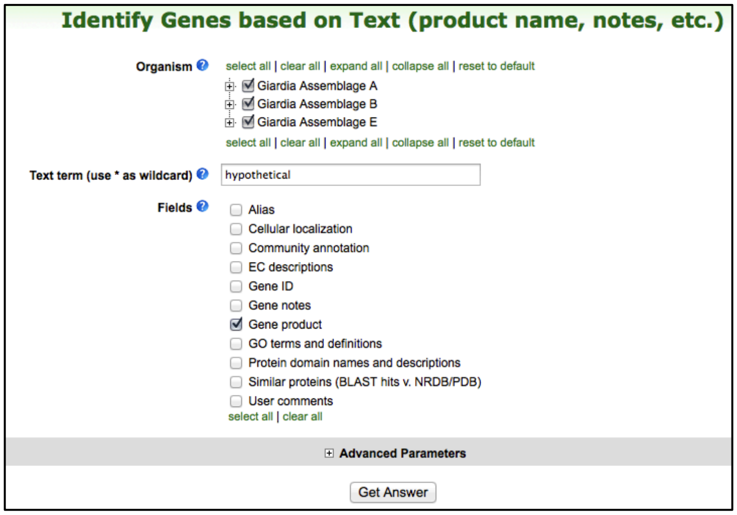
Motif Searches and Regular Expressions
Exercise 6

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

a. 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?

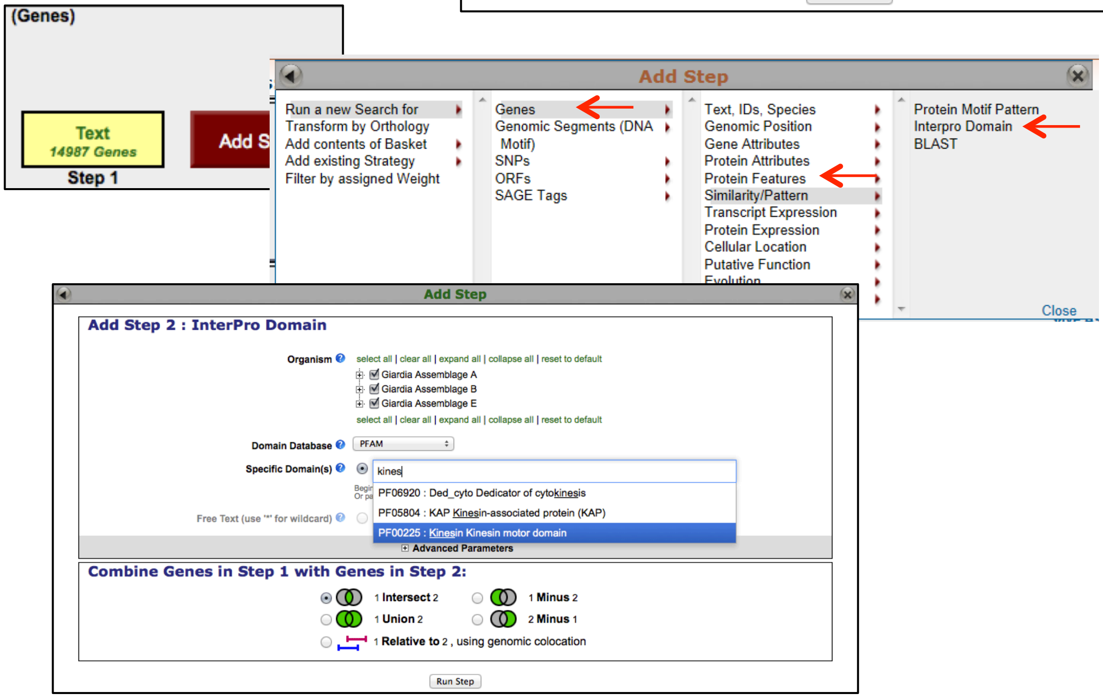


Identify Genes based on Text (product name, notes, etc.)

Organism: select all | clear all | expand all | collapse all | reset to default
☑ Giardia Assemblage A
☑ Giardia Assemblage B
☑ Giardia Assemblage E
select all | clear all | expand all | collapse all | reset to default

Text term (use * as wildcard): hypothetical

Fields:
□ Alias
□ Cellular localization
□ Community annotation
□ EC descriptions
□ Gene ID
□ Gene notes
☑ Gene product
□ GO terms and definitions
□ Protein domain names and descriptions
□ Similar proteins (BLAST hits v. NRDB/PDB)
□ User comments
select all | clear all



Add Step 2 : InterPro Domain

Organism: select all | clear all | expand all | collapse all | reset to default
☑ Giardia Assemblage A
☑ Giardia Assemblage B
☑ Giardia Assemblage E
select all | clear all | expand all | collapse all | reset to default

Domain Database: PFAM

Specific Domain(s): ● kines

PF06920 : Ded_cyto Dedicator of cytokinesis
PF05804 : KAP Kinesin-associated protein (KAP)
PF00225 : Kinesin Kinesin motor domain

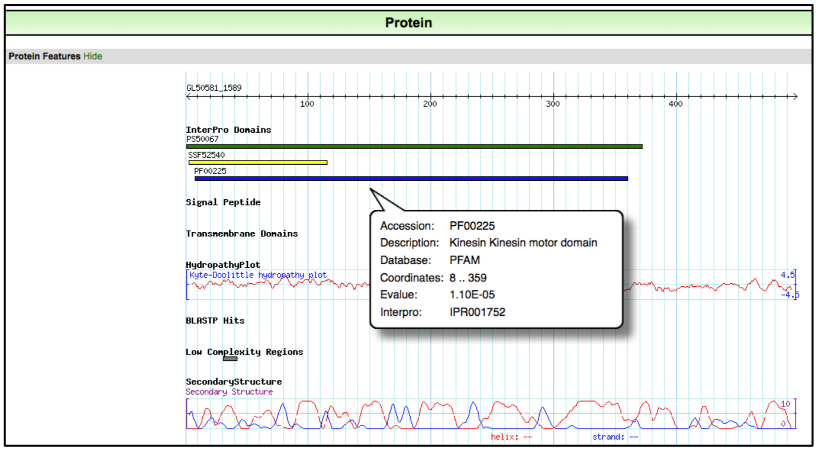
Free Text (use '*' for wildcard): ○

Combine Genes in Step 1 with Genes in Step 2:

● 1 Intersect 2
○ 1 Minus 2
○ 1 Union 2
○ 2 Minus 1
○ 1 Relative to 2 , using genomic colocation

(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.)



Accession: PF00225
Description: Kinesin Kinesin motor domain
Database: PFAM
Coordinates: 8 .. 359
Evalue: 1.10E-05
Interpro: IPR001752

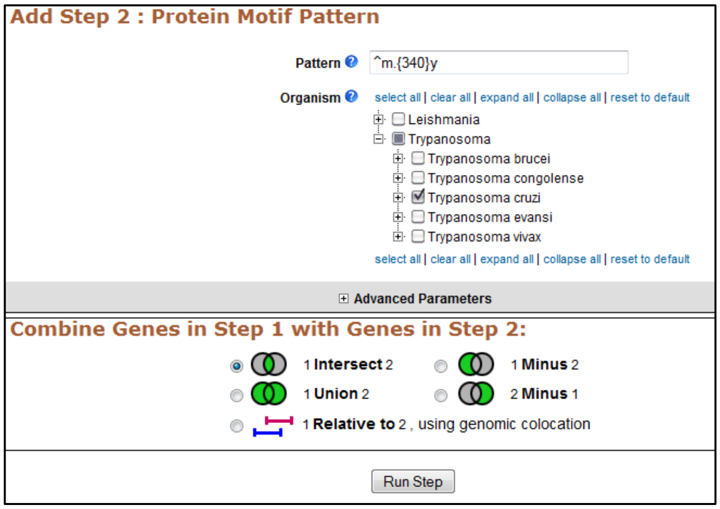
2. Using regular expressions to find motifs in TriTypDB: finding active trans-sialidases in *T. cruzi.*

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

a. *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 3500 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.



Add Step 2 : Protein Motif Pattern

Pattern: ^m.{340}y

Organism: select all | clear all | expand all | collapse all | reset to default

- ⊞ □ Leishmania
- ⊟ ▣ Trypanosoma
  - ⊞ □ Trypanosoma brucei
  - ⊞ □ Trypanosoma congolense
  - ⊞ ☑ Trypanosoma cruzi
  - ⊞ □ Trypanosoma evansi
  - ⊞ □ Trypanosoma vivax

select all | clear all | expand all | collapse all | reset to default

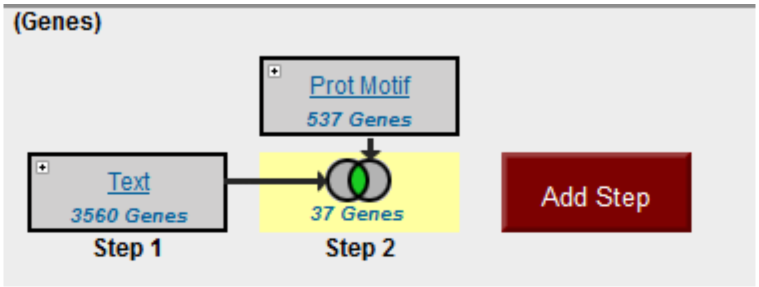
⊞ Advanced Parameters

Combine Genes in Step 1 with Genes in Step 2:

● 1 **Intersect** 2   ○ 1 **Minus** 2
○ 1 **Union** 2   ○ 2 **Minus** 1
○ 1 **Relative to** 2 , using genomic colocation

Run Step

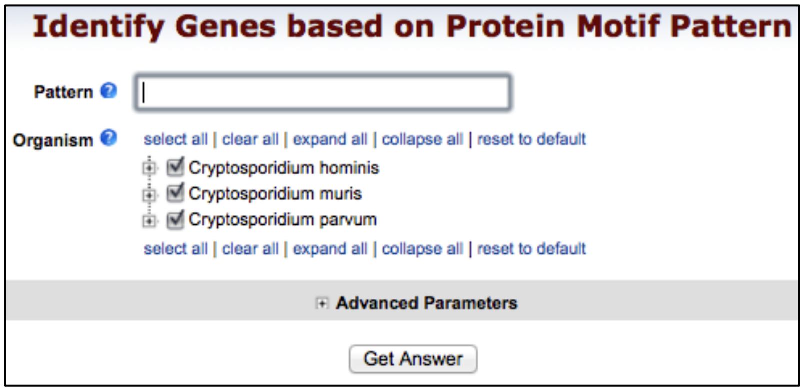
If you need help, you can go to this sample strategy below to see the answer:
http://tritrypdb.org/tritrypdb/im.do?s=a905e36f634f7b42



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

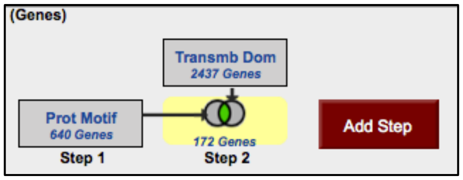
   a. The YXXΦ (Y=tyrosine, X=any amino acid, Φ=bulky hydrophobic [phenylalanine, tyrosine, threonine]) motif is conserved in many eukaryotic membrane proteins that are recognized by adaptor proteins for sorting in the endosomal/lysosomal pathway. This motif is typically located in the c-terminal end of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

   b. Use the "protein motif pattern" search to find all *Cryptosporidium* proteins that contain this motif anyware 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).

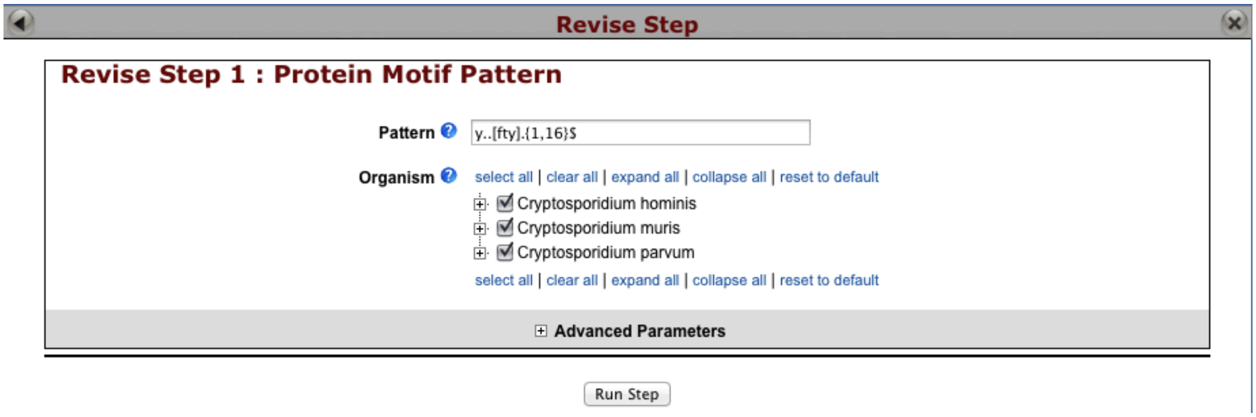

Identify Genes based on Protein Motif Pattern

Pattern
Organism
select all | clear all | expand all | collapse all | reset to default
☑ Cryptosporidium hominis
☑ Cryptosporidium muris
☑ Cryptosporidium parvum
select all | clear all | expand all | collapse all | reset to default
Advanced Parameters
Get Answer

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



d. 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 fist step and modify your regular expression).



Revise Step 1 : Protein Motif Pattern

Pattern y..[fty].{1,16}$

Organism select all | clear all | expand all | collapse all | reset to default

☑ Cryptosporidium hominis

☑ Cryptosporidium muris

☑ Cryptosporidium parvum

select all | clear all | expand all | collapse all | reset to default

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

http://cryptodb.org/cryptodb/im.do?s=f8b92af87d10013f