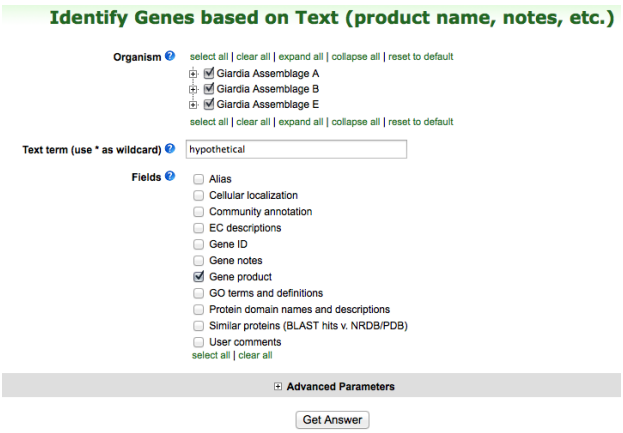
Protein Motif Searches and Regular Expressions

1. **Using InterPro domain searches to identify unannotated kinesin motor proteins.**
   Note: For this exercise use http://giardiadb.org

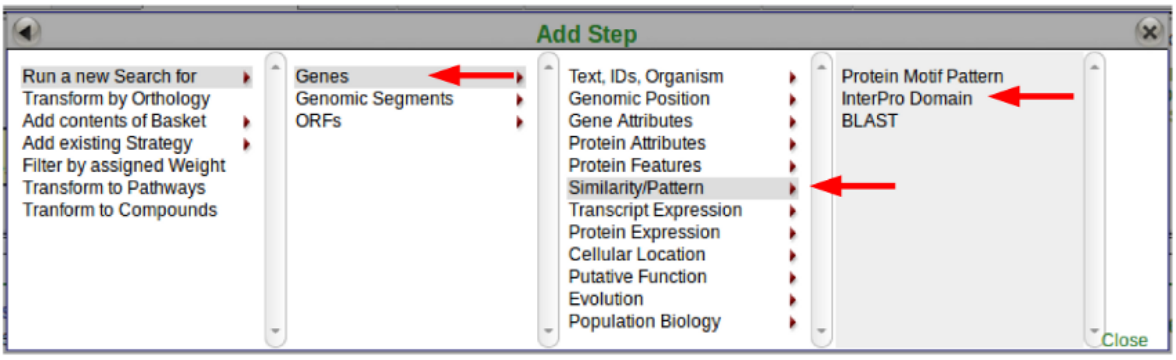
   Identify all genes annotated as hypothetical in all *Giardia* assemblages.

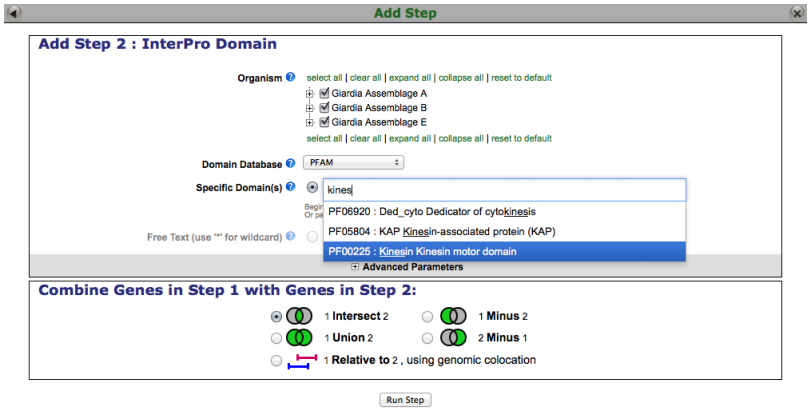
   (Hint: use the full text search and look for genes with the word "hypothetical" in their product names)



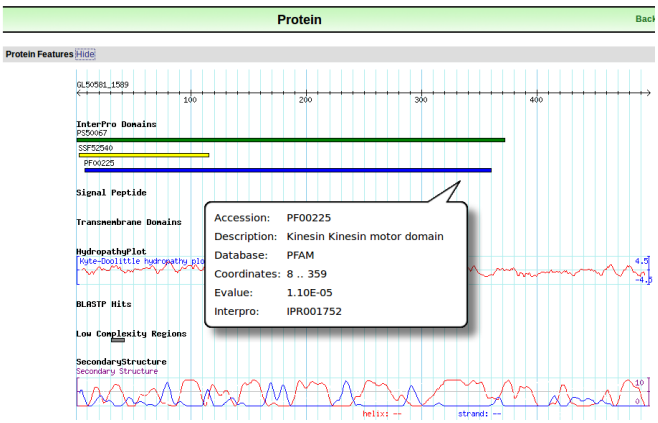
   b. How many of these hypothetical genes have a kinesin-motor protein PFAM domain?

   (Hint: add a step to the strategy. Go to the "Interpro Domain" search under similarity/pattern, start typing the work kinesin and it should autocomplete.)





c. Go to the gene page for GL50581_1589 and look at the protein feature section. Does this look like a possible motor protein?

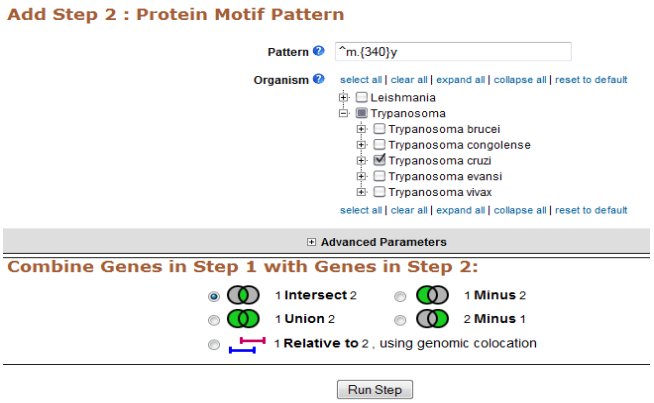


Hint: click on the ID for GL50581_1589 in the result table to go to the gene page. Scroll down to the protein section and mouse over the glyphs in the Protein Features graphic.

2. Using regular expressions to find motifs in TriTrypDB: finding active trans-sialidases in *T. cruzi*.
Note: for this exercise use http://tritrypdb.org

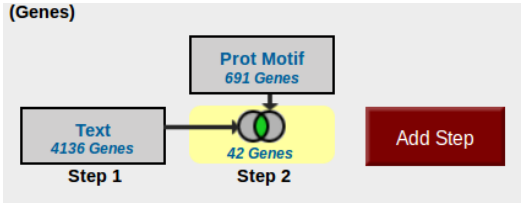
T. cruzi has an expanded family of trans-sialidases. In fact, if you run a text search for any gene with the word "trans-sialidase", you return over 4000 genes among the strains in the database!!! Try this and see what you get.

b. However, not all of these are predicted to be active. It is known that active trans-sialidases have a signature tyrosine (Y) at position 342 in their amino acid sequence. Add a motif search step to the text search in 'a' to identify only the active trans-sialidases.



Hint: for your regular expression, remember that you want the first amino acid to be a methionine, followed by 340 of any amino acid, followed by a tyrosine 'Y'. Refer to [regular expression tutorial] if you need to.

If you need help, you can go to this sample strategy below to see the answer:



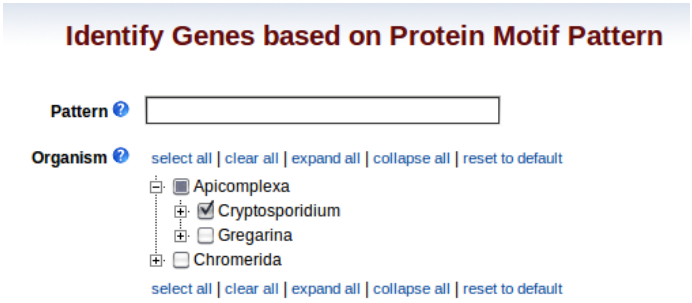
http://tritrypdb.org/tritrypdb/im.do?s=a905e36f634f7b42

3. **Using regular expressions to find motifs in CryptoDB: finding genes with the YXXΦ receptor signal motif**

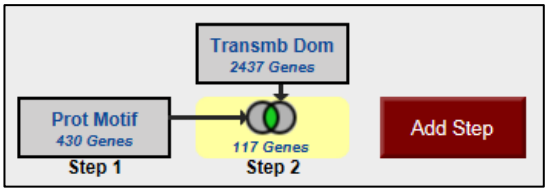
Note: for this exercise use http://cryptodb.org

The YXXΦ (Y=tyrosine, X=any amino acid, Φ=bulky hydrophobic [phenylalanine, tyrosine, threonine]) motif is conserved in many eukaryotic membrane proteins that are recognized by adaptor proteins for sorting in the endosomal/lysosomal pathway. This motif is typically located in the c-terminal end of the protein.

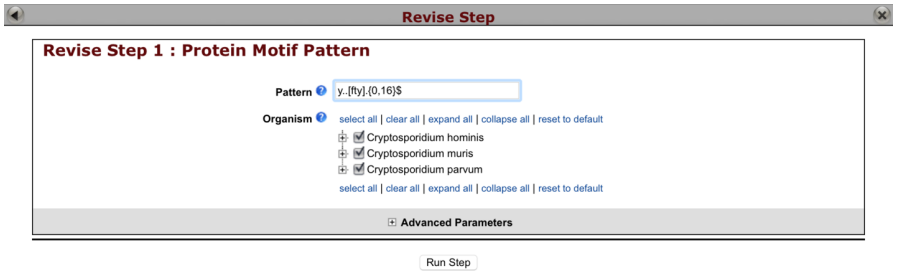
a. Use the "protein motif pattern" search to find all *Cryptosporidium* proteins that contain this motif anywhere in the terminal 10 amino acids of proteins. (hint: for your regular expression, remember that you want the first amino acid to be a tyrosine, followed any two amino acids, followed by any bulky hydrophobic amino acid (phenylalanine, tyrosine, threonine). Refer to [regular expression tutorial](#) if you need to).



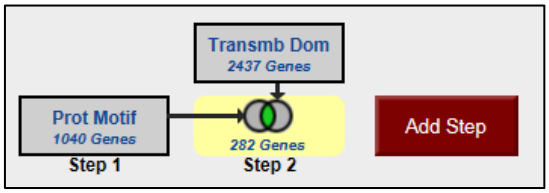
b. How many of these proteins also contain at least one transmembrane domain.



c. What would happen if you revise the first step (the motif pattern step) to include genes with the sorting motif in the C-terminal 20 amino acids? (hint: edit the first step and modify your regular expression)



Here is a saved strategy that provides you with the results of the above search:

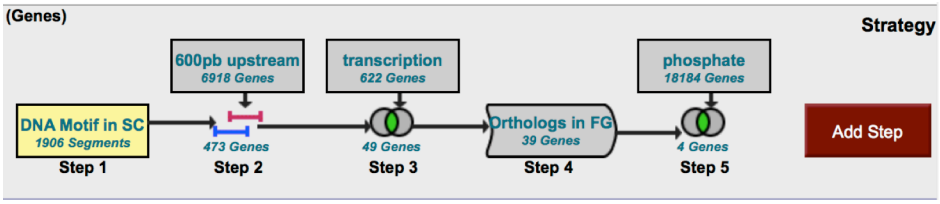


http://cryptodb.org/cryptodb/im.do?s=928309b4c1b9ef3f

Exercise: Identify genes downstream of regulatory DNA motifs.

We will start our search with *S. cerevisiae* as a model organism where motifs and regulatory factors have been studied in great detail.

Transcriptional start sites are often located within a certain distance upstream of genes or gene clusters that they regulate. In some fungi, DNA motifs are important for regulation of processes linked to host cell invasion. Readily available genomic data facilitate the discovery of regulatory motifs and allow identification of novel genes via examination of orthologous sequences.



1. Initiate a search for CACGTG DNA motif

Click on "Search for genomic segments (DNA motif)" in the search menu at the top of the page.

Enter CACGTG in the search box and perform search in *Saccharomyces cerevisiae.*

Your search should return 1906 segments containing CACGTG motif.
Next, we will look for putative regulatory targets of this motif by searching for genes located 600bp downstream of this sequence.

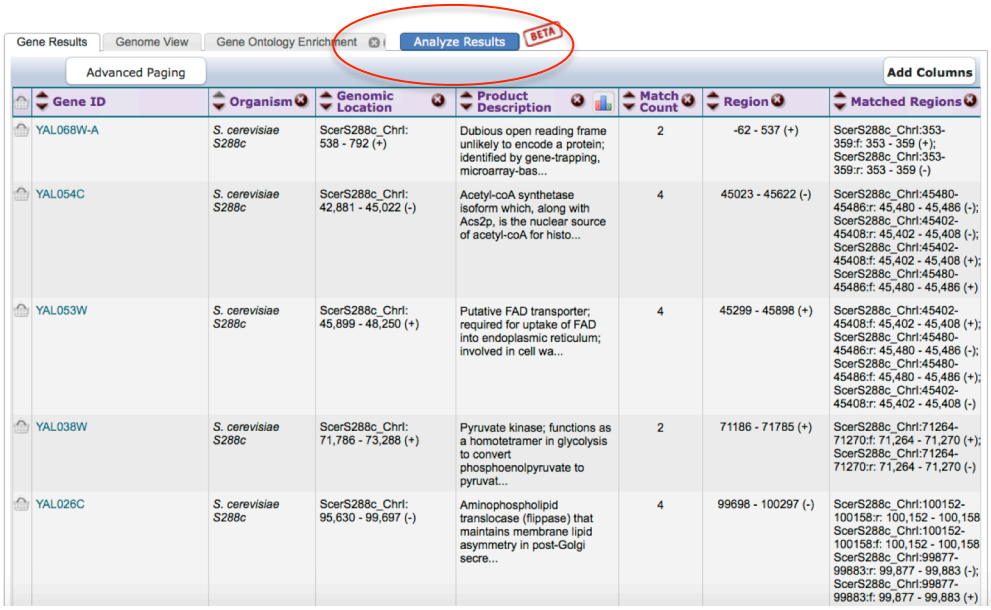
2. Identify genes with CACGTG motif being 600bp upstream of an open reading frame.

Click "Add Step".
Choose Run a new search for Gene > Text > Organism and select "Relative to genomic location". In the next screen you will set parameters of your search.

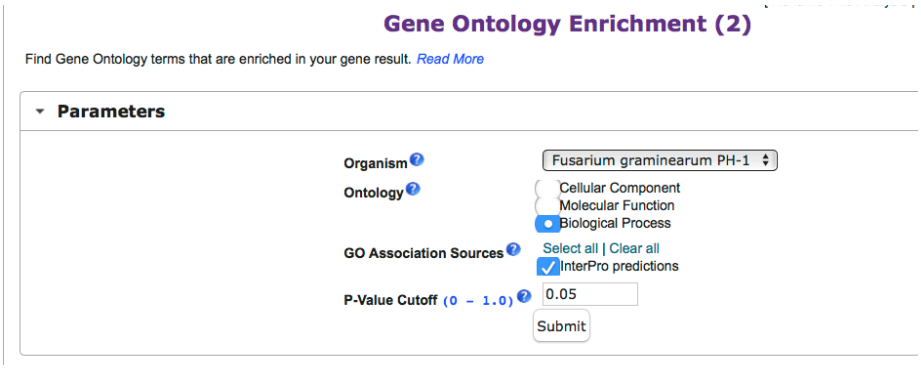
Set up your search using the following guideline:
Return each gene from step 2 whose upstream region (600bp) overlaps the exact region of a Genomic Segment in Step1 (CACGTG) and is on either strand.

Let's look a little closer at the 6918 genes we have identified and determine if there is an enrichment for certain biological processes.

Click on Analyze Results –BETA tab and view Gene Ontology based on biological processes.



Gene ID	Organism	Genomic Location	Product Description	Match Count	Region	Matched Regions
YAL068W-A	*S. cerevisiae S288c*	ScerS288c_Chrl: 538 - 792 (+)	Dubious open reading frame unlikely to encode a protein; identified by gene-trapping, microarray-bas...	2	-62 - 537 (+)	ScerS288c_Chrl:353-359:f: 353 - 359 (+); ScerS288c_Chrl:353-359:r: 353 - 359 (-)
YAL054C	*S. cerevisiae S288c*	ScerS288c_Chrl: 42,881 - 45,022 (-)	Acetyl-coA synthetase isoform which, along with Acs2p, is the nuclear source of acetyl-coA for histo...	4	45023 - 45622 (-)	ScerS288c_Chrl:45480-45486:r: 45,480 - 45,486 (-); ScerS288c_Chrl:45402-45408:r: 45,402 - 45,408 (-); ScerS288c_Chrl:45402-45408:f: 45,402 - 45,408 (+); ScerS288c_Chrl:45480-45486:f: 45,480 - 45,486 (+)
YAL053W	*S. cerevisiae S288c*	ScerS288c_Chrl: 45,899 - 48,250 (+)	Putative FAD transporter; required for uptake of FAD into endoplasmic reticulum; involved in cell wa...	4	45299 - 45898 (+)	ScerS288c_Chrl:45402-45408:f: 45,402 - 45,408 (+); ScerS288c_Chrl:45480-45486:r: 45,480 - 45,486 (-); ScerS288c_Chrl:45480-45486:f: 45,480 - 45,486 (+); ScerS288c_Chrl:45402-45408:r: 45,402 - 45,408 (-)
YAL038W	*S. cerevisiae S288c*	ScerS288c_Chrl: 71,786 - 73,288 (+)	Pyruvate kinase; functions as a homotetramer in glycolysis to convert phosphoenolpyruvate to pyruvat...	2	71186 - 71785 (+)	ScerS288c_Chrl:71264-71270:f: 71,264 - 71,270 (+); ScerS288c_Chrl:71264-71270:r: 71,264 - 71,270 (-)
YAL026C	*S. cerevisiae S288c*	ScerS288c_Chrl: 95,630 - 99,697 (-)	Aminophospholipid translocase (flippase) that maintains membrane lipid asymmetry in post-Golgi secre...	4	99698 - 100297 (-)	ScerS288c_Chrl:100152-100158:r: 100,152 - 100,158 ScerS288c_Chrl:100152-100158:f: 100,152 - 100,158 ScerS288c_Chrl:99877-99883:r: 99,877 - 99,883 (-); ScerS288c_Chrl:99877-99883:f: 99,877 - 99,883 (+)



You can explore gene ontology enrichment for cellular components and molecular function.

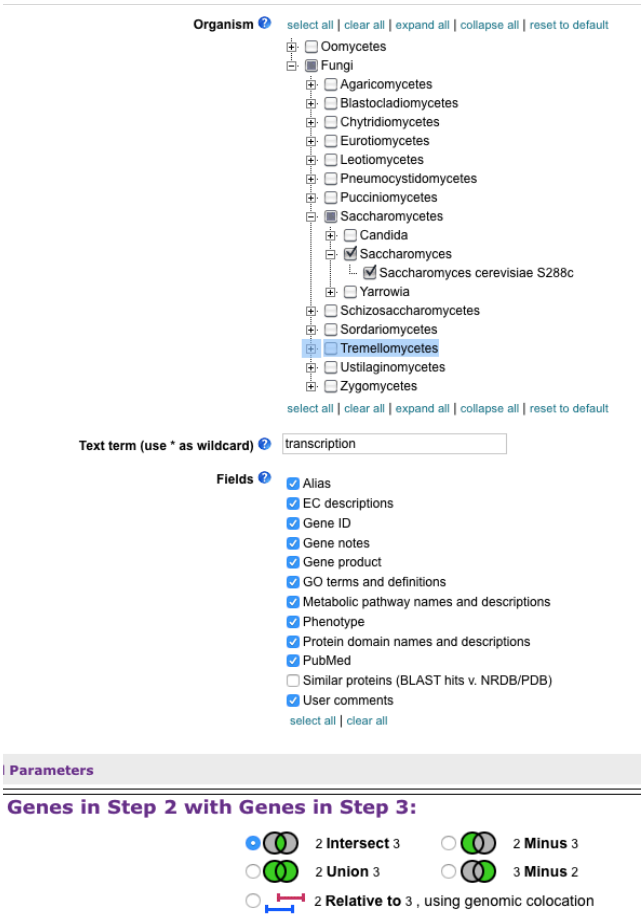
It looks like this motif is located upstream of genes that belong to various biological processes. Let's narrow down our selection and look at those genes that are important for transcription.

Click “Add Step” and select Gene ID > Text and search for “transcription”.

Your search should return 39 genes that are assigned to 35 biological pathways.

How do you find which biological categories are enriched?

Hint: refer to Analyze Results tab > Gene Ontology > Biological Process



This search strategy can be extended to identify orthologous genes in less studied model systems. For example, let’s identify orthologs of the 49 *S. cerevisiae* genes in *Fusarium graminearum*, a plant pathogen.

Initiate search for orthologs in *F. graminearum* using “Add Step” tool. To determine which cellular components are enriched within our search click on “Analyze Results-BETA” tab and choose GO > Cellular Component

It is known that cellular components are affected by various growth conditions. For example, phosphate deprivation leads to production of phosphorus-free membranes in F. *graminearum*, which is an important process during host invasion.

In order to identify genes that have CACGTG DNA motif upstream of their reading frame and may participate in metabolism of phosphate, perform a text search for “phosphate”.

You should identify 4 genes that comprise a conserved PHO system in filamentous fungi, and have important roles during metabolic and cellular response to phosphate starvation.