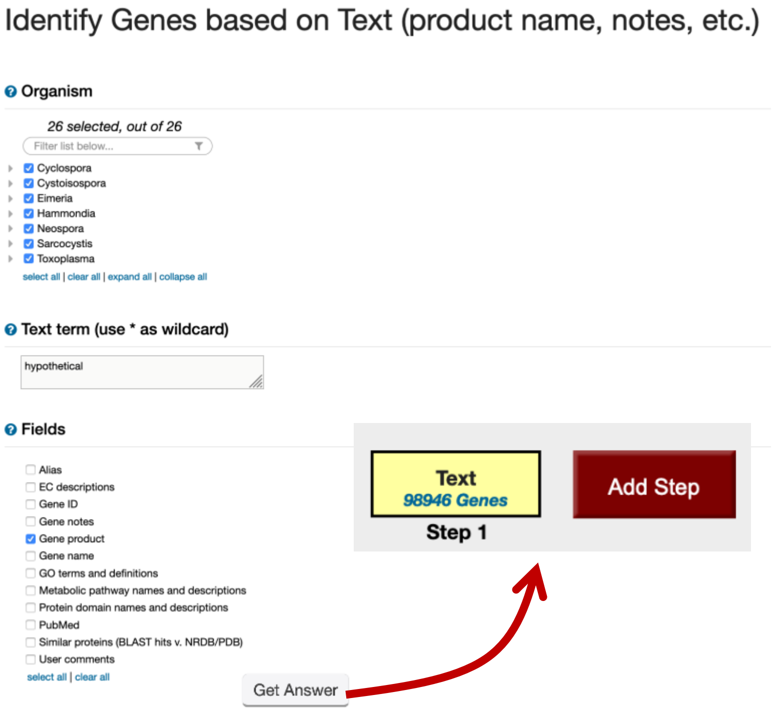
Sequence Exercises: Motifs, Domains and Colocation

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

   a. Identify all genes annotated as hypothetical in organisms in ToxoDB (select the gene product field). Use the full text search and look for genes with the word *hypothetical* in their Gene products.



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

Organism

26 selected, out of 26

- ☑ Cyclospora
- ☑ Cystoisospora
- ☑ Eimeria
- ☑ Hammondia
- ☑ Neospora
- ☑ Sarcocystis
- ☑ Toxoplasma

select all | clear all | expand all | collapse all

Text term (use * as wildcard)

hypothetical

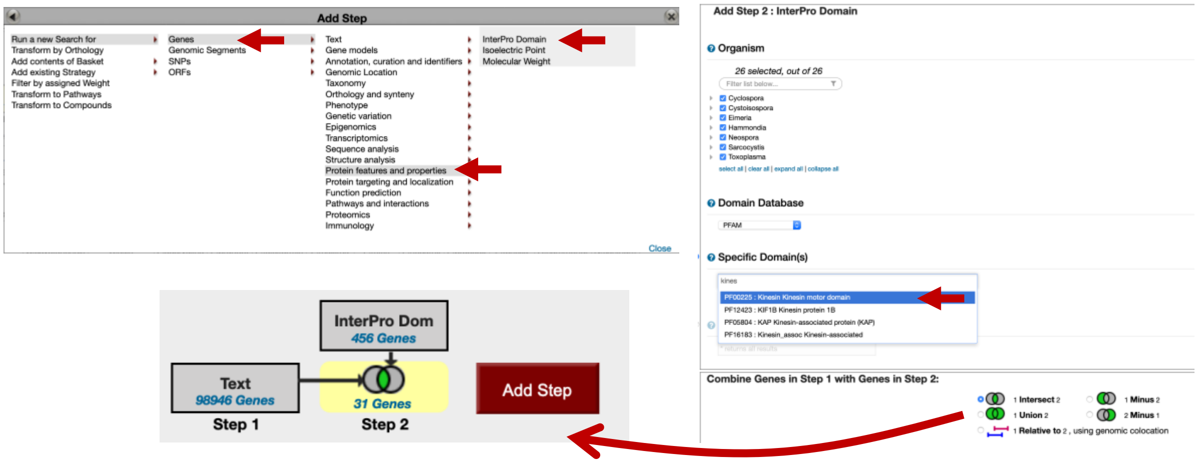
Fields

- ☐ Alias
- ☐ EC descriptions
- ☐ Gene ID
- ☐ Gene notes
- ☑ Gene product
- ☐ Gene name
- ☐ GO terms and definitions
- ☐ Metabolic pathway names and descriptions
- ☐ Protein domain names and descriptions
- ☐ PubMed
- ☐ Similar proteins (BLAST hits v. NRDB/PDB)
- ☐ User comments

select all | clear all

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

   - Add a step to the strategy. Go to the “Interpro Domain” search under ‘Protein features and properties’ similarity/pattern, start typing the work kinesin and it should autocomplete.



Add Step 2 : InterPro Domain

Organism

26 selected, out of 26

- ☑ Cyclospora
- ☑ Cystoisospora
- ☑ Eimeria
- ☑ Hammondia
- ☑ Neospora
- ☑ Sarcocystis
- ☑ Toxoplasma

select all | clear all | expand all | collapse all

Domain Database

PFAM

Specific Domain(s)

kines

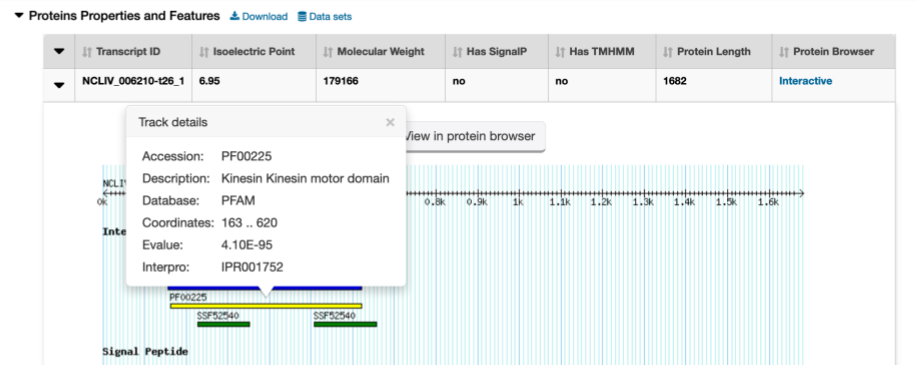
- PF00225 : Kinesin Kinesin motor domain
- PF12423 : KIF1B Kinesin protein 1B
- PF05804 : KAP Kinesin-associated protein (KAP)
- PF16183 : Kinesin_assoc Kinesin-associated

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

c. Go to the gene page for NCLIV_006210 and look at the protein feature section. Does this look like a possible motor protein?

- Click on the ID for NCLIV_006210 in the result table to go to the gene page. Scroll down to the Protein Properties and Features section and mouse over the glyphs in the InterPro domain section.



▼ Proteins Properties and Features Download Data sets

	Transcript ID	Isoelectric Point	Molecular Weight	Has SignalP	Has TMHMM	Protein Length	Protein Browser
▼	NCLIV_006210-t26_1	6.95	179166	no	no	1682	Interactive

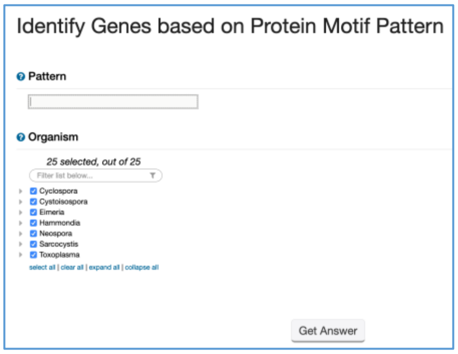
Track details

Accession: PF00225
Description: Kinesin Kinesin motor domain
Database: PFAM
Coordinates: 163 .. 620
Evalue: 4.10E-95
Interpro: IPR001752

- What other evidence on the gene page supports your conclusion?

2. Find Cryptosporidium genes with the YXXΦ receptor signal motif.

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. ****Note: do not look for the Φ symbol on your keyboard – this will not work. Rather you should use the amino acid symbols.*



Identify Genes based on Protein Motif Pattern

Pattern

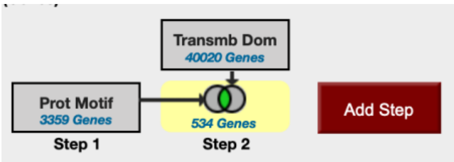
Organism

25 selected, out of 25

- ☑ Cyclospora
- ☑ Cystoisospora
- ☑ Eimeria
- ☑ Hammondia
- ☑ Neospora
- ☑ Sarcocystis
- ☑ Toxoplasma

a. Use the "protein motif pattern" search to find all proteins in ToxoDB 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) https://toxodb.org/toxo/im.do?s=e036d471ad1ccc7c