FINDING GENES AND BUILDING SEARCH STRATEGIES
(Exercise 1)

1.1 Finding a gene using text search.
   Note: For this exercise use http://www.plasmodb.org

a. Find all possible kinases in *Plasmodium*.

   Hint: use the keyword “kinase” (without quotations) in the “Gene Text Search” box.

   - How many genes did you get?
   - Look closely at the sections of the result page. How many of those are in *P. falciparum*? How did you find this out?
   - What happens if you search using “kinases”? How many results were returned?

b. Find only the kinases that specifically have the word “kinase” in the gene product name.

   Use the full text search, the specific page where you can define the fields to be searched. There are several ways to navigate to the Text Search page.

   The search you did in ‘a’ will miss plurals and compound words like “kinases” or “6-phosphofructokinase”. Using a wild card in your search term will broaden your search – try kinase kinase* *kinase *kinase*

   - Try giving each new search a name to help you keep track of the searches.
   - How did you get to the Text Search page?
   - Did you remember to use the wild card?
   - How many genes have the word kinase in their product names?
c. How can you quickly examine the genes that were identified using the key word *kinase* but not with the word kinase?

- Hint: build a search strategy that combines 2 of your searches. Click on “Add Step” then select “existing strategy”:
- Select the right strategy from your list of Gene Strategies and combine the strategies with the correct operation:
- Do the results make sense? Do all the product names contain the word kinase?
d. Find *P. berghei* ANKA genes that have between 2 and 5 transmembrane domains.

In this example we will search for genes based on attributes other than text annotations. There are many attributes associated with genes that can be used to identify genes. A gene’s genomic location, Gene Ontology assignments, number of exons, number of transmembrane domains are all examples of gene attributes that form the basis of a search for genes within EuPathDB.

- How many *P. berghei* ANKA genes have between 2 and 5 transmembrane domains?
- Hint – use the Transmembrane Domain Count search under Protein Features.
- There are two options for setting the Organism parameter. Which option did you use? Using Option 1 takes advantage of the filter table to easily toggle between search results of different species.
1.2 Combing text search results with results from other searches

a. In exercise 1.1b you identified genes that have the word “kinase” somewhere in their product name. Can you now find out how many of these kinases are likely secreted?

Hint: grow your search strategy by adding a step. Choose a search that identifies genes with likely secretory signal peptides. How did you combine the search results?
Now that you have a list of possible secreted kinases, expand this strategy even further.

There is no wrong answer here!!

- From a biological standpoint what else would be interesting to know about these kinases? Add more searches to grow this strategy. Open the categories under 'Identify Genes By' on the home page and explore the types of searches that are available. You can reduce (or expand) your result set by adding searches that are based on many types of data.

- For example, how many of these secreted kinases also have transmembrane domains?

c. In the above example, how can you define kinases that have either a secretory signal peptide AND/OR a transmembrane domain(s)?

Hint: to do this properly you will have to employ the “Nested Strategy” feature. Why?
1.3 Finding a gene by BLAST Similarity.

Note: For this exercise use http://www.toxodb.org
Imagine that you generated an insertion mutant in *Toxoplasma* that is providing you with some of the most interesting results in your career! You sequence the flanking region and you are only able to get sequence from one side of the insertion (the sequence shown below). You immediately go to ToxoDB to find any information about this sequence. What do you do?

```
- aaaggagagaagataaaatatataaaaaagtctcccaagacacgatagtgttactgacaacatacagaatcaggtcgagcaatggaagaaccaagcaccggcgccagagattgaactcgcctggattgcgtcgagcaatggaagaaccaagcaccggcgccagagattgaactcgc
```

- Try using the BLAST search with this sequence (hint: you can get to the BLAST tool by clicking on the BLAST link under tools on the home page).

Which blast program should you use? (hint: try different combinations, just keep in mind that you have a nucleotide sequence so you have to use an appropriate BLAST program).

**Note on BLAST programs:**
- blastp compares an amino acid sequence against a protein sequence database;
- blastn compares a nucleotide sequence against a nucleotide sequence database;
- blastx compares the six-frame conceptual translation products of a nucleotide sequence (both strands) against a protein sequence database;
- tblastn compares a protein sequence against a nucleotide sequence database dynamically translated in all six reading frames (both strands);
• `tblastx` compares the six-frame translations of a nucleotide sequence against the six-frame translations of a nucleotide sequence database.

- Are you getting any results from `blastx`? `tblastn`? What about `blastn`?
- What is your gene? (hint: after running a blastn against *Toxoplasma* ME49 genomic sequence, click on the “link to the genome browser”. In the genome browser zoom out to see what gene is in the area).

1.5 More BLASTing in EuPathDB (optional).
Note: for this exercise use [http://www.eupathdb.org](http://www.eupathdb.org)

a. The first thing we will need to do is get a sequence to use for BLAST. Search for the keyword "dihydrofolate" (without quotations). (Hint: use the Gene Text Search on the upper right hand side of the EuPathDB home page).

b. You should get multiple hits. Find the first one that is annotated as "dihydrofolate reductase-thymidylate synthase" (Look in the product description column).

c. Once you find one, click on the gene ID and go to the gene page. It might be helpful to open the gene page in a new window or tab.

d. Scroll down to the bottom of the page to the “Sequences” section.

e. Copy the amino acid sequence and go back to EuPathDB (if you have not done so
already, it might be helpful to open EuPathDB in a new window or tab).


g. Paste the amino acid sequence into the input window.

h. Select target data type (start with "Proteins").

i. Select BLAST Program. (Hint: BLASTP).

   Expectation value : 10
   Maximum descriptions/alignments (V=B) : 100

j. Select the target organism. Click on "Get Answer".

Based on the results you should have identified excellent hits in almost all pathogens in EuPathDB but can you find good hits in Giardia or Trichomonas? Let's try a different BLAST method:

k. Go back to the BLAST window. Change the target data type to Genome.

l. Select the BLAST Program. Notice you cannot select BLASTP anymore. Try the other options. Notice how your input sequence type has to change when you select a different program. (Hint: TBLASTN is the one you need).

m. Select all target organisms. Click on "Get Answer".

Note that the results are still missing a dihydrofolate from Giardia and Trichomonas. Let's try a different BLAST method.

n. Go to your gene page window (in CryptoDB) and copy the nucleotide coding sequence.

o. Go to the BLAST window and paste the nucleotide sequence into the input window.

p. Select the target data type (try different ones) and the BLAST program. Notice you can only select TBLASTX or BLASTN when your input sequence is nucleotide. (Hint: select TBLASTX).

q. Select the target organisms. This time let's specifically only select Giardia and Trichomonas. Click on "Get Answer".

Getting frustrated?

Not getting a hit for Giardia in this case is actually the correct answer! This organism does not have dihydrofolate reductase or thymidylate synthetase activity.