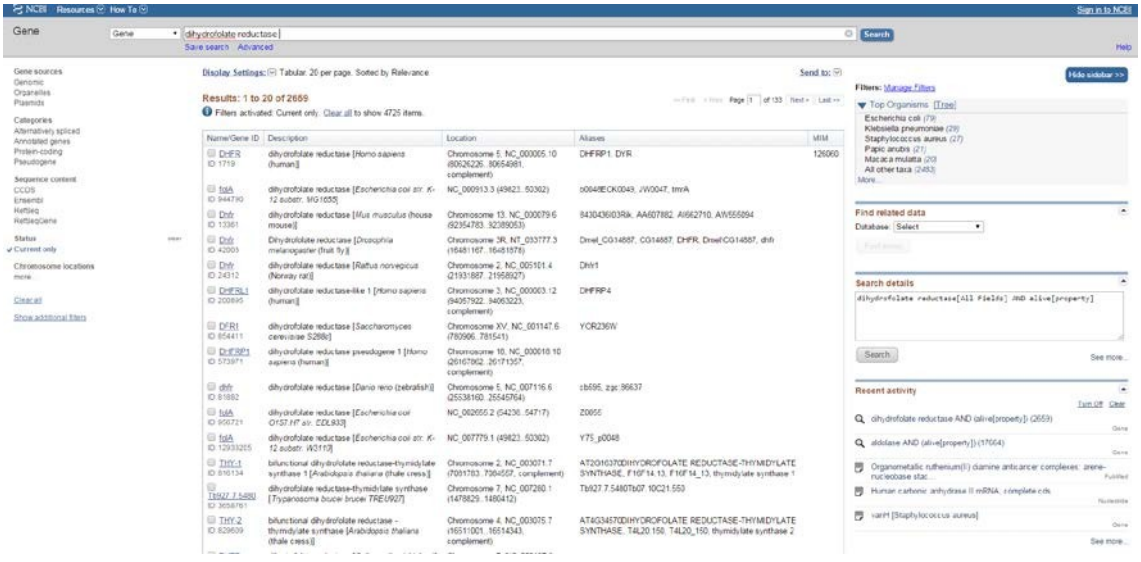
Exercise: In this introductory exercise, image that you are a biochemist looking for information about the enzyme dihydrofolate reductase (DHFR). Your goal is to compare the enzymes from two different organisms: *Streptococcus pneumoniae* (a prokaryote) and *Saccharomyces cerve* (yeast).

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.

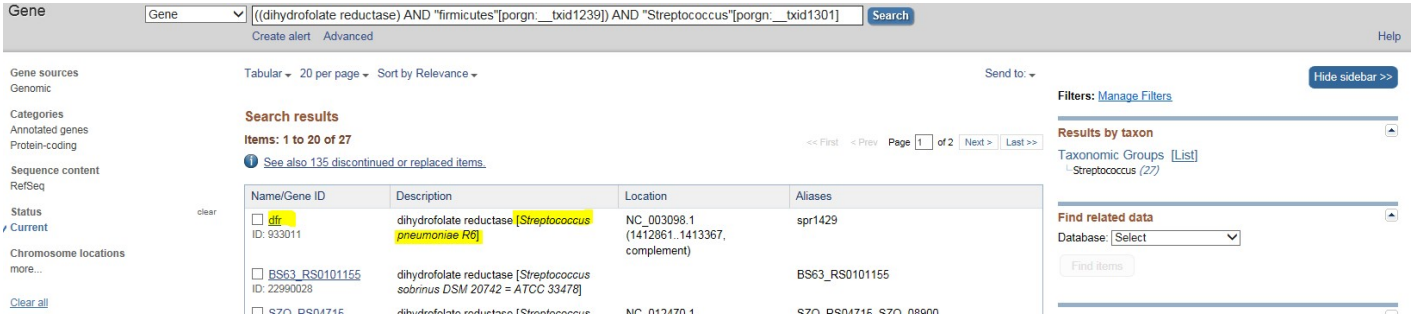
Task 1.Find the exact number of entries. Take a snapshot of the screen.

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.

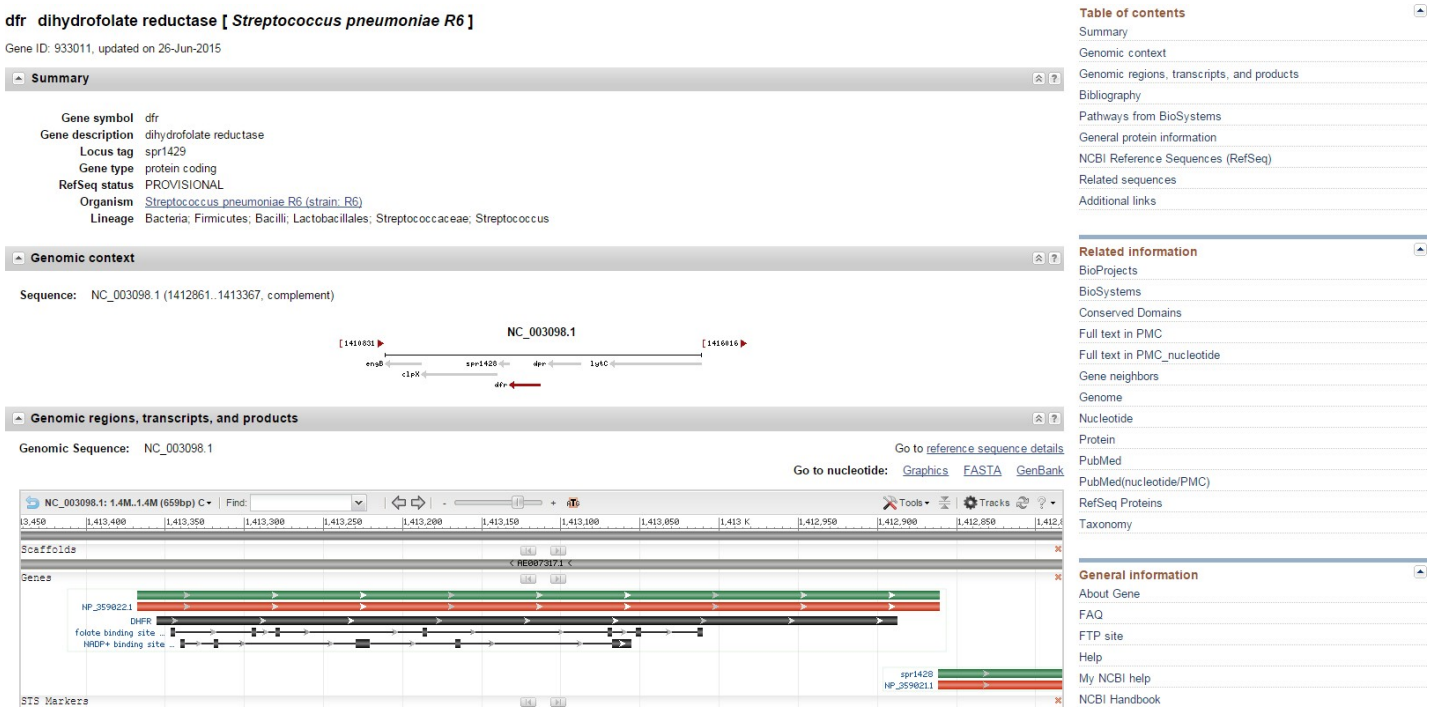
a. Use Boolean characters in the search bar to distinguish an organism:
   i. Search for: dihydrofolate reductase AND saccharomyces. You should end up with a list that has *Saccharomyces cerevisiae S288c* as the top hit. Go ahead and select the GeneID "DFR1". Open this in a new tab. We will come back to this one below.
   ii. Similarly, find the DFR1for *Streptococcus pneumoniae* click on it.

Task 2. Take screenshots of the screens wherever appropriate.

   iii. 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.



iv. 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.

Task 3. Explain the different bars and what information you gather. Also compare the representation between the yeast [*Saccharomyces*] and the bacteria DFR1gene.

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.

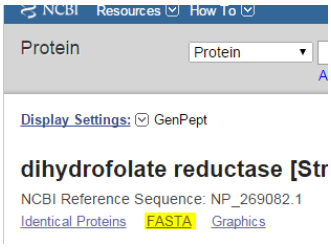
```
FEATURES             Location/Qualifiers
     source          1..168
                     /organism="Streptococcus pneumoniae R6"
                     /strain="R6"
                     /db_xref="taxon:171101"
     Protein         1..168
                     /product="dihydrofolate reductase"
                     /EC_number="1.5.1.3"
                     /calculated_mol_wt=19630
     Region          5..160
                     /region_name="DHFR"
                     /note="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;
                     cd00209"
                     /db_xref="CDD:238127"
     Site            order(8,25,30,61,100,106,119)
                     /site_type="other"
                     /note="folate binding site [chemical binding]"
                     /db_xref="CDD:238127"
     Site            order(10,17,47..49,68,101..104)
                     /site_type="other"
                     /note="NADP+ binding site [chemical binding]"
```

Task4: How many amino acids [represented as aa] are there in the two DFR1 proteins [yeast and bacteria]?

Task 5: Identify the two main features.

Task 6: List the aminoacid residues involved in the two features and compare them between the two organisms.

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.



Task 7: 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.

Hints:

The Yeast tab that you opened above. 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?

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.